

UNIVERSITY OF UDINE

DEPARTMENT OF MATHEMATICS, COMPUTER SCIENCE AND PHYSICS
PH.D. COURSE IN COMPUTER SCIENCE, MATHEMATICS AND PHYSICS



IN CROWD VERITAS
LEVERAGING HUMAN INTELLIGENCE
TO FIGHT MISINFORMATION

Supervisor:
Prof. STEFANO MIZZARO

Candidate:
MICHAEL SOPRANO

Co-Supervisor:
Ph.D. KEVIN ROITERO

YEAR 2023

We can only see a short distance ahead,
but we can see plenty there that needs to be done.

– *Alan Turing*

Abstract

The spread of online misinformation has important effects on the stability of democracy since the information that is consumed every day influences human decision-making processes. Fact-checking is a complex process that involves several activities. The sheer size of digital content on the web and social media and the ability to immediately access and share it has made it difficult to perform timely fact-checking at scale. This thesis copes with the misinformation-spreading problem by leveraging human intelligence along three main research directions.

Truthfulness judgments are a fundamental step in the process of fighting misinformation. Usually, such judgments are made by experts, like journalists for political statements or medical doctors for health-related statements. A different approach to cope with the misinformation spreading problem can be relying on a (non-expert) crowd of human judges to perform the fact-checking activity instead of expert judges. This of course leads to the following research question: can such human judges detect and objectively categorize online (mis)information? To provide an answer, several extensive studies based on crowdsourcing are performed. Thousands of truthfulness judgments over two datasets are collected by recruiting a crowd of workers from crowdsourcing platforms and the expert judgments are compared with the crowd ones, expressed on scales with various granularity levels. Also, the political bias and cognitive background of the workers are measured and used to quantify the effect on the reliability of the data provided by the crowd.

The recent COVID-19 pandemic brought up the need of understanding whether crowdsourcing is an effective and reliable method to judge the truthfulness of (mis)information that is both related to a sensitive and personal issue and very recent as compared to when the judgment is done. Crowd workers are thus asked to judge the truthfulness of statements related to the pandemic, and to provide evidence for them. Besides showing that the crowd is indeed able to accurately judge the truthfulness of the statements, results on workers' behavior, agreement, the effect of aggregation functions, scale transformations, and workers' background and bias are reported. Also, a longitudinal study is performed by re-launching the task multiple times with both novice and experienced workers, deriving important insights on how the behavior and quality change over time. The results obtained allow for concluding that the workers are indeed able to detect and objectively categorize online recent (mis)information, such as the one related to COVID-19. Both crowdsourced and expert judgments can be transformed and aggregated to improve quality, and workers' background and other signals (e.g., source of information, behavior) impact the quality of the data. The longitudinal study demonstrates that the time span has a major effect on

the quality of the judgments, for both novice and experienced workers. Also, an extensive failure analysis of the statements misjudged by the crowd workers is provided.

Despite the prevalence of longitudinal studies, there is a limited understanding of factors that influence worker participation in them across different crowdsourcing marketplaces. Thus, a large-scale survey aimed at understanding how longitudinal studies are performed using crowdsourcing is run across multiple platforms. The answers collected are analyzed from both a quantitative and a qualitative point of view to report on crowd workers' experiences with (and their perception of) longitudinal studies. A list of recommendations for task requesters to conduct these studies effectively is provided together with a list of best practices for crowdsourcing platforms to adequately support longitudinal studies are thus provided.

Under certain conditions, crowdsourcing is indeed a viable tool for judging the truthfulness of publicly available statements and the crowd can provide reliable identification of disputed statements. However, truthfulness is a subtle matter: statements can be just biased ("cherry-picked"), imprecise, wrong, etc. and a unidimensional truth scale cannot account for such differences. Thus, a multidimensional notion of truthfulness is proposed. This time, the crowd workers are asked to judge seven different dimensions of truthfulness selected based on existing literature: Correctness, Neutrality, Comprehensibility, Precision, Completeness, Speaker's Trustworthiness, and Informativeness. A comprehensive analysis of the newly collected crowdsourced judgments shows that the workers are indeed reliable when compared to an expert-provided gold standard. Also, the proposed dimensions of truthfulness capture independent pieces of information and the crowdsourcing task can be easily learned by the workers. Indeed, the resulting judgments provide a useful basis for a more complete estimation of statement truthfulness.

Judges, as humans, are subject to limitations that might interfere with their ability to judge the truthfulness of an information item. Among such limitations, cognitive biases are human processes that often help minimize the cost of making mistakes but keep assessors away from an objective judgment of information. Such processes are particularly frequent and critical. These biases can cause errors that have a huge potential impact as they propagate not only in the community but also, for instance, in the datasets used to train automatic and semi-automatic machine learning models to fight misinformation. A review of the cognitive biases which might manifest during the fact-checking process inspired by PRISMA – a methodology used for systematic literature reviews – is thus presented. The list of cognitive biases that may affect humans is manually derived and those that might manifest during the fact-checking process are selected and grouped into categories. Then, a list of countermeasures that can be adopted by researchers, practitioners, and organizations to limit the effect of the identified cognitive biases on the fact-checking process is presented. However, characterizing cognitive biases is not enough. Indeed, also identifying these systematic biases in crowdsourced judgments is a relevant matter. Unveiling them would support, for instance, a more reliable collection of crowdsourced training data for automatic approaches and enable bias mitigation methods for existing data sets. An exploratory study on the previously collected data set containing crowdsourced truthfulness judgments for political statements is thus performed. The findings from these exploratory analyses are used to formulate specific hypotheses concerning which individual characteristics of statements or judges and what cognitive biases may affect the accuracy of crowd

workers' truthfulness judgments. To test these hypotheses, a new crowdsourcing task is thus conducted. The findings suggest that crowd workers' degree of belief in science has an impact, that they generally overestimate truthfulness, and that their judgments can be biased due to various cognitive biases. Also, exploratory evidence shows that the different dimensions of truthfulness may be affected by these biases to different degrees.

Automated fact-checking (AFC) systems to combat misinformation spreading exist, however, their complexity usually makes them opaque to the end user, making it difficult to foster trust in the system. Thus, the E-BART model is introduced with the hope of making progress on this front. E-BART can provide a truthfulness prediction for a statement, and jointly generate a human-readable explanation for this decision. It is competitive with the state-of-the-art on the e-FEVER and e-SNLI tasks. Additionally, its joint-prediction architecture is validated by showing that generating explanations does not significantly impede the model from performing well in its main task of truthfulness prediction, and that predicted truthfulness and explanations are more internally coherent when generated jointly than separately. E-BART is also calibrated, allowing the output of the final model to be correctly interpreted as the confidence of correctness. Finally, an extensive human evaluation on the impact of generated explanations is conducted, showing that the explanations increase the human ability to spot misinformation, make people more sceptical about statements and that they are competitive with ground truth explanations. The whole set of data collected and analyzed in this thesis is publicly released to the research community at [392]:

<https://doi.org/10.17605/OSF.IO/JR6VC>

Acknowledgments

I can confidently say that choosing the rightmost words to express gratitude towards the people that allowed me to join, pursue and conclude the PhD course in Computer Science, Mathematics and Physics at the University of Udine is both the hardest and the nicest task to accomplish.

I would like to start by thanking my supervisor, Stefano Mizzaro, for believing in me since 2017. Back then, I was just a master's student in Computer Science who was looking for a thesis project. He convinced me to embark on this adventure which, I would say, has proven fruitful, during 2019. His guidance has been fundamental in guiding the research work in the right direction. Then, I have to spend some words about my co-supervisor, Kevin Roitero. I am sure that without his relentless capacity to push further each goal set while also motivating myself and several other people, I would have achieved way much less than what I managed to. So thanks, Kevin, for being such a great friend, researcher, collaborator and, most notably, mentor. I have to thank also my lab-mates Marc Donada and Mihai Horia Popescu, with whom I shared countless coffees, work hours and laughs. They are probably the people who listened to my complaints most of the time when the research work did not proceed as intended. Last but not least, I would like to thank also David La Barbera for being one of my best friends for 17 years and counting, and now also a lab-mate and collaborator.

Next, I would like to thank other well-known researchers with whom I had the honour of cooperating on several projects during the latest three years and a half. I start with Gianluca Demartini, from the University of Queensland, and Damiano Spina from RMIT University. Since they both work in Australia, I had the chance to finally meet them only in December 2022. Not even the global pandemic stopped us from working together. I would like to thank also Davide Ceolin. Besides working together, he also provided me with the opportunity to spend three months at the Centrum Wiskunde & Informatica, in Amsterdam. This allows me to thank also Tim Draws, from the Delft's University of Technology, whom I had the chance to meet while being there. Living in such an international and renowned city has been a remarkable experience for the future of my research and career.

Michael Soprano

Udine, April 24th, 2023

Reading Order

This thesis is about leveraging human intelligence to fight misinformation. If you are interested in a subset of topics, there are four (4) main reading approaches:

- If you are interested in crowdsourcing-based approaches to judge the truthfulness of (mis)information items, you can read Chapter 4, Chapter 5, Chapter 6, and Chapter 7.
- If you are interested in a characterization of the cognitive biases that might manifest during the fact-checking process, you can read Chapter 8 and Chapter 9.
- If you are interested in a machine learning-based architecture to predict the truthfulness of an information item and jointly generate an explanation, you can read Chapter 10.
- If you are interested in understanding the mechanisms of Crowd_Frame, a software system that allows you to easily design and deploy crowdsourcing tasks, you can read Appendix A.

Contents

Abstract	iii
Acknowledgments	vii
Reading Order	ix
List of Figures	xix
List of Tables	xxiv
List of Equations	xxvii
List of Algorithms	xxxi
1 Introduction	1
1.1 The Rise Of Misinformation	1
1.2 The Process Of Fact-Checking	2
1.3 The Impact Of Biases	4
1.4 Automated Fact-Checking	6
1.5 The Crowdsourcing Activity Workflow	6
1.6 Meta-Research Questions	7
1.7 Terminology	8
1.8 Synopsis	9
1.9 Publications	10
2 Related Work	13
2.1 Fact-Checking Using Crowdsourcing-Based Approaches	13
2.2 The Effect of Information Recency	15

2.3	Crowdsourcing-Based Longitudinal Studies	15
2.4	The Multidimensionality Of Truthfulness	17
2.5	Bias, Echo Chambers, And Filter Bubbles In User Generated Data	17
2.6	Argument Mining For Fact-Checking	19
2.7	Automated Fact-Checking Using Machine Learning Techniques	20
2.8	Supporting Crowdsourcing-Based Approaches	21
3	Dataset	25
3.1	Politifact	25
3.2	ABC Fact Check	26
3.3	FEVER	27
3.4	e-FEVER	28
3.5	e-SNLI	29
4	The Effect of Judgment Scales and Workers' Background	31
4.1	Research Questions	31
4.2	Experimental Setting	32
4.2.1	Crowdsourcing Task	32
4.2.2	Judgment Scales And Collections	34
4.3	Descriptive Analysis	35
4.3.1	Worker Demographics	35
4.3.2	Task Abandonment	36
4.3.3	Crowdsourced Judgments Distributions	36
4.4	Results	38
4.4.1	RQ1: Crowd Workers Accuracy	38
4.4.1.1	External Agreement	38
4.4.1.2	Internal Agreement	41
4.4.2	RQ2: Judgment Scales Adequacy	43
4.4.2.1	Alternative Aggregation Functions	43
4.4.2.2	Merging Assessment Levels	43
4.4.3	RQ3: Sources Of Information	46
4.4.4	RQ4: Effect Of Worker Background and Bias	46
4.4.4.1	Cognitive Reflection Tests	46
4.4.4.2	Political Background	49
4.5	Summary	50
5	A Longitudinal Study On Misinformation About COVID-19	53
5.1	Research Questions	53
5.2	Experimental Setting	55
5.2.1	Crowdsourcing Task	55
5.2.2	Longitudinal Study	56
5.3	Descriptive Analysis	57
5.3.1	Worker Demographics	57

5.3.2	Task Abandonment	58
5.4	Results	58
5.4.1	RQ5: Crowd Workers Accuracy	58
5.4.1.1	External Agreement	58
5.4.1.2	Internal Agreement	60
5.4.2	RQ6: Transforming Judgments Scales	61
5.4.2.1	Merging Ground Truth Levels	61
5.4.2.2	Merging Crowd Levels	62
5.4.2.3	Merging Both Ground Truth And Crowd Levels	62
5.4.3	RQ7: Worker Background And Bias	64
5.4.4	RQ8: Worker Behavior	65
5.4.4.1	Time And Queries	65
5.4.4.2	Exploiting Worker Signals to Improve Quality	66
5.4.5	RQ9: Sources Of Information	66
5.4.5.1	URLs Analysis	66
5.4.5.2	Justifications	68
5.4.6	RQ10: Repeating The Experiment With Novice Workers	69
5.4.6.1	Worker Background, Behavior, Bias, And Abandonment	70
5.4.6.2	Agreement Across Batches	70
5.4.6.3	Crowd Workers Accuracy: External Agreement	71
5.4.6.4	Crowd Workers Accuracy: Internal Agreement	77
5.4.6.5	Worker Behavior: Time and Queries	77
5.4.6.6	Sources of Information: URL Analysis	78
5.4.6.7	Sources of Information: Justifications	79
5.4.7	RQ11: Analysis Of Returning Workers	81
5.4.8	RQ12: Qualitative Analysis Of Misjudged Statements	83
5.5	Summary	87
6	The Barriers To Longitudinal Studies On Crowdsourcing Platforms	89
6.1	Research Questions	89
6.2	Experimental Setting	90
6.2.1	Survey And Crowdsourcing Task Design	90
6.2.2	Statistical Testing	91
6.2.3	Qualitative Analysis Of Workers' Response	92
6.3	Worker Demographics	93
6.4	Results	95
6.4.1	RQ12: Quantitative Analysis Of Workers' Responses	95
6.4.1.1	Initial Remarks	95
6.4.1.2	Previous Experiences With Longitudinal Studies	97
6.4.1.3	P1: Spreading Of Longitudinal Studies	97
6.4.1.4	P2: Design Of Future Longitudinal Studies	103
6.4.1.5	Summary	109
6.4.2	RQ13: Key Findings From Qualitative Analysis	111
6.4.2.1	Initial Remarks	114

6.4.2.2	1.1.X.7.2: Worker Loyalty And Commitment	114
6.4.2.3	2: Crowdsourcing Platforms Suitability	116
6.4.2.4	11: Suggestions About Longitudinal Study Design	118
6.4.3	RQ14: Recommendations For Researchers And Practitioners	120
6.4.4	RQ15: Best Practices For Crowdsourcing Platforms	123
6.5	Summary	125
7	The Multidimensionality Of Truthfulness	127
7.1	Research Questions	127
7.2	Experimental Setting	128
7.2.1	The Seven Dimensions Of Truthfulness	129
7.2.2	Crowdsourcing Task	130
7.3	Descriptive Analysis	131
7.3.1	Worker Demographics	131
7.3.2	Task Abandonment	131
7.4	Results	133
7.4.1	RQ16: Reliability Of Multidimensional Judgment	133
7.4.1.1	Distributions Of Judgments	133
7.4.1.2	External Agreement	134
7.4.1.3	Internal Agreement	137
7.4.1.4	Behavioral Data	139
7.4.1.5	Summary	140
7.4.2	RQ17: Independence of the Dimensions	140
7.4.3	RQ18: Worker Behavior	144
7.4.4	RQ19: Dimension Informativeness	144
7.4.5	RQ20: Learning Truthfulness from Multidimensional Judgment	146
7.4.5.1	Supervised Approach	146
7.4.5.2	Unsupervised Approach	150
7.5	Summary	151
8	Characterizing Cognitive Biases In Fact-Checking	155
8.1	Research Questions	155
8.2	Methodology	156
8.2.1	Preferred Reporting Items For Systematic Reviews and Meta-Analyses	156
8.2.2	Eligibility Criteria, Information Sources, And Search Strategy	156
8.2.3	Data Collection And Selection Process	157
8.3	Results	158
8.3.1	RQ21: List Of Cognitive Biases	158
8.3.2	RQ22: Categorization Of Cognitive Biases	162
8.3.3	RQ23: List Of Countermeasures	164
8.3.4	RQ24: Towards A Bias-Aware Judgment Pipeline	166
8.4	Summary	168

9	The Effect Of Cognitive Biases In Fact-Checking Tasks	169
9.1	Research Questions	169
9.2	Exploratory Study	170
9.2.1	Data Preprocessing	170
9.2.1.1	Scale Transformations	170
9.2.1.2	Judgment Bias Metrics	171
9.2.1.3	Worker Bias Metrics	171
9.2.2	Exploratory Analyses	171
9.2.2.1	Exploring Worker's eME	172
9.2.2.2	Exploring Worker's eMAE	172
9.2.2.3	Exploring Worker's iME	174
9.2.3	Hypotheses For The Novel Data Collection	174
9.2.3.1	RQ25: General Worker Traits	175
9.2.3.2	RQ26: Cognitive Biases	175
9.3	Experimental Setting	176
9.3.1	Crowdsourcing Task	176
9.3.2	Variables	177
9.4	Descriptive Statistics	178
9.4.1	Worker Demographics	178
9.4.2	Task Abandonment	178
9.4.3	Agreement	180
9.5	Results	180
9.5.1	Hypothesis Tests	180
9.5.2	Exploratory Analyses	182
9.5.3	RQ25: The Role Of Workers' And Statements' Political Affiliations	182
9.5.4	RQ26: Predicting eMAE	183
9.5.5	RQ27: Looking At Individual Truthfulness Dimensions	183
9.6	Summary	184
10	A Neural Model To Jointly Predict and Explain Truthfulness	187
10.1	Research Questions	187
10.2	RQ28: E-BART Definition	188
10.3	Experimental Setting	191
10.3.1	Training Methodology	191
10.3.2	Evaluation Methodology	191
10.4	Results	192
10.4.1	RQ29: E-BART Evaluation And Validation	192
10.4.1.1	Evaluation: Original FEVER	192
10.4.1.2	Evaluation: e-FEVER	193
10.4.1.3	Evaluation: e-SNLI	194
10.4.1.4	Validation: Experiment 1	195
10.4.1.5	Validation: Experiment 2	196
10.4.2	RQ30: Testing The Impact Of The Explanations Generated	197

10.4.2.1	Crowdsourcing Task	197
10.4.2.2	External Agreement	198
10.4.2.3	Internal Agreement	201
10.4.2.4	Summary	201
10.4.3	RQ31: Network Calibration And Generation Of Confidence Scores	201
10.4.4	Summary	203
11	Discussion	207
11.1	Contributions	207
11.1.1	MRQ1: Information Truthfulness Judgment	207
11.1.2	MRQ2: Cognitive Biases	208
11.1.3	MRQ3: Predict And Explain Truthfulness	208
11.2	Practical Implications	209
11.2.1	MRQ1: Information Truthfulness Judgment	209
11.2.2	MRQ2: Cognitive Biases	210
11.2.3	MRQ3: Predict And Explain Truthfulness	211
11.2.4	Multidimensional Reviews Quality Judgment	211
11.3	Limitations	214
11.3.1	MRQ1: Information Truthfulness Judgment	215
11.3.2	MRQ2: Cognitive Biases	216
11.4	Future Directions	217
11.4.1	MRQ1: Information Truthfulness Judgment	217
11.4.2	MRQ2: Cognitive Biases	220
11.4.3	MRQ3: Predict And Explain Truthfulness	221
11.4.4	Statistical Power In Crowdsourcing	221
11.5	Conclusions	222
11.6	Acknowledgments	223
A	Crowd_Frame: Design and Deploy Crowdsourcing Tasks	225
A.1	Crowsourcing Platforms	225
A.1.1	Amazon Mechanical Turk	225
A.1.2	Toloka	228
A.1.3	Prolific	232
A.1.4	Discussion	234
A.2	Aims	237
A.3	System Design	237
A.3.1	General Architecture	237
A.3.2	Generator	238
A.3.2.1	Use Cases	238
A.3.2.2	Architecture	239
A.3.2.3	Case Study	240
A.3.3	Skeleton	241
A.3.3.1	Use Cases	241

A.3.3.2	Architecture	243
A.3.3.3	Wrapper	244
A.3.3.4	Data Format	246
A.3.3.5	Cost Estimation	247
A.3.4	Search Engine	254
A.3.4.1	Use Cases	254
A.3.4.2	Architecture	255
A.3.4.3	Microsoft Search API	256
A.3.4.4	Entrez Programming Utilities	258
A.3.4.5	fakeJSON	260
A.3.4.6	Cost Estimation	262
A.3.5	Logger	264
A.3.5.1	Architecture	264
A.3.5.2	Event Handling	265
A.3.5.3	Performance Evaluation	267
A.3.5.4	Pilot Experiment	268
A.3.5.5	Cost Estimation	270
A.3.6	Log Events	275
A.3.6.1	Context	275
A.3.6.2	Mouse Movements	275
A.3.6.3	Mouse Clicks	276
A.3.6.4	Button Click	276
A.3.6.5	Shortcuts	277
A.3.6.6	Keypress	277
A.3.6.7	Selection	278
A.3.6.8	Before Unload, Focus, and Blur	278
A.3.6.9	Scroll	278
A.3.6.10	Resize	279
A.3.6.11	Copy, Cut, and Paste	279
A.3.6.12	Text Input Backspace and Blur	279
A.3.6.13	Radio Group Input	280
A.3.6.14	Search Engine Queries and Results	280
A.3.6.15	System Usage	280
A.4	Getting Started	281
A.4.1	Environment Variables	282
A.4.2	Build Output	285
A.4.2.1	build/task/	285
A.4.2.2	build/environments/	286
A.4.2.3	build/config/	286
A.4.2.4	build/skeleton/	286
A.4.2.5	build/deploy/	286
A.4.2.6	build/mturk/	287
A.4.2.7	build/toloka/	287
A.4.3	Task Configuration	287
A.4.4	HITs Allocation	288

A.4.4.1	Manual Approach	288
A.4.4.2	Automatic Approach	289
A.4.5	Quality Checks	292
A.4.6	Local Development	294
A.5	Task Performing	295
A.5.1	Manual Recruitment	295
A.5.2	Amazon Mechanical Turk	296
A.5.3	Toloka	297
A.5.4	Prolific	298
A.6	Task Results	299
A.6.1	result/Task/	299
A.6.2	result/Data/	300
A.6.3	result/Resources/	300
A.6.4	result/Crawling/	301
A.6.5	result/Dataframe/	303
A.7	Conclusions	306
A.8	Future Work	307
B	Chapter 4: Questionnaires And Statement List	309
B.1	Demographic Questionnaire	309
B.2	Cognitive Reflection Test (CRT)	311
C	Chapter 5: Statements List	313
D	Chapter 6: Survey Questions	317
D.1	P1: Current Perception Of Longitudinal Studies	317
D.2	P2: Possible Participation And Commitment To Longitudinal Studies	319
E	Chapter 7: Task Instructions	323
F	Chapter 8: PRISMA Checklists And List Of Cognitive Biases	325
F.1	PRISMA 2020 Checklist For Abstracts	326
F.2	PRISMA 2020 Checklist	327
F.3	The 220 Cognitive Biases	329
G	Chapter 9: Questionnaires	333
G.1	Citizen Trust in Government Organizations (CTGO)	333
G.2	Belief in Science Scale (BISS)	334
H	Section 11.2.4: Multidimensional Scale For Reviews Quality Judgment	335

Bibliography	337
Analytical Index	383

List of Figures

4.1	From left to right: S_3, S_6, S_{100} . From top to bottom: individual scores distribution (first row), gold judgments distribution (second row), and aggregated judgments distribution (third row).	37
4.2	External agreement with PolitiFact and ABC Fact Check statements, separated by the vertical dashed line.	39
4.3	Pairwise agreement and relative frequency by looking at each HIT of the task.	40
4.4	Agreement between judgment scales. From left to right: S_6 vs. S_3, S_{100} vs. S_3 , and S_{100} vs. S_6	41
4.5	α cuts sorted by decreasing values.	42
4.6	Agreement with the ground truth while using the median as aggregation function (highlighted by the red diamond). Compare with Figure 4.2.	44
4.7	Agreement between S_3 (first row) and S_6 (second row), and ABC Fact Check (left) and PolitiFact (right). Aggregation function: majority vote.	45
4.8	Agreement with ground truth for merged categories for PolitiFact. From top to bottom: three and two resulting categories. The median is highlighted by the red diamond.	47
5.1	Agreement between the PolitiFact experts and the crowd judgments.	59
5.2	Agreement between the PolitiFact experts and the crowd judgments. First row: E_6 to E_3 . Second row: E_6 to E_2 . Compare with Figure 5.1.	61
5.3	Crowd judgments merged into groups and then aggregated with the mean function. Comparison with E_6 . Compare with Figure 5.1.	63
5.4	Expert judgments merged into groups and then aggregated with the mean function. Compare with Figure 5.2, Figure 5.3, and Figure 5.3.	64
5.5	Distribution of the ranks of the URLs selected by workers.	67
5.6	The effect of the origin of a justification on worker accuracy. Text copied/not copied from the selected URL.	69
5.7	Correlation values between the judgments (aggregated using the mean) across Batch1, Batch2, Batch3, and Batch4.	72
5.8	Agreement between the PolitiFact experts and crowd judgments. Figure 5.8a is the same as Figure 5.1b.	74
5.9	Distribution of the ranks of the URLs selected by workers for all the batches.	78
5.10	Effect of the origin of a justification on the labelling error. Text copied/not copied from the selected URL.	80

5.11	MAE and CEM ^{ORD} for individual judgments for returning workers. Green indicates that the set of workers on the y-axis is better than the one on the x-axis. Red indicates the opposite.	82
5.12	Relative ordering of statements across batches according to MAE for each PolitiFact category. Rank 1 represents the highest MAE.	84
5.13	MAE (aggregated by statement) against the number of days elapsed (from when the statement was made to when it was evaluated). Each point is the MAE of a single statement in a batch. Dotted lines are the trend of MAE in time for the batch, straight lines are the mean MAE for the batch. The black dashed line is the global trend of MAE across all batches.	86
6.1	Continent of provenance of the 300 workers recruited, breakdown across every crowdsourcing platform.	93
6.2	Worker provenance distribution across the whole world, breakdown across each crowdsourcing platform.	96
6.3	Number of previous experiences as reported by workers.	98
6.4	Months elapsed since each experience.	98
6.5	Number of sessions of each experience.	99
6.6	Time interval between sessions of each experience.	99
6.7	Average session duration of each experience (minutes and hours).	100
6.8	Crowdsourcing platforms where each experience took place.	100
6.9	Model used to pay the workers during each experience (i.e., when the reward was provided).	101
6.10	Workers willingness to participate again in each experience reported.	101
6.11	Incentives that drive workers to participate in each experience.	102
6.12	Completion of the longitudinal studies in which the worker participated.	102
6.13	Incentives that drive workers to complete each experience.	103
6.14	Reasons that the limit popularity of longitudinal studies in crowdsourcing according to workers.	103
6.15	Ideal amount of commitment (days) for a longitudinal study according to workers.	104
6.16	Reasons that drive workers to refuse participation in longitudinal studies.	104
6.17	Ideal participation frequency according to workers.	105
6.18	Ideal session duration for longitudinal studies according to workers.	105
6.19	Ideal hourly payment (USD\$) for the participation in a longitudinal study according to the workers.	106
6.20	Ideal amount of daily time to allocate to participate in longitudinal studies according to the workers.	106
6.21	Most important incentives for participating in longitudinal studies according to the workers.	107
6.22	Ideal task type to be performed in longitudinal studies according to the workers.	108
6.23	Benefits provided by involvement in longitudinal studies according to the workers.	108
6.24	Downsides of being involved in longitudinal studies according to the workers.	109

7.1	Abandonment rate shown as number of workers that reached a certain number of steps in the task. The abandonment monotonically decreases as the step number increases.	132
7.2	Correlation between dimensions: individual in the lower triangle and diagonal, aggregated in the upper triangle, aggregated distribution in the last row; breakdown on PolitiFact (in blue) and ABC Fact Check (in orange) categories (better on screen and using the zoom feature). Workers values are skewed towards positive values, i.e., Agree (+1) and Completely Agree (+2) (diagonal and bottom plots), and different dimensions have different correlation values (upper and lower triangle).	135
7.3	Correlation with the ground truth of the Overall Truthfulness and behavior of the Correctness and Precision dimensions with a breakdown on PolitiFact and ABC Fact Check labels. Mean as aggregation function.	136
7.4	Correlation with the ground truth of Overall Truthfulness and a sample of the other dimensions. PolitiFact has been grouped into 3 bins. Mean as aggregation function. The binning allows seeing more clearly the increasing median trends. Compare to Figure 7.3.	138
7.5	Average time (in seconds) spent by workers to judge the Overall Truthfulness for each statement position.	139
7.6	Principal components for the statements \times judgments matrix.	141
7.7	Truthfulness dimensions first aggregated using the mean function then combined. Compare with Figure 7.3.	143
7.8	Overall Truthfulness judgments aggregated with the weighted mean function. Compare with Figure 7.3 and Figure 7.4.	145
7.9	Effectiveness over the 3 folds for the ABC Fact Check 30 levels case. The dashed line represents the best baseline.	149
7.10	Visualization of the embeddings space. All dimensions are compared to average. The coloring in the first visualization follows the legend used in the other plots.	153
8.1	Data collection and selection process of the PRISMA-inspired approach.	158
9.1	Mean eME in the dataset (Section 7.2). Four workers who considered themselves something else other than Democrat, Independent, or Republican are excluded.	173
9.2	Comparison of workers' abandonment and failure distribution. The orange lines represent the task described in Section 9.3.1. The blue lines represent the task described in Section 7.2.2	179
9.3	Correlation with the ground truth of three dimensions with a breakdown on PolitiFact and ABC Fact Check labels. Compare with Figure 7.3	181
9.4	Scatter plots showing the relationships between workers' eMAE and their <i>trust in politics</i> (H1a, left-hand plot), <i>belief in science</i> (H1b, center-left plot), <i>cognitive reasoning abilities</i> (H1c, center-right plot), and <i>mean confidence</i> (H2d, right-hand plot).	183
10.1	The Joint Prediction Head of the E-BART architecture.	188

10.2	The inference process of the E-BART architecture.	189
10.3	The training configuration of the E-BART architecture.	190
10.4	External agreement between ground truth and crowd for raw (first column) and aggregated (second column) truthfulness judgments. Each cell represents either the count of judgments (first column), or statements (second column). Correctly classified statements lay on the main diagonal.	199
10.5	External agreement between ground truth and crowd for raw (left chart) and aggregated (center chart) truthfulness judgments. The worker preferences are shown in the right chart. Each cell represents either the count of judgments (left and right charts), or statements (center chart).	200
10.6	Confidence as a function of accuracy for the E-BARTFull model. The calibration error is corrected using Temperature Scaling [167].	203
11.1	Example of labeling of reviews following the standard argumentation theory adopted.	213
A.1	Type selection for a task on Amazon Mechanical Turk.	226
A.2	Parameters configuration for a task on Amazon Mechanical Turk.	227
A.3	Task type selection for a project on Toloka.	229
A.4	Interface design for a project on Toloka using its template builder.	230
A.5	Input and output data specification for a project on Toloka.	230
A.6	Worker attributes configuration for a pool of a project on Toloka.	233
A.7	General parameters set up for a study on Prolific.	234
A.8	Data collection set up for a study on Prolific.	235
A.9	Audience configuration for a study on Prolific.	236
A.10	Study cost configuration for a study on Prolific.	236
A.11	General architecture of Crowd_Frame.	238
A.12	Use case diagram for a requester that designs and deploy a task using Crowd_Frame.	239
A.13	Worker interface of a misinformation assessment task deployed by Roitero et al. [362] using Crowd_Frame.	242
A.14	Use case diagram for a worker that performs a task deployed using Crowd_Frame.	243
A.15	Crowd_Frame Skeleton execution flow.	244
A.16	Wrapper interface for to a task configured Crowd_Frame that relies on Amazon Mechanical Turk for recruiting workers.	245
A.17	Use case diagram for a worker that performs a task deployed using Crowd_Frame.	254
A.18	Interface of the Search Engine component of Crowd_Frame.	255
A.19	Interface of the Search Engine component of Crowd_Frame.	256
A.20	Infrastructure of the Logger component of Crowd_Frame.	265
A.21	Requests sent to evaluate the performances of the infrastructure during each round of the test.	268
A.22	Requests managed per second by the dedicated server solution during each round of the test.	268

A.23	CPU usage of the dedicated server solution during each round of the test.	269
A.24	Memory usage of the dedicated server solution during each round of the test.	269
A.25	Event type distribution of the pilot test deployed to gather real data using the Logger component.	270
A.26	Worker behavior while evaluating a HIT's element reconstructed using the log data produced by the Logger component.	271
A.27	Worker behavior while answering a questionnaire reconstructed using the log data produced by the Logger component.	272
A.28	Deploy diagram of the infrastructure used to allow Crowd_Frame and the solver communicating.	292
A.29	Sample configuration for the solver used to automatically allocate elements to be evaluated in a set of HITs.	293

List of Tables

3.1	Statements fact-checked during 2022 by PolitiFact.	26
3.2	Statements fact-checked during 2019 by ABC Fact Check.	27
3.3	Statements sampled from the FEVER dataset.	28
3.4	Statements sampled from the e-FEVER dataset.	29
3.5	Statements sampled from the e-SNLI dataset.	30
4.1	Example of statements sampled from the PolitiFact and ABC Fact Check datasets.	32
4.2	Worker completion, abandonment, and failure rates during the crowdsourcing tasks for the S_3 , S_6 , and S_{100} collections.	36
4.3	Websites from which workers chose URLs to justify their judgments without considering gold questions for S_3 , S_6 , and S_{100} . Only websites with percentage $\geq 1\%$ are shown.	48
4.4	Distribution of the ranks in search results for the URLs chosen by workers in S_3 , S_6 , and S_{100}	48
4.5	Correlation between Cognitive Reflection Test (CRT) performance and z-scores for each truthfulness level and the correlation between worker age and z-scores.	49
5.1	Experimental setting for the longitudinal study. All dates refer to 2020. The values reported are absolute numbers.	57
5.2	Statement position in the task versus: time elapsed, cumulative on each single statement (first row), CEM^{ORD} (second row), number of queries issued (third row), and number of times the statement has been used as a query (fourth row).	66
5.3	Websites from which workers chose URLs to justify their judgments.	67
5.4	Abandonment data for each batch of the longitudinal study.	71
5.5	ρ (lower triangle) and τ (upper triangle) correlation values among batches for the aggregated scores of Figure 5.8.	75
5.6	RBO bottom-heavy (lower triangle) and RBO top-heavy (upper triangle) correlation values among batches for the aggregated scores of Figure 5.8. Document sorted by increasing aggregated score.	76
5.7	Correlation between α and Φ values. ρ in the lower triangle, τ in the upper triangle.	77
6.1	Breakdown of the country of provenance of the 300 workers recruited, grouped by continent.	94

6.2	Previous experiences with longitudinal studies reported by the workers recruited.	97
6.3	Summary of the key findings for the P1 part of the survey presented in the quantitative analysis.	111
6.4	Summary of the key findings for the P2 part of the survey presented in the quantitative analysis.	112
6.5	Summary of statistical tests comparing answer groups of each platform. Questions without statistically significant comparisons are not reported. Statistical significance is computed using adjusted p-values according to Section 6.2.2.	113
6.6	Themes emerged while reading each text-based answer provided by the workers.	115
6.7	Distribution across each theme of the answers collected for question 1.1.X.7.2.	116
6.8	Sample of answers provided by workers concerning loyalty to longitudinal studies.	117
6.9	Distribution across each theme of the answers collected for question 2.	117
6.10	Sample of answers provided by workers concerning the adequacy of crowdsourcing platforms in supporting longitudinal studies.	119
6.11	Distribution across each theme of the answers collected for question 11.	120
6.12	Sample of suggestions provided by workers concerning longitudinal studies.	121
7.1	Example of statements sampled from the PolitiFact and ABC Fact Check datasets.	128
7.2	Effectiveness metrics when predicting the expert judgment. Baselines above the dashed line.	148
8.1	Taxonomy of biases, adapted from Dimara et al. [108]. The biases are classified according to their task (rows) and each task's subcategory, called "flavor" (columns).	163
8.2	Constituting elements of a bias-aware assessment pipeline.	167
10.1	Statement where the ground truth label is NOT ENOUGH INFO and the one predicted by E-BART is REFUTES.	192
10.2	Effectiveness of E-BART on the e-FEVER dataset.	193
10.3	Statement where the ground truth label is SUPPORTS and the one predicted by E-BART is SUPPORTS.	194
10.4	Statement where the ground truth label is REFUTES and the one predicted by E-BART is REFUTES (E-BARTSmall).	194
10.5	Example where the ground truth label is Entailment and the one predicted by E-BART is Entailment.	195
10.6	Example where the ground truth label is Neutral and the one predicted by E-BART is Neutral.	195
10.7	Effectiveness of E-BART and Separate-Bart on the e-FEVERSmall dataset.	196
10.8	Internal consistency of E-BART and Separate-Bart on the e-FEVERSmall dataset.	197
10.9	Statement where the ground truth label is SUPPORTS and the one predicted by E-BART is SUPPORTS. Confidence score is 0.851, original confidence was 0.987 (E-BARTFull).	204

10.10	Statement where the ground truth label is REFUTES and the one predicted by E-BART is SUPPORTS. Confidence score is 0.432, original confidence was 0.550 (E-BARTFull).	205
A.1	Summary of each configuration step of the Generator component.	241
A.2	Sample cost estimation parameters for the usage of the data table by the Skeleton component of Crowd_Frame.	251
A.3	Headers provided while sending requests to the Bing Web Search API.	257
A.4	Headers captured while receiving responses from the Bing Web Search API.	257
A.5	Query parameters provided when sending queries to the Bing Web Search API.	257
A.6	General description the of the nine E-utilities provided by the Entrez system.	259
A.7	Query parameters provided when sending requests to the ESearch utility of the Entrez system.	260
A.8	Payload parameters used to customize the behavior of the responses generated by the fakeJSON service.	261
A.9	Environment variables used to customize Crowd_Frame.	283
A.10	Folder structure of the output of a Crowd_Frame build.	285
A.11	Configuration files of a task deployed using Crowd_Frame.	286
A.12	Source files of a task deployed using Crowd_Frame.	287
A.13	Content of the build folder to deploy a task on Toloka.	287
A.14	Structure of the results folder of Crowd_Frame.	299
A.15	DataFrame produced when downloading the final results of a task.	304

List of Equations

7.1	Word Mover Distance computation between dimensions scores and ground truth rationales.	146
10.1	Confidence prediction using Temperature Scaling [167].	202
A.1	Sample estimation of the storage cost for the usage of S3 by the Skeleton component of Crowd_Frame.	249
A.2	Sample estimation of the storage cost for the usage of S3 by the Skeleton component of Crowd_Frame.	249
A.3	Sample estimation of the data transfer cost for the usage of S3 by the Skeleton component of Crowd_Frame.	249
A.4	Cost estimation for the usage of Amazon S3 for the Skeleton component.	249
A.5	Sample estimation of the WRUs cost for the usage of the data table of the Skeleton component.	251
A.6	Sample estimation of the RRUs cost for the usage of the data table of the Skeleton component.	251
A.7	Sample estimation of the data storage cost for the usage of the data table by the Skeleton component of Crowd_Frame.	252
A.8	Sample overall estimation of the cost for the usage of the data table by the Skeleton component of Crowd_Frame.	252
A.9	Sample estimation of the WRUs cost for the usage of the access control list table of the Skeleton component.	252
A.10	Sample estimation of the RRUs cost for the usage of the access control list table of the Skeleton component.	253
A.11	Sample estimation of the data storage cost for the usage of the access control list table by the Skeleton component of Crowd_Frame.	253
A.12	Sample overall estimation of the cost for the usage of the data table by the Skeleton component of Crowd_Frame.	253
A.13	Sample overall estimation of the cost for the usage of the data table by the Skeleton component of Crowd_Frame.	254
A.14	Sample estimation of the cost for the usage of the Bing Web Search API.	263
A.15	Sample estimation of the API Gateway cost for the usage of the Logger component.	273
A.16	Sample estimation of Simple Queue cost for the usage of the Logger component.	273
A.17	Sample estimation of the Lambda cost for the usage of the Logger component.	274
A.18	Sample estimation of the DynamoDB cost for the usage of the Logger component.	274
A.19	Sample estimation of the overall cost for the usage of the Logger component.	275

List of Listings

A.1	Interface of a sample task on Amazon Mechanical Turk built using Crowd Elements.	228
A.2	Image classification JSON configuration for the interface of a project on Toloka built using the template builder.	231
A.3	Input data specification for a project on Toloka.	232
A.4	Output data specification for a project on Toloka.	232
A.5	Input data initialization for two HITs within a pool of a project on Toloka. . .	232
A.6	Format for a DynamoDB table that contains the access control list of a task deployed using Crowd_Frame.	246
A.7	Format for a DynamoDB table that contains the data produced by the worker during a task deployed using Crowd_Frame.	247
A.8	JSON data stored for the element of a HIT evaluated by a worker during a task deployed using Crowd_Frame.	248
A.9	Sample request that contains headers and query parameters sent to the Bing Web Search API.	258
A.10	Sample initial request sent to the ESearch utility of the Entrez system.	259
A.11	Sample initial request sent to the ESummary utility of the Entrez system. . . .	260
A.12	Sample request sent to the fakeJSON service using CURL.	261
A.13	Payload of the request sent to the fakeJSON service to generate fake search results.	262
A.14	Payload of each log request sent by the Logger component of Crowd_Frame. .	266
A.15	Input data specification for a project on Toloka.	282
A.16	Sample <code>credentials.json</code> file to store IAM user access key.	282
A.17	Subset of environment variables required by Crowd_Frame.	282
A.18	Python packages required to initialize Crowd_Frame.	283
A.19	Valid set of one HIT for a task designed and deployed using Crowd_Frame. .	289
A.20	Fragment of elements to evaluated in the task published by Roitero et al. [361].	293
A.21	Fragment of the allocation built automatically using the solver integrated with Crowd_Frame.	294
A.22	Valid set of one HIT with two elements, where one is used within a custom quality check.	295

A.23 Default implementation of the static method that performs custom quality checks in Crowd_Frame.	296
A.24 Sample development environment of Crowd_Frame.	297
A.25 Snapshot of a worker who participates in a task with a single batch deployed using Crowd_Frame.	300
A.26 Subset of the information obtained by perform the reverse lookup of the IP address of a worker.	301
A.27 Subset of the information obtained by analyzing the user agent string of a worker.	302
A.28 Metadata produced by the download script while trying to crawl a webpage retrieved by the search engine of Crowd_Frame.	303
A.29 Example of the workers_acl dataframe produced by Crowd_Frame.	305
A.30 Example of the workers_answers dataframe produced by Crowd_Frame.	306

List of Algorithms

A.1	Procedure to allocate a dataset into HITs using the format required	290
A.2	Procedure to sample elements without duplicates for a single HIT	291

Introduction

1.1 The Rise Of Misinformation

The rise of (online) misinformation is a problem that harms society, and the information we consume every day influences our decision-making process [429]. Thus, understanding what information should be trusted and which should not is crucial for democratic processes to function as supposed to, since it is often done with the intended mean of deceiving people towards a certain political agenda. The sheer size of digital content on the web and social media and the ability to immediately access and share it has made it difficult to perform timely fact-checking at scale [408]. The rate at which such a problem propagates continues to increase, largely aided by the increasing popularity of social media platforms [325].

The task of checking the truthfulness of published information has been traditionally performed by expert fact checkers, that is, journalists who perform the task by verifying information sources and searching for evidence that supports the claims made by the document or statement they are verifying. Indeed, it is infeasible for journalists to provide fact-checking results for all news which are being continuously published. Also, relying on fact-checking results requires trusting those who performed the fact-checking job. This is something the average web user may not be willing to accept. Even worse, fact-checking might actually decrease trust in news outlets [55]. Significant efforts have been made by different research communities on developing techniques and datasets to automatize fact-checking, also defined as the information credibility assessment task [85, 19, 129, 176, 436]. Key approaches to automatically differentiate between false and valid statements also include neural models [366, 387, 438].

The spread of misinformation is further exacerbated by events such as the COVID-19 pandemic. The problem is (and was) so serious that the World Health Organization (WHO) used the neologism “infodemic” to refer to the problem of misinformation, during the peak of the COVID-19 pandemic [6].

“We’re concerned about the levels of rumours and misinformation that are hampering the response. [...] we’re not just fighting an epidemic; we’re fighting an infodemic. Fake news spreads faster and more easily than this virus, and is just as dangerous.”

That's why we're also working with search and media companies like Facebook, Google, Pinterest, Tencent, Twitter, TikTok, YouTube and others to counter the spread of rumours and misinformation. We call on all governments, companies and news organizations to work with us to sound the appropriate level of alarm, without fanning the flames of hysteria."

These are the alarming words used by Dr. Tedros Adhanom Ghebreyesus, the WHO (World Health Organization) Director General during his speech at the Munich Security Conference on 15 February 2020.¹ Such words tell us that the WHO Director General chooses to target explicitly misinformation-related problems. Indeed, all of us have experienced mis- and dis-information during the COVID-19 health emergency. The research community has focused on several COVID-19 related issues [52], ranging from machine learning systems aiming to classify statements and claims based on their truthfulness [440], search engines tailored to the COVID-19 related literature, as in the TREC-COVID Challenge [352], topic-specific workshops like the NLP for COVID-19 workshop at ACL 2020 [428] and evaluation initiatives like the TREC Health Misinformation Track [80].² Besides the academic research community, commercial social media platforms also have looked at this issue.

These considerations show that it is still necessary to involve humans in the fact-checking process. A more scalable and decentralized approach that relies on a (large) crowd of non-expert would allow fact-checking to be more widely available. As it is well known, *crowdsourcing* means to outsource a task – which is usually performed by a limited number of experts – to a large mass (the “crowd”) of unknown people (the “crowd workers”), using an open call. The idea that the crowd can identify misinformation might sound implausible at first – isn't the crowd the very means by which misinformation is spread? – However, recent research has shown that people can reliably perform fact-checking using crowdsourcing-based approaches [232, 356] and assess information quality across multiple truthfulness dimensions or quality aspects [266, 419], provided that adequate countermeasures and quality assurance techniques are employed. The recent works mentioned specifically crowdsource the task of misinformation identification, or rather the judgment of the truthfulness of statements made by public figures (e.g., politicians), usually on political, economical, and societal issues. Even though experts are still considered the most reliable when it comes to truthfulness judgments, leveraging them to judge and render a verdict on the truthfulness of news becomes too expensive and impractical if performed at scale. Crowdsourced fact-checking is indeed widely used in academic research [326, 332, 436, 376, 378] and has already found applications in industry [9, 340].

1.2 The Process Of Fact-Checking

Fact-checking is a complex process that involves several activities [283, 430]. An abstract and general pipeline for fact-checking might include the following steps (not necessarily in this order): check-worthiness (i.e., ensure that a piece of information includes a claim that is of great interest for a possibly large audience), evidence retrieval (i.e., retrieve

¹<https://www.who.int/dg/speeches/detail/munich-security-conference>

²<https://trec-health-misinfo.github.io/>

the evidence needed to fact-check the statement), veracity classification or truthfulness assessment, discussion among the assessors to reach a consensus, and assignment and publication of the final veracity/truthfulness score for the information item inspected. It is thus interesting to briefly examine the fact-checking processes adopted in practice by three famous organizations, namely *FactCheck.org*, *PolitiFact*, and *RMIT ABC Fact Check* – verified signatories to the International Fact-Checking Network (IFCN, <https://www.poynter.org/ifcn/>) – given that they set a de-facto standard for the pipeline required to perform fact-checking at scale.

PolitiFact fact-checks statements by US Politicians (see [337] for the detailed description of the process). The reporter in charge of running the fact-checking proposes a rating using a six-level scale to perform the truthfulness judgment step. Such assessment is reported to an editor. The reporter and the editor work together to reach a consensus on the rating proposed by adding clarifications and details if needed. Then, the statement is shown to two additional editors, which review the work of the editor and the reporter by providing an answer to a set of four questions. The questions are:

1. Is the statement literally true?
2. Is there another way to read the statement?
3. Did the speaker provide evidence? Did the speaker prove the statement to be true?
4. How have we handled similar statements in the past? What is PolitiFact 's jurisprudence?

Then, the definitive rating of the statement is decided upon using the majority vote of the score submitted by the editors, final edits are made to make sure everything is consistent, and the report is finally published.

FactCheck.org, similarly, fact-checks statements dealing with US Politicians (see [137] for a detailed description of the process). As for the check-worthiness step, they select statements said by the president of the United States and important politicians, focusing on statements made by presidential candidates during public appearances, top senate races, and congress actions. To perform evidence retrieval, they seek through video transcripts or articles to identify possible misleading or false statements and ask the organization or the person making the claim to prove the veracity of the statement by providing supporting documentation. If no evidence is provided, *FactCheck.org* searches trusted sources for evidence confirming or refusing the claim. Finally, the verdict about the claim is published. At *FactCheck.org*, each statement is revised in most cases by four people (see [137, Section Editing]): a line editor (reviewing content), a copy editor (reviewing style and grammar), a fact-checker (in charge of the fact-checking process), and the director of the Annenberg Public Policy Center.

ABC Fact Check, on the other hand, focuses on statements made by Australian public figures, advocacy groups, and institutions (see [351] for a detailed description of the process). The statement to be checked needs to be approved by the director who assesses its checkworthiness. Then, one of the researchers at *ABC Fact Check* contacts experts in the field and occasionally the speaker to retrieve evidence and get back data which can be helpful in the fact-checking process. The researcher writes the data and the information.

An expert fact-checker inspects and reviews them. In this stage, the expert fact-checker identifies possible problems and questions the researcher on anything that they might have missed (e.g., missing or not exhaustive evidence retrieved). The expert fact-checker and the researcher revise the draft until the fact-checker is satisfied with the outcome; Then, the whole team discusses the final verdict for the statement. The final verdict of the statement is expressed on a fine-grained categorical scale, which is used in their publications. For documentation purposes, the verdict is also refined into a three-level scale defining its truthfulness value: False, In-Between, True.

In summary, the fact-checking processes of the three organizations share similarities and differences. All three organizations are committed to upholding the principles of the International Fact-Checking Network (IFCN) and focus on checking statements made by politicians and public figures. However, they differ in the specific process followed for evidence retrieval, truthfulness assessment, and rating of the statements. FactCheck.org focuses on US politicians, seeking evidence from statements and trusted sources, and has a four-person team to review each statement. PolitiFact also concentrates on US politicians, using a six-level rating scale and a consensus-based process among editors and reporters to determine the final rating. ABC Fact Check targets Australian public figures, engaging field experts and a collaborative review process with the whole team to decide the final verdict. Despite these differences, all three organizations demonstrate a strong commitment to accuracy, transparency, and thoroughness in their fact-checking processes, providing valuable resources for the public to access reliable information on political statements. Moreover, it is worth mentioning that since all three organizations rely exclusively on human judgment for their evaluations, their processes are potentially susceptible to systematic errors due to the limits of human cognition.

1.3 The Impact Of Biases

During an experiment that took place in 1974, Tversky et al. [420] showed to a group of people brief personality descriptions of several individuals, allegedly sampled at random from a group of 100 professional engineers and lawyers. The subjects of the experiment were asked to assess, for each description, whether it referred to an engineer or a lawyer. The odds of any particular description belonging to an engineer rather than to a lawyer were roughly the same for each of the two groups. Subjects were split into two experimental groups; the former was told that the group of professionals from which the descriptions were drawn consisted of 70 engineers and 30 lawyers, while the latter was told the opposite (i.e., those descriptions were drawn from a group of 30 engineers and 70 lawyers). According to Bayes probabilities, the two groups should have reported unbalanced annotations while, in reality, the subjects in the two conditions reported roughly the same probability judgments, ignoring the prior probabilities of the two categories and relying only on the degree to which the description was representative of the two stereotypes.

Tversky et al. use their example to illustrate that people often rely on a limited number of heuristic principles in their cognitive processes, such as the “judgment by representativeness” detailed in the example. These heuristics, despite being useful, sometimes lead to severe and systematic errors and they are usually known as “biases”. A general definition

of bias is, according to the Oxford Dictionary:

A strong feeling in favour of or against one group of people, or one side in an argument, is often not based on fair judgment.

People such as fact-checkers, being they experts [137, 337, 351] or crowd workers [99, 361, 363, 395], can thus be subject to errors that can harm the information assessment process. Indeed, crowdsourcing often relies on contributions from large groups of laypeople with different backgrounds, expertise, and skills. Systematic errors among those workers may reduce the quality of their annotations [112, 126, 192]. In fact-checking tasks, factors such as workers' political affiliation or their general trust in politics may affect their ability to correctly identify misinformation.

Systematic errors due to the limits of human cognition are called "cognitive biases". There exist different types of biases: cognitive biases, conflicts of interest, statistical biases, and prejudices. Cognitive biases must be focused on because they are systematic biases due to limits in human cognition that can unintentionally affect the effectiveness of fact-checking processes. From a psychological point of view, evolutionary studies suggest that humans developed behavioral biases to minimize the cost of making mistakes in the long period, as they can improve decision-making processes [198]. According to error management studies that aim at explaining human processes in decision-making, cognitive biases, defined as "the ones that skew our assessments away from an objective perception of information" [198], have been favored by nature in order to minimize whichever error that caused a great cost [179, 302, 303]. In fact, decision-making processes are often complex, and we are not always capable of keeping up to date – and statistically correct – the estimations of the error probabilities involved in such processes; thus, natural selection might have favored cognitive biases to simplify the overall decision process [87, 413]. To summarize, cognitive biases evolved because of the intrinsic limitations of humans when making a decision. Cognitive biases play a major role in the way (mis)information and verified content are consumed, and different debiasing strategies have been proposed in relation to cognitive factors such as people's memory for misinformation [243].

It is important to remark that biases can have far-fetched consequences. Keeping the focus on fact-checking, machine learning approaches are an interesting potential solution to address the obvious scalability issues of the approach based on human experts [76, 261, 436, 444]. In this respect, biases not only interfere with the human fact-checking activity in practice, but they also create issues for automatic approaches as they creep into the datasets that are then used to train the machine learning systems, in some cases leading to blatant errors, such as the famous "Gorilla case" [450]. Moreover, such biases might affect the accuracy (or even question the feasibility) of human-in-the-loop hybrid systems that try to identify misinformation at scale by combining experts, crowd, and automatic machine learning systems [99]. Since biases introduce errors due to systematic limits in human cognition that are potentially shared among several individuals, unveiling these systematic biases would support a more reliable collection of crowdsourced training data and enable bias mitigation methods for existing data sets.

1.4 Automated Fact-Checking

Automated fact-checking (AFC) uses natural language processing (NLP) techniques to determine the truthfulness of a claim. The problem is defined in the following way: given a statement (claim) and some evidence, determine whether the statement is true with respect to the evidence [396]. The automated approaches previously hinted at can fall under this category. This is a challenging task for a human, let alone an autonomous system [162]. However, AFC systems can approximate this process of evidence retrieval and synthesis with some degree of success [396, 430]. The benefits and applications of an AFC system are numerous. Indeed, they are starting to become a critical tool in combating the sheer quantity of claims that need to be verified.

AFC systems have been unable to supplement traditional fact-checkers due to a limitation in their design, even though they are accurate [339, 396]. A user may not accept to believe in a statement without first understanding the concepts and facts underpinning that statement. Such justifications are expected when reading journalistic fact-checking outcomes such as on PolitiFact. The fact-check outcome is accompanied by an explanation informing the reader of how the decision was reached. Without providing users with an explanation, the decision provided by an automated system is far less likely to be trusted [414], especially as it is not generated by humans. Automated systems have recently been developed to this effect, and have demonstrated promising initial results [162]. While these initial results are unquestionably impressive, critical evaluation of the work reveals that many of these systems use separate models for veracity prediction and explanation generation. It can be argued that systems such as these are not describing their own actions and decision processes and that the truthfulness prediction model is not made any more transparent.

1.5 The Crowdsourcing Activity Workflow

In recent years, crowdsourcing has become a popular method for collecting human work on a large scale. Typically, platforms host the tasks to be performed. These tasks are then allocated to crowd workers in a first-come, first-served approach. Several crowdsourcing platforms such as Amazon Mechanical Turk emerged to support the ever-increasing need for crowd-powered data gathered using such outsourced tasks. These platforms aim to help human task requesters to outsource their tasks to a diverse, distributed, and large workforce able to perform each task. Indeed, although an increasing amount of published studies show the usage of Amazon Mechanical Turk [211], alternatives are also used [322]. Popular alternatives include Prolific, a platform dedicated to the scientific community where the crowd workers are explicitly recruited for participation in research tasks [318], or Toloka, a platform mainly focused on data labelling tasks.

A task requester is an individual who wants to deploy a crowdsourcing task on a chosen platform. The workflow involves several phases. Initially, the requester sets some parameters such as the number of crowd workers required, the time allowed for each worker to perform the task, and so on. Then, they design the task layout. Usually, a markup language is provided to help to build the user interface. The requester must write

the logic to handle the task using client-side programming. Once the project is finalized the task can be deployed an arbitrary amount of times using a support file to vary the input data. Each instance of a task assigned to a worker is usually called HIT (Human Intelligence Task). Paolacci et al. [319] define a set of HITs as a batch. When a worker completes the HITs assigned, the requester can approve and thus pay the worker or reject the final submission. This workflow shows several difficulties across all platforms: the requester must have advanced programming skills; the user interface is built by mixing the presentation and business logic; the input data passing mechanism is often cumbersome, and the responsibility to store the data produced by each worker lies on the requester for non-trivial experiments.

Moreover, the requesters sometimes need to run studies that require a specific worker to perform new chunks of work over multiple days, weeks, or months: namely longitudinal studies (LS). Longitudinal studies aim at observing changes that may occur with respect to a chosen subject over a given or extended period. Running longitudinal studies on crowdsourcing platforms has become popular. Litman et al. [257], for instance, presents a tool for longitudinal study functions on the top of Amazon Mechanical Turk, mainly because of the convenience and easy means that crowdsourcing platforms provide to access potential crowd workers. Little is currently understood about how crowd workers perceive longitudinal studies, despite the growing popularity of turning to the crowd as opposed to carrying out lab studies [151].

1.6 Meta-Research Questions

A possible solution to the misinformation spreading problem is to rely on machine learning systems to detect and judge the truthfulness level of information. However, they require a great amount of data for the training phase and their reliability, effectiveness, and explainability are often not adequate. Therefore, to address such issues, one may think of relying on the large number of non-expert people that consume information and ask them to perform the fact-checking activity. Indeed, it is possible to use crowdsourcing [189] based approaches to collect truthfulness labels provided by non-expert people on statements. Such a decision leads to several opportunities but also difficulties. While it is possible to collect a large amount of data in a considerably short time, there is not any guarantee of the quality of the data collected. A long-term approach could be building a human-in-the-loop system [99] to cope with (mis)information by measuring truthfulness in real-time (e.g., as they appear on some social media using crowd-powered data, human intelligence, and machine learning techniques).

This thesis focuses mostly on crowdsourcing-based approaches that target the misinformation problem, but a machine learning-based approach is also proposed. There is still much work to do before achieving the long-term goal of building a human-in-the-loop system, yet this thesis represents a step towards the design and development of systems that are robust, trustworthy, explainable, and transparent. In more detail, three (3) meta-research questions (detailed below) are proposed and further expanded into thirty one (31) different research questions. Each chapter focuses on a subset of specific research questions.

- MRQ1 Are human assessors able to detect and objectively categorize online (mis)information? Can their judgment be compared and related to those of expert people? Which is the environment that allows obtaining the best results when judging information truthfulness? Can a multidimensional notion of truthfulness be defined?
- MRQ2 What is the impact of cognitive biases on human assessors while judging information truthfulness? Is it possible to detect this kind of bias? Are there countermeasures to combat their effects? Is it possible to define a bias-aware judgment pipeline for fact-checking?
- MRQ3 Can the truthfulness judgments collected be leveraged using machine learning-based approaches? Can an approach able to predict information truthfulness and, at the same time, generate a natural language explanation supporting the prediction itself be designed? Are machine-generated explanations useful for human assessors to better judge the truthfulness of information items?

1.7 Terminology

This thesis uses a set of nouns and technical terms belonging to the field of crowdsourcing. A definition of all these concepts could be convenient for the reader to grasp and understand the remaining chapters. Part of the definitions reported and expanded in the following is originally proposed by Howe [189] and Paolacci et al. [319].

- *Crowdsourcing*: the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call.
- *Platforms*: marketplaces that allow individuals and businesses to outsource their processes and jobs to a distributed workforce who can perform these tasks virtually.
- *Human Intelligence Task (HIT)*: a single, self-contained, virtual work unit performed by an individual.
- *Element*: an item that a worker evaluates, uses, and addresses within a HIT. A Human Intelligence Task is composed of a set of elements.
- *Batch*: a set composed of multiple HITs published by a single individual.
- *Requester*: an employer who recruits employees (usually called *workers* or *participants*) from an online labor marketplace for the execution of HITs in exchange for a wage (usually called *reward*).
- *Worker*: an individual who joins a crowdsourcing platform to perform and complete HITs published by requesters.
- *Session*: a specific activity, composed of a set of actions performed by a worker.
- *Interval Between Sessions*: the time that elapses between the completion of a session and the beginning of the following one.
- *Session Duration*: time employed by a worker to complete a session.
- *Longitudinal Study (LS)*: a series of HITs from the same requester which is published regularly over time and requires the same workers to participate. A longitudinal study is made of a collection of subsequent sessions, with some temporal delay between them. We thus define two more terms specific to the LS:

- *Duration (of the LS)*: the length of time required to complete a longitudinal study, from the beginning of the first session to the completion of the last one, including all the intervals.
- *Frequency (of the LS)*: the number of sessions that a longitudinal study requires a worker to complete over a time span.

1.8 Synopsis

This thesis features eleven (11) chapters and seven (7) appendices. Chapter 2 recall several works related to the research questions addressed in the remaining chapters. The topics of the related works can be coarsely grouped into three main categories: fact-checking and information truthfulness judgments, the impact of (cognitive) bias, and automated approaches for fact-checking. The crowdsourcing activity is involved to various extents in many of the works considered. Chapter 3 describes the five (5) sources of data employed in the experiments described in the remaining chapters.

Chapter 4 focuses on collecting truthfulness judgments for publicly available fact-checked statements distributed over two datasets, using different judgment scales, where each scale has a different granularity level. Also, the political bias and the cognitive background of the workers are measured to quantify their effect on the reliability of the data provided. Chapter 5 addresses recent (mis)information about the COVID-19 pandemic. A crowd of workers judges publicly available fact-checked statements using a six-level truthfulness scale and a longitudinal study is performed by re-launching the crowdsourcing task multiple times with both novice and experienced workers, deriving important insights on how the behavior and quality change over time. Chapter 6 aims to understand the barriers to running longitudinal studies on crowdsourcing platforms by running a large-scale survey across multiple popular commercial platforms. Detailed quantitative and qualitative analyses are performed. A list of recommendations for researchers and practitioners who wish to conduct longitudinal studies is provided, together with a list of best practices for crowdsourcing platforms to better support such kinds of studies. Chapter 7 proposes a multidimensional notion of truthfulness. The crowd workers judge publicly available fact-checked statements using seven different dimensions of truthfulness selected based on the existing literature.

Chapter 8 presents a characterization of the cognitive biases which might manifest while fact-checking information items, performed using a PRISMA-inspired methodology. A list of countermeasures that can be adopted to limit the effect of the cognitive biases identified is presented, together with a bias-aware judgment pipeline. Chapter 9 investigates which systematic biases may decrease data quality for crowdsourced truthfulness judgments. An exploratory study on the previously collected truthfulness judgments for publicly available fact-checked statements is performed. The findings are used to formulate specific hypotheses which are then tested using a novel crowdsourcing experiment. Chapter 10 describes a machine learning-based architecture which can provide a truthfulness prediction for a statement and jointly generate a human-readable explanation for it. The architecture is competitive with state-of-the-art approaches. The architecture is calibrated and validated and an extensive human evaluation of the impact of generated explanations is conducted.

Chapter 11 summarizes the main contributions. Then, it outlines the practical implications, sketches the future work, and concludes the thesis.

Appendix A provides a detailed description of Crowd_Frame, the software system used to perform and support the crowdsourcing experiments. Appendix B reports the demographic questionnaire and the CRT tests used in several of the crowdsourcing experiments performed. Appendix C provides the list of statements employed to perform the crowdsourcing experiment and the related longitudinal study. Appendix D provides the survey employed to investigate the barriers to running longitudinal tasks on crowdsourcing platforms. Appendix E shows the instructions for the crowdsourcing experiment performed to study the multiple dimensions of truthfulness. Appendix F provides the PRISMA checklists used to characterize the cognitive biases that can manifest while performing the fact-checking process and the full list of cognitive biases found in the literature. Appendix G presents additional questionnaires used to evaluate the impact of a subset of cognitive biases.

1.9 Publications

This thesis is based on 13 articles. Ten out of thirteen (10/13) articles are already published. Three out of thirteen articles (3/13) are under review. Each work sets the basis for a given chapter or section. The following list shows the articles sorted according to their chronological order.

1. Chapter 4: Roitero, Kevin and **Soprano, Michael** and Fan, Shaoyang and Spina, Damiano, and Mizzaro, Stefano and Demartini, Gianluca. (2020). Can The Crowd Identify Misinformation Objectively? The Effects of Judgment Scale and Assessor's Background. In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2020)*. Pages: 439–448. Xi'an, China (Virtual Event). July 25-30, 2020. Conference Ranks: GGS A++, Core A*. DOI: 10.1145/3397271.3401112. Reference Number: [361]. Status: Published.
2. Chapter 5: Roitero, Kevin and **Soprano, Michael** and Portelli, Beatrice and De Luise, Massimiliano and Spina, Damiano and Mea, Vincenzo Della and Serra, Giuseppe and Mizzaro, Stefano and Demartini, Gianluca (2021). Can The Crowd Judge Truthfulness? A Longitudinal Study On Recent Misinformation About COVID-19. In: *Personal and Ubiquitous Computing*. ISSN: 1617-4917. Journal Ranks: Journal Citation Reports (JCR) Q2 (2020), Scimago (SJR) Q1 (2021). DOI: 10.1007/s00779-021-01604-6. Reference Number: [362]. Status: Published.
 - Journal extension of: Roitero, Kevin and **Soprano, Michael** and Portelli, Beatrice and Spina, Damiano and Della Mea, Vincenzo and Serra, Giuseppe and Mizzaro, Stefano and Demartini, Gianluca. (2020). The COVID-19 Infodemic: Can the Crowd Judge Recent Misinformation Objectively? In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM2020)*. Pages: 1305–1314. Galway, Ireland (Virtual Event). October 19-23, 2020. Conference Ranks: GGS A+, Core A. DOI: 10.1145/3340531.3412048. Reference Number: [363]. Status: Published.

3. Chapter 7: **Soprano, Michael** and Roitero, Kevin and La Barbera, David and Ceolin, Davide and Spina, Damiano and Mizzaro, Stefano and Demartini, Gianluca (2021). The Many Dimensions of Truthfulness: Crowdsourcing Misinformation Assessments on a Multidimensional Scale. In: *Information Processing & Management*, 58(6). Journal Ranks: Journal Citation Reports (JCR) Q1 (2021), Scimago (SJR) Q1 (2021). DOI: 10.1016/j.ipm.2021.102710. Reference Number: [395]. Status: Published.
4. Appendix A: **Soprano, Michael** and Roitero, Kevin and Bombassei De Bona, Francesco and Mizzaro, Stefano (2022). Crowd Frame: A Simple and Complete Framework to Deploy Complex Crowdsourcing Tasks Off-the-Shelf. In: *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining (WSDM '22)*. Pages: 1605–1608. Virtual Event, Arizona, USA. ISBN: 9781450391320. Conference Ranks: GGS A+, Core A*. DOI: 10.1145/3488560.3502182. Reference Number: [393]. Status: Published.
5. Chapter 9: Draws, Tim and La Barbera, David and **Soprano, Michael** and Roitero, Kevin and Ceolin, Davide and Checco, Alessandro and Mizzaro, Stefano (2022). *The Effects of Crowd Worker Biases in Fact-Checking Tasks*. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. Pages: 2114–2124. DOI: 10.1145/3531146.3534629. Reference Number: [111]. Status: Published.
6. Chapter 10: Brand, Erik and Roitero, Kevin and **Soprano, Michael** and Rahimi, Afshin and Demartini, Gianluca (2022). A Neural Model to Jointly Predict and Explain Truthfulness of Statements. *ACM Journal of Data and Information Quality*, 58(6). Journal Ranks: Journal Citation Reports (JCR) Q3 (2021), Scimago (SJR) Q2 (2021). DOI: 10.1145/3546917. Reference Number: [48]. Status: Published.
 - Journal extension of: Brand, Erik and Roitero, Kevin and **Soprano, Michael** and Demartini, Gianluca (2021). E-BART: Jointly Predicting and Explaining Truthfulness. In: Augenstein, Isabelle and Papotti, Paolo and Wright, Dustin (Eds.) *Proceedings of the 2021 Truth and Trust Online Conference (TTO 2021)*. Virtual Event. October 7-8, 2021 (pp. 18–27). Url: https://truthandtrustonline.com/wp-content/uploads/2021/10/TTO2021_paper_16-1.pdf. Reference Number: [47]. Status: Published.
7. Section 11.2.4: Ceolin, Davide and Primiero, Giuseppe and **Soprano, Michael** and Wielemaker, Jan. Transparent Assessment of Information Quality of Online Reviews Using Formal Argumentation Theory. In: *Information Systems* (2022). ISSN: 0306-4379. Journal Ranks: Journal Citation Reports (JCR) Q2 (2021), Scimago (SJR) Q1 (2021). DOI: 10.1016/j.is.2022.102107. Reference Number: [61]. Status: Published.
 - Journal extension of: Ceolin, Davide and Primiero, Giuseppe and **Soprano, Michael** and Wielemaker, Jan. Assessing the Quality of Online Reviews Using Formal Argumentation Theory. In Brambilla, Marco and Chbeir, Richard and Frasinca, Flavius and Manolescu, Ioana (Ed.), *20th International Conference on Web Engineering*. Pages: 71–87. Springer International Publishing. Conference Ranks (2021): GGS B-, Core B. DOI: 10.1007/978-3-030-74296-6_6. Reference Number: [62]. Status: Published.

8. Chapter 8: Roitero, Kevin and **Soprano, Michael** and Ceolin, Davide and La Barbera, David and Spina, Damiano and Demartini, Gianluca and Mizzaro, Stefano (2023). Characterizing Cognitive Biases in Fact-Checking and Its Countermeasures. In: *Information Processing & Management*. Journal Ranks: Journal Citation Reports (JCR) Q1 (2021), Scimago (SJR) Q1 (2021). Reference Number: [360]. Status: under review.
9. Section 11.4.4: Roitero, Kevin and La Barbera, David and Soprano, Michael and Demartini, Gianluca and Mizzaro, Stefano and Sakai, Tetsuya (2023). How Many Assessors Do I Need? On Statistical Power When Crowdsourcing Relevance Judgments. In: *ACM Transactions on Information Systems*. Journal Rank: Scimago Q1 (2021). Reference Number: [357]. Status: Under Review.
10. Chapter 6: **Soprano, Michael** and Roitero, Kevin and Gadiraju, Ujwal and Maddalena, Eddy and Demartini, Gianluca (2023). Longitudinal Loyalty: Understanding the Barriers to Running Longitudinal Studies on Crowdsourcing Platforms. In: *Information Processing & Management*. Journal Ranks: Journal Citation Reports (JCR) Q1 (2021), Scimago (SJR) Q1 (2021). Reference Number: [394]. Status: Under Review.

Related Work

2.1 Fact-Checking Using Crowdsourcing-Based Approaches

The research community has been looking at automatic check-worthiness predictions. Gencheva et al. [154] create a corpus of political debates containing fact-checked claims to train machine learning models and predict which claims should be prioritized for fact-checking. Related to this, Vasileva et al. [426] propose a deep learning-based approach for estimating such check-worthiness. Atanasova et al. [19] propose a different model to detect check-worthy claims and fact-checks them using a neural network. Other researchers focus on describing the truthfulness of information items. Kim et al. [215] model true and false news spread in social networks by considering their topic. Vo et al. [431] develop a recommender system that allows users to correct misinformation by referring to fact-checking URLs. Later, they propose a machine learning model to perform a fact-checking URL recommendation task You et al. [463].

Other researchers focus on the credibility and trust of sources of information. Epstein et al. [132] conduct a survey experiment with about 1,000 Americans to understand their perceived trust in numerous news sites. Their results show that participants tend to trust mainstream sources more than hyper-partisan or fake news sources. Bhuiyan et al. [40] collect credibility annotations on the topic of climate change from both crowd workers and students with journalism or media programs. They study and compare the two sets of annotations against expert-provided ones.

The researchers also look at how to use crowdsourcing to collect reliable truthfulness labels in order to scale up and help study the manual fact-checking effort [98, 99, 332, 429]. For example, Kriplean et al. [223] analyze volunteer crowdsourcing when applied to fact-checking. Zubiaga et al. [474] investigate using crowdsourcing the reliability of tweets in the setting of disaster management. Their results show that it is difficult for crowd workers to properly assess information truthfulness, but also that the source reliability is a good indicator of trustworthy information. Related to this, the CLEF initiative develop a Fact-Checking Lab [32, 129, 299, 300, 301] to address the issue of ranking sentences according to some fact-checking property. The SemEval-2019 Task 8 [285] requires providing

truthfulness labels for factual information needs. Maddalena et al. [266] look at assessing news quality along eight different quality dimensions using crowdsourcing. Pennycook et al. [326] crowdsourced news source quality labels. Giachanou et al. [158] introduce a tutorial on online harmful information that includes social media and fake news. Ghenai et al. [155] use crowdsourcing and machine learning to track misinformation on Twitter.

Fact-checking websites collect a large number of high-quality labels generated by experts. However, each fact-checking site and dataset defines its own labels and rating system used to describe the truthfulness of the content. Therefore converging to a common rating scale becomes very important to integrate multiple datasets.

Vlachos et al. [430] align labels from Channel 4 and PolitiFact to a five-level scale: False, Mostly-False, Half-True, Mostly-True, and True. Nakov et al. [299] retrieve evaluations of different articles at `factcheck.org` to assess claims made in American political debates. Then, they generate labels on a three-level scale: False, Half-True, and True. Vosoughi et al. [432] check the consistency between multiple fact-checking websites on three levels: False, Mixed, and True. Tchechmedjiev et al. [407] look at rating distributions over different datasets and define a standardized rating scheme using four basic categories: False, Mixed, True, and Other.

Samples of statements from the PolitiFact dataset, originally published by Wang [436], are used to analyze the agreement of workers with labels provided by experts in the dataset itself. Workers are asked to provide the truthfulness of the selected statements using different fine-grained rating scales. Roitero et al. [356] compare two fine-grained scales: one in the $[0, 100]$ range and one in the $(0, +\infty)$ range, on the basis of Magnitude Estimation [292]. They find that both scales allow the collection of reliable truthfulness judgments that are in agreement with the ground truth. Furthermore, they show that the scale with one hundred levels leads to slightly higher agreement levels with the expert judgments. La Barbera et al. [232] ask workers to use the original scale proposed by the PolitiFact experts and the scale in the $[0, 100]$ range on a larger sample of PolitiFact statements. They find that aggregated judgments have a high level of agreement with expert judgments. They also find evidence of differences in the way workers provide judgments, influenced by the sources they examine. In more detail, La Barbera et al. [232] find that the majority of workers use indeed the PolitiFact website to provide judgments. These works allow concluding that different datasets use different scales and that meta-analyses try to merge scales and aggregate ratings together. While no clear preferred scale has yet emerged, there seems to be a preference towards coarse-grained scales with just a few (e.g., from three to six) levels as they may be more user-friendly when labels need to be interpreted by information consumers.

This thesis, as compared to previous work, analyzes in Chapter 4 the impact of assessors' background and judgment scales on the quality of the truthfulness judgments they provide. The quality judgments are collected using a six-level scale with judgments collected using a more coarse-grained scale (i.e., three levels) and a more fine-grained scale (i.e., a hundred levels). Chapter 5 studies the effect of information recency by addressing recent COVID-19 related information items. Furthermore, Chapter 7 investigates the effect of asking crowd workers to judge truthfulness along multiple dimensions and observes if doing so has an impact on the quality of the judgments collected.

2.2 The Effect of Information Recency

The number of initiatives that apply for Information Access and, more generally, Artificial Intelligence techniques to combat the COVID-19 infodemic has been rapidly increasing (see Bullock et al. [52, p. 16] for a survey). There is a significant effort by researchers like Cinelli et al. [77], Gallotti et al. [153] and Yang et al. [458] on analyzing COVID-19 information on social media and linking to data from external fact-checking organizations to quantify the spread of misinformation. Mejova et al. [281] analyze Facebook advertisements related to COVID-19, and find that around 5% of them contain errors or misinformation. Desai et al. [101] use a crowdsourcing-based methodology to collect and analyze data from patients with cancer who are affected by the COVID-19 pandemic. Mendes et al. [284] propose a system based on NLP methods for human-in-the-loop fact-checking of tweets in the domain of COVID-19 treatments. The 2021 edition of the CheckThat! laboratory [301] at the CLEF initiative focused on COVID-19 related statements.

This thesis, as compared to previous work, studies the effect of recent information about the COVID-19 infodemic on truthfulness judgments collected using crowdsourcing (Chapter 5). It investigates whether the health domain makes a difference in the ability of crowd workers to identify and correctly classify (mis)information. It focuses on a single truthfulness scale, given the evidence that the scale used does not make a significant difference (Chapter 4). Moreover, the experiments described involved asking workers to provide a textual justification for their judgments. The justifications are analyzed to better understand the process followed by workers to verify the information and if they can be exploited to derive useful information. Lastly, a longitudinal study that includes 3 crowdsourcing experiments is performed over a period of 4 months. It allows the collection of additional data and evidence that include new responses from new and old crowd workers.

2.3 Crowdsourcing-Based Longitudinal Studies

The research community studies the crowd worker experience, including current barriers to a good experience as well as tools and methods aiming at improving it. Several works focus on the crowd workers' needs and experience on micro-task platforms. For example, Wu et al. [454] looks at the impact of task design choices on workers' experience and performance. Irani et al. [194] and Williams et al. [449] look at the impact of the use of tools to support crowd work showing how they introduce task switching and multi-tasking while improving productivity. Another way to improve crowd work experience is using coaching by fellow workers, as described by Chiang et al. [73]. Self-organization may also help to obtain stronger negotiation power with platforms and requesters [368]. Hara et al. [177] took a quantitative approach to analyze earnings on crowdsourcing platforms showing how underpaid crowd workers are on average. Related to their experience and earnings, Toxtli et al. [416] analyzed the time spent by workers on non-rewarded activities, which further decrease hourly wages.

The original definition of longitudinal study has been proposed in the past by researchers in the fields of psychology and medicine. Bauer [34] described the types of longitudinal

designs along with practical considerations on how to conduct them. Ployhart et al. [335] propose an answer to a list of 12 questions that typically researchers must address when designing and conducting longitudinal studies. More recently, researchers have run longitudinal studies on crowdsourcing platforms. Fan et al. [139] deploys a crowdsourcing task multiple times inviting the same group of participating workers each day for 20 days observing a sharp decline in return rates over time. Strickland et al. [400] run a study on alcohol use involving a first task taking 21 minutes to complete followed by regular 2-minute follow-up tasks. They run a study with a weekly survey over 18 weeks. High response rates (64.1%-86.8%) were observed across the 18 weeks demonstrating the feasibility of longitudinal studies on crowdsourcing platforms. Active participation was incentivized by entry into a raffle for one of five \$50 bonuses if participants completed 14 or more weekly surveys. Wang et al. [437] present a game with a purpose to be used over two weeks; in their longitudinal studies they observe how individuals are subject to hedonic and social factors during early stages of use, and hedonic, social, and usability-related factors during later stages. Daly et al. [95] present a study that examines a two-month re-response rate (study 1, $n=752$; 75%) of a US Amazon Mechanical Turk sample. A second study ($n=373$) investigates the four- and eight-month re-response rate (56% and 38%, respectively) of a US immigrant sample. A third study examines the thirteen-month re-response rate (47%). These involved a 23-minute task. Hata et al. [182] look at longitudinal crowdsourcing platform data to observe how work quality is stable over time for the same worker and conclude it is possible to predict long-term work quality after the first five tasks. Auer et al. [22] compare traditional work to crowd work in terms of the effect of performance payment. They observe that while performance is not different, it is important for the experimenter to reward crowd participants ethically given their lack of power to negotiate pay. Strickland et al. [401] overview the use of Amazon Mechanical Turk to run longitudinal studies for addiction science. They show how the number of papers making use of this platform to recruit participants has increased fourfold from 2014 to 2017. Leung et al. [242] run a survey with 1000 Amazon Mechanical Turk workers to understand the triggers to continued participation. They find out that the two key factors are external regulation (i.e., monetary rewards) and workers' intrinsic motivation.

Retention rates vary significantly across longitudinal studies. Different studies use different reward schemes and incentives to increase retention rates. Holden et al. [187] run a study over three weeks observing a 69% retention rate after three weeks from the initial task. Buhrmester et al. [51] run a study over three weeks observing a 60% retention rate after three weeks from the initial task. Shapiro et al. [379] have 80% rate over one week. Lanaj et al. [233] observe a response rate of 61% for surveys completed over 10 consecutive workdays. Strategies observed in the literature to increase retention and participation in such types of studies are mainly based on payment schemes. A common approach is to incentivize participant retention using extra payments. Auer et al. [22] shows that pay has a significant effect on attrition (i.e., single task abandonment) but not on retention in the second wave of experiments in longitudinal studies. Difallah et al. [107] show that using a bonus to reach a milestone is the most effective to retain workers up to the pre-defined milestone within a continuous series of tasks with no interruptions.

This thesis, as compared with previous work, addresses in Chapter 6 the crowd worker experience with a particular focus on longitudinal studies that require workers to commit for

a considerable amount of time to the same task and requester as compared to standard one-off micro-tasks. The findings complement that of previous work by providing guidelines and recommendations for task designers and crowdsourcing requesters on how to best design longitudinal studies tasks and how to best engage workers in longitudinal studies.

2.4 The Multidimensionality Of Truthfulness

The research community looks at how assessors perform judgments when using multiple dimensions and at comparing experts and non-experts. Multidimensional scales proved to be effective in the setting of information retrieval when dealing with relevance. Barry et al. [33] and Xu et al. [457] list the different relevance criteria used to perform relevance evaluation. Zhang et al. [468] extend the psychometric framework for multidimensional relevance proposed by Zuccon et al. [476] by using crowdsourcing, detailing its limitations, and describing various quality control methods derived from psychometrics which can be applied to the information retrieval context. Jiang et al. [196] collect multidimensional relevance along with contextual feedback from users and correlate their judgments with user metrics. Furthermore, they investigate two variants of TREC-style relevance judgments used in information retrieval. They also study contextual judgments and collect multidimensional judgments using four different dimensions. Uprety et al. [423] define multidimensional relevance using a quantum-inspired structure. It seems natural to try and apply the same approach to truthfulness judgments, given the amount of research done and the demonstrated effectiveness of multidimensional relevance judgments. There is indeed some preliminary work in this direction. Ceolin et al. [60] collect multidimensional truthfulness judgments on web documents dealing with vaccines, where few experts provided the assessments. Their results showed that experts manifest a high level of agreement, but also that the task is very demanding, and that the availability of experts online is rather limited. Maddalena et al. [266] extend the work by Ceolin et al. [60] by comparing crowd and expert truthfulness assessment for a small dataset of 20 selected documents dealing with vaccines. Results show that experts are inclined to use lower values than crowd workers (i.e., they are more critical) and that the agreement between crowd and experts is high, but not total.

This thesis, as compared to previous work, describes the collection of a large number of truthfulness judgments using a multidimensional scale inspired by the literature, thus making it available to the research community (Chapter 7).

2.5 Bias, Echo Chambers, And Filter Bubbles In User Generated Data

The activities of the fact-checking process are driven by humans, both when explicit human judgments are used and when human-labelled data are used to train machine-learning models. Human bias is often reflected in manually labelled datasets and therefore in supervised systems that make use of such data. Thus, fact-checking is prone to suffer from biases of various kinds, including cognitive biases. According to the literature, more than

200 cognitive biases exist [56, 180, 186, 205]. Standard conceptualization or classification of such biases is a debated problem [159, 186], yet many works confirmed their presence in many domains using reproducible studies [409], for example in information seeking and retrieval [24]. The effect of cognitive biases has been widely studied in multiple disciplines. Ehrlinger et al. [125] and several other researchers [28, 96, 186, 402] study the effect of cognitive biases in decision processes and planning. Fisher et al. [145] focus on market forecasting.

Recent work presents surveys on potential effects derived from biases on the web in general [26] and search and recommendation systems [70, 25]. For example, Otterbacher et al. [315] show that human bias and stereotypes are reflected in search engine results, while Yue et al. [464] investigate presentation bias in click-through data generated by a search engine. Kiesel et al. [214] study biases related to presentation format using conversational interfaces in the context of systems for argument search. Furthermore, several works investigate the role of commonly occurring cognitive biases in a web search on debated topics [113, 131, 336, 349, 447].

Many researchers focus specifically on issues related to bias management in user-generated and crowdsourced data. Love [263] studies different user biases in peer assessment methods. Chandar et al. [65] estimate click-through bias in cascade models for information retrieval. Eickhoff [126] and shows in the context of crowdsourced relevance judgments how common types of bias can impact the collected judgments and the results of information retrieval evaluation initiatives. Yildirim et al. [461] and Lee [238] study bias in user-generated data dealing with news media, while Muchnik et al. [295] focused on social influence bias. Sylvia Chou et al. [403] study the role of cognitive biases in social media platforms. Hube et al. [192] analyze the effect of workers' opinions in subjective tasks. There is also evidence of differences in the way workers provide judgments, influenced by the impact of worker bias. La Barbera et al. [232] find that political background has an impact on how workers provide truthfulness judgments. In more detail, workers are more tolerant and moderate when judging statements from their very own political party. Draws et al. [112] create a checklist to cope with common cognitive biases. Pennycook et al. [326, 327, 328] evaluate the ability of humans in identifying true and false news and find a positive with cognitive skills usually measured using cognitive reflection tests [148]. Lim et al. [252] propose a news bias dataset to facilitate the development and evaluation of approaches for debiasing news articles. The biases present in datasets made using user-generated data may also impact machine learning models [53].

Other researchers study the role of specific cognitive biases in relation to the misinformation topic. Zollo [473] study how information spread across communities on Facebook, focusing on echo chambers and confirmation bias. Wesslen et al. [446] consider the role of visual anchors in decision-making processing related to Twitter misinformation. Karduni et al. [208] focus on uncertainty on truthfulness assessment when using visual analysis, Acerbi [2] consider a cognitive attraction phenomenon in online misinformation. Traberg et al. [417] study perceived source credibility to mitigate the effect of political bias. Zhou et al. [472] consider confirmation bias on misinformation related to the topic of climate change. The way information spreads through social media and, in general, the Web has been widely studied, leading to the discovery of a number of phenomena that were not so evident in the pre-Web world. Among those, echo chambers and epistemic bubbles seem

to be central concepts [305, 321]. Eady et al. [122] investigate the extent of ideological echo chambers on social media using well-known media organizations and political actors as anchors.

Flaxman et al. [146] examine the browsing history of US-based users who read news articles. They find that both search engines and social networks increase the ideological distance between individuals and that they increase the exposure of the user to the material of opposing political views. These effects can be exploited to spread misinformation. Törnberg [422] models how echo chambers contribute to the virality of misinformation, by providing an initial environment¹ in which misinformation is propagated up to some level that makes it easier to expand outside the echo chamber. This helps to explain why clusters, usually known to restrain the diffusion of information, become central enablers of spread. On the other side, acting against misinformation seems not to be an easy task, at least due to the backfire effect. It is the effect for which someone's belief hardens when confronted with evidence opposite to its opinion. Sethi et al. [374] study the backfire effect and presented a collaborative framework aimed at fighting it by making the user understand her/his emotions and biases. However, the paper does not discuss the ways techniques for recognizing misinformation can be effectively translated into actions for fighting it in practice.

This thesis, as compared with previous work, described in Chapter 4 the collection of assessors' background and bias data to then identify patterns in their judgment behaviors. Chapter 8 presents a systematic review of the whole set of cognitive biases that may manifest while performing a fact-checking activity. Chapter 9 investigates which systematic biases may decrease data quality for crowdsourced truthfulness judgments.

2.6 Argument Mining For Fact-Checking

Dung [117] abstract argumentation framework emerged as a central formalism in formal argumentation. Throughout the years, it has been extended by the research community and different families of argumentations frameworks exist [31]. Such families include Preferential Argumentation Frameworks [12, 13, 287] and Value-based Argumentation Frameworks [37, 38]. Dunne et al. [119] propose a specific approach represented by systems defining preferences based on weighted attacks, establishing that some inconsistencies are tolerated in the set of arguments, provided that the sum of the weights of attacks does not exceed a given value. Weights can be used to provide a total order of attacks [272]. This approach can be generalized in several ways. Coste-Marquis et al. [89, 88] present a different approach for relaxing the admissibility condition and strengthening the notion of defence. Furthermore, they propose different selections of extensions based on the order of weights.

Truthfulness classification and the fact-checking activity are strongly related to the scrutiny of factual information extensively studied in argumentation theory [21, 236, 375, 390, 415, 429]. Argument mining, which is the automatic identification and extraction of the structure of inference and reasoning expressed as arguments presented in natural language, is also related. Lawrence et al. [236] survey the techniques used for argument mining and detail how crowdsourcing-based approaches can be used to overcome the limitations of manual analysis. Sethi [375] proposes a prototype social argumentation framework to curb

the propagation of fake news where the argumentation structure is crowdsourced and reviewed/moderated by a set of experts in a virtual community. Sethi et al. [377] develop a recommender system that makes use of argumentation and pedagogical agents to fight misinformation. Snaith et al. [390] presents a platform based on a modular architecture and distributed open source for argumentation and dialogue. Visser et al. [429] shows how to use argument mining to increase the skills of workers that assess media reports.

This thesis, as compared with previous work, studies in Chapter 7 the usage of an argumentation framework to leverage the quality of crowdsourced items, e.g., by providing the crowd workers with some tools to better assess the argument structure of statements.

2.7 Automated Fact-Checking Using Machine Learning Techniques

The research community investigate the usage of machine learning techniques to cope with disinformation besides human-powered systems [181, 411]. These techniques rely on training a machine learning algorithm on a labelled dataset which is usually built using human assessors. Vlachos et al. [430] define the setting and the challenges needed to create a benchmark dataset for fact-checking. Ferreira et al. [140] describe a dataset for stance classification. Wang [436] creates the LIAR dataset which contains a large collection of fact-checked statements. Several works focus on the algorithms which can be employed to build a fully automatic methodology to fact-check information. Weiss et al. [444] develop a method based on adversarial networks. Alhindi et al. [8] leverage justification modeling. Reis et al. [346] and Wu et al. [452] discuss explainable machine learning algorithms that can be employed for fake news detection. Oeldorf-Hirsch et al. [312] and Evans et al. [136] consider information sources and their metadata.

Automated fact-checking aims at replacing experts, i.e., usually journalists, in performing the fact-checking process. As an example of such methods, Liu et al. [261] propose a deep neural network model to detect misinformation statements. Their model is based on a feature extractor which works both at the textual and at the user level, an attention layer used to detect important and specific user responses, and a pooling algorithm to do feature aggregation. Their results on two datasets show that the developed model reaches an accuracy level higher than 0.9 within 5 minutes of the spread of the misinformation statement. Lim et al. [252] use crowdsourcing to gather bias labels on news articles and propose an automatic approach for analyzing and detecting them. Li et al. [245] propose to identify possible misinformation on Twitter by learning a topic-based model from expert-provided assessment. However, fact-checking still requires manual effort, as evidenced by the approaches that exploit machine learning to build completely automatic classifiers. Such an effort is usually performed by expert fact-checkers to generate labels that can eventually lead to the training of supervised methods like the ones described.

The research community propose various techniques for generating explanations to accompany fact-checking decisions. Saliency-based methods, such as those proposed by Shu et al. [385] and Wu et al. [453], use attention mechanisms to highlight the input that is most useful in determining the veracity prediction and present this information to the end user as a form of explanation. Logic-based approaches make use of graphs [100], rule

mining, and probabilistic answer set programming [4] to output a series of logical rules that result in a veracity prediction. Summarisation techniques provide an explanation by summarising the evidence retrieved. Atanasova et al. [20] propose a system that uses DistilBERT [370] to pass contextual representations of the claim and evidence to two task-specific feed-forward networks which produce a classification and an extractive summary.

Stammbach et al. [396] proposes a framework that also produces abstractive explanations but places a higher emphasis on the evidence retrieval process. The framework consists of two components. These components are an evidence retrieval and veracity prediction module, and an explanation generation module. The first component is an enhanced version of the DOMLIN system [397], which uses separate BERT-based models for evidence retrieval and veracity prediction. GPT-3 [50], a large pretrained multi-purpose NLP model based on the Transformer, is used in “few-shots” mode to generate a summary of the evidence with respect to the claim provided as an explanation.

BART [244] is a transformer [427] model that aims to generalise the capabilities of both BERT [103] and GPT-style models. It consists of a bi-directional encoder, similar to BERT, as well as an auto-regressive decoder, similar to GPT. BART is pre-trained on a de-noising task whereby input text is corrupted and the model aims to reconstruct the original document, minimising the reconstruction loss. In contrast to existing de-noising models, BART is more flexible in that it is not trained to rectify a specific type of input corruption, but rather any arbitrarily corrupted document. The pre-trained BART model can be fine-tuned for a number of downstream tasks. Lewis et al. [244] note that BART performs comparably to other models, such as RoBERTa [262], on natural language inference tasks. They also note that BART outperforms current state-of-the-art models on natural language generation tasks, such as summarisation [244, 383]. Its ability to perform well on these two contrasting tasks made it an attractive choice as the base model for a system that can jointly predict the truthfulness of a claim (an inference task) and provide an explanation (a generative task).

This thesis, as compared with previous work, describes in Chapter 4 how assessors’ background and bias data are collected to then identify patterns in their assessment behaviours. The work described in Chapter 7 is complementary to those that require manual effort from expert fact-checkers to generate labels that can eventually lead to the training of supervised methods used to automate the fact-checking process. Furthermore, the work presented in Chapter 10 aims to support the activity performed by expert fact-checkers to generate labels usable by machine-automated fact-checking approaches. The approach presented differs from the existing literature as rather than using two separate models for the truthfulness prediction and explanation generation, a single model is used to output both a truthfulness prediction and an abstractive summarization.

2.8 Supporting Crowdsourcing-Based Approaches

Individuals and organizations who need to gather data of some kind using crowdsourcing-based approaches may rely on different crowdsourcing platforms. These platforms help task requesters access the global human workforce available. Amazon Mechanical Turk is one of the most well-known platforms. Paolacci et al. [319] presented, in the past, demographic data about the Mechanical Turk worker population, reviewing the strengths of

Mechanical Turk relative to other online and offline methods of recruiting subjects, and comparing the magnitude of effects obtained using Mechanical Turk and traditional worker pools. More recently, Mellis et al. [282] reviews different research conducted using Mechanical Turk, provides examples and discusses the limitations and best practices of the platform.

In the last years, researchers are arguing that the quality of data collected by recruiting workers from Mechanical Turk is declining. According to Peer et al. [322], already in 2017, Mechanical Turk workers were becoming less naive. In more detail, they define workers as “professional survey-takers”. Related to this, Chmielewski et al. [75] conducted before, during and after the summer of 2018 an experiment related to psychological research, finding empirical evidence of a substantial decrease in data quality. Kennedy et al. [213] show the presence of many fraudulent responders that provide low-quality data. Chandler et al. [66] explain that a substantial number of participants misrepresent relevant characteristics to meet the eligibility criteria expected in the studies. Webb et al. [442] explain that the data collected for their research were valid for the 2.6% of humans recruited and claim that a call for caution is needed while using Mechanical Turk. However, other crowdsourcing platforms seem to be a viable alternative. Peer et al. [323] compare 5 crowdsourcing platforms and panels by examining aspects of data quality for online behavioral research. They conduct two studies and only the platform [318] provides high data quality on all measures for both studies, while CloudResearch (formerly known as TurkPrime [257]) only for the second study. Litman et al. [255] analyze the claims of Peer et al. [323] and point out the presence of methodological decisions undisclosed in their work that limit the inference that can be drawn from the data collected. In more detail, they assert that Peer et al. [323] chose to turn off the recommended data quality filters while using the CloudResearch platform. They thus replicate the studies with CloudResearch using the recommended options, finding that the final data are of better quality. Litman et al. [255] point out in their work that Peer et al. [323] are members of Prolific . However, Litman et al. [255] clarify that they are part of the CloudResearch team. While other researchers suggest that using Prolific leads to some extent to results of better quality [405], Litman et al. [256] advocates that the platform should be selected by matching the study’s goals and the platform’s strengths and weaknesses.

Several tools that aid requesters during the whole crowdsourcing activity exist. Vukovic [433] proposes a taxonomy for the categorization of crowdsourcing platforms and evaluates a set of systems with respect to such taxonomy. Erickson [134] proposes a conceptual framework for the design of systems to support crowdsourcing and human computation. Liu [259] proposes a set of best practices to develop crowdsourcing systems designed to support the articulation work needed to facilitate spontaneous volunteer effort during emergencies. Clark et al. [79] develops a framework to allow governments to use crowdsourcing bases approaches to solve problems when interacting with their citizens. Brito et al. [49] propose a conceptual framework to guide the design of gamification-based approaches within crowdsourcing platforms. Li et al. [247] conceptualize a blockchain-based decentralized framework in which a requester can propose a task without relying on any third trusted institution.

Ye et al. [460] introduce a crowdsourcing framework to support the annotation of medical data sets. Hamrouni et al. [170] propose a framework for spatial mobile crowdsourcing,

where workers are required to be physically present at a particular location. The requesters solicit workers to provide photos of ongoing events for event reporting purposes. CloudResearch [257] is a research platform that integrates with Amazon Mechanical Turk that aims to improve the quality of the crowdsourcing data collection process. CrowdForge [221] is a general-purpose framework for accomplishing complex human computation tasks. It is based on the idea that complex work can be broken up into small and independent pieces while the system manages its coordination dependencies. iCrowd [138] is an adaptive crowdsourcing framework that estimates in real-time the accuracy of workers by evaluating their performances on completed tasks. It can be used before the task's launch to choose the best workers available. CrowdTruth 2.0 [116] is a method to aggregate crowdsourcing responses after the task using disagreement-aware metrics. It can be used to leverage workers' data processing after the task. CrowdEIM [382] is a tool based on mobile social media platforms which allow crowdsourcing information during emergencies.

Researchers proposed in the past approaches to model and predict user behaviour when interacting with web applications. Historically, these approaches relied heavily on Markov models, which have been widely popular for this type of task. Ching et al. [74] study the usage of such models to analyze categorical data sequences. The research community later focused on different challenges. Borges et al. [44] describe user web navigation sessions up to a given length. Shukla et al. [386] model the behaviour of users that chooses to switch the browser used to surf the web. Manavoglu et al. [271] aims to generate probabilistic browsing behaviour for users on the web. Benevenuto et al. [39] analyze user behaviour on social networks. Li et al. [248] integrate user behaviour into contextual advertising. Awad et al. [23], Deshpande et al. [102], and Dongshan et al. [109] aim predicting web page accesses. Anderson et al. [17] take advantage of user behaviour to personalize websites. Ju et al. [201] and Ren et al. [347] use behavioural data to build systems that can detect and prevent access to malicious users. More recent approaches involve the usage of behavioural data to represent user interaction using embeddings. Tran et al. [418], for instance, propose a recommendation model that learns user and item attributes represented using embeddings in the context of a recommender system. Learning vector representations for texts have been studied by the research community in depth. For example, Le et al. [237] and Mikolov et al. [286] analyze the usage of sentences as a better way to learn word semantics. Embedding can be helpful also when considering information retrieval tasks such as query rewriting. Grbovic et al. [164] propose a query rewriting method based on a query embedding algorithm. Other approaches include content advertising, as the one by Grbovic et al. [163]. They propose a neural language-based algorithm specifically tailored for delivering effective product recommendations. The embedding representations learnt can be used to predict and understand user behaviour across different scenarios, as done by Chen et al. [69] and Han et al. [171].

Crowdsourcing-based approaches can provide a massive amount of behavioural data due to the availability of a large human workforce. Robinson et al. [353] find out that there are more than 250.000 workers around the whole world whose potential is largely untapped. Stewart et al. [398] explains that a task requester can reach more than 7.000 workers each quarter year. Difallah et al. [105] analyze the population dynamics and demographics of Amazon Mechanical Turk workers based on the results of a survey that they conducted over 28 months. They also discover that during the day the peak of active US workers is 90%

around 11 PM UTC. Han et al. [172] perform a data-driven analysis of logs collected during a large-scale relevance judgment experiment to study the phenomenon of crowdsourcing task abandonment.

This thesis, as compared with previous work, describes in Appendix A, describes software that supports diverse crowdsourcing platforms and allows task requesters to recruit general-purpose crowd workers. Its configuration mechanism allows one to easily decompose a complex task into several intermediate steps without the need for coordination mechanisms. All the data produced during task performance are securely stored. Furthermore, it allows for gathering detailed behavioral data while a user performs a crowdsourcing task deployed on a platform that provides access to a marketplace of the human workforce. The data produced can be leveraged using one of the approaches proposed to learn and predict future user behaviour to improve the task's structure.

The next chapter provides a description of all the sources of data involved ABC Fact Check in the experiments described in the remaining chapters.

Dataset

The experiments described in this thesis use either different sets of elements from the same dataset for different experiments or multiple datasets within a single experiment. It is thus useful to describe all the datasets used before detailing each experimental setup. The PolitiFact dataset is presented in Section 3.1, while the ABC Fact Check one in Section 3.2. The FEVER dataset is introduced in Section 3.3, while Section 3.4 describes its extended version e-FEVER. Lastly, Section 3.5 shows the e-SNLI dataset.

3.1 Politifact

The PolitiFact dataset [436] is maintained and updated by the Poynter Institute¹ (Section 1.2). It is built and described as a “benchmark dataset for fake news detection”. It contains 21,340 statements produced by public appearances of US politicians, by other well-known people, and by social media posts. The publication dates of the statements range from 2007 to 2022. The PolitiFact website can be updated with special sections relative to events or ongoing crises. For instance, the organization added during 2020 a specific section related to the COVID-19 pandemic (which is still active today).² Also, a section devoted to the United States 2020 presidential elections was added.³ Table 3.1 shows a sample of two statements fact-checked by PolitiFact during 2022 and published on the organization’s website.

The statements are labeled by expert judges on a six-level truthfulness scale (referred in Chapter 5 also as E_6): *Pants-On-Fire*, *False*, *Mostly-False*, *Half-True*, *Mostly-True*, and *True*. The *Mostly-False* originally was *Barely-True*. PolitiFact chose to replace it in 2011⁴. Over the years, many readers complained about the emphasis that the old label put on the “true” component, when the actual judgment described something without much truth. The samples of the PolitiFact dataset used in this thesis and the results produced

¹<https://www.poynter.org/>

²<https://www.politifact.com/coronavirus/>

³<https://www.politifact.com/2020/>

⁴<https://www.politifact.com/article/2011/jul/27/-barely-true-mostly-false/>

may contain the Barely-True label instead of False, since statements published before 2011 were used when the experiments took place. However, the semantics behind the truthfulness level did not change. However, researchers and practitioners should be aware of that while exploring and using the outputs of this thesis.

Table 3.1: Statements fact-checked during 2022 by PolitiFact.

Statement	Speaker	Party	Date	Ground Truth
The United States spends “almost three times per capita what they spend in the U.K.” on health care and “50 percent more than they pay in France.”	Bernie Sanders	Democrat	2022-12-19	Half-True
“There hasn’t been a single of these mass shootings that have been purchased at a gun show or on the internet.”	Marco Rubio	Republican	2022-05-25	False

3.2 ABC Fact Check

The dataset ABC Fact Check⁵ consist of 561 verified statements covering the time span from 2013 to 2022. It is a partnership between RMIT University⁶ and the Australian Broadcasting Corporation⁷ that aims “combining academic excellence and the best of Australian journalism to inform the public through an independent non-partisan voice”. Professional fact-checkers seek expert opinions and collect evidence before a team makes a collective decision on how to label each claim (Section 1.2). A fine-graded truthfulness scale is used to such an end. The final verdicts can be labeled in various ways, such as: Correct, Checks Out, Misleading, Not The Full Story, Overstated, Wrong, and many others. These verdicts are then grouped and refined using a three-level scale: True, In-Between, and False. In the experiments described in this thesis, the three-level scale is used and considered as ground truth. Table 3.2 shows a sample of two statements fact-checked within the ABC Fact Check dataset during 2019 and published on the organization’s website.

Similarly to the case of the PolitiFact dataset, sometimes the False and True labels become Negative and Positive in the dataset. There is no clear explanation of the underlying rationale behind such choices. However, the semantics of these labels does not change. Again, researchers and practitioners should be aware of that while exploring and using the outputs of this thesis.

⁵<https://apo.org.au/collection/302996/rmit-abc-fact-check>

⁶<https://www.rmit.edu.au/>

⁷<https://www.abc.net.au/>

Table 3.2: Statements fact-checked during 2019 by ABC Fact Check.

Statement	Speaker	Party	Date	Ground Truth
“Labor has more than double the number of women the Liberals have in the Parliament and about twice the number of women on our front bench — that speaks for itself.”	Tanya Plibersek	Labor	2019-02-07	True
“Labor’s proposal is to dismantle offshore detention and will essentially give the ability for two doctors — as has been pointed out, doctors including Dr Brown, Bob Brown, and Dr Richard Di Natale — potentially can provide the advice”	Peter Dutton	Liberal	2019-02-12	In-Between

3.3 FEVER

The FEVER dataset⁸ [412] consists of 185,445 statements, associated evidence, and truthfulness judgments. The examples for the training are 165,447, while those for the development are 19,998. The statements have been generated by human annotators in 2017. The annotators extract sentences from Wikipedia thus mutating them in a variety of ways, some of which are meaning-altering. They are subsequently verified without knowledge of the sentence they were derived from. They are labelled with either SUPPORTS, REFUTES, or NOT ENOUGH INFO based on whether the evidence entails the statement. The annotators also record the sentence(s) forming the necessary evidence for their judgment, for the first two classes.

The data is distributed using the JSONL⁹ format. Such a format enforces the usage of the UTF-8 encoding to obtain a valid file. Furthermore, each line separator must be the `\n` character and, most importantly, each line must contain a valid JSON value, such as objects or arrays. Thus, each line of the dataset contains a single example of statements generated by relying on Wikipedia. In more detail, the training and development data contain four different fields. The word “claim” is used in the dataset instead of “statement”. In this thesis, the latter word is used to employ a consistent notation. The four fields are:

- `id`: the ID of the statement;
- `claim`: the text of the statement;
- `label`: the label associated to the statement (SUPPORTS, REFUTES, NOT ENOUGH INFO);
- `evidence`: a list of evidence sets, where each element of the list is in the form:
 - Annotation ID, Evidence ID, Wikipedia URL, Sentence ID, if the label is either SUPPORTS or REFUTES;

⁸<https://fever.ai/dataset/fever.html>

⁹<https://jsonlines.org/>

– Annotation ID, Evidence ID, null, null, if the label is NOT ENOUGH INFO.

Table 3.3 shows a sample of three statements, one for each class. Each statement has an evidence set made of a single element. For each piece of evidence, the first attribute identifies the human annotator activity, while the second is the internal identifier of the overall set. The third attribute indicates the Wikipedia page, while the fourth and last attribute identifies a sentence within the current piece of evidence. Thorne et al. provide also a dump containing all the Wikipedia pages processed.¹⁰

Table 3.3: Statements sampled from the FEVER dataset.

ID	Statement	Label	Evidence
77712	Newfoundland and Labrador is the most linguistically homogeneous of Canada.	SUPPORTS	(94661, 107645, "Newfoundland_and_Labrador", 4)
73170	Puerto Rico is not an unincorporated territory of the United States.	REFUTES	(89957, 102650, "Puerto_Rico", 0)
210010	Afghanistan is the source of the Kushan dynasty.	NOT ENOUGH INFO	(248748, null, null, null)

3.4 e-FEVER

The e-FEVER dataset [396] augments the original FEVER dataset (Section 3.3) with explanations generated by their framework. The underlying motivation is that in 16.82% of cases in the FEVER dataset, a statement requires the combination of more than one sentence to be able to support or refute it. Furthermore, they find that sometimes the evidence is not only conditioned by the statement but also by the evidence already retrieved.

In light of this, Stambach et al. propose a two-staged selection process based on the “two-hop” evidence enhancement process [309]. The documents are retrieved by re-using the ukpathene [175] system. The component that performs the final statement verification step employs two strategies one used by Thorne et al. [412]. The statements are labelled with either the supports or refutes label. The NOT ENOUGH INFO label, present in the original fever dataset, is mapped into one of the other labels. The document retrieval system predicts relevant pages and uses the two-staged process to select relevant evidence for these uncertain statements.

The resulting dataset consists of a total amount of 67.687 total examples. The examples for the training are 50.000, while those for the development are 17.687. The resulting dataset thus provides a resource with statements, retrieved evidence, truthfulness labels, and explanations. Table 3.4 shows a sample of two statements, one for each class. Each statement is provided together with the “gold” evidence found, the label assigned and a summary written by a human assessor. The first example shows that the evidence retrieved

¹⁰<https://fever.ai/download/fever/wiki-pages.zip>

might not suffice to fully explain the statement, according to the human annotator’s opinion. The dataset is available upon request to Stambach et al. It contains some explanations due to the evidence retrieval policy.

Table 3.4: Statements sampled from the e-FEVER dataset.

Statement	Label	Gold Evidence	Summary
The Bahamas is a state that’s recognized by other states that includes a series of islands that form an archipelago.	SUPPORTS	The Bahamas, known officially as the Commonwealth of The Bahamas, is an archipelagic state within the Lucayan Archipelago. An archipelagic state is any internationally recognized state or country that comprises a series of islands that form an archipelago.	The relevant information about the claim is lacking in the context.
Scandinavia does not contain Greenland.	REFUTES	The remote Norwegian islands of Svalbard and Jan Mayen are usually not seen as a part of Scandinavia, nor is Greenland, an overseas territory of Denmark.	Greenland is not a part of Scandinavia.

3.5 e-SNLI

The e-SNLI dataset¹¹ [54] extends the SNLI dataset [45] by generating human explanations for 543.950 out of the 570.152 examples of the dataset. The SNLI task is to take two sentences and predict whether one entails, contradicts, or is neutral with respect to the other. The examples for the training set are 550.152, while those for the development set are 20.000.

Camburu et al. collect the data to build the e-SNLI dataset by publishing a crowdsourcing task on the Amazon Mechanical Turk platform. The main goal for the workers is to answer a question asking to assess why a pair of sentences are in a relation of entailment, neutrality or contradiction. The workers are asked to focus on non-obvious elements that induce the given relation and not on parts of the premises which are repeated identically in the hypotheses. Camburu et al. recruit 6325 workers who provide 86 explanations on average. One explanation for each pair of sentences in the training set is collected. Three explanations for each pair of sentences in the validation and test sets are collected. Each sentence pair is provided with the following attributes:

- `pairID`: the identifier of the sentence pair;
- `gold_label`: the ground truth relation of the sentence pair (contradiction, neutral, entail);
- `Sentence1`: the first sentence of the pair;

¹¹<https://github.com/OanaMariaCamburu/e-SNLI>

- **Sentence2**: the second sentence of the pair;
- **Explanation1**: the explanation generated by the worker recruited.

There are 5 additional attributes not included, for a total of 10 attributes. Table 3.5 shows a sample of three sentence pairs, one for each type of relation.

Table 3.5: Statements sampled from the e-SNLI dataset.

Pair ID	Sentence 1	Sentence 2	Gold Label	Explanation
3636329461 .jpg#0r1e	The school is having a special event in order to show the american culture on how other cultures are dealt with in parties.	A school is hosting an event.	entail	An event is a special occasion so all event is a special event.
3636329461 .jpg#0r1n	The school is having a special event in order to show the american culture on how other cultures are dealt with in parties.	A high school is hosting an event.	neutral	The school was never described as a high school
3636329461 .jpg#0r1n	The school is having a special event in order to show the american culture on how other cultures are dealt with in parties.	A school hosts a basketball game.	contradiction	Basketball is american culture.

The next chapter starts to describe the experiments performed in this thesis. It studies whether crowd workers are able to detect and objectively categorize online (mis)information related to political statements.

The Effect of Judgment Scales and Workers' Background

This chapter is based on the article published at the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval [361]. Section 2.1 and Section 2.5 describe the relevant related work. Section 4.1 details the research questions. Section 4.2 presents the experimental setup. Section 4.4 describes the results obtained. Finally, Section 4.5 summarizes the main findings and concludes the chapter.

4.1 Research Questions

This chapter studies how non-expert fact-checkers identify misinformation online. A very large crowdsourcing experiment is set up to such an end. Crowd workers are asked to fact-check statements given by politicians and search for evidence of statement validity using a custom web search engine. The PolitiFact and ABC Fact Check datasets (Section 1.2) are used to sample statements related to USA and Australian politics. US-based crowd workers are recruited to perform the fact-checking task. The experiment involves collecting data on the workers' political background and cognitive abilities. For each dataset, the available expert judgements are compared against the non-expert ones provided by the crowd of workers. This allows also observing how each crowd worker's bias is reflected in the data they generated. In more detail, US-based workers might know about US politics but are less likely would have knowledge of Australian politics in terms of political figures and topics of discussion. The following research questions are investigated:

- RQ1 Which is the relationship and the agreement between the crowd and the expert judgments? And between the judgments collected using different scales?
- RQ2 Are the judgment scales used suitable to gather truthfulness judgments on political statements using crowdsourcing?

RQ3 Which are the sources of information that crowd workers use to identify online misinformation?

RQ4 Which is the effect and the role of assessors' background in objectively identify online misinformation?

4.2 Experimental Setting

The experimental setup involves statements sampled from PolitiFact (Section 3.1) and ABC Fact Check (Section 3.2). In more detail, a subset of 20 statements for each truth level from the PolitiFact dataset covering the time span 2007 to 2015 is used. The sample includes statements by politicians belonging to the two main US parties (Democratic and Republican). The sample from ABC Fact Check includes 60 randomly selected statements (20 statements for each truth level) by politicians belonging to the two main Australian parties (i.e., Liberal and Labor) covering the timespan from 2007 to 2015. For both PolitiFact and ABC Fact Check datasets, a balanced number of statements per class and per political party is included in the sample. A total of 180 statements is thus sampled. Table 4.1 shows an example of PolitiFact and ABC Fact Check statements. Appendix B reports the demographic questionnaire and the CRT tests used.

Table 4.1: Example of statements sampled from the PolitiFact and ABC Fact Check datasets.

Dataset	Label	Statement	Speaker	Year
PolitiFact	Mostly-True	Florida ranks first in the nation for access to free prekindergarten.	Rick Scott	2014
ABC Fact Check	In-Between	Scrapping the carbon tax means every household will be \$550 a year better off.	Tony Abbott	2014

4.2.1 Crowdsourcing Task

The Amazon Mechanical Turk platform has been used to collect truthfulness judgments.¹ Each worker accepting a Human Intelligence Task (HIT) receives a unique input token, which identifies uniquely both the MTurk HIT and the worker. Then, they is redirected to an external application (Appendix A) where to complete the task. The task is designed as follows: in the first part, the workers are asked to provide some details about their background, such as age, family income, political views, the party in which they identify themselves, their opinion on building a wall along the southern border of United States, and on the need for environmental regulations to prevent climate change. Then, workers are asked to answer three modified Cognitive Reflection Test (CRT) questions to assess their

¹The experimental setup was reviewed and approved by the Human Research Ethics Committee at The University of Queensland.

cognitive abilities. In more detail, CRT questions are used to measure whether a person tends to overturn the incorrect “intuitive” response, and further reflect based on their own cognition to find the correct answer. Frederick [148] proposed the original version of the CRT test in 2005. These modified questions are:

- *If three farmers can plant three trees in three hours, how long would it take nine farmers to plant nine trees?* (correct answer = 3 hours; intuitive answer = 9 hours)
- *Sean received both the 5th highest and the 5th lowest mark in the class. How many students are there in the class?* (correct answer = 9 students; intuitive answer = 10 students)
- *In an athletics team, females are four times more likely to win a medal than males. This year the team has won 20 medals so far. How many of these have been won by males?* (correct answer = 4 medals; intuitive answer = 5 medals)

Workers are asked to provide truthfulness judgments after the initial survey. They judge 11 statements: 6 from PolitiFact, 3 from ABC Fact Check, and 2 which serve as gold questions, one obviously true and the other obviously false. All the PolitiFact statements used are sampled from the most frequent five *contexts* (i.e., the circumstance or media in which the statement was said/written) available in the dataset. To avoid bias, a balanced amount of data from each context is selected. Workers are presented with the following information about each statement to judge its truthfulness:

- *Statement*: the text of the statement itself.
- *Speaker*: the name and surname of whom said the statement.
- *Year*: the year in which the statement was made.

Each worker is asked to provide both the truthfulness level of the statement and a URL that serves both as justification for their judgment as well as a source of evidence for fact-checking. In order to avoid workers finding and using the original expert labels (which are available on the Web) as the primary source of evidence, workers must use a provided custom web search engine to look for supporting evidence. The custom search engine uses the Bing Web Search API (Section A.3.4.3) to filter out from the retrieved results from any page from the websites that contain the collection of expert judgments used in the experiment. Workers are allowed to submit the HIT after judging the whole set of 11 statements. In order to increase the quality of collected data, the following quality check is embedded in the crowdsourcing task:

- *Gold Questions*: the worker must assign to the obviously false statement a truthfulness value lower than the one assigned to the obviously true statement.
- *Time Spent*: the worker must spend at least two seconds on each statement and cognitive question.

The HIT reward is set to \$1.5 after measuring the time and effort taken to successfully complete it. This was computed based on the expected time to complete it and targeting to pay at least the US federal minimum wage of \$7.25 per hour. Several small pilots of the task are performed. The task allows only US-based workers to participate, given the aim of the experiment. Each worker was allowed to complete only one of the HITs for only one experimental setting (i.e., one judgment scale) to avoid the learning effect. Overall, not including pilot runs whose data was then discarded, the task allowed collecting judgments

for 120 (PolitiFact) + 60 (ABC Fact Check) = 180 statements, each one of them judged by 10 distinct workers. Such a setup is repeated over 3 different judgment scales. In total, 1800 (for each scale) * 3 = 5400 judgments are collected. Workers provide a total of 6600 assessments if also gold questions are considered.

4.2.2 Judgment Scales And Collections

The experimental design involves three truthfulness scales and five generated collections: two ground truths labeled by experts (for PolitiFact and ABC Fact Check), and three created using the crowdsourcing task (S_3 , S_6 , and S_{100}).

- PolitiFact: uses a six-level scale, with labels *Pants-On-Fire*, *False*, *Mostly-False*, *Half-True*, *Mostly-True*, and *True*.
- ABC Fact Check: uses a three-level scale, with labels *False*, *In-Between*, and *True*.
- S_3 : uses a three-level scale, with the same labels as the ABC Fact Check scale.
- S_6 : uses a six-level scale, with the same labels as the PolitiFact scale, but replacing *Pants-On-Fire* with *Lie*.²
- S_{100} : uses a one-hundred-and-one level scale, with values in the $[0, 100]$ range.³

The nature and usage of these scales deserve some discussion. The scales we use are made of different levels, i.e., categories, but they are not nominal scales. They would be nominal if such categories were independent, which is not the case because they are ordered. This can be seen immediately by considering, for example, that misclassifying a *True* statement as *Mostly-True* is a smaller error than misclassifying it as *Half-True*. Indeed all of them are ordinal scales. However, they are not mere rankings, as the output of an information retrieval system. Statements are assigned to categories, besides being ranked. Let us suppose having two statements with ground truth labels *True* and *Mostly-True* respectively. Misclassifying them as *Half-True* and *Mostly-False* is an error. It is a smaller error than misclassifying them as *False* and *Pants-On-Fire*. However, the original ranking has been preserved in both two cases. These scales are sometimes named Ordinal Categorical Scales [3].

For ordinal categorical scales, it cannot be assumed that the categories are equidistant. For example, misclassifying a *Pants-On-Fire* statement as *False* cannot be assumed to be a smaller error than a misclassifying a *Mostly-False* statement as *True*, generally. In light of this, taking the arithmetic mean to aggregate individual worker judgments for the same statement into a single label is not correct, since this would assume equidistant categories. On the contrary, the aggregation function for nominal scales (called *majority vote* [329] by the crowdsourcing community) would discard important information, even if correct. For example, the aggregation of four *Pants-On-Fire* with six *False* judgments should be rather different from –and lower than– six *False* and four *True*, though the aggregation function is the same. The orthodox and correct aggregation function for this kind of scale is the median. However, the situation is not so clear-cut. In the last example, the median would give the exact same result as the mode, thus discarding useful information. A reasonably defined ordinal categorical scale would feature labels which are approximately equidistant.

²The assumption is that *Lie* would be more clear than the colloquial *Pants-On-Fire* expression.

³The number of levels of this scale is 101 but it is called S_{100} for simplicity.

This is particularly true for S_{100} , since it involves numerical labels in the $[0, 100]$ interval for the categories, and the crowd workers had to use a slider to select the values. This makes S_{100} (at least) very similar to an interval scale, for which the usage of the mean is correct. Indeed, it has been already used for S_{100} [358]. In the field of Information Retrieval people are used to interpreting ordinal scales as interval ones (e.g., when assigning arbitrary gains in the NDCG effectiveness metrics) and/or (ab)using the arithmetic mean (e.g., when taking the mean of ranks in the Mean Reciprocal Rank metric) [149]. In many practical cases, using the arithmetic mean turns out to be not only adequate but even more useful than the correct aggregation function [161, 7, 273]. Even worse, there are no metrics for tasks defined on ordinal categorical scales, like predicting the number of “stars” in a recommendation scenario. Sometimes Accuracy is used, like in NTCIR-7 [207], but it is a metric for nominal scales. In some other cases, like in RepLab 2013 [14], Reliability and Sensitivity [16] are used, which consider only ranking information and no category membership. Even metrics for interval scales, like Mean Average Error (MAE), have been used [157].

For these reasons, in the following the (aggregated) truthfulness labels provided by workers are sometimes used as if they were expressed using an interval scale. Such a decision allows treating the various scales in a homogeneous way, and thus use the same aggregation used for S_{100} also for S_3 , S_6 , PolitiFact, and ABC Fact Check. Accordingly, in the following the labels of ABC Fact Check and S_3 are denoted with \emptyset , 1, and 2, as if they were in the $[0, 2]$ range. Moreover, the labels of PolitiFact and S_6 are denoted with \emptyset , 1, \dots , 5, as if they were expressed in the $[0, 5]$ range. Finally, the truthfulness labels of S_{100} are denoted with \emptyset , 1, \dots , 100.

4.3 Descriptive Analysis

The crowdsourcing task has been published on Amazon Mechanical Turk. Overall, about 600 US resident⁴ crowd workers participate in it.

4.3.1 Worker Demographics

The majority of workers (46.33%) are between 26 and 35 years old, across all three experiments. Furthermore, the workers are well-educated as more than 60.84% of workers have a four years college degree at least. Around 67.66% of the workers earn less than \$75,000 a year. Nearly half (47.33%) of the workers think their views are more democratic. Only about 22.5% of workers select the Republican Party as a voting preference. As for political views, Liberal and Moderate account for the most substantial proportion of workers, 29.5% and 28.83% respectively. The Very Conservative answer accounts for the least, only 5.67%. In response to the border issue, 52.33% of US-based workers are against the construction of a wall on the southern border, and 36.5% of the workers support it. For environmental protection, 80% of the workers support the government strengthening environmental regulation to prevent climate change, while 11.33% of the workers are against such a decision.

⁴Amazon Mechanical Turk workers based in the US must provide evidence they are eligible to work.

4.3.2 Task Abandonment

Table 4.2 shows the ratio of workers who complete the task, abandon the task, and fail the quality checks, by relying on the behavioral actions logged as workers proceed through the HITs of the task. Abandonment numbers are in line with previous studies [174]. Higher failure and lower completion rates can be observed for S_{100} . This may show a slight lack of comfort for workers in using the most fine-grained scale.

Table 4.2: Worker completion, abandonment, and failure rates during the crowdsourcing tasks for the S_3 , S_6 , and S_{100} collections.

Collection	Completion	Abandonment	Failure
S_3	35	53	12
S_6	33	52	14
S_{100}	25	53	22

4.3.3 Crowdsourced Judgments Distributions

Figure 4.1 shows the distribution (and the cumulative distribution in red) for the individual judgments provided by workers over the three crowd collections (i.e., S_3 , S_6 , and S_{100}) for all the statements considered in the experiment. The behavior is consistent when considering separately PolitiFact and ABC Fact Check statements.

The first row of the figure shows the raw judgments distributions for the three collections. The distribution is skewed towards the right part of the scale representing higher truthfulness values, for the whole set of collections. This can be also seen by looking at the cumulative distribution, which is steeper on the right-hand side of the charts. It can also be seen that all three distributions are multimodal. The S_{100} collection shows a mild round number tendency that is, the tendency of workers to provide truthfulness judgments which are multiple of 10 (35% of S_{100} scores are multiple of 10; 23% are 0, 50, or 100); such behavior was already noted by Maddalena et al. [267], and Roitero et al. [358]. The behavior is consistent when considering separately PolitiFact and ABC Fact Check statements.

The gold judgments are those provided for the special High and Low statements used to perform quality checks during the task. The second row of Figure 4.1 shows the distribution of the scores for the three crowd collections considered. The large majority of workers (44% for Low and 45% for High in S_3 , 34% for Low and 39% for High in S_6 , and 27% for Low and 24% for High in S_{100}) provide as truthfulness judgment for the gold statements the extreme value of the scales (respectively the lower bound of the scale for Low and the upper bound of the scale for High). This can be interpreted as a signal that the data gathered is of good quality. However, some workers provide judgments inconsistent with the gold labels.

The third row of Figure 4.1 shows the distributions of S_3 , S_6 , and S_{100} judgments aggregated by taking the average of the 10 scores obtained independently for each statement. The distribution of the judgments aggregated for S_3 , S_6 , and S_{100} are similar, they are no longer multimodal, and they are roughly bell-shaped. It is worth noting that the judgments for S_3 and S_{100} (bottom-mid and bottom-right plots in Figure 4.1) are skewed to higher/positive scores. On the other hand, the judgments aggregated for S_3 (bottom-left plot in Figure 4.1)

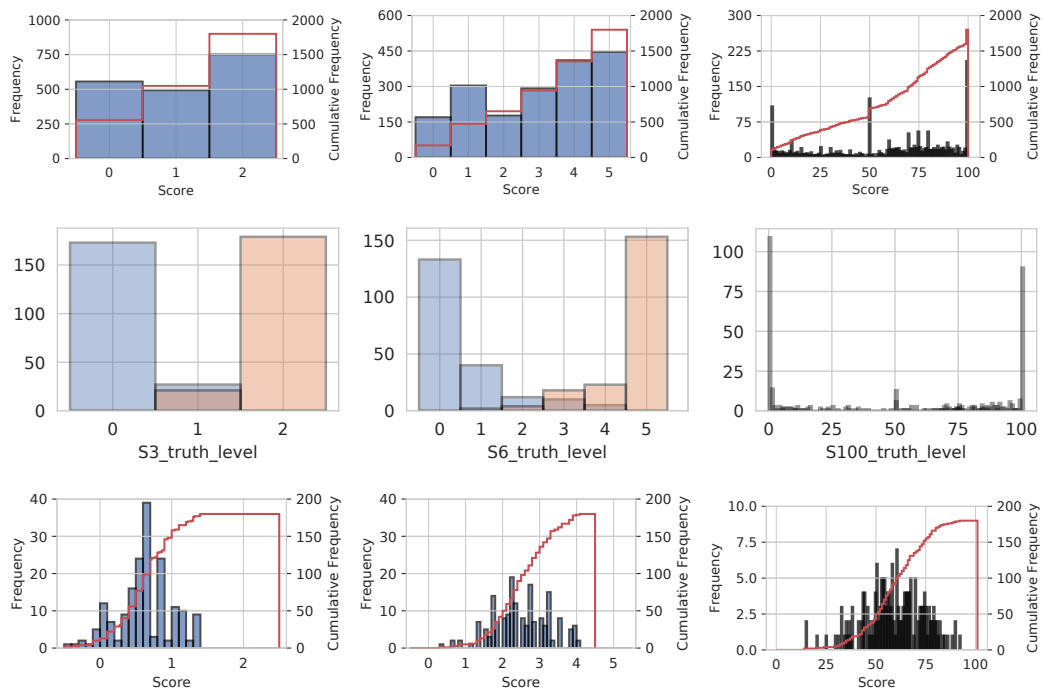


Figure 4.1: From left to right: S_3 , S_6 , S_{100} . From top to bottom: individual scores distribution (first row), gold judgments distribution (second row), and aggregated judgments distribution (third row).

are skewed towards lower/negative (i.e., False and In-Between scores). This shows how different judgment scales are used differently by the crowd. For S_{100} the round number tendency effect also disappears when judgments from different workers are aggregated together, as expected [267, 356, 358].

4.4 Results

Section 4.4.1 discusses the external agreement between the crowd judgments and the expert labels as well as the internal agreement among workers. Section 4.4.2.2 analyzes the aggregation of the judgments collected into more coarse-grained scales. Section 4.4.3 studies the sources of information used by the workers. Section 4.4.4 discusses the relationships that exist between workers' background and their performances.

4.4.1 RQ1: Crowd Workers Accuracy

The quality of the judgments provided by the crowd workers can be studied by analyzing the external agreement, i.e., the agreement between the crowd-collected judgments and the experts' ground truth (Section 4.4.1.1). Another standard way to address that is to analyze the quality of the work by the crowd workers is to compute the internal agreement, i.e., the agreement among the workers (Section 4.4.1.2).

4.4.1.1 External Agreement

Figure 4.2 shows the agreement between the crowd judgments aggregated and the ground truth (i.e., the expert labels provided for PolitiFact and ABC Fact Check) for the S_3 , S_6 , and S_{100} crowd collections. The behavior over all three scales is similar, both on PolitiFact and ABC Fact Check statements.

Focusing on PolitiFact statements (shown on the left-hand side of each chart) allows seeing that the 0 and 1 boxplots are very similar. This can point out a difficulty for workers to distinguish between the *Pants-On-Fire* and *False* labels. The same behavior, even if less evident, is present between the *False* and *Mostly-False* labels; this behavior is consistent across all the scales. On the contrary, focusing on the remaining PolitiFact labels and the ABC Fact Check ones shows that the median lines of each boxplot are increasing while going towards labels representing higher truthfulness values (i.e., going towards the right-hand side of each chart), indicating that workers have a higher agreement with the ground truth for those labels. Again, this behavior is consistent and similar for all the S_3 , S_6 , and S_{100} scales.

The statistical significance of the differences between the judgments aggregated using the mean for the categories of the S_6 , S_3 , and S_{100} collections are measured according to the Mann-Whitney rank test and the t-test. Concerning ABC Fact Check, adjacent categories are significantly different in 5 cases out of 12, while the difference between non-adjacent categories are all significant to the $p < .01$ level. Concerning PolitiFact, the differences between the judgments aggregated using the mean for adjacent categories and not adjacent ones by the distance of 2 (e.g., 0 and 2) are never significant with only one exception (distance 2). Differences for not adjacent categories of distance 3 are significant in 4/18

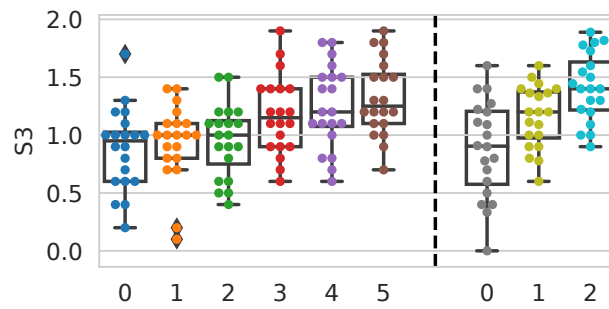
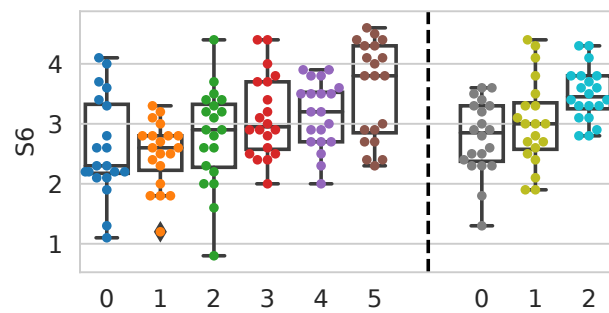
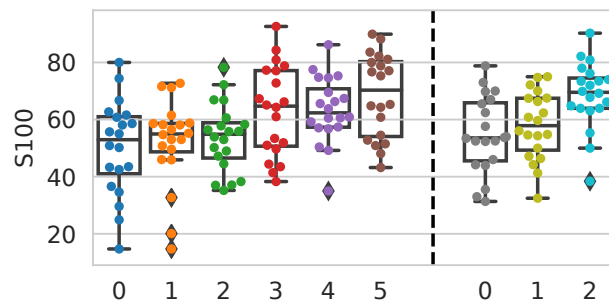
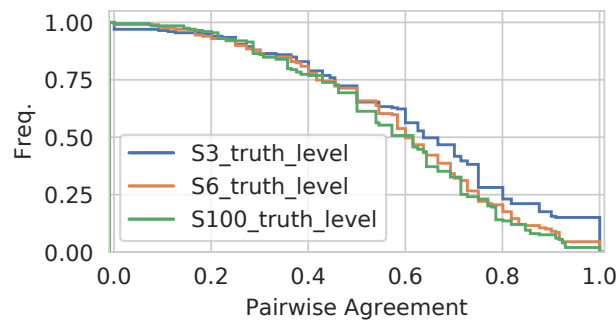
(a) S_3 (b) S_6 (c) S_{100}

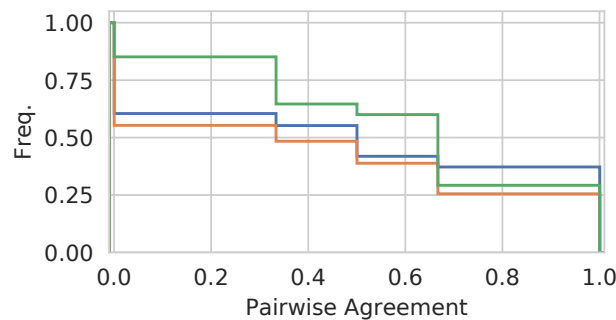
Figure 4.2: External agreement with PolitiFact and ABC Fact Check statements, separated by the vertical dashed line.

cases. Differences for categories of distance 4 are significant in 5/12 of the cases. Finally, categories of distance 5 (i.e., 0 and 5) are significant in 4/6 cases. Although there is some signal, it is clear that the answer to RQ1 cannot be positive on the basis of these results. This is further detailed in Section 4.4.2.2.

Figure 4.3 inspects the agreement between the workers and the ground truth by looking at each HIT. The pairwise agreement [269] between the truthfulness judgments expressed by workers and the ground truth labels is computed for all the S_3 , S_6 , and S_{100} collections, with a breakdown over PolitiFact (Figure 4.3a) and ABC Fact Check (Figure 4.3b) statements. A slightly modified version of the pairwise agreement measure [269] is used. In the attempt to make the pairwise agreement measure fully comparable across the different scales, all the ties are removed. Intuitively, the pairwise agreement described by Maddalena et al. [269] measures the fraction of pairs in the agreement between a “ground truth” scale and a “crowd” scale. Specifically, a pair of crowd judgments (crowd-judgment_1 , crowd-judgment_2) is considered to be in agreement if $\text{crowd-judgment}_1 \leq \text{crowd-judgment}_2$ and the ground truth for crowd-judgment_1 is $<$ the ground truth for crowd-judgment_2 . In this thesis' measurement⁵ all the ties (i.e., $\text{crowd-judgment}_1 = \text{crowd-judgment}_2$) are removed and $<$ is used in place of \leq .



(a) PolitiFact



(b) ABC Fact Check

Figure 4.3: Pairwise agreement and relative frequency by looking at each HIT of the task.

⁵The code used to compute the pairwise agreement as defined is available at <https://github.com/KevinRoit ero/PairwiseAgreement>.

In more detail, Figure 4.3 shows the CCDF (Complementary Cumulative Distribution Function) of the relative frequencies of the agreement among HITs. The S_3 , S_6 , and S_{100} scales show a very similar level of external agreement. Such behavior is consistent across the PolitiFact and ABC Fact Check datasets. Again, this result confirms that all the considered scales present a similar level of external agreement with the ground truth, with the only exception of S_{100} for the ABC Fact Check dataset. This is probably due to the treatment of ties in the measure, which removes a different number of units for the three scales.

4.4.1.2 Internal Agreement

The Krippendorff's α coefficient [224] is computed as a metric to measure the level of internal agreement in a dataset. All α values within each of the three scales S_3 , S_6 , S_{100} and on both PolitiFact and ABC Fact Check collections are in the 0.066–0.131 range. These results show that there is a rather low agreement among the workers [68, 224]. Furthermore, all the possible transformations of judgments from one scale to another are performed to investigate whether the low agreement found depends on the specific scale used to judge the statements, following the methodology described by Han et al. [173].

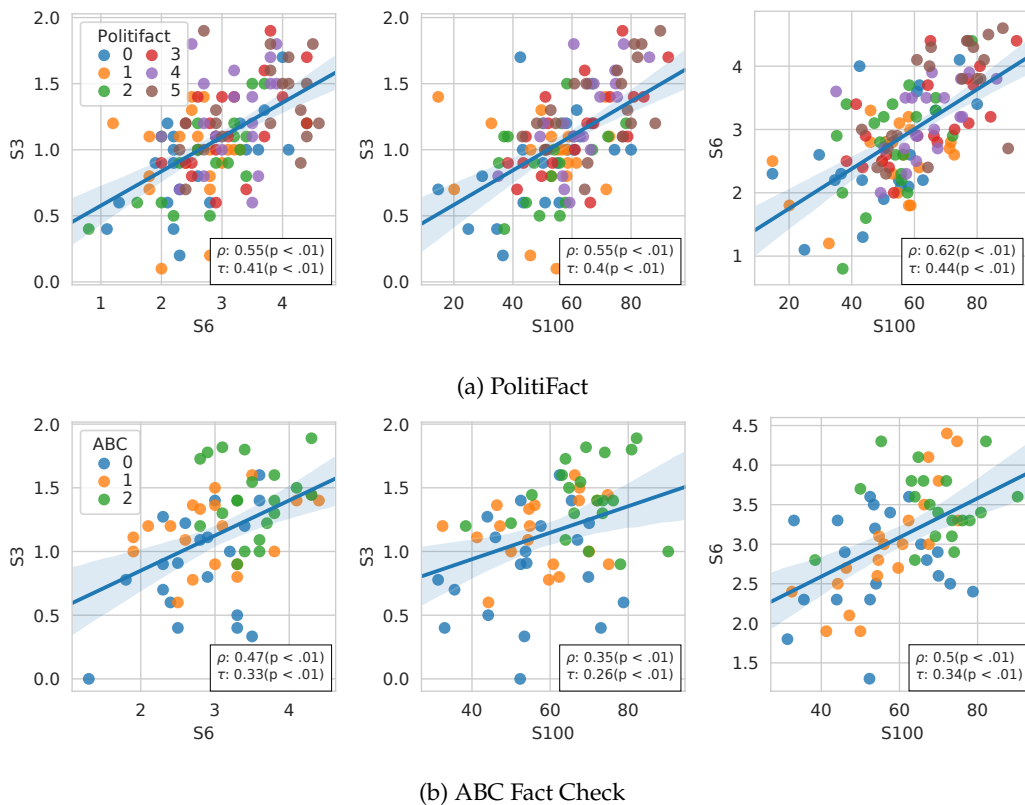


Figure 4.4: Agreement between judgment scales. From left to right: S_6 vs. S_3 , S_{100} vs. S_3 , and S_{100} vs. S_6 .

Figure 4.4 shows the scatterplots, as well as the correlations, between the different

scales on the PolitiFact and ABC Fact Check statements. The correlation values are around $\rho = 0.55\text{--}0.6$ for PolitiFact and $\rho = 0.35\text{--}0.5$ for ABC Fact Check, for all the scales. The rank correlation coefficient τ is around $\tau = 0.4$ for PolitiFact and $\tau = 0.3$ for ABC Fact Check. These values indicate a low correlation between all the scales. This is an indication that the same statements on different scales tend to be judged differently, both when considering their absolute value (i.e., ρ) and their relative ordering (i.e., τ).

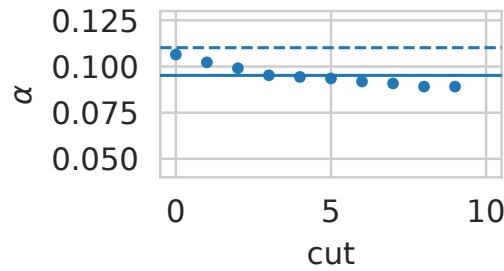
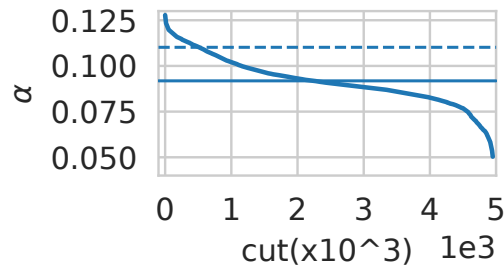
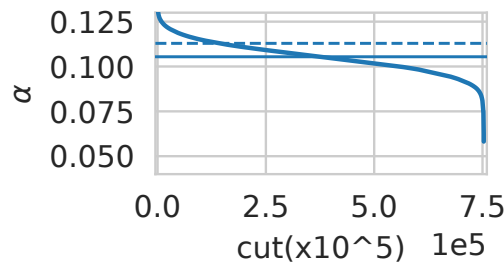
(a) S_6 cut into S_3 (b) S_{100} cut into S_3 (c) S_{100} cut into S_6 (1% stratified sampling)Figure 4.5: α cuts sorted by decreasing values.

Figure 4.5 shows the distribution of the α values when transforming one scale into another. The total number of possible cuts from S_{100} to S_6 is 75,287,520. Thus, a sub-sample of all the possible cuts is selected. Both stratified and random sampling has been used, getting indistinguishable results. The dotted horizontal line in the plot represents α on the original dataset, the dashed line is the mean value of the (sampled) distribution. The values on the y-axis are very concentrated and all α values are close to zero ($[0, 0.15]$ range). It can

be concluded that across all collections there is a low level of internal agreement among workers, both within the same scale and across different scales.

4.4.2 RQ2: Judgment Scales Adequacy

Section 4.4.2.1 studies the effect of aggregation functions alternative to the arithmetic mean and workers are compared using different scales. Section 4.4.2.2 studies.

4.4.2.1 Alternative Aggregation Functions

Figure 4.6 shows the results of using the median. In this case, the final truthfulness judgment for each statement has been computed by considering the median of the judgments expressed by the workers. It is clear that the median produces the worst results, by comparing the charts to those in Figure 4.2 attempts grouping of adjacent categories to improve results quality.

The heatmaps in Figure 4.7 show the results of using the majority vote (i.e., the mode) as the alternative aggregation function. The mode is more difficult to compare with the mean, but it is again clear that the overall quality is rather low. Although the squares around the diagonal tend to be darker and contain higher values, there are many exceptions. These are mainly in the lower-left corners, indicating false positives. In other words, statements whose truthfulness value is over-evaluated by the crowd. This tendency to false positives is absent when using the mean (see Figure 4.2). Overall, these results confirm that the choice of mean as aggregation function seems the most effective.

4.4.2.2 Merging Assessment Levels

The answer to RQ1 cannot be completely positive, in light of the results presented in Section 4.4.1, and Section 4.4.2.1 (with a particular focus on Figure 4.2, Figure 4.6 and Figure 4.7, but also the rather low agreement and correlation values). There is a clear signal that aggregated values resemble the ground truth, but there are also several exceptions and statements that are misjudged. However, there are some further considerations that can be made. First, results seem better for ABC Fact Check than PolitiFact. Second, it is not clear which specific scale should be used. The two expert collections used as ground truth use different scales. The experimental setting, however, involves the crowd on S_3 , S_6 , and S_{100} . Also, comparisons across different scales are possible. Finally, a binary choice (true/false) seems also meaningful, and in many real applications, it is what may really be needed. Third, the above-mentioned possible confusion between *Pants-On-Fire* and *False* suggest that these two categories could be fruitfully merged. This has been done, for example, by Tchechmedjiev et al. [407].

All these remarks suggest attempting some grouping of adjacent categories, to check if by looking at the data on a more coarse-grained ground truth the results improve. Therefore, the six PolitiFact categories are grouped into either three (i.e., 01, 23, and 45) or two (i.e., 012 and 345) resulting ones, adopting the approach discussed by Han et al. [173]. Figure 4.8 shows the results. The agreement with the ground truth can now be seen more clearly. The boxplots also seem quite well separated, especially when using the mean (the first

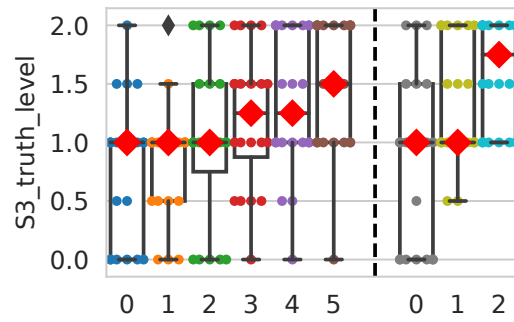
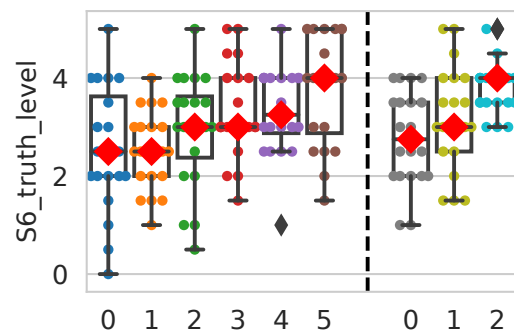
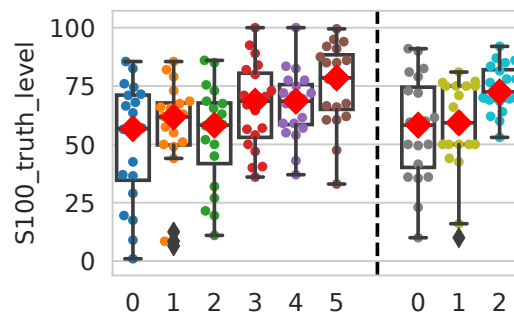
(a) S_3 (b) S_6 (c) S_{100}

Figure 4.6: Agreement with the ground truth while using the median as aggregation function (highlighted by the red diamond). Compare with Figure 4.2.

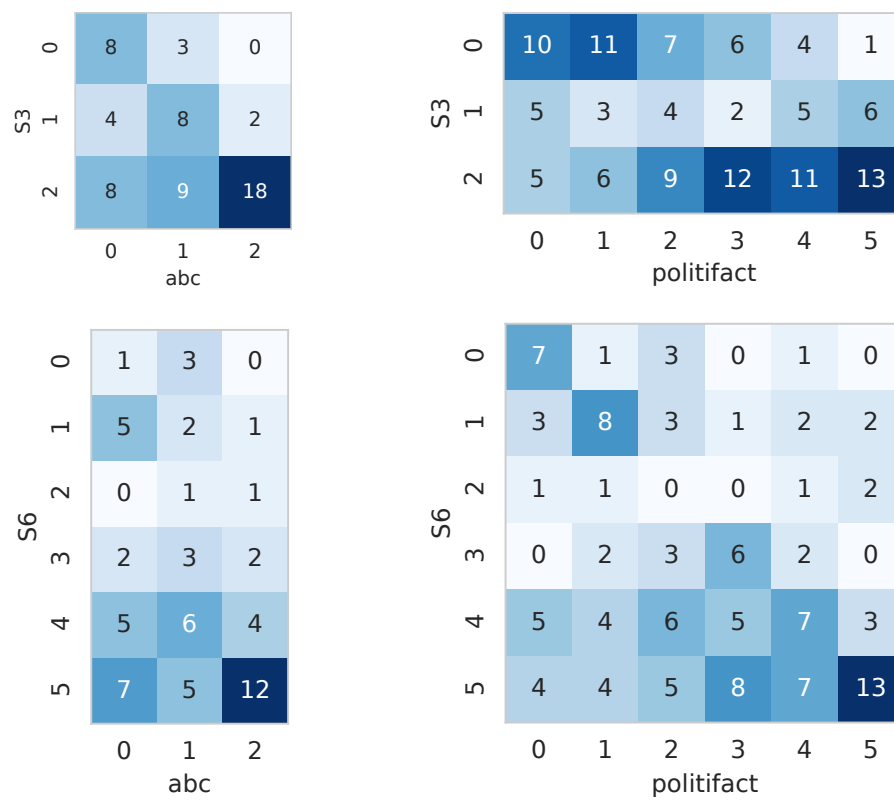


Figure 4.7: Agreement between S_3 (first row) and S_6 (second row), and ABC Fact Check (left) and PolitiFact (right). Aggregation function: majority vote.

three charts on the left). This is confirmed by analyses of statistical significance. All the differences in the boxplots on the bottom row are statistically significant at the $p < .01$ level for both the t-test and Mann–Whitney, both with Bonferroni correction. The same holds for all the differences between 01 and 45 (the not adjacent categories) in the first row. For the other cases (i.e., concerning the adjacent categories), further statistical significance is found at the $p < .05$ level in 8 out of 24 possible cases. These results are much stronger than the previous ones. It is thus possible to state that the crowd is able to single out true from false statements with good accuracy. For statements with an intermediate degree of truthfulness/falsehood, the accuracy is lower.

4.4.3 RQ3: Sources Of Information

Table 4.3 shows the distribution of websites used by workers to justify the truthfulness judgment they provide for each statement. The most used sources are “Wikipedia” and “YouTube”, for all the scales, followed by popular news websites such as “The Guardian” and “The Washington Post”. Furthermore, among the most popular sources, there is one fact-checking website (i.e., FactCheck). The `abc.com.au` and `politifact.com` URLs have been intentionally removed during the crowdsourcing task from those which could be selected. This shows that workers, supported by the search engine, tend to identify trustworthy information sources to support their judgment decisions.

Table 4.4 shows the distribution of the ranks within the search engine results of the URLs chosen by workers to justify their judgments (without considering the gold questions), for S_3 , S_6 and S_{100} . The majority of workers tend to click on the first results shown by the search engine, as expected [90, 197, 212]. Nevertheless, the results also show that workers explore the first ten documents as ranked by the search engine and do not simply click on the first returned URL, thus putting some effort to find a reliable source and/or justification. Finally, all the workers stop at the first page of results as returned by the search engine over all the scales. Nobody investigates search results with a rank greater than 10.

4.4.4 RQ4: Effect Of Worker Background and Bias

The role of assessors' background in objectively identifying online misinformation can be understood by assessing the relationships that exist between workers' cognitive performances (Section 4.4.4.1) and their background (Section 4.4.4.2).

4.4.4.1 Cognitive Reflection Tests

The Cognitive Reflection Test (CRT) performances are measured as the percentage of correct answers given by them to estimate workers' cognitive skills. Thus, a higher CRT score is associated with higher analytical thinking ability [148]. The performances are compared across the three scales using the standardized calculation of the z-score for each worker and each truthfulness level. The z-score for each statement represents the performance of crowd workers as compared to others. The lower the z-score for false statements, the stronger the ability of the crowd to identify lies and the higher the z-score for true statements, the higher the ability to identify accurate information. “Discernment” is then calculated

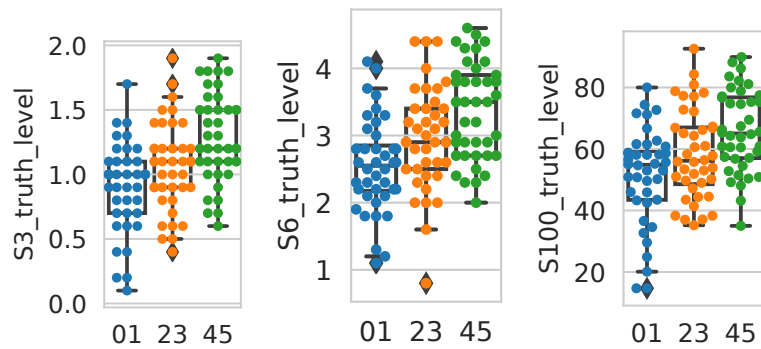
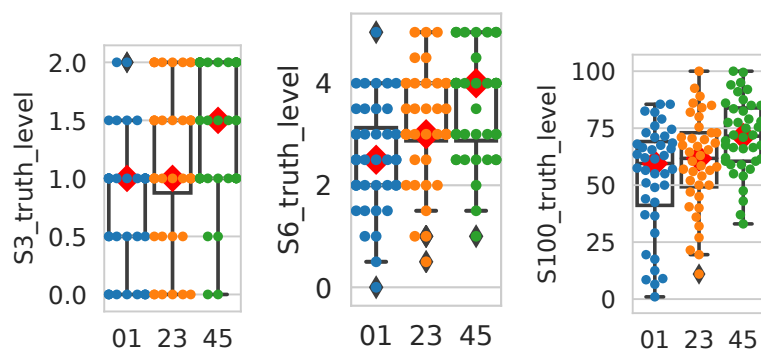
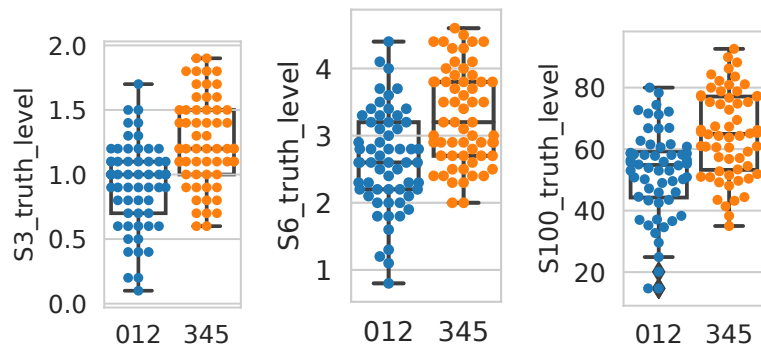
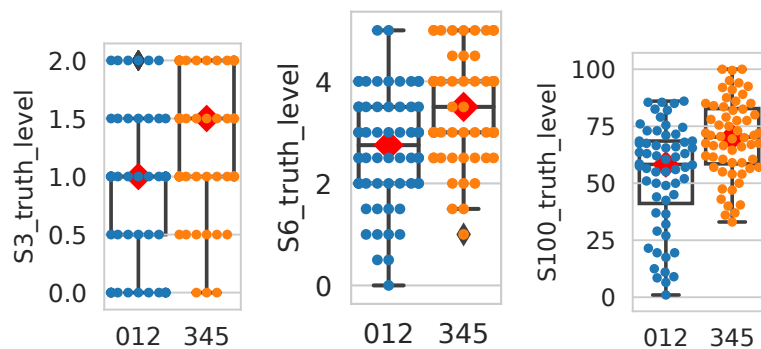
(a) Mean for S_6 , S_3 , and S_{100} (b) Median for S_6 , S_3 , and S_{100} (c) Mean for S_6 , S_3 , and S_{100} (d) Median for S_6 , S_3 , and S_{100}

Figure 4.8: Agreement with ground truth for merged categories for PolitiFact. From top to bottom: three and two resulting categories. The median is highlighted by the red diamond.

Table 4.3: Websites from which workers chose URLs to justify their judgments without considering gold questions for S_3 , S_6 , and S_{100} . Only websites with percentage $\geq 1\%$ are shown.

Source	S_3	S_6	S_{100}
Wikipedia	17%	19%	23%
YouTube	13%	13%	12%
The Guardian	11%	12%	13%
FactCheck	8%	8%	9%
Smh	6%	3%	6%
Cleveland	6%	6%	5%
Washington Post	6%	6%	6%
News	5%	4%	5%
Blogspot	5%	6%	5%
On The Issues	4%	0%	0%
Quizlet	4%	3%	4%
New York Times	3%	3%	0%
CBS News	3%	4%	5%
Forbes	3%	3%	0%
House	3%	3%	3%
Madison	3%	0%	0%
The Australian	0%	4%	0%
Milwaukee Journal	0%	3%	0%
Yahoo	0%	0%	3%

Table 4.4: Distribution of the ranks in search results for the URLs chosen by workers in S_3 , S_6 , and S_{100} .

Rank	S_3	S_6	S_{100}	Avg.
1	17%	13%	15%	15%
2	12%	13%	15%	13%
3	13%	16%	15%	15%
4	14%	12%	12%	11%
5	12%	11%	8%	10%
6	9%	9%	12%	10%
7	7%	8%	7%	7%
8	6%	7%	6%	6%
9	4%	5%	5%	5%
10	3%	4%	2%	3%
11	1%	1%	1%	1%

by deducting the z -score for false statements from the z -score for true statements. This represents the ability of the crowd to distinguish truthfulness from falsehood [327]. This analysis focuses on statements with extreme true or false ground truth levels and discards the In-Between statements as they do not provide additional evidence on the ability of the crowd to distinguish between true/false information.

Table 4.5: Correlation between Cognitive Reflection Test (CRT) performance and z -scores for each truthfulness level and the correlation between worker age and z -scores.

Dataset	Correlation With	Age	CRT Performance
PolitiFact	Lie	-0.038	-0.098*
	False	-0.022	-0.072
	True	0.127*	0.062
	Discernment	0.113**	0.128**
ABC Fact Check	False	-0.075	-0.021
	True	0.048	0.110**
	Discernment	0.088*	0.110**
Total	Discernment	0.125**	0.154**

** : $p < 0.01$, * : $p < 0.05$.

Table 4.5 shows the results. First, there is a statistically significant (Spearman's rank-order test), moderate positive correlation between Discernment and CRT score on statements from PolitiFact and ABC Fact Check ($r_s(598) = 0.128$, $p = 0.002$ and $r_s(598) = 0.11$, $p = 0.007$ respectively). This shows that workers who ponder more perform better in identifying Lie statements of US (local) politicians ($r_s(598) = -0.098$, $p = 0.017$), and identifying true statements of AU (not local) politicians ($r_s(598) = 0.11$, $p = 0.007$). In general, people with strong analytical abilities (as determined by the CRT test) can better recognize true statements from false ($r_s(598) = 0.154$, $p < 0.0005$). Besides, the ability to distinguish truthfulness from falsehood increases with age ($r_s(598) = 0.125$, $p = 0.002$). Indeed, older workers perform better in recognizing true statements by US politicians ($r_s(598) = 0.127$, $p = 0.02$). The level of education and income do not have a statistically significant correlation with their judgments.

4.4.4.2 Political Background

Concerning the effect of the political background, the Shapiro-Wilk test [380] confirms that discernment scores are normally distributed ($p > 0.05$) for groups with diverse political views. In light of this, a one-way ANOVA analysis can be performed to determine whether the ability to distinguish true from false statements is different across groups. Levene's test [373] shows that the homogeneity of variances is violated ($p = 0.034$). Thus, a Welch-Satterthwaite correction [371] is used to calculate the degrees of freedom and the Games-Howell posthoc test [367] to show multiple comparisons. Discernment score is statistically significantly different between different political views (Welch's $F(4, 176.735) = 3.451$, $p = 0.01$). A Games-Howell posthoc test confirms that the increase of discernment score (0.453,

95% CI (0.028 to 0.879)) from conservative (-0.208 ± 1.293) to liberal (0.245 ± 1.497) is statistically significant ($p = 0.03$). In light of these results, the crowd workers who have liberal views can better differentiate between false and true statements. Furthermore, there is no statistically significant difference in discernment scores based on the political party with which crowd workers explicitly identified themselves ($\chi^2(3) = 3.548, p = 0.315$). Since a Shapiro-Wilk test [380] shows a non-normal distribution $p < 0.05$, a Kruskal-Wallis H test [225] is used. This shows there is no difference in judgment quality based on workers' explicit political stances.

However, an analysis of workers' implicit political views rather than their explicit party identification shows a different result. The partisan gap on the immigration issue is apparent in the US. According to the survey conducted by Pew Research Center in January 2019, [59], about 82% of Republicans and Republican-leaning independents support developing the barrier along the southern border of the United States, while 93% of Democrats and Democratic leaners oppose it. Therefore, asking workers' opinions on this matter can be a way to know their implicit political orientation. A Kruskal-Wallis H test [225] is used –Shapiro-Wilk test's $p < 0.05$ – in workers' discernment between immigration policy groups defined based on their answer to the wall question. Statistically significant differences are observed for PolitiFact statements ($\chi_2(2) = 10.965, p = 0.004$) and for the whole set of statements ($\chi_2(2) = 11.966, p = 0.003$). A posthoc analysis using Dunn's procedure with a Bonferroni correction [118] reveals statistically significant differences in discernment scores on PolitiFact statements between agreeing (-0.335) and disagreeing (0.245) ($p = 0.007$) on building a wall. Similar results are obtained when discernment scores are compared on the whole set of statements. There are statistically significant differences in discernment scores between agreeing (-0.377) and disagreeing (0.173) ($p = 0.002$) on building a wall along the southern border of the USA. These results show how, in this experimental setup, crowd workers who do not want a wall on the southern US border perform better in distinguishing between true and false statements. Lastly, there are no significant differences concerning workers' stances on climate issues.

4.5 Summary

This chapter presents an extensive crowdsourcing experiment that aims at studying crowd workers that identify misinformation online. The dataset used in the research includes statements given by the US and Australian politicians. The experiment asks US-based crowd workers to perform the fact-checking task by using the customized and controllable Internet search engine to find evidence of the validity of the statements. The experiment allows collecting and analyzing data on the workers' political background and cognitive abilities. Furthermore, it allows controlling for the politically-consistent statements to be fact-checked, the geographical relevance of the statements, the judgment scale granularity, and the truthfulness level. The answers to the research questions can be summarized as follows.

RQ1 The behavior over all the three scales is similar both on PolitiFact and ABC Fact Check statements, in terms of agreement w.r.t. the ground truth. There is a low level of internal agreement among workers, on all the $S_3, S_6,$ and S_{100} scales, across all the judgments

collected using crowdsourcing.

RQ2 The grouping of adjacent categories reveals that crowdsourced truthfulness judgments are useful to accurately single out true from false statements.

RQ3 The workers put effort to find a reliable source to justify their judgments, and tend to choose a source found in the first search engine result page, but not necessarily the first search result.

RQ4 The assessors' background affects in objectively identifying online misinformation.

The next chapter analyzes whether crowd workers can detect and objectively categorize recent (mis)information. Statements related to the COVID-19 pandemic are employed to such an end and a longitudinal study is performed.

A Longitudinal Study On Misinformation About COVID-19

This chapter is based on the article published in the “Personal and Ubiquitous Computing” journal [362]. It is an extension of the one published at the 29th ACM International Conference on Information and Knowledge Management [363]. Section 2.1, Section 2.2, and Section 2.5 describe the relevant related work. Section 5.1 details the research questions, addressed using the experimental setting described in Section 5.2. Section 5.4 presents the results obtained and presents the longitudinal study conducted. Finally, Section 5.5 summarizes the main findings and concludes the chapter.

5.1 Research Questions

This chapter studies how non-expert fact-checkers identify misinformation online. This might look similar to the work presented in Chapter 4, but the experiments focus on statements about COVID-19. This is motivated by several reasons.

The pandemic was a hot topic in 2020 and there were no studies yet using crowdsourcing to assess the truthfulness of related statements, even though there is a great number of research efforts worldwide devoted to its study. Moreover, the health domain is particularly sensitive. It is thus interesting to understand if crowdsourcing-based approaches are adequate also in such a particular domain. In the previous work [232, 356, 361] the statements judged by the crowd were not recent. This means that evidence of statements’ truthfulness was often available on the Web. It cannot thus be excluded that the workers did find it, although the experimental design prevented easily finding that evidence. They might be familiar with a given statement because, for instance, it had been discussed in the press. Focusing on COVID-19 related statements allowed us to naturally target more statements. In some cases, the evidence could have been still out there, but that would happen more rarely. Furthermore, an almost ideal tool to address misinformation would be a crowd able to judge truthfulness in real-time, immediately after the statement becomes public. Targeting recent statements is a step forward in such a direction, although much work must

still be done. The experimental design differs in some details and allows to address of novel research questions. Lastly, a longitudinal study is performed. It involves collecting the data multiple times and launching the task at different timestamps, considering both novice workers – i.e., workers who have never done the task before – and experienced workers – i.e., workers who have performed the task in previous batches and were invited to do the task again. This allows us to study the multiple behavioral aspects of workers that judge the truthfulness of judgments.

The experiments focus on statements about COVID-19, which are recent and interesting for the research community, and arguably deal with a more relevant/sensitive topic for the workers. They investigate whether the health domain makes a difference in the ability of crowd workers to identify and correctly classify (mis)information and whether the very recent nature of COVID-19 related statements has an impact as well. focus on a single truthfulness scale, given the evidence that the scale used does not make a significant difference. Another important difference with respect to the setting described in Section 4 and previous work [232, 356] is that workers to provide are asked to provide a textual justification for their decision. The justifications are analyzed to better understand the process followed by workers to verify the information. This allows for investigating whether they can be exploited to derive useful information. The longitudinal study includes 3 crowdsourcing experiments over a period of 4 months and allows for collecting additional data and evidence that include novel responses from novice and returning crowd workers. The workers' behavior is also analyzed, as done in Section 4.4. The following research questions are investigated:

- RQ5 Are the crowd workers able to detect and objectively categorize online (mis)information related to the medical domain and more specifically to COVID-19? What are the relationship and the agreement between the crowd and the expert labels?
- RQ6 Can the crowdsourced and/or the expert judgments be transformed or aggregated in a way that improves the ability of workers to detect and objectively categorize online (mis)information?
- RQ7 What is the effect of workers' political bias and cognitive abilities?
- RQ8 What are the signals provided by the workers while performing the task that can be recorded? To what extent are these signals related to workers' accuracy? Can these signals be exploited to improve accuracy and, for instance, aggregate the judgments in a more effective way?
- RQ9 Which sources of information does the crowd consider when identifying online mis-information? Are some sources more useful? Do some sources lead to more accurate and reliable judgments by the workers?
- RQ10 What is the effect of re-launching the experiment and re-collecting all the data at different time spans? Are the findings from all previous research questions still valid?
- RQ11 How does considering the judgments from workers which did the task multiple times change the findings of RQ6? Do they show any difference when compared to workers who did the task only once?

RQ12 Which are the statements for which the truthfulness judgments provided using crowdsourcing fail? Which are the features of the statements that are misjudged by the crowd workers?

5.2 Experimental Setting

The experimental setup involves statements sampled from PolitiFact (Section 3.1). In more detail, 10 statements for each of the six PolitiFact categories are selected. Such statements belong to the COVID-19 section and with dates ranging from February 2020 to early April 2020. The sample includes statements by politicians belonging to the two main US parties (Democratic and Republican). A balanced number of statements per class and per political party is included in the sample. Appendix C contains the full list of the statements used.

5.2.1 Crowdsourcing Task

The design of the task design to collect truthfulness judgments about COVID-19 and then used to perform the longitudinal study is similar to the one described in Section 4.2.1. The crowdsourcing platform Amazon Mechanical Turk is used to collect the judgments. Each worker is assigned a unique pair of values (input token, output token). Then, they is redirected to an external website (Appendix A) where to complete the task.

The task itself is as follows. First, a (mandatory) questionnaire is shown to the worker, to collect background information such as age and political views. Then, the worker needs to provide answers to three Cognitive Reflection Test (CRT) questions. These questionnaires are those already described in Section 4.2.1. The worker is then asked to judge the truthfulness of 8 statements: 6 from the dataset described in Section 3.1 (one for each of the six considered PolitiFact categories) and 2 special statements called “gold questions” (one clearly true and the other clearly false) manually written and used as quality checks. A randomization process is used when building the HITs to avoid all the possible sources of bias, both within each HIT and considering the overall task. The worker is shown: the *Statement*, the *Speaker/Source*, and the *Year* in which the statement was made to judge its truthfulness.

The workers are asked to provide the following information:

- the truthfulness judgment for the statement using the six-level scale adopted by PolitiFact, from now on referred to as C_6 . The scale is presented to the worker using a radio button containing the label description for each category as reported in the original PolitiFact website;
- the URL that they use as a source of information for the fact-checking;
- a textual motivation for her/his judgment. The motivation can not include the URL and should contain at least 15 words.

In order to prevent the user from using PolitiFact as the primary source of evidence, its domain is filtered out from the returned search results. The following quality checks are implemented in the task:

- the judgments provided for the gold questions have to be coherent (i.e., the judgment of the clearly false question should be lower than the one assigned to the true question);
- the cumulative time spent to perform each judgment should be of at least 10 seconds.

The CRT and the questionnaire answers were not used for quality checks, although the workers were not aware of that.

If the worker successfully completes the assigned HIT, they are shown the output token. The output token is used to submit the HIT and receive the payment. The payment is \$1.5 for a set of 8 statements. The average time spent to complete the task was investigated before publishing the task to relate it to the minimum US hourly wage.

Overall, the task involves 60 statements in total and each statement is evaluated by 10 distinct workers. Thus, 100 MTurk HITs have been deployed for the main experiment leading to the collection of 800 judgments in total (600 judgments plus 200 gold question answers). Over 4300 judgments from 542 workers over a total of 7 batches of the crowdsourcing task are collected, considering the main experiment and the longitudinal study altogether. The choice of making each statement evaluated by 10 distinct workers deserves a discussion. Such a number is aligned with previous studies using crowdsourcing to judge truthfulness [232, 356, 361, 363] and other concepts like relevance [267, 358]. This number is a reasonable trade-off between having fewer statements judged by many workers and more statements judged by few workers.

5.2.2 Longitudinal Study

The longitudinal study is based on the same dataset and experimental setting of the main experiment Roitero et al. [363]. The data for the main experiment (from now denoted as Batch1) has been collected on May 2020. The HITs from Batch1 have been published again with a novel set of workers (i.e., the workers of Batch1 were prevented from performing the experiment again) on June 2020. The resulting dataset is denoted as Batch2. Additional judgments have been collected in July 2020. The HITs from Batch1 have been published again with a novel set of workers (i.e., the workers of Batch1 and Batch2 were prevented to perform the experiment again). The resulting dataset is denoted as Batch3. Finally, the last set of judgments has been collected in August 2020. Again, the workers from the previous three batches were prevented from performing the experiment again. The resulting dataset is denoted as Batch4.

Then, another set of experiments is considered. For a given batch, the workers that participated in the previous one have been contacted, sending them a \$0.01 bonus and asking them to perform the task again. Table 5.1 describes the resulting datasets, where $\text{BatchX}_{\text{fromY}}$ denotes the subset of workers that performed BatchX and had previously participated in BatchY. Note that an experienced (returning) worker who does the task for the second time gets generally a new HIT assigned. In other words, a HIT is different from the one performed originally. At that time, there was no control on that matter, since HITs were assigned to workers by the Amazon Mechanical Turk platform. Finally, the union of the data from Batch1, Batch2, Batch3, and Batch4; is also considered. The resulting dataset is denoted as $\text{Batch}_{\text{all}}$.

Table 5.1: Experimental setting for the longitudinal study. All dates refer to 2020. The values reported are absolute numbers.

Date	Acronym	Number of Workers				Total
		Batch1	Batch2	Batch3	Batch4	
May	Batch1	100	–	–	–	100
June	Batch2	–	100	–	–	100
	Batch2 _{from1}	29	–	–	–	29
July	Batch3	–	–	100	–	100
	Batch3 _{from1}	22	–	–	–	22
	Batch3 _{from2}	–	20	–	–	20
	Batch3 _{from1or2}	22	20	–	–	42
August	Batch4	–	–	–	100	100
	Batch4 _{from1}	27	–	–	–	27
	Batch4 _{from2}	–	11	–	–	11
	Batch4 _{from3}	–	–	33	–	33
	Batch4 _{from1or2or3}	27	11	33	–	71
	Batch _{all}	100	100	100	100	400

5.3 Descriptive Analysis

Overall, 334 workers resident in the United States participated in the main experiment.

5.3.1 Worker Demographics

In each HIT, workers are first asked to complete a demographics questionnaire with questions about their gender, age, education and political views. The following demographic statistics are derived by analyzing the answers to the questionnaire of the workers who successfully completed the experiment. The majority of workers are in the 26–35 age range (39%), followed by 19–25 (27%), and 36–50 (22%). The majority of the workers are well educated: 48% of them have a four-year college degree or a bachelor’s degree, 26% have a college degree, and 18% have a postgraduate or professional degree. Only about 4% of the workers have a high school degree or less. Concerning political views, 33% of workers identify themselves as liberals, 26% as moderate, 17% as very liberal, 15% as conservative, and 9% as very conservative. Moreover, 52% of workers identify themselves as being Democrat, 24% as being Republican, and 23% as being Independent. Finally, 50% of workers disagree with building a wall on the southern US border, and 37% of them agree. Overall, the sample is well-balanced.

The analysis of the CRT scores shows that: 31% of workers do not provide any correct answer, 34% answer correctly to 1 question, 18% answer correctly to 2 questions and only 17% answer correctly to all 3 questions. The results of the CRT tests and the worker quality are correlated to answer RQ7.

5.3.2 Task Abandonment

The abandonment rate is measured according to the definition provided by Han et al. [174]. 100/334 workers (about 30%) successfully complete the task, 188/334 (about 56%) abandon it (i.e., voluntarily terminate the task before completing it), and 46/334 (about 7%) fail (i.e., terminate the task due to failing the quality checks too many times). Furthermore, 115/188 workers (about 61%) abandon the task before judging the first statement (i.e., before truly starting it). Furthermore, it is aligned to those of other tasks (Section 4.3 and Section 7.3).

5.4 Results

Crowd accuracy (RQ5) is addressed in Section 5.4.1. Section 5.4.2 addresses transforming judgments scales (RQ6). Section 5.4.3 studies the effect of workers' background and bias (RQ7), while their behavior (RQ8) is analyzed in Section 5.4.4. The information sources (RQ9) are studied in Section 5.4.5. The following sections focus on the longitudinal study. Section 5.4.6 addresses the impact of repeating the experiment recruiting novice workers (RQ10). Section 5.4.7 studies the effect provided by returning workers (RQ11). Finally, Section 5.4.8 studies whether the statements misjudged are the same across different batches (RQ12).

5.4.1 RQ5: Crowd Workers Accuracy

Understanding if the workers can identify misinformation about COVID-19 involves studying the quality of the judgments provided. A standard way to perform such an activity is by analyzing the external (Section 5.4.1.1) and internal agreement (Section 5.4.1.2).

5.4.1.1 External Agreement

Figure 5.1 shows the agreement between the PolitiFact experts (x-axis) and the crowd judgments (y-axis). In Figure 5.1a, each point is a judgment by a worker on a statement, i.e., there is no aggregation of the workers working on the same statement. In the remaining charts all workers redundantly working on the same statement are aggregated using the mean (Figure 5.1b), median (Figure 5.1c), and majority vote (Figure 5.1d). Focusing on Figure 5.1a (i.e., the chart with no aggregation function applied), allows seeing that the individual judgments are in agreement with the expert labels, as shown by the median values of the boxplots, which are increasing as the ground truth truthfulness level increases. Concerning the aggregated values, for all the aggregation functions the *Pants-On-Fire* and *False* categories are perceived in a very similar way by the workers. This behavior was already shown by Roitero et al. [361] and La Barbera et al. [232], and suggests that indeed workers have clear difficulties in distinguishing between the two categories. This is even more evident considering that the interface presented to the workers contained a textual description of the categories' meaning on every page of the task.

It can be seen by looking at the charts as a whole that within each chart the median values of the boxplots increase when going from *Pants-On-Fire* to *True* (i.e., going from

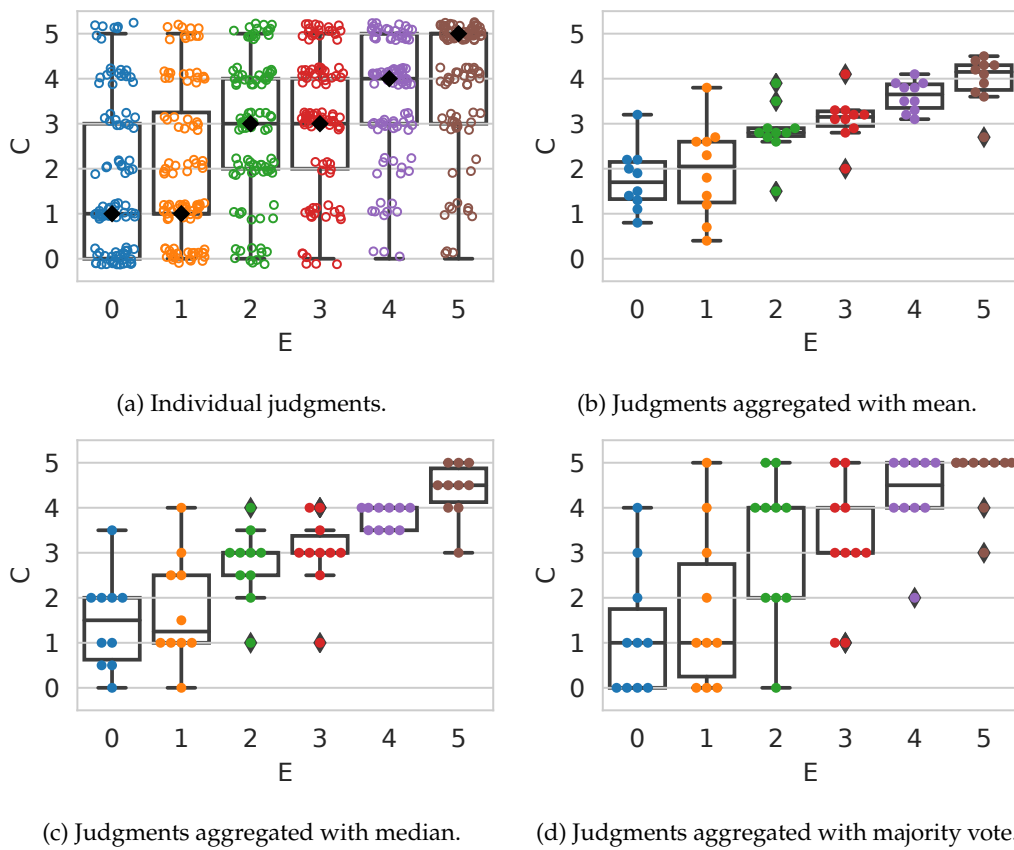


Figure 5.1: Agreement between the PolitiFact experts and the crowd judgments.

left to right of the x-axis of each chart). This indicates that the workers are overall in agreement with the PolitiFact ground truth, thus indicating that they are indeed capable of recognizing and correctly classifying misinformation statements related to the COVID-19 pandemic. This is a very important and not obvious result: in fact, the crowd (i.e., the workers) is the primary source and cause of the spread of disinformation and misinformation statements across social media platforms [71]. It appears thus evident that the mean (Figure 5.1b) is the aggregation function which leads to higher agreement levels, followed by the median (Figure 5.1c) and the majority vote (Figure 5.1d). Again, this behavior was already highlighted in previous work [232, 358, 361]. All the works cited, indeed, use the mean as the primary aggregation function.

The statistical significance between the aggregated judgments for all the six PolitiFact categories is measured to validate the external agreement. Both the Mann-Whitney rank test and the t-test are considered. The Bonferroni correction is applied to account for multiple comparisons. The difference between categories is never significant, for both tests and for all the three aggregation functions, when considering adjacent categories (e.g., *Pants-On-Fire* and *False*). The differences are never significant when considering categories of distance 2 (e.g., *Pants-On-Fire* and *Mostly-False*), apart from the median aggregation function, where there is statistical significance to the $p < .05$ level in 2/4 cases for both Mann-Whitney and t-test. The differences are significant (for the Mann-Whitney and the t-test respectively) when considering categories of distance 3 in the following cases: for the mean, 3/3 and 3/3 cases; for the median, 2/3 and 3/3 cases; for the majority vote, 0/3 and 1/3 cases. The differences are always significant to the $p > 0.01$ level for all the aggregation functions and for all the tests when considering categories of distance 4 and 5. The only exception is the majority vote function and the Mann-Whitney test, where the significance is at the $p > .05$ level. In the following, the mean is used as an aggregation function as it is the most commonly used approach for this type of data.

5.4.1.2 Internal Agreement

The internal agreement is measured using α [224] and Φ [68], two popular measures often used to compute workers' agreement in crowdsourcing tasks [269, 356, 358, 361]. The overall agreement always falls in the [0.15, 0.3] range. Agreement levels measured with the two scales are very similar for the PolitiFact categories, with the only exception of Φ , which shows higher agreement levels for the *Mostly-True* and *True* categories. This is confirmed by the fact that the α measure always falls in the Φ confidence interval, and the little oscillations in the agreement value are not always an indication of a real change in the agreement level, especially when considering α [68]. Nevertheless, Φ seems to confirm the finding derived from Figure 5.1 that workers are most effective in identifying and categorizing statements with a higher truthfulness level. This remark is also supported by Checco et al. [68] who show that Φ is better in distinguishing agreement levels in crowdsourcing than α , which is more indicated as a measure of data reliability in non-crowdsourced settings.

5.4.2 RQ6: Transforming Judgments Scales

In light of the results discussed in Section 5.4.1, the answer to RQ5 is overall positive, even if with some exceptions. There are several remarks that can be made. First, there is a clear issue that affects the *Pants-On-Fire* and *False* categories, which are very often misclassified by workers. Moreover, while *PolitiFact* used a six-level judgment scale, the usage of a scale with only two or three levels is common when judging the truthfulness of statements, as done in Chapter 4. Finally, categories can be merged together to improve accuracy, as done for example by Tchechmedjiev et al. [407]. All these remarks lead to RQ6, addressed in the following.

5.4.2.1 Merging Ground Truth Levels

The six *PolitiFact* categories (i.e., E_6) are grouped together into three (referred to as E_3) or two (E_2) categories, referred respectively as 01, 23, and 45 for the three-level scale, and 012 and 345 for the two-level scale.

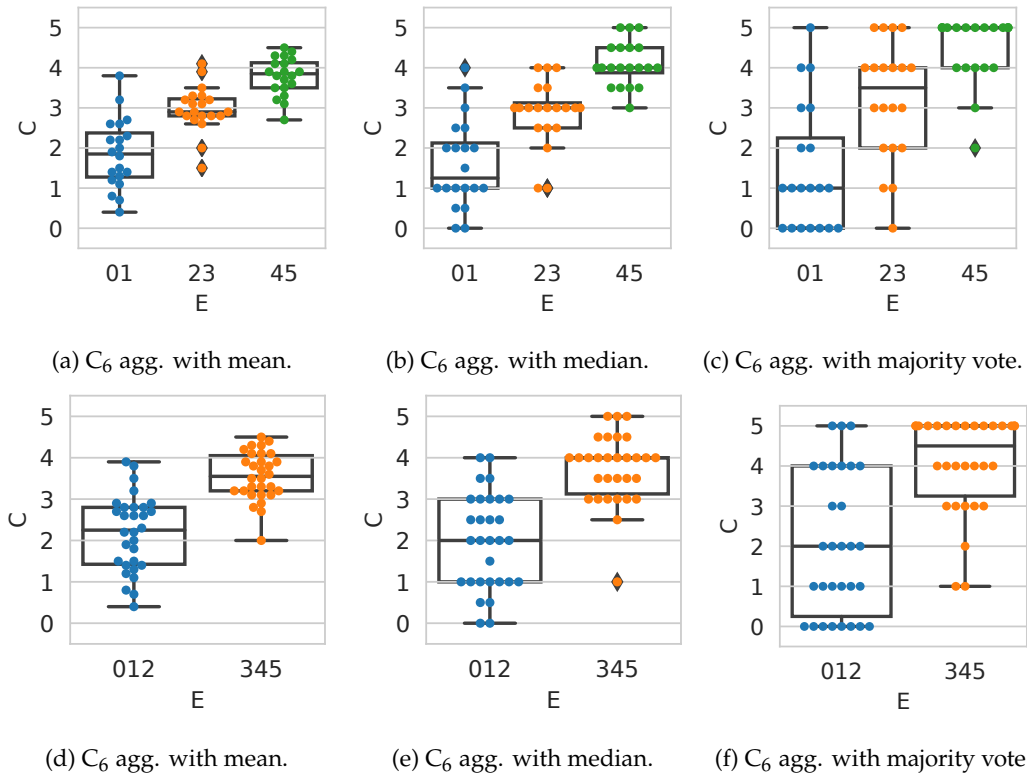


Figure 5.2: Agreement between the *PolitiFact* experts and the crowd judgments. First row: E_6 to E_3 . Second row: E_6 to E_2 . Compare with Figure 5.1.

Figure 5.2 shows the result of such a process. The agreement between the crowd and the expert judgments can be seen in a more neat way. As for Figure 5.1, the median values for all the boxplots increase when going towards higher truthfulness values (i.e., going from

left to right within each plot). This holds for all the aggregation functions considered. It is valid for both transformations of the E_6 scale, into three and two levels.

Also, in this case, the statistical significance between categories is computed by applying the Bonferroni correction to account for multiple comparisons. For the case of three groups, both the categories at distance one and two are always significant to the $p < 0.01$ level, for both the Mann-Whitney and the t-test, for all three aggregation functions. The same behavior holds for the case of two groups, where the categories of distance 1 are always significant to the $p < 0.01$ level. Summarizing, it can be concluded that merging the ground truth levels allows for obtaining a much stronger signal: the crowd can effectively detect and classify misinformation statements related to the COVID-19 pandemic.

5.4.2.2 Merging Crowd Levels

The crowd judgments (i.e., C_6) can be merged as done for the ground truth labels. In more detail, the judgments are merged into either three (referred to as C_3) or two (C_2) categories. The transformation process relies on the approach detailed by Han et al. [173]. This approach has many advantages. It allows simulating the effect of having the crowd judgments in a more coarse-grained scale (rather than C_6), and thus simulating new data without running the whole experiment on Amazon Mechanical Turk again. The following experiment is thus performed.

All the possible cuts¹ from C_6 to C_3 and from C_6 to C_2 are performed. Then, the internal agreement is measured (using α and Φ) both on the source and on the target scale. Those values are thus compared. In such a way, it is possible identifying among all the possible cuts the cut which lead to the highest possible internal agreement. For the C_6 to C_3 transformation, there is a single cut which leads to higher agreement levels with the original C_6 scale both for α and Φ . On the contrary, for the C_6 to C_2 transformation, there is a single cut for α which leads to similar agreement levels as in the original C_6 scale, and there are no cuts with such property when using Φ . Having identified the best possible cuts for both transformations and for both agreement metrics, the external agreement is measured between the crowd and the expert judgments, using the selected cut.

Figure 5.3 shows such a result when considering the judgments aggregated with the mean function. It is again the case that the median values of the boxplots are always increasing, for all the transformations. Nevertheless, inspecting the plots allows stating that the overall external agreement appears to be lower than the one shown in Figure 5.1. Moreover, even using these transformed scales the categories *Pants-On-Fire* and *False* are still not separable. Summarizing, it is feasible to transform the judgments collected on a C_6 level scale into two new scales, namely C_3 and C_2 . The judgments obtained have a similar internal agreement as the original ones and a slightly lower external agreement with the expert judgment.

5.4.2.3 Merging Both Ground Truth And Crowd Levels

It is now natural to combine the two approaches. Figure 5.4 shows the comparison between C_6 transformed into C_3 and C_2 , and E_6 transformed into E_3 and E_2 . Also in this

¹ C_6 can be transformed into C_3 in 10 different ways, and C_6 can be transformed into C_2 in 5 different ways.

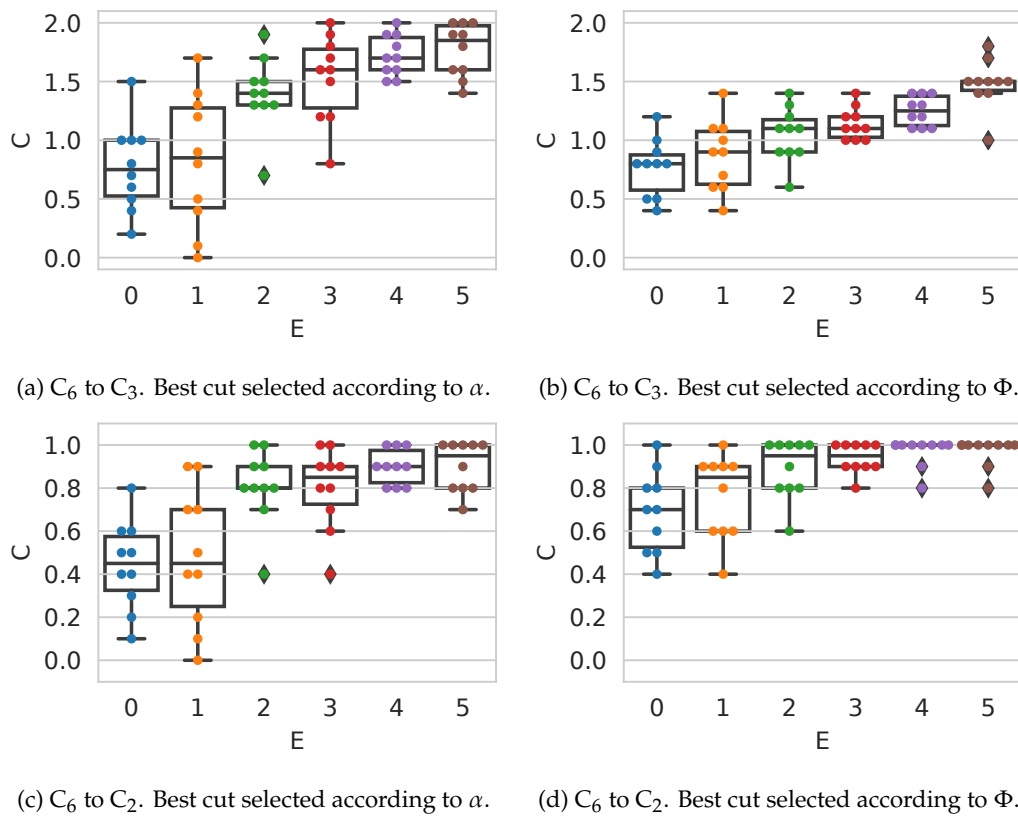


Figure 5.3: Crowd judgments merged into groups and then aggregated with the mean function. Comparison with E_6 . Compare with Figure 5.1.

case, the median values of the boxplots are increasing, especially for the E_3 case (Figure 5.4a and Figure 5.4b). Furthermore, the external agreement with the ground truth is present, even if for the E_2 case (Figure 5.4c and Figure 5.4d) the classes appear to be not separable. Summarizing, all these results show that it is feasible to successfully combine the aforementioned approaches, and transform them into a three-level and two-level scale for both the crowd and the expert judgments.

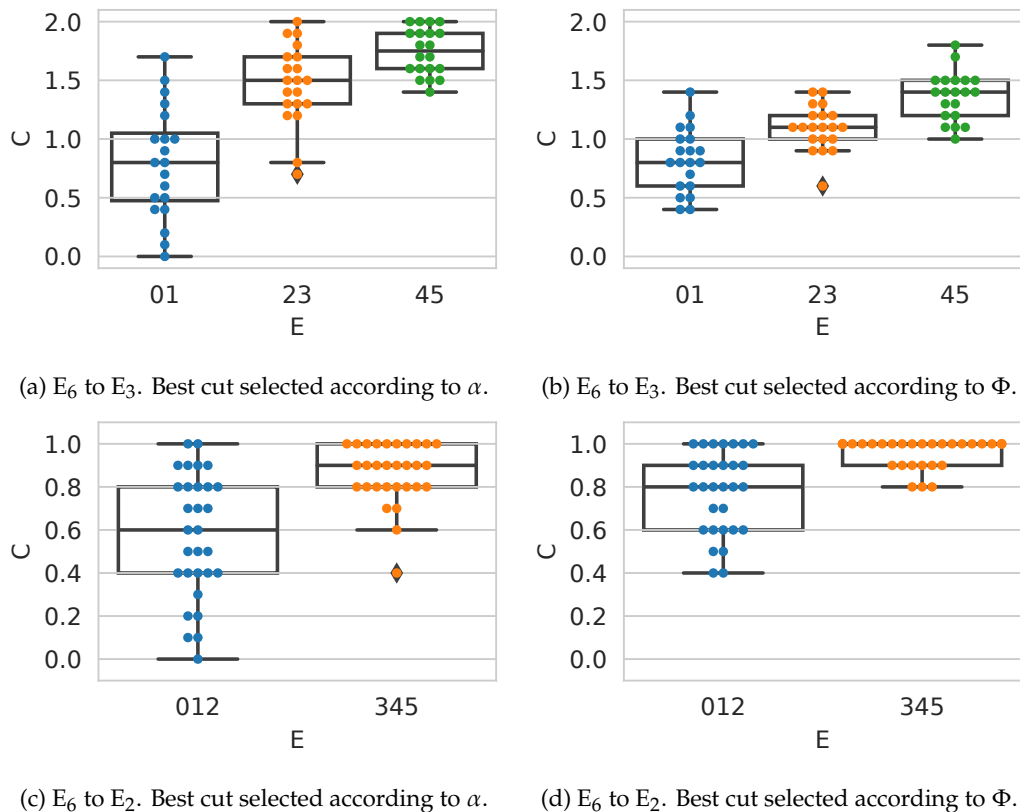


Figure 5.4: Expert judgments merged into groups and then aggregated with the mean function. Compare with Figure 5.2, Figure 5.3, and Figure 5.3.

5.4.3 RQ7: Worker Background And Bias

Previous work has shown that political and personal biases as well as cognitive abilities have an impact on the workers' quality [232, 361]. Recent articles have shown that the same effect might apply also to fake news [406]. For this reason, it is reasonable to investigate if workers' political biases and cognitive abilities influence their quality in the setting of misinformation related to COVID-19.

When looking at the questionnaire answers, there is a relation with the workers' quality only when considering the answer to the workers' political views. In more detail, the quality of workers in each group is measured using Accuracy (i.e., the fraction of exactly classified

statements). The number and fraction of correctly classified statements are however rather crude measures of worker's quality, as small misclassification errors (e.g., *Pants-On-Fire* in place of *False*) are as important as more striking ones (e.g., *Pants-On-Fire* in place of *True*). Therefore, to measure the ability of workers to correctly classify the statements, the Closeness Evaluation Measure (CEM^{ORD}) is also computed. It is an effectiveness metric proposed for the specific case of ordinal classification [15] (see Section 4.2.2 for a more detailed discussion of these issues). The accuracy and CEM^{ORD} values are respectively of 0.13 and 0.46 for "Very conservative", 0.21 and 0.51 for "Conservative", 0.20 and 0.50 for "Moderate", 0.16 and 0.50 for "Liberal", and 0.21 and 0.51 for "Very liberal". By looking at both Accuracy and CEM^{ORD} , it is clear that "Very conservative" workers provide lower quality judgments. The Bonferroni corrected two-tailed t-test on CEM^{ORD} confirms that "Very conservative" workers perform statistically significantly worse than both "Conservative" and "Very Liberal" workers. The workers' political views affect the CEM^{ORD} score, even if in a small way and mainly when considering the extremes of the scale. An initial analysis of the other answers to the questionnaire (not shown) does not seem to provide strong signals.

The effect of the CRT tests on worker quality is also investigated. Although there is a small variation in both Accuracy and CEM^{ORD} (not shown), this is never statistically significant. It appears that the number of correct answers to the CRT tests is not correlated with worker quality.

5.4.4 RQ8: Worker Behavior

The workers provide multiple behavioral signals while performing the misinformation assessment task. Such signals can be derived from the time spent by the workers on each statement and their query patterns (Section 5.4.4.1). It is thus worth understanding whether these signals can be used to improve the work performed (Section 5.4.4.2).

5.4.4.1 Time And Queries

Table 5.2 (first two rows) shows the amount of time spent on average by the workers on the statements and their CEM^{ORD} score. The time spent on the first statement is considerably higher than on the last statements, and overall the time spent by the workers almost monotonically decreases while the statement position increases. This, combined with the fact that the quality of the judgment provided by the workers (measured with CEM^{ORD}) does not decrease for the last statements indicates a learning effect: the workers learn how to judge truthfulness in a faster way.

Table 5.2 (third and fourth row) shows query statistics for the 100 workers who finished the task. The total and average number of queries are respectively 2095 and 262, while the total and average number of statements as a query are respectively 245 and 30.6. The higher the statement position, the lower the number of queries issued: 3.52% on average for the first statement down 2.30% for the last statement. This can indicate the attitude of workers to issue fewer queries the more time they spend on the task, probably due to fatigue, boredom, or learning effects. Nevertheless, it can be seen that on average, for all the statement positions each worker issues more than one query. In other words, workers often reformulate their initial query. This provides further evidence that they put the effort into

performing the task and that suggests the overall high quality of the collected judgments. The third row of the table shows the number of times the worker uses to query the whole statement. The percentage is rather low (around 13%) for all the statement positions, indicating again that workers put some effort when providing their judgments.

Table 5.2: Statement position in the task versus: time elapsed, cumulative on each single statement (first row), CEM^{ORD} (second row), number of queries issued (third row), and number of times the statement has been used as a query (fourth row).

Statement Position	1	2	3	4	5	6	7	8
Time (sec)	299	282	218	216	223	181	190	180
CEM^{ORD}	0.63	0.618	0.657	0.611	0.614	0.569	0.639	0.655
Number of Queries	352 16.8%	280 13.4%	259 12.4%	255 12.1%	242 11.6%	238 11.3%	230 11.0%	230 11.4%
Statement as Query	22 9%	32 13%	31 12.6%	33 13.5%	34 13.9%	30 12.2%	29 11.9%	34 13.9%

5.4.4.2 Exploiting Worker Signals to Improve Quality

The workers provide many signals that to some extent correlate with the quality of their work while performing the task. These signals could in principle be exploited to aggregate the individual judgments in a more effective way (i.e., giving more weight to workers that possess features indicating a higher quality). For example, the relationships between worker background/bias and worker quality (Section 5.4.3) could be exploited to this aim.

The following experiment is thus performed. The C_6 individual scores are aggregated using a weighted mean. The weights are either represented by the political views or the number of correct answers to CRT, both normalized in $[0.5, 1]$. There is a behavior very similar to the one observed in Figure 5.4b. Leveraging quality-related behavioral signals like questionnaire answers or CRT scores to aggregate the judgments collected does not provide a noticeable increase in the external agreement, although it does not have any negative effect.

5.4.5 RQ9: Sources Of Information

The sources of information used by the workers while performing the task are analyzed by considering the URLs distribution (Section 5.4.5.1) and the textual justifications written to support the evidence found (Section 5.4.5.2).

5.4.5.1 URLs Analysis

Figure 5.5 shows the distribution of the ranks of the URL selected as evidence by the worker when providing each judgment. The URLs selected less than 1% times are filtered out from the results. About 40% of workers select the first result retrieved by the custom

search engine and select the remaining positions less frequently, with an almost monotonic decreasing frequency (rank 8 makes the exception). Furthermore, 14% of workers inspect up to the fourth page of results (i.e., rank= 40). The breakdown on the truthfulness levels of PolitiFact does not show any significant difference.

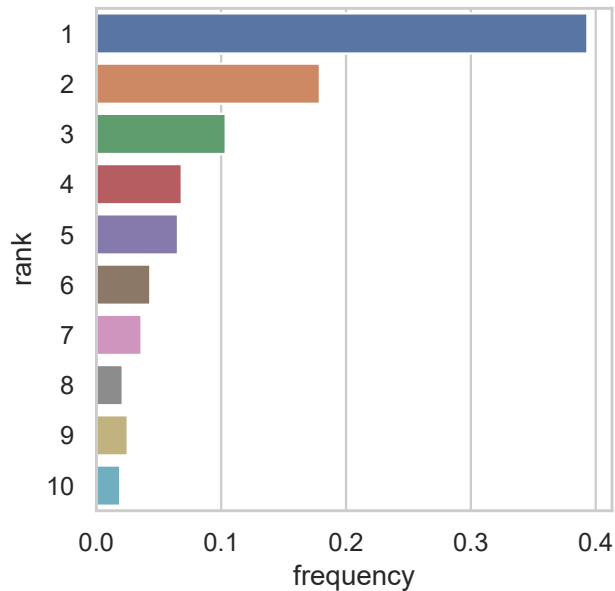


Figure 5.5: Distribution of the ranks of the URLs selected by workers.

Table 5.3 reports the top 10 websites from which the workers choose the URL to justify their judgments. Websites with percentage $\leq 3.9\%$ are filtered out. There are many fact-check websites among the top 10 URLs (e.g., snopes: 11.79%, factcheck 6.79%). Medical websites are also present (cdc: 4.29%). This indicates that workers use various kinds of sources as URLs from which they take the information. Thus, they put effort into finding evidence to provide a reliable truthfulness judgment.

Table 5.3: Websites from which workers chose URLs to justify their judgments.

URL	Percentage
snopes.com	11.79%
msn.com	8.93%
factcheck.org	6.79%
wral.com	6.79%
usatoday.com	5.36%
statesman.com	4.64%
reuters.com	4.64%
cdc.gov	4.29%
mediabiasfactcheck.com	4.29%
businessinsider.com	3.93%

5.4.5.2 Justifications

The textual justifications provided by the workers, their relations with the web pages at the selected URLs, and their links with worker quality are also analyzed. 54% of the justifications provided contains text copied from the web page available at the URL selected as evidence, while 46% do not. Furthermore, 48% of the justification include some “free text” (i.e., text generated and written by the worker), and 52% do not. Considering all the possible combinations:

- 6% of the justifications use both free text and text from web pages.
- 42% of the justifications use free text but no text from the web page.
- 48% of the justifications use no free text but only text from web pages.
- 4% use neither free text nor text from the web page, and either insert text from a different (not selected) web page or insert part of the instructions provided to perform the task or text from the user interface.

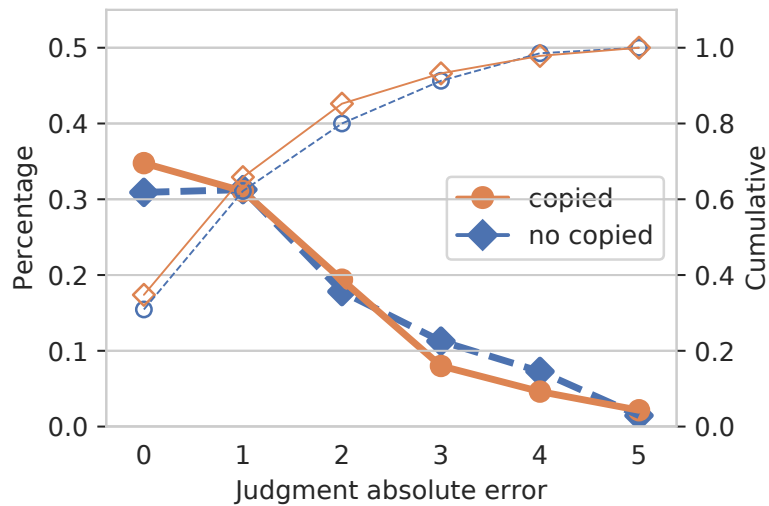
Each worker seems to have a clear attitude concerning the preferred way to provide justifications:

- 48% of the workers use only text copied from the selected web pages.
- 46% of the workers use only free text.
- 4% of the workers use both free and copied text.
- 2% of the workers consistently provide text from the user interface or random internet pages.

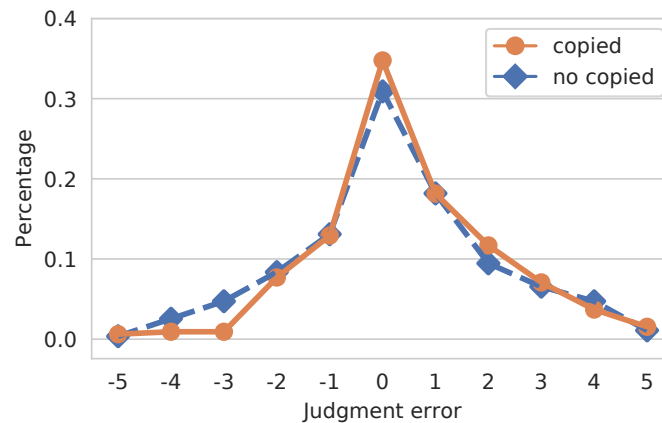
Such a behavior can be correlated with worker quality. Figure 5.6 shows the relations between different kinds of justifications and the worker accuracy. Figure 5.6a shows the absolute value of the prediction error. Figure 5.6b shows the prediction error itself. The figure shows whether the text inserted by the worker was copied or not from the web page selected. The same analysis is performed by considering if the worker used or not free text, but the results are almost identical to the former analysis.

As it can be seen, statements on which workers make fewer errors (i.e., where x-axis = 0) tend to use a text copied from the web page selected. On the contrary, statements on which workers make more errors (values close to 5 in Figure 5.6a, and values close to +/- 5 in Figure 5.6b) tend to use text not copied from the selected web page. The differences are small but might indicate that workers of higher quality tend to read the text from the selected web page and report it in the justification box. To confirm this result, the CEM^{ORD} scores are computed for the two classes considering the individual judgments. The class “copied” has $CEM^{ORD} = 0.640$, while the class “not copied” has a lower value, $CEM^{ORD} = 0.600$. Such behavior is consistent with what concerns the usage of free text. Statements on which workers make fewer errors tend to use more free text than the ones that make more errors. This indicates that workers which add free text as a justification, possibly reworking the information present in the selected URL, are of a higher quality. In this case, the CEM^{ORD} measure confirms that the two classes are very similar. The class “free text” has $CEM^{ORD} = 0.624$, while the class “not free text” has $CEM^{ORD} = 0.621$. The right part of Figure 5.6 shows that the distribution of the prediction error is not symmetrical, as the frequency of the errors is higher on the positive side of the x-axis ([0,5]). These errors correspond to workers overestimating the truthfulness value of the statement (with 5 being the result of labeling

a Pants-On-Fire statement as True). It is also noticeable that the justifications containing text copied from the selected URL have a lower rate of errors in the negative range, meaning that workers which directly quote the text avoid underestimating the truthfulness of the statement.



(a) Absolute value of the prediction error (cumulative distributions shown with thinner lines and empty markers).



(b) Prediction error.

Figure 5.6: The effect of the origin of a justification on worker accuracy. Text copied/not copied from the selected URL.

5.4.6 RQ10: Repeating The Experiment With Novice Workers

Studying the effect of repeating the experiment in the longitudinal study by recruiting novice workers involves addressing the whole set of analyses done for the base crowdsourc-

ing experiment (Section 5.2.1) and described in the previous subsections across different batches. Section 5.4.6.1 studies the variation in the composition of the worker population. Section 5.4.6.2 studies the quality of both individual and aggregated judgments. Section 5.4.6.3 analyzes the external agreement, while Section 5.4.6.4 the internal one. Section 5.4.6.5 addresses the time spent by the novice workers and their querying behavior. Section 5.4.6.6 analyzes the URLs distribution. Finally, Section 5.4.6.7 studies the effect of using different kinds of justifications on worker accuracy.

5.4.6.1 Worker Background, Behavior, Bias, And Abandonment

The variation in the composition of the worker population across different batches is studied using a General Linear Mixture Model (GLMM) [277] together with the Analysis Of Variance (ANOVA) [291]. This allows for analyzing how worker behavior changes across batches and measuring the impact of such changes. In more detail, the ANOVA effect size ω^2 is considered. It is an unbiased index used to provide insights into the population-wide relationship between a set of factors and the studied outcomes [141, 142, 143, 355, 466]. A linear model is fitted using such a setting to measure the effect of age, school, and all other possible answers to the questions in the questionnaire. Inspecting the ω^2 index allows finding that the largest effects are provided by workers' answers to the taxes and southern border questions, while all the effects are either small or non-present [314]. The effect of the batch is small but not negligible and is on the same order of magnitude as the effect of other factors. The interaction plots (see for example [97]) are computed considering the variation of the factors from the previous analysis on the different batches. Results suggest a small or not significant [130] interaction between the batch and all the other factors. This analysis suggests that, while the difference among different batches is present, the population of workers that performs the task is homogeneous, and thus the different datasets (i.e., batches) are comparable.

Table 5.4 shows the abandonment data for each batch of the longitudinal study, indicating the number of workers which complete, abandon, or fail the task (due to not satisfying the quality checks). Overall, the abandonment ratio is quite well balanced across batches, with the only exception of Batch3, which shows a small increase in the number of workers who failed the task. Nevertheless, such a small variation is not significant and might be caused by a slightly lower quality of workers which started Batch3. On average, Table 5.4 shows that 31% of the workers complete the task, 50% abandon it, and 19% fail the quality checks; these values are aligned with those of Section 4.3.

5.4.6.2 Agreement Across Batches

Measuring the correlation between individual judgments shows rather low correlation values:

- The correlation between Batch1 and Batch2 is of $\rho = 0.33$ and $\tau = 0.25$.
- The correlation between Batch1 and Batch3 is of $\rho = 0.20$ and $\tau = 0.14$, between Batch1 and Batch4 is of $\rho = 0.10$ and $\tau = 0.074$.
- The correlation between Batch2 and Batch3 is of $\rho = 0.21$ and $\tau = 0.15$, between Batch2 and Batch4 is of $\rho = 0.10$ and $\tau = 0.085$.

Table 5.4: Abandonment data for each batch of the longitudinal study.

Acronym	Number of Workers			Total
	Complete	Abandon	Fail	
Batch1	100 (30%)	188 (56%)	46 (14%)	334
Batch2	100 (37%)	129 (48%)	40 (15%)	269
Batch3	100 (23%)	220 (51%)	116 (26%)	436
Batch4	100 (36%)	124 (45%)	54 (19%)	278
Average	100 (31%)	165 (50%)	64 (19%)	1317

- The correlation values between Batch3 and Batch4 is of $\rho = 0.08$ and $\tau = 0.06$.

Overall, the most recent batch (Batch4) is the batch which achieves the lowest correlation values w.r.t. the other batches, followed by Batch3. The highest correlation is achieved between Batch1 and Batch2. This preliminary result suggests that it might be the case that the time span in which the judgments of the different batches have been collected has an impact on their similarity across batches, and batches which have been launched in time spans close to each other tend to be more similar than other batches.

The aggregated judgments are analyzed to study if such a relationship is still valid when individual judgments are aggregated. Figure 5.7 shows the agreement between the aggregated judgments of Batch1, Batch2, Batch3, and Batch4. The figure shows in the diagonal the distribution of the aggregated judgments. The lower triangle shows the scatterplot between the aggregated judgments of the different batches, while the upper triangle the corresponding ρ and τ correlation values. The charts show that the correlation values of the aggregated judgments are greater than the ones measured for individual judgments. This is consistent for all the batches. In more detail, the agreement between Batch1 and Batch2 ($\rho = 0.87$, $\tau = 0.68$) is greater than the agreement between any other pair of batches. Furthermore, the correlation values between Batch1 and Batch3 are similar to the agreement between Batch2 and Batch3. Furthermore, it is again the case the Batch4 achieves lower correlation values with all the other batches. Overall, these results show that:

1. individual judgments are different across batches, but they become more consistent across batches when they are aggregated;
2. the correlation seems to show a trend of degradation, as early batches are more consistent with each other than more recent batches
3. it also appears that batches which are closer in time are also more similar.

5.4.6.3 Crowd Workers Accuracy: External Agreement

Concerning the external agreement, Figure 5.8 shows the agreement between the PolitiFact experts (x-axis) and the crowd judgments (y-axis) for Batch1, Batch2, Batch3, Batch4, and Batch_{all}. The judgments are aggregated using the mean. Overall, the individual judgments are in agreement with the expert labels, as shown by the median values of the

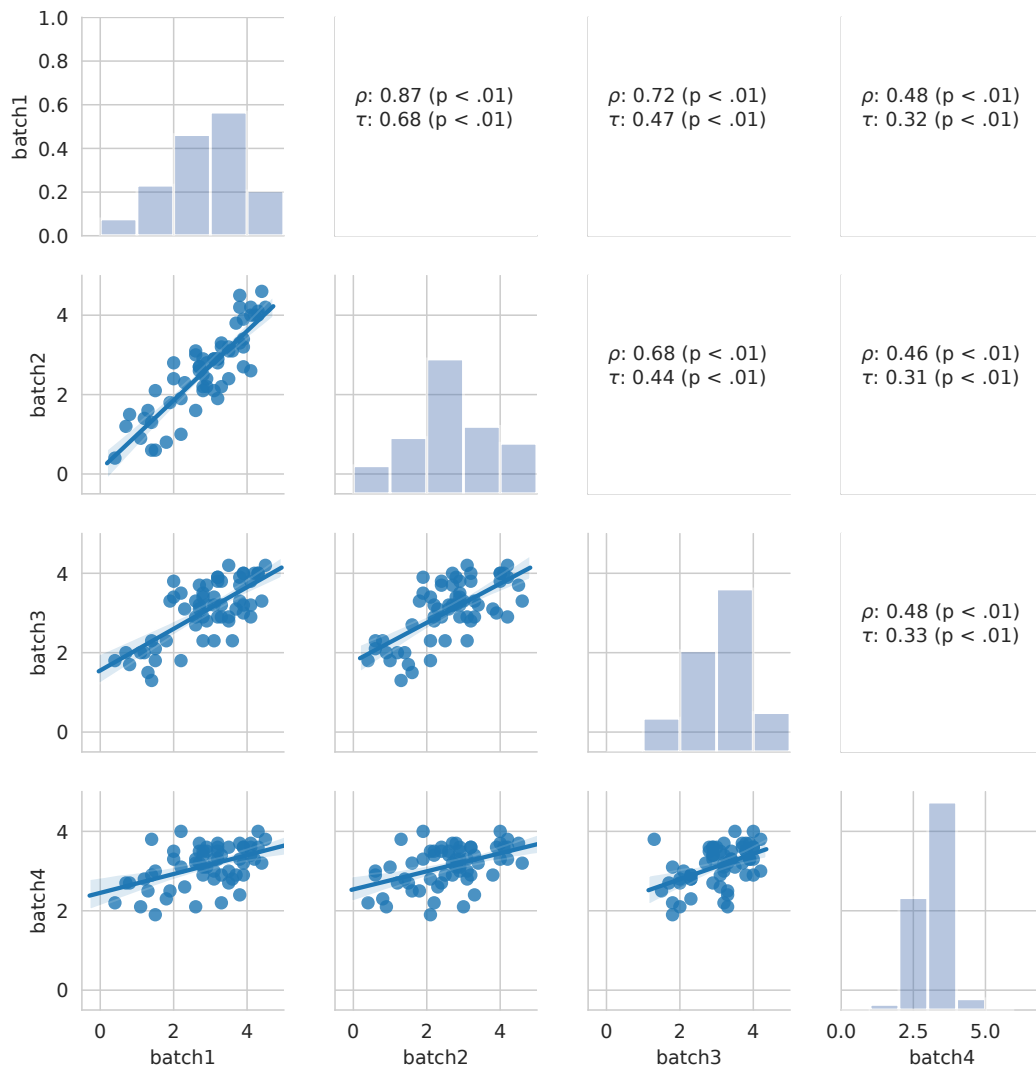


Figure 5.7: Correlation values between the judgments (aggregated using the mean) across Batch1, Batch2, Batch3, and Batch4.

boxplots, which are increasing as the ground truth truthfulness level increases. Nevertheless, Batch1 and Batch2 show higher agreement levels with the expert labels than Batch3 and Batch4.

Furthermore, as already noted in Figure 5.1b, the *Pants-On-Fire* and *False* categories are perceived in a very similar way by the workers for all the aggregation functions. This again suggests that workers have clear difficulties in distinguishing between the two categories. Furthermore, the median values of each boxplot are increasing when going from *Pants-On-Fire* to *True* (i.e., going from left to right of the x-axis of each chart), except Batch3 and in a more evident way Batch4. This indicates that, overall, the workers are in agreement with the PolitiFact ground truth and that this is true when repeating the experiment at different time spans. Nevertheless, there is an unexpected behavior: the data for the batches is collected across different time spans. Thus, it seems intuitive that the more time passes, the more the workers should be able to recognize the true category of each statement (for example by seeing it online or reported on the news). Figure 5.8 however tells a different story. It appears that the more time passes, the less agreement is found between the crowd-collected labels and the experts' ground truth. This behavior can be caused by many factors, which are discussed in the next sections.

Finally, Figure 5.8e allows understanding that $\text{Batch}_{\text{all}}$ shows a behavior which is similar to Batch1 and Batch2. This indicates that the median values of the boxplots are increasing going from left to right of the x-axis of each chart, apart from the *Pants-On-Fire* and *False* categories. Thus, also in this case, the workers are in agreement with the PolitiFact ground truth.

The presence of differences in how the statements are evaluated across different batches can be seen from the previous analyses. The ρ , τ , and rank-biased overlap (RBO) [443] correlation coefficients are computed between the scores aggregated using the mean as aggregation function, among batches, for the PolitiFact categories. This allows for investigating if the same statements are ordered in a consistent way over the different batches. The RBO parameter such as the top-5 results get about 85% of the weight of the evaluation. Table 5.5 shows the ρ and τ correlation scores, while Table 5.6 shows the bottom- and top-heavy RBO correlation scores. Given that statements are sorted by their aggregated score in decreasing order, the top-heavy version of RBO emphasizes the agreement on the statements which are misjudged for the *Pants-On-Fire* and *False* categories. On the contrary, the bottom-heavy version of RBO emphasizes the agreement on the statements which are misjudged for the *True* category.

As it can be observed by inspecting Table 5.5 and Table 5.6, there is a rather low agreement between how the same statements are judged across different batches, both when considering the absolute values (i.e., when considering ρ), and their relative ranking (i.e. when considering both τ and RBO). Focusing on the RBO metric allows seeing that, in general, the statements which are misjudged are different across batches, with the exceptions of the ones in the *False* category for Batch1 and Batch2 (RBO top-heavy = 0.85), and the ones in the *True* category, again for the same two batches (RBO bottom-heavy = 0.92). This behavior holds also for statements which are correctly judged by workers: in fact, we observe an RBO bottom-heavy correlation value of 0.81 for *False* and an RBO top-heavy correlation value of 0.5 for *True*. This is another indication of the similarities between Batch1 and Batch2.

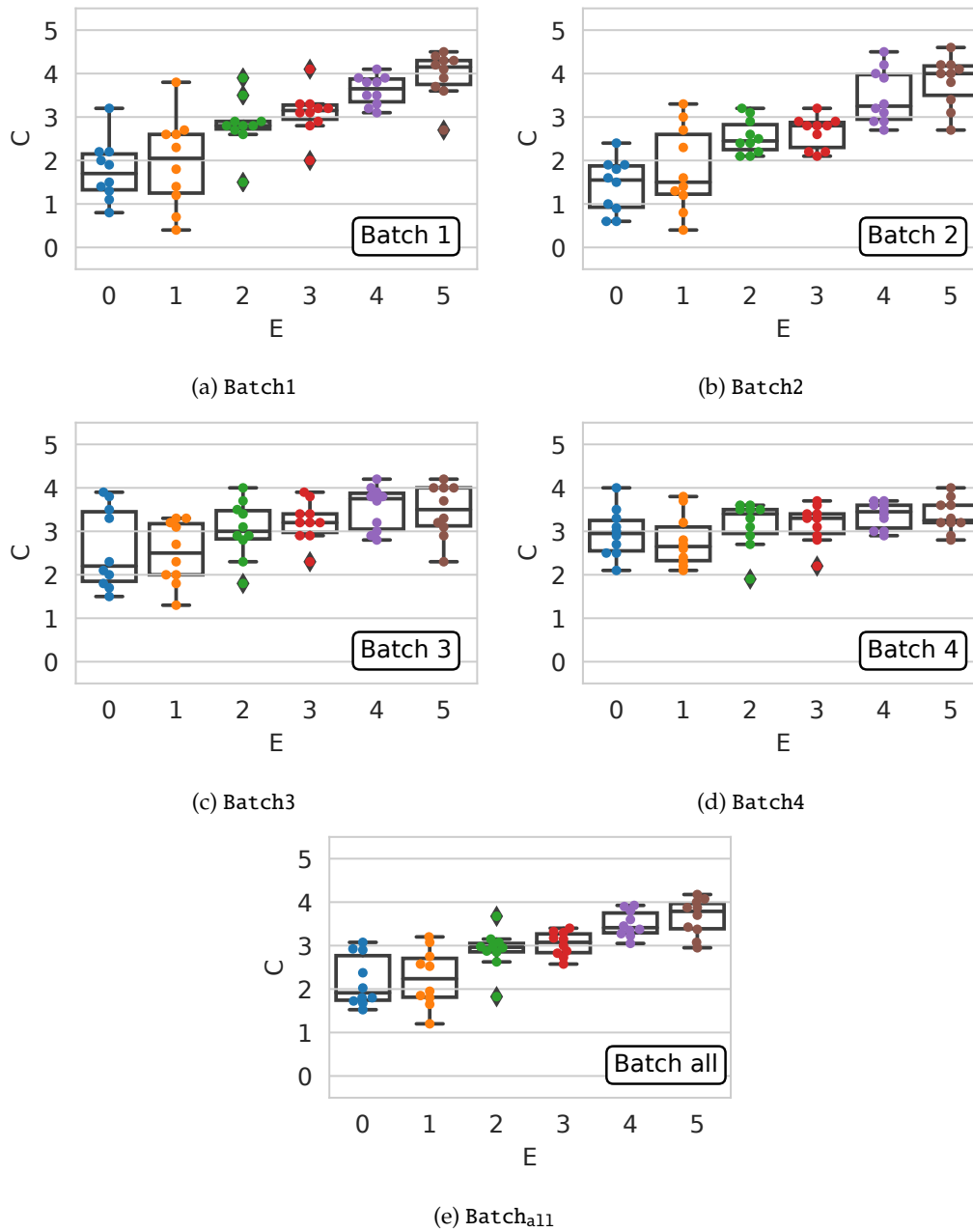


Figure 5.8: Agreement between the PolitiFact experts and crowd judgments. Figure 5.8a is the same as Figure 5.1b.

Table 5.5: ρ (lower triangle) and τ (upper triangle) correlation values among batches for the aggregated scores of Figure 5.8.

Pants-On-Fire (0)					False (1)				
	B1	B2	B3	B4		B1	B2	B3	B4
B1	–	0.37	0.58	0.54	B1	–	0.72	0.74	0.04
B2	0.44	–	0.3	0.25	B2	0.87	–	0.75	0.02
B3	0.74	0.69	–	0.42	B3	0.84	0.85	–	-0.2
B4	0.58	0.24	0.46	–	B4	-0.01	-0.07	-0.29	–

Mostly-False (2)					Half-True (3)				
	B1	B2	B3	B4		B1	B2	B3	B4
B1	–	0.07	0.47	0.51	B1	–	0.12	0.12	0
B2	0.46	–	0.37	0.09	B2	-0.03	–	0.52	0.22
B3	0.72	0.49	–	0.58	B3	0.01	0.7	–	0.1
B4	0.82	0.36	0.83	–	B4	0.09	0.28	0.2	–

Mostly-True (4)					True (5)				
	B1	B2	B3	B4		B1	B2	B3	B4
B1	–	0.35	0.16	0.24	B1	–	0.74	0.51	0.48
B2	0.6	–	-0.07	0.69	B2	0.9	–	0.26	0.28
B3	0.31	0.03	–	-0.28	B3	0.33	0.31	–	0.67
B4	0.24	0.62	-0.22	–	B4	0.51	0.45	0.69	–

Table 5.6: RBO bottom-heavy (lower triangle) and RBO top-heavy (upper triangle) correlation values among batches for the aggregated scores of Figure 5.8. Document sorted by increasing aggregated score.

Pants-On-Fire (0)					False (1)				
	B1	B2	B3	B4		B1	B2	B3	B4
B1	–	0.47	0.79	0.51	B1	–	0.85	0.86	0.36
B2	0.31	–	0.54	0.6	B2	0.81	–	0.98	0.24
B3	0.49	0.27	–	0.51	B3	0.53	0.47	–	0.23
B4	0.5	0.28	0.32	–	B4	0.34	0.41	0.33	–
Mostly-False (2)					Half-True (3)				
	B1	B2	B3	B4		B1	B2	B3	B4
B1	–	0.62	0.7	0.43	B1	–	0.26	0.26	0.22
B2	0.62	–	0.74	0.34	B2	0.26	–	0.47	0.25
B3	0.71	0.74	–	0.59	B3	0.29	0.75	–	0.64
B4	0.71	0.64	0.76	–	B4	0.22	0.51	0.36	–
Mostly-True (4)					True (5)				
	B1	B2	B3	B4		B1	B2	B3	B4
B1	–	0.33	0.28	0.43	B1	–	0.5	0.79	0.49
B2	0.48	–	0.22	0.78	B2	0.92	–	0.29	0.38
B3	0.28	0.18	–	0.15	B3	0.49	0.41	–	0.49
B4	0.39	0.88	0.17	–	B4	0.49	0.44	0.79	–

Table 5.7: Correlation between α and Φ values. ρ in the lower triangle, τ in the upper triangle.

	α					Φ			
	B1	B2	B3	B4		B1	B2	B3	B4
B1	–	0.49	0.61	0.52	B1	–	0.25	0.13	-0.03
B2	0.72	–	0.42	0.39	B2	0.38	–	0.15	0.04
B3	0.79	0.67	–	0.57	B3	0.19	0.23	–	0.06
B4	0.67	0.55	0.78	–	B4	-0.06	0.05	0.09	–

5.4.6.4 Crowd Workers Accuracy: Internal Agreement

Concerning the internal agreement, Table 5.7 shows the agreement measured with α [224] and Φ [68] for the different batches. The lower triangular part of the table shows the correlation measured using ρ , and the upper triangular part shows the correlation obtained with τ . α and Φ values are used to compute the correlation values on all PolitiFact categories. For the sake of computing the correlation values on Φ consider only the mean value and not the upper 97% and lower 3% confidence intervals.

Table 5.7 shows that the highest correlation values are obtained between Batch1 and Batch3 when considering α , and between Batch1 and Batch2 when considering Φ . Furthermore, it can be seen that Φ leads to obtaining in general lower correlation values, especially for Batch4, which shows a correlation value of almost zero with the others batches. This is an indication that Batch1 and Batch2 are the two most similar batches (at least according to Φ) and that the other two batches (i.e., Batch3) and especially Batch4, are composed of judgments made by workers with different internal agreement levels.

5.4.6.5 Worker Behavior: Time and Queries

Analyzing the amount of time spent by the workers for each position of the statement in the task confirms something already highlighted in Section 5.4.4. The amount of time spent on average by the workers on the first statements is considerably higher than the time spent on the last statements, for all the batches. This is a confirmation of a learning effect: the workers learn how to assess truthfulness in a faster way as they spend time performing the task. Furthermore, the average time spent on all documents decreases substantially as the number of batches increases. The average time spent for the four batches on each document is respectively of 222, 168, 182 and 140 seconds. A statistical test is performed between each pair of batches it is significant at each comparison, with the only exception of Batch2 when compared against Batch3. Such decreasing time might indeed be a cause for the degradation in quality observed while the number of batches increases. If workers spend on average less time on each statement, it can be assumed that they spend less time thinking before judging each statement, or that they spend less time searching for an appropriate and relevant source of evidence.

To further investigate the cause for such quality decrease in more recent batches, the querying behavior of the workers is inspected for the different batches. The number of queries issued by each worker shows that the trend to use a decreasing number of queries

as the statement position increases is still present, although less evident (but not in a significant way) for Batch2 and Batch3. Thus, it is still possible to state that the attitude of workers to issue fewer queries the more time they spend on the task holds, probably due to fatigue, boredom, or learning effects.

Furthermore, it is again the case that on average, for all the statement positions, each worker issues more than one query. In other words, workers often reformulate their initial query. This provides further evidence that they put the effort into performing the task and suggests an overall high quality of the collected judgments. Finally, only a small fraction of queries (i.e., less than 2% for all batches) correspond to the statement itself. This suggests that the vast majority of workers put significant effort into the task of writing queries, which might be assumed as an indication of their willingness to perform high-quality work.

5.4.6.6 Sources of Information: URL Analysis

Figure 5.9 shows the rank distributions of the URLs selected as evidence by the workers when performing each judgment. As for Figure 5.5 URLs selected less than 1% of the times are filtered out from the results. The trend is similar for Batch1 and Batch2, while Batch3 and Batch4 display different behavior. For Batch1 and Batch2 about 40% of workers select the first result retrieved by the search engine, and select the results down the rank less frequently: about 30% of workers from Batch2 and less than 20% of workers from Batch3 select the first result retrieved by the search engine. The behavior of workers from Batch3 and Batch4 is more oriented towards a model where the user clicks randomly on the retrieved list of results; moreover, the spike which occurs in correspondence of the ranks 8, 9, and 10 for Batch4 can be caused by the fact that workers from such batch scroll directly down the user interface with the aim of finishing the task as fast as possible, without putting any effort in providing meaningful sources of evidence.

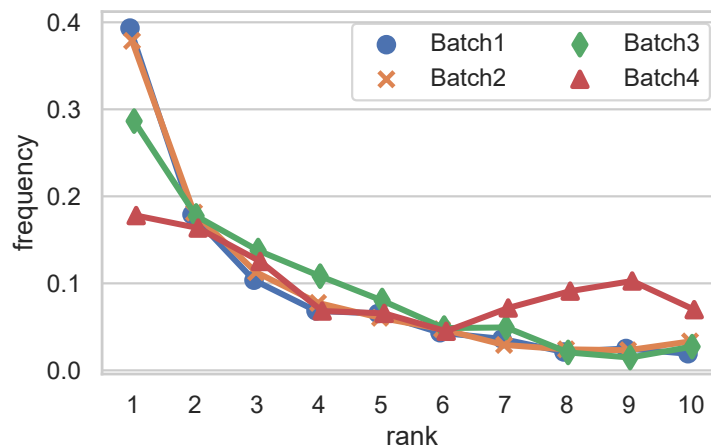


Figure 5.9: Distribution of the ranks of the URLs selected by workers for all the batches.

To provide further insights into the observed change in worker behavior associated with the usage of the custom search engine, the sources of information provided by the workers

as justification for their judgments are investigated. Investigating the top 10 websites from which the workers choose the URL to justify their judgments shows that, similarly to Table 5.3, it is again the case that there are many fact check websites among the top 10 URLs: `snopes` is always the top ranked website, and `factcheck` is always present within the ranking. The only exception is Batch4, in which each fact-checking website appears in lower rank positions. Furthermore, medical websites such as `cdc` are present only in two batches out of four (i.e., Batch1 and Batch2) and the Raleigh area news website `wral` is present in the top positions in all batches apart from Batch3: this is probably caused by the location of workers which is different among batches and they use different sources of information. Overall, such analysis confirms that workers tend to use various kinds of sources as URLs from which they take information, confirming that it appears that they put the effort into finding evidence to provide reliable truthfulness judgments.

As further analysis, the amount of change in the URLs as retrieved by the custom search engine is investigated, focusing in particular on the inter- and intra-batch similarity. To such an end, the subset of judgments for which the statement is used as a query is selected. The remaining judgments can not be considered because the difference in the URLs retrieved is caused by the different queries issued. The MAE of the two populations of workers (i.e., the ones who used the statements as queries and the ones who do not) is computed to ensure that a representative and unbiased subset of workers is selected. In both cases, the MAE is almost the same: 1.41 for the former case and 1.46 for the latter. Then, for each statement, all possible pair of workers which used the statement as a query is considered. For each pair, the overlap among the lists of results considering the top 10 URLs retrieved, is measured. Three different metrics are used: the rank-based fraction of documents which are the same on the two lists, the number of elements in common between the two lists, and RBO. A number in the $[0, 1]$ range is obtained, indicating the percentage of overlapping URLs between the two workers. Note that since the query issued is the same for both workers, the change in the ranked list returned is only caused by some internal policy of the search engine (e.g., to consider the IP of the worker who issued the query, or load balancing policies). Both the complete URL or the domain only are considered when measuring the similarities between the lists. The latter option is focused. In this way, if an article moved for example from the landing page of a website to another section of the same website, such a behavior can be captured. The findings are consistent also when considering the full URL. Then, the average of the similarity scores for each statement among all the workers is computed to normalize the fact that the same queries can be issued by a different number of workers. Note that this normalization process is optional and findings do not change. After that, the average similarity score for the three metrics is computed. The similarity of lists of the same batch is greater than the similarity of the lists from different batches; in the former case, the similarity scores are respectively 0.45, 0.64, and 0.72, while in the latter the scores are 0.14, 0.42, and 0.49.

5.4.6.7 Sources of Information: Justifications

The textual justifications provided, their relations with the web pages at the selected URLs, and their links with worker quality are analyzed to study the effect of using different kind of justifications on the worker accuracy, as done in the main analysis (Section 5.4.5.2).

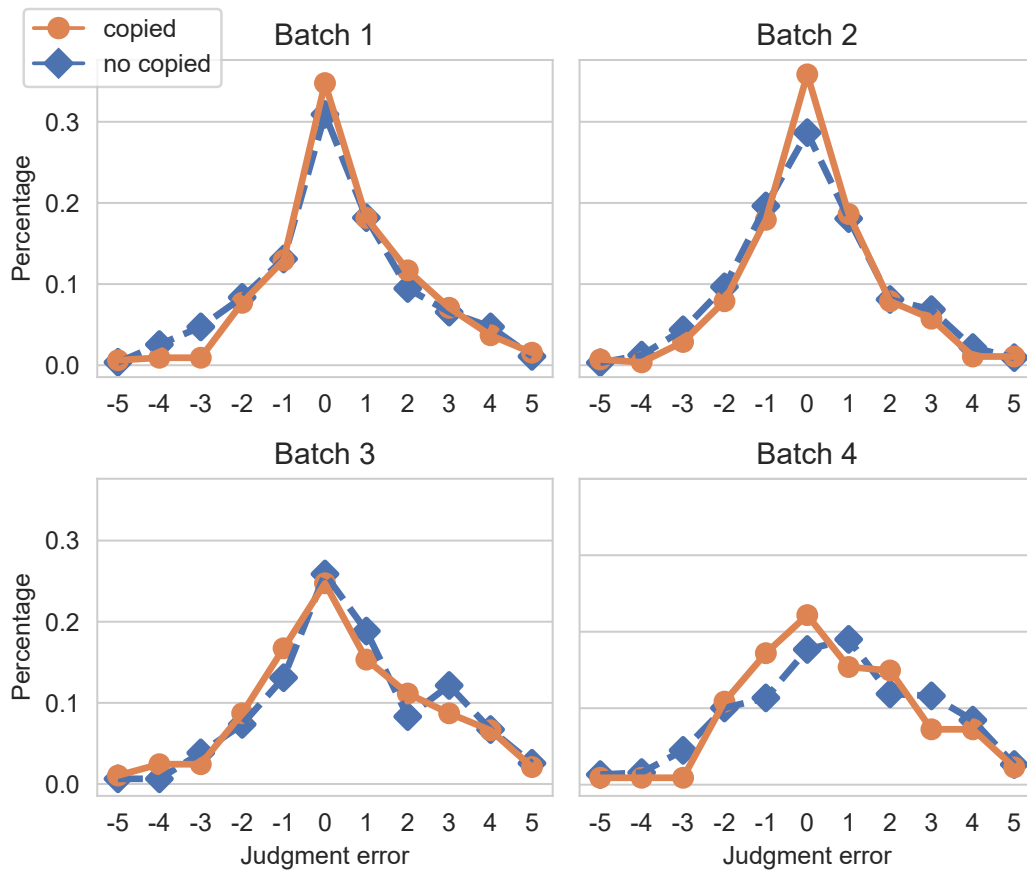


Figure 5.10: Effect of the origin of a justification on the labelling error. Text copied/not copied from the selected URL.

Figure 5.10 shows the relations between different kinds of justifications and the worker accuracy, as done for Figure 5.6. The charts show the prediction error for each batch, calculated at each point of difference between expert and crowd judgments. Furthermore, they show if the text inserted by the worker was copied or not from the selected web page. While Batch1 and Batch2 are very similar, Batch3 and Batch4 present important differences. Statements on which workers make fewer errors (i.e., where $x\text{-axis} = 0$) tend to use a text copied from the web page selected. On the contrary, statements on which workers make more errors (i.e., values close to $+/- 5$) tend to use text not copied from the selected web page. Overall, workers of Batch3 and Batch4 tend to make more errors than workers of Batch1 and Batch2. Similarly to Figure 5.6, the differences between the two groups of workers are small, but this might indicate that workers of higher quality tend to read the text from the selected web page and report it in the justification box. Furthermore, the distribution of the prediction error is not symmetrical, as the frequency of the errors is higher on the positive side of the $x\text{-axis}$ ($[0,5]$) for Batch1, Batch2, and Batch3; Batch4 shows a different behavior. These errors correspond to workers overestimating the truthfulness level of the statements. We can see that the right part of the chart is way higher for Batch3 with respect to Batch1 and Batch2, confirming that workers of Batch3 are of lower quality.

5.4.7 RQ11: Analysis Of Returning Workers

The effect of the returning workers on the dataset is investigated. In more detail, if workers which performed the task more than one time are of higher quality than the workers who performed the task only once. To such an end, each possible pair of datasets where the former contains returning workers and the latter contains workers who performed the task only once is considered. For each pair, only the subset of HITs performed by returning workers is considered. For such a set of HITs, the MAE and CEM^{ORD} scores of the two sets of workers are compared.

Figure 5.11 shows on the $x\text{-axis}$ the four batches, while on the $y\text{-axis}$ the batch containing returning workers ($2f1$ denotes $Batch2_{from1}$, and so on). Each value represents the difference in MAE (Figure 5.11a) and CEM (Figure 5.11b). A cell is colored green if the set of workers on the $y\text{-axis}$ has a higher quality than the one on the $x\text{-axis}$, and red otherwise. The behavior is consistent across the two metrics considered. Apart from a few cases involving Batch4 (and with a small difference), each set of returning workers has similar or higher quality than the other ones. This is more evident when the reference batch is Batch3 or Batch4 and the returning workers are either from Batch1 or Batch2, indicating the high quality of the data collected for the first two batches. This is somehow an expected result and reflects the fact that people gain experience by doing the same task over time; in other words, they learn from experience. At the same time, such a behavior can not be taken for granted, especially in a crowdsourcing setting. Another possible phenomenon that could have happened is that returning workers focused on passing the quality checks to get the reward without caring about performing the task well. The findings show that this is not the case and that the quality checks are well-designed.

The average time spent on each statement position for all the batches is also investigated. The average time spent for $Batch2_{from1}$ is 190 seconds (169 seconds for Batch2), 199 seconds for $Batch3_{from1or2}$ (was 182 seconds for Batch3) and 213 seconds for $Batch4_{from1or2or3}$ (140

batch	2f1	-0.06	-0.11	-0.45	-0.57
	3f1	-0.15	0.04	-0.4	-0.52
	3f1or2	-0.17	-0.09	-0.47	-0.7
	3f2	-0.19	-0.21	-0.53	-0.88
	4f1	-0.33	-0.06	-0.3	-0.44
	4f1or2or3	-0.07	0.01	-0.21	-0.35
	4f2	0.09	-0.07	-0.43	-0.36
	4f3	0.06	0.09	-0.08	-0.28
			1	2	3

batch-reference

(a) MAE

batch	2f1	0.01	0.03	0.09	0.13
	3f1	0.04	0.02	0.11	0.13
	3f1or2	0.04	0.03	0.11	0.16
	3f2	0.04	0.05	0.1	0.2
	4f1	0.08	0.04	0.08	0.14
	4f1or2or3	0.02	0.01	0.04	0.08
	4f2	-0.02	0.02	0.08	0.04
	4f3	-0.02	-0.03	-0.01	0.05
			1	2	3

batch-reference

(b) CEM^{ORD}

Figure 5.11: MAE and CEM^{ORD} for individual judgments for returning workers. Green indicates that the set of workers on the y-axis is better than the one on the x-axis. Red indicates the opposite.

seconds for Batch4). Overall, the returning workers spend more time on each document respect to the novice workers of the corresponding batch. A statistical test between each pair of batches of new and returning workers is also performed. The tests report statistical significance ($p < 0.05$) in 12 cases out of 24.

5.4.8 RQ12: Qualitative Analysis Of Misjudged Statements

The statements are sorted according to their MAE (i.e., the absolute difference between the expert and the worker judgment) for each PolitiFact category, to investigate if the statements which are misjudged by the workers are the same across all batches and if such ordering is consistent across batches. In other words, if the most misjudged statement is the same across different batches.

Figure 5.12 shows, for each PolitiFact category, the relative ordering of its statements sorted according to decreasing MAE (the document with rank 1 is the one with highest MAE). Some statements are consistently misjudged for all the PolitiFact categories. In more detail, those statements are the following (sorted according to MAE):

- Pants-On-Fire: S2, S8, S7, S5, S1;
- False: S18, S14, S11, S12, S17;
- Mostly-False: S21, S22, S25;
- Half-True: S31, S37, S33;
- Mostly-True: S41, S44, S42, S46;
- True: S60, S53, S59, S58.

The 24 statements (Appendix C) selected are manually inspected to investigate the cause of failure, and the justifications provided are manually checked. For all the statements analyzed, most of the errors in Batch3 and Batch4 are given by workers who answer randomly, generating noise. Answers are categorized as noise when the following two criteria are met: (i) the chosen URL is unrelated to the statement (e.g. a Wikipedia page defining the word “truthfulness” or a website to create flashcards online); (ii) the justification text does not explain the truthfulness level chosen (neither personal nor copied from a URL which is different from the selected one). Noisy answers become more frequent with every new batch and account for almost all the errors in Batch4. The number of judgments with a noisy answer for the four batches is respectively 27, 42, 102, and 166; conversely, the number of non-noisy answers for the four batches are respectively 159, 166, 97, and 54. The non-noise errors in Batch1, Batch2 and Batch3 seem to depend on the statement. The following main reasons for failure in identifying the correct label are found by manually inspecting the justifications provided by the workers.

In four cases (S53, S41, S25, S14), the statements are objectively difficult to evaluate. This is because they either require extreme attention to the detail in the medical terms used (S14), address highly debated points (S25), or require knowledge of legislation (S53).

In four cases (S42, S46, S59, S60), the workers could not find relevant information, so they decided to guess. The difficulty in finding information is justified: the statements are either too vague to find useful information (S59), others have few official data on the matter (S46) or the issue has already been solved and other news on the same topic had taken its place, making the web search more difficult (S60, S59, S42) (e.g. truck drivers had trouble getting

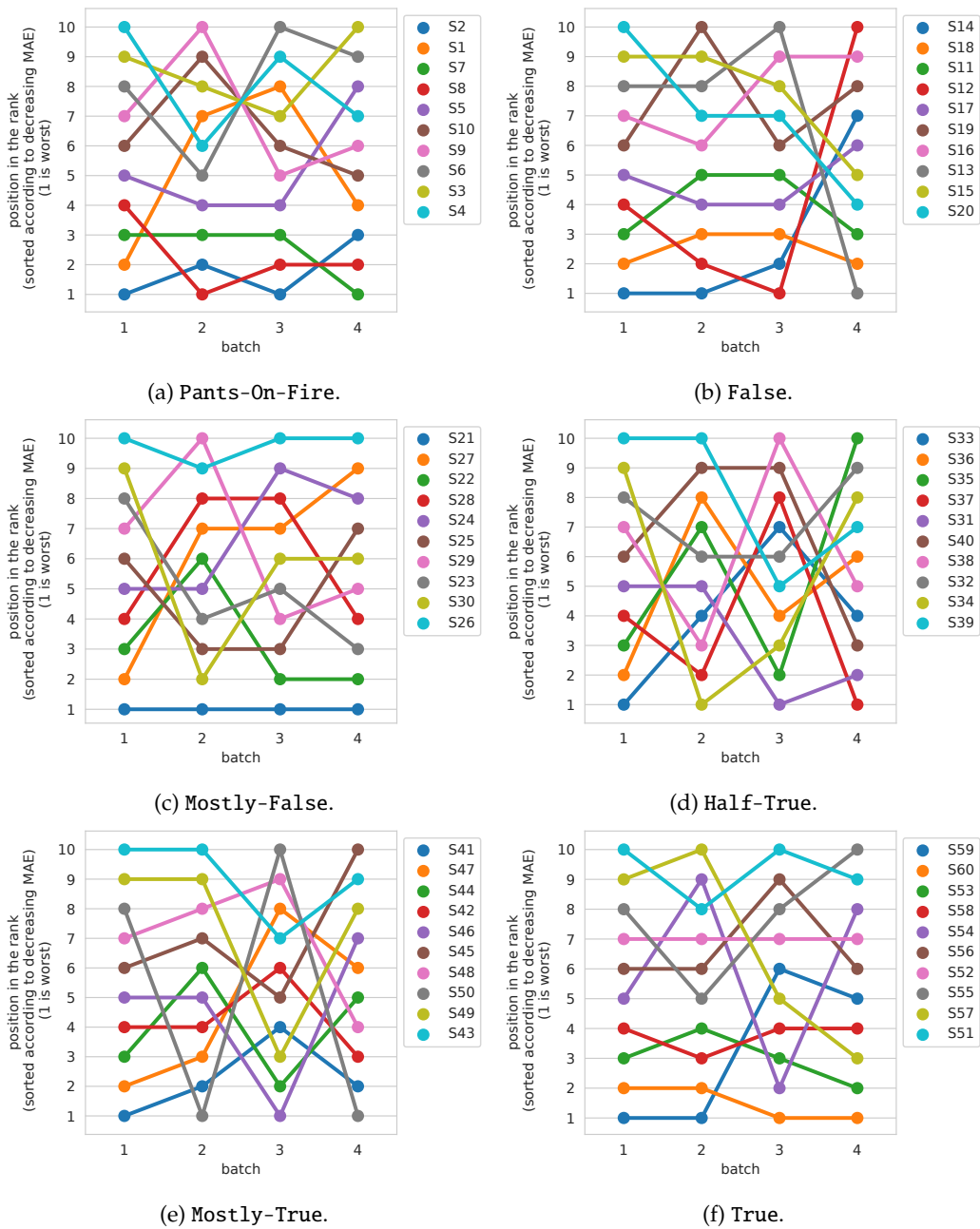


Figure 5.12: Relative ordering of statements across batches according to MAE for each Politifact category. Rank 1 represents the highest MAE.

food in fast food restaurants, but the issue was solved and news outlets started covering the new problem “lack of truck drivers to restock supermarkets and fast food chains”).

In four cases (S33, S37, S59, S60), the workers retrieve information which covers only part of the statement. Sometimes this happens by accident (S60, information on Mardi Gras 2021 instead of Mardi Gras 2020) or because the workers recover information from generic sites, which allow them to prove only part of the statement (S33, S37).

In four cases (S2, S8, S7, S1), Pants-On-Fire statements are judged as true (probably) because they have been stated by the person. In these cases, the workers use a fact-checking site as the selected URL, sometimes even explicitly writing that the statement was false in the justification, but select True as the label.

In thirteen cases (S7, S8, S2, S18, S22, S21, S33, S37, S31, S42, S44, S58, S60), the statements are deemed as more true (or more false) than they are by focusing on part of the reasoning on how plausible they sounded. In most cases, the workers find a fact-checking website which reports the ground truth judgment, but they decide to modify their judgment based on their personal opinion. True statements from politics are doubted (S60, about nobody suggesting to cancel Mardi Gras) and false statements are excused as exaggerations used to frame the gravity of the moment (S18, about church services not resuming until everyone is vaccinated).

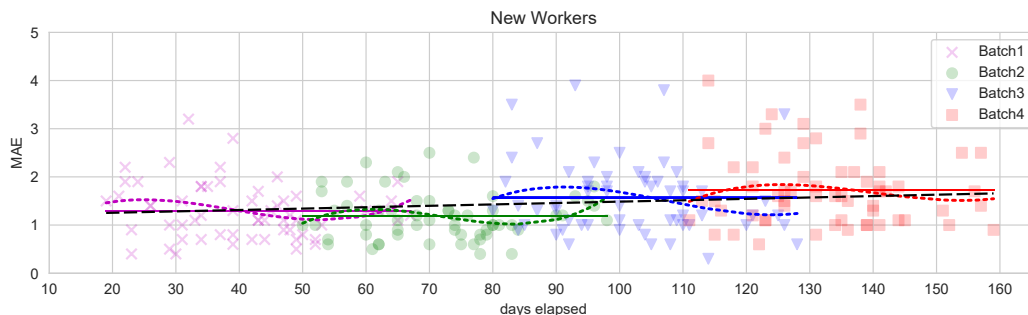
In five cases (S1, S5, S17, S12, S11), the statements are difficult to prove/disprove (lack of trusted articles or test data) and they report concerning information (mainly on how the coronavirus can be transmitted and how long it can survive). Most of the workers retrieve fact-checking articles which label the statements as False or Pants-On-Fire, but they choose an intermediate rating. In these cases, the written justifications contain personal opinions or excerpts from the selected URL which enforce uncertainty (e.g. tests being not definitive enough, lack of knowledge on the behavior of the virus). They also suggest it is safe to act under the assumption of being in the worst-case scenario (e.g. avoid buying products from China, leave packages in the sunlight to try and kill the virus).

Following the results from the failure analysis, the worst individual judgments (i.e., the ones with noise) are removed according to the failure analysis. The effect on aggregated judgments is minimal, and the resulting boxplots are very similar to the ones obtained in Figure 5.1 without removing the judgments.

The way the judgments’ correctness correlates with the attributes of the statement (namely position, speaker, and context) and the passing of time are investigated. To such an end, the absolute distance from the correct truthfulness value is computed for each judgment in a batch and then aggregated by the values of the statement, obtaining the mean absolute error (MAE) and standard deviation (STD) for each statement. For each batch, the statements are sorted in descending order according to MAE and STD. The top 10 statements are selected and their attributes are analyzed. When considering the position (of the statement in the task), the wrong statements are spread across all positions, for all the batches; thus, this attribute does not have any particular effect. When considering the speaker and the context most of the wrong statements have “Facebook User” as the speaker, which is also the most frequent source of statements in our dataset. The MAE of each statement is plotted against the time passed from the day the statement was made to the day it was evaluated by the workers, to investigate the effect of time. This is done for all the batches of novice workers (Batch1 to Batch4) and returning workers (Batch2_{from1},

Batch3_{from1or2}, Batch4_{from1or2or3}).

Figure 5.13 shows that the trend of MAE for each batch (dotted lines) is similar for all batches: statements made in April (leftmost ones for each batch) have more errors than the ones made at the beginning of March and in February (rightmost ones for each batch), regardless of how much time has passed since the statement was made. Figure 5.13a shows that the MAE tends to grow with each new batch of workers (black dashed trend line). The previous analyses suggest that this is probably not an effect of time but of the decreasing quality of the workers. This is also suggested by Figure 5.13b, which shows that MAE tends to remain stable in time for returning workers (which were shown to be of higher quality). Furthermore, the trend of every batch remains the same for returning workers. Statements made in April and at end of March keep being the most difficult to assess. Overall, the time elapsed since the statement was made seems to have no impact on the quality of the workers' judgments.



(a) Novice workers.



(b) Returning workers.

Figure 5.13: MAE (aggregated by statement) against the number of days elapsed (from when the statement was made to when it was evaluated). Each point is the MAE of a single statement in a batch. Dotted lines are the trend of MAE in time for the batch, straight lines are the mean MAE for the batch. The black dashed line is the global trend of MAE across all batches.

5.5 Summary

This chapter presents a comprehensive investigation of the ability and behavior of crowd workers when asked to identify and assess the truthfulness of recent health statements related to the COVID-19 pandemic. The workers perform a task consisting of judging the truthfulness of 8 statements using a customized search engine, which allows for controlling worker behavior. Workers' backgrounds and biases are analyzed, as well as workers' cognitive abilities. Such information is correlated to workers' quality. The experiment is repeated in four different batches, each of them a month apart, with both novice/new and experienced/returning workers.

It must be remarked that the longitudinal study aims to study two different phenomena: (i) how novice workers address the truthfulness of COVID-19-related news over time; (ii) how returning workers address the truthfulness of the same set of news after some time. The underlying hypothesis is that with time workers became more aware of the truthfulness of COVID-19-related news. This does not hold when considering a set of novice workers. This result is in line with other works [341]. Nevertheless, a batch launched considering only returning workers leads to an increase in agreement, showing how workers tend to learn by experience. Furthermore, returning workers did not focus only on passing the quality checks, thus confirming the high quality of collected data. An increase in quality over time by running an additional batch with returning workers can be expected. The answers to the research questions can be summarized as follows.

RQ5 There is evidence which shows that the workers can detect and objectively categorize online (mis)information related to the COVID-19 pandemic. The aggregated workers judgments show high levels of agreement with the expert labels, with the only exception of the two truthfulness categories at the lower end of the scale (Pants-On-Fire and False). The agreement among workers does not provide a strong signal.

RQ6 Both crowdsourced and expert judgments can be transformed and aggregated to improve label quality.

RQ7 The effectiveness of workers is slightly correlated with their answers to the questionnaire, although this is never statistically significant.

RQ8 The relationship between the workers' background/bias and their quality has been used to improve the effectiveness of the aggregation methods used on individual judgments. However, it does not provide a noticeable increase in external agreement. It might be the case that such signals effectively inform new ways of aggregating crowd judgments. This should be further addressed by using more complex methods in the future.

RQ9 Workers use multiple sources of information, and they consider both fact-checking and health-related websites. There are interesting relations between the justifications provided by them and the judgment quality.

RQ10 Re-collecting all the data at different time spans has a major effect on the quality of the judgments. When considering novice workers, early batches produced are more consistent with each other than more recent batches. Also, batches which are closer in time to each other are more similar in terms of workers' quality. Novice workers also put effort into the

task to look for evidence using different sources of information and to write queries, since they often reformulate it.

RQ11 Experienced/returning workers spend more time on each statement w.r.t. novice workers in the corresponding batch. Also, experienced workers have similar or higher quality w.r.t. to other workers. Furthermore, as the number of batches increases, the average time spent on all statements decreases substantially.

RQ12 An extensive analysis of features and peculiarities of the statements that are misjudged by the crowd-workers across all datasets is provided. The time elapsed since the statement was made seems to have to impact on the quality of the workers' judgments.

The next chapter aims to understand the barriers to running longitudinal studies on crowdsourcing platforms such as the one described in this chapter. A large-scale survey is deployed to such an end on three different commercial crowdsourcing platforms.

The Barriers To Longitudinal Studies On Crowdsourcing Platforms

This chapter is based on the article under review in the “Information Processing & Management” journal [394]. Section 2.3 describe the relevant related work. Section 6.1 details the research questions. Section 6.2 details the methodology used to capture workers’ perception of longitudinal studies. Finally, Section 6.5 summarizes the main findings and concludes the chapter. The rest of this chapter relies heavily on the terminology defined in Section 1.7.

6.1 Research Questions

This chapter addresses the current research gap in workers’ perception of longitudinal studies (Section 1.5). Since little is currently known about workers’ perceptions of longitudinal studies, several questions shall be investigated. What exactly encourages workers to participate in such studies and what dissuades them from doing so? Why do workers drop out from or abandon longitudinal studies? What can we learn from crowd worker experiences that can improve how longitudinal crowdsourcing studies are carried out? How can crowdsourcing platforms better support the execution of longitudinal studies? The longitudinal study described in Section 5.2.2 reported a not negligible abandonment rate across the various batches, month after month (Section 5.4.6.1). A better understanding of the workers’ needs and perception of longitudinal studies could have helped in designing a better experimental environment.

A study that aims at uncovering the challenges and obstacles faced when running online longitudinal studies by recruiting workers via crowdsourcing platforms is thus needed. To such an end, a crowd of workers is surveyed over three distinct crowdsourcing platforms which are popular among workers on different continents. Such platforms are

Amazon Mechanical Turk, Prolific, indexProlific and Toloka (Appendix A.1). The survey is designed with the goal of eliciting crowd workers' experience with longitudinal studies to surface existing obstacles and identify ways to overcome them. We analyzed the collected answers using a mixed-methods approach. Initially, the results of a quantitative analysis of workers' contributions highlighting common patterns in longitudinal studies participation are reported. Then the main findings from the analysis of worker comments are presented, which were coded by using a conventional qualitative content analysis approach. These results not only allow an understanding of what barriers have been experienced by crowd workers who participated in longitudinal studies but also propose a list of recommendations for practitioners and researchers who need to run longitudinal studies by recruiting workers on crowdsourcing platforms. It must thus be remarked that the final focus is on task requesters and crowdsourcing platforms which are the actors that, respectively, design and allow the publishing of such kinds of studies. The questions proposed to the workers aimed to elicit the aforementioned list of recommendations and best practices to improve the effectiveness of conducting longitudinal studies. The following research questions are investigated:

RQ12 Can longitudinal studies be characterized quantitatively from the perspective of crowd workers?

RQ13 Does collecting qualitative responses from the workers provide additional insights about their perception of longitudinal studies?

RQ14 Which are the recommendations that researchers and practitioners who want to conduct longitudinal studies over crowdsourcing platforms should follow?

RQ15 Which are the best practices that commercial crowdsourcing platforms should employ to support longitudinal studies?

6.2 Experimental Setting

A systematic survey is used to characterize longitudinal studies from the perspective of crowd workers. The survey is designed and deployed as a crowdsourcing task on three popular commercial crowdsourcing platforms, namely Amazon Mechanical Turk, Prolific [318], and Toloka.

The complete survey and the responses provided by workers together with the dataset related to the quantitative and qualitative analysis of the answers provided is released and available publicly. The qualitative part employs a thematic analysis and includes a complete account of the coding scheme, codes, and themes.

6.2.1 Survey And Crowdsourcing Task Design

The survey is composed of two parts, namely P1 and P2. The first part of the survey (P1) aims at exploring the popularity of longitudinal studies in crowdsourcing, focusing on workers' prior experience, the perceived suitability of platforms for longitudinal studies, and possible reasons limiting the popularity of longitudinal crowdsourcing studies. The

aim is to synthesize through P1 the perceived characteristics and measure the popularity of longitudinal studies. The second part (P2) investigates workers' thoughts, opinions and ideas about the design of and their underlying motivations to participate in future longitudinal studies.

The crowdsourcing task involves recruiting 300 workers with previous experiences with longitudinal studies from three popular crowdsourcing platforms (100 each): Amazon Mechanical Turk (Appendix A.1.1), Toloka (Appendix A.1.2), and Prolific (Appendix A.1.3). The participation is restricted to workers who complete at least 4000 tasks on Amazon Mechanical Turk and 2000 tasks on Prolific informed by the pilot run on each platform to increase the odds of recruiting workers with previous experience with longitudinal studies. Since Toloka does not directly support such a filter, the workers recruited from Toloka are explicitly asked about their prior experience with longitudinal studies. The responses are gathered upon obtaining 100 workers with prior experience (i.e., ≥ 1 longitudinal study) on each platform. Each worker received 2 USD\$ for participating. This is based on task completion time during a pilot. In the actual study, the task completion time (avg=700s, $\sigma=593$, median=548s) leads to \$10-13/hour median reward. The crowdsourcing task is designed and run using Crowd_Frame (Appendix A).

The task is as follows. Initially, the workers are first presented with task instructions alongside the context of the study, which included the definition of longitudinal studies as detailed in Section 1.7. Workers are then asked to respond to the first part of the survey (P1), followed by the second part (P2). The whole survey is reported in Appendix D. Within P1, workers are asked to report their experiences pertaining to up to 3 longitudinal studies they had completed by responding to a subset of 11-13 questions in each case. The worker's responses and conditional logic are used to determine whether or not certain sub-questions have to be asked, as described in Appendix D.1. Overall, if a worker reports $0 \leq n \leq 3$ experiences the number of questions shown within P1 ranges from $1 + (11 * n) + 2$ to $1 + (13 * n) + 2$. P2 (Appendix D.2) comprises of 11 questions. Hence, the number of questions in the entire survey ranges from $1 + (11 * n) + 13$ to $1 + (13 * n) + 13$. The survey consists of 9 multiple-choice questions, 4 text-based questions (i.e., questions with a mandatory textual answer), 7 checkbox-based questions, and 6 numerical questions. Moreover, 11 questions allow the workers to write a custom-free text to detail their answers. Upon completing P1 and P2, workers are able to submit their responses and receive the corresponding payment. To ensure the quality of the responses, a single criterion which verifies that workers had spent a minimum of 3 seconds on each question is implemented. Additionally, upon successful completion of the task, the workers have the opportunity to provide a final comment to the task requesters.

6.2.2 Statistical Testing

Statistical significance tests are conducted on the data collected from the survey for the questions that do not require a custom text-based answer to investigate the relationship between the different variables of interest. In the six cases where the answer provided by the workers is numeric such as for the question 1.1.X.1 of the P1 part (Appendix D.1), ANOVA [314] is used to determine if there was a statistically significant difference (to the 0.05 level) between the means of the groups. Specifically, a one-way ANOVA is used to

compare the means of the three groups of workers (i.e. Amazon Mechanical Turk, Prolific, and Toloka). In the cases where a statistically significant difference to the $p < 0.05$ level is found, posthoc tests are performed using Tukey's HSD method [1] to determine which groups differed significantly from each other, which is a multiple comparison test that controls for Type I error rate by adjusting the significance level based on the number of pairwise comparisons.

For the nine multiple-choice questions, i.e., questions that require choosing a mutually-exclusive answer among a predefined set, such as question 1.1.X.5 of the P1 part, chi-squared tests are used to determine if there are statistically significant differences between the groups. Specifically, the observed contingency table of frequencies is calculated and the chi-squared test is used to compare it to the expected contingency table under the null hypothesis of no difference between the groups. The false discovery rate (FDR) correction [364] is used to account for and correct multiple comparisons. Such correction controls for the expected proportion of false discoveries among the rejected null hypotheses. If zero expected frequencies are encountered while performing the chi-squared test, the comparison is excluded from the analysis.

Similarly to the previous case, for the seven questions that allowed choosing multiple alternatives as an answer among a predefined set, such as question 7 of the P2 part (Appendix D.2), chi-squared tests are used to determine if there are statistically significant differences between the groups. However, in this case, a respondent could select multiple options, leading to overlapping categories; to account for this, the observed contingency table of frequencies is calculated using a modified approach that allows for overlapping categories. The chi-squared test and FDR correction are then used as in the previous case to determine if there are significant differences between the groups.

6.2.3 Qualitative Analysis Of Workers' Response

A conventional qualitative content analysis approach [190] is followed to analyze the open-ended responses from workers on Amazon Mechanical Turk, Prolific, and Toloka. Such an inductive approach is used to describe a phenomenon for which there is limited existing research or theory, as opposed to deductive qualitative analysis, which builds on predetermined themes from previous literature. In this case, two assessors act as expert researchers, reading all of the responses to the open-ended mandatory questions and those that allowed providing a free text. For each answer, they generate a custom "code" by highlighting the key phrases that seem to capture the most significant insights using a custom keyword. To provide a simple example, the answer provided by a worker when asked about why they participated in the longitudinal study is "It was interesting, I learned something about myself". In this case, the initial code chosen by the two assessors is the keyword *task_interest*. Multiple core concepts emerged as the analysis progressed, thus setting the basis of the initial overall coding scheme.

The next phase of the analysis involved merging the codes that were initially identified based on their inter-dependencies. This process is carried out through multiple iterations and discussions. To provide a simple example, the initial codes of *task_interest*, *task_payment*, and *task_easiness* have been merged into the overall theme of *task_features*. Ultimately, this process results in the emergence of seven themes. The internal agreement is not reported

due to the involvement of expert researchers and the multiple iterations that we ran on the coding scheme, The interested reader can refer to McDonald et al. [278] for further details.

6.3 Worker Demographics

Initially, some demographic details about the 300 workers recruited are provided. The three crowdsourcing platforms used share demographic information about each worker to varying extents. In more detail, Amazon Mechanical Turk does not share any kind of worker attribute, while Prolific shares their IP addresses along with their age, gender and ethnicity. Lastly, Toloka shares their age, country, level of education and known languages. Even though Amazon Mechanical Turk does not provide natively the IP addresses of their workers, Crowd_Frame fetches the IP addresses for every worker who participates and performs the reverse lookup using well-known geolocation services (Section A.6). Workers' geolocation data are used only to analyse their approximate provenance since some other kind of demographic information is provided by 2 out of 3 platforms only. An additional survey sheet to capture additional information about worker's background is left for future work.

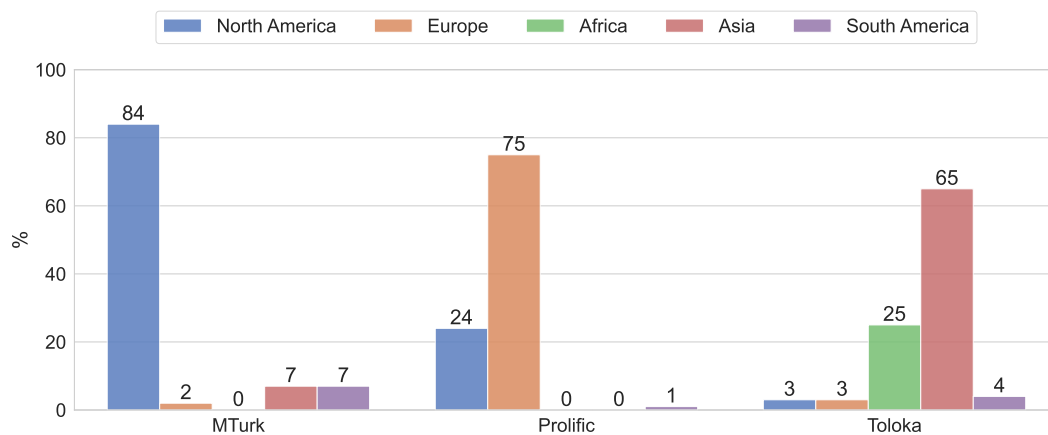



































Figure 6.1: Continent of provenance of the 300 workers recruited, breakdown across every crowdsourcing platform.

Overall, 26.7% of them live in Europe, 37% in North America, 24% in Asia, 8.3% in Africa and, lastly, 4% in South America. Figure 6.1 shows the continent of the provenance of the workers broken down across each crowdsourcing platform. As it can be seen, the vast majority of North American workers (84%) are recruited from Amazon Mechanical Turk, while for the Europeans 75% of them are recruited from Prolific. Interestingly, Toloka provides many workers (90%) from Africa and Asia, which are underrepresented by the other platforms. In more detail, 3% of Toloka workers live in Europe, while 25% in Africa, and 65% in Asia. As a final remark, we can say that recruiting workers from South America is difficult on every platform, with percentages lower than 7%.

Table 6.1 is presented to provide greater clarity regarding the origin of European workers. It provides a detailed breakdown of their country of provenance across the continent. As

Table 6.1: Breakdown of the country of provenance of the 300 workers recruited, grouped by continent.

Continent	Country		Workers	Percentage
Europe	United Kingdom		58	19.33%
Europe	Russian Federation		35	11.67%
Europe	Portugal		9	3.00%
Europe	Italy		5	1.67%
Europe	Belarus		2	0.67%
Europe	Poland		1	0.33%
Europe	Netherlands		1	0.33%
Europe	Bulgaria		1	0.33%
Europe	France		1	0.33%
Europe	Spain		1	0.33%
Europe	Germany		1	0.33%
North America	United States		110	36.67%
North America	Canada		1	0.33%
Asia	India		10	3.33%
Asia	Pakistan		10	3.33%
Asia	Turkey		5	1.67%
Asia	Vietnam		4	1.33%
Asia	Kazakhstan		2	0.67%
Asia	Sri Lanka		2	0.67%
Asia	Uzbekistan		1	0.33%
Asia	Philippines		1	0.33%
Asia	Bangladesh		1	0.33%
Asia	Jordan		1	0.33%
Africa	Kenya		12	4.00%
Africa	Nigeria		8	2.67%
Africa	Morocco		3	1.00%
Africa	Tunisia		1	0.33%
Africa	South Africa		1	0.33%
South America	Brazil		8	2.67%
South America	Venezuela		1	0.33%
South America	Peru		1	0.33%
South America	Colombia		1	0.33%
South America	Chile		1	0.33%
			300	100%

can be seen, the majority of European workers (11.67% of the total) are actually Russian (or even Belarusian) people. The geolocation service used to perform the reverse lookup considers Russia as a European country, even though it stretches also across Asia. One thus may argue by looking at Figure 6.1 and Table 6.1 that the workers recruited from Prolific include several Russian people.

Figure 6.2 shows the location of each worker on a world map. As can be seen, all the Russian and Belarusian workers are recruited from Toloka, together with the vast majority of African and Asian workers (Figure 6.2c). On the other hand, the vast majority of the remaining European workers are recruited from Prolific (Figure 6.2b). Turning to the remaining continents, the workers recruited from the United States are more evenly distributed between Amazon Mechanical Turk (Figure 6.2a) and Prolific. Such data allows concluding that relying on the different crowdsourcing platforms allows for obtaining a more distributed population of workers, with different provenances, cultural and political backgrounds and, arguably, different experiences and underlying motivations for performing and participating in longitudinal studies.

6.4 Results

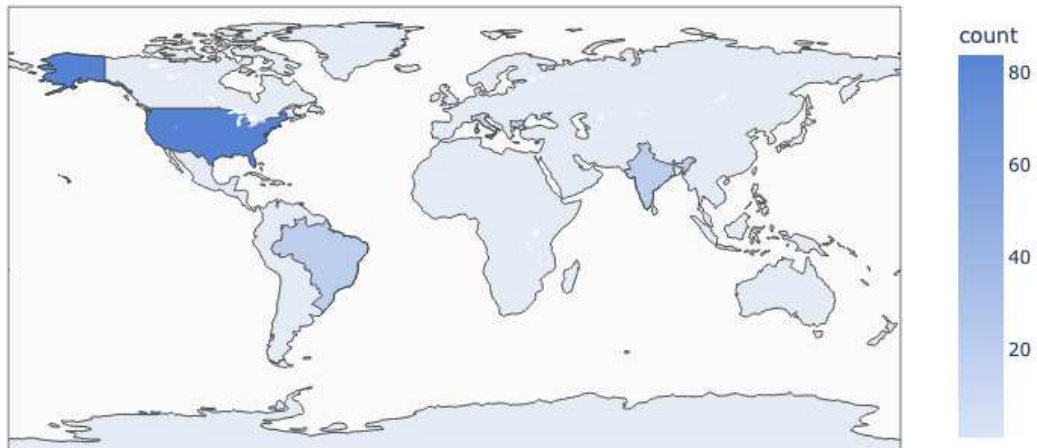
Section 6.4.1 provides a quantitative analysis of the answers provided by the workers', for each survey part. Section 6.4.2 outlines the key findings from the qualitative analysis. Section 6.4.3 details the recommendations for practitioners and researchers to conduct longitudinal studies. Finally, Section 6.4.4 sketches the best practices for crowdsourcing platforms to support such kind of crowdsourcing experiment.

6.4.1 RQ12: Quantitative Analysis Of Workers' Responses

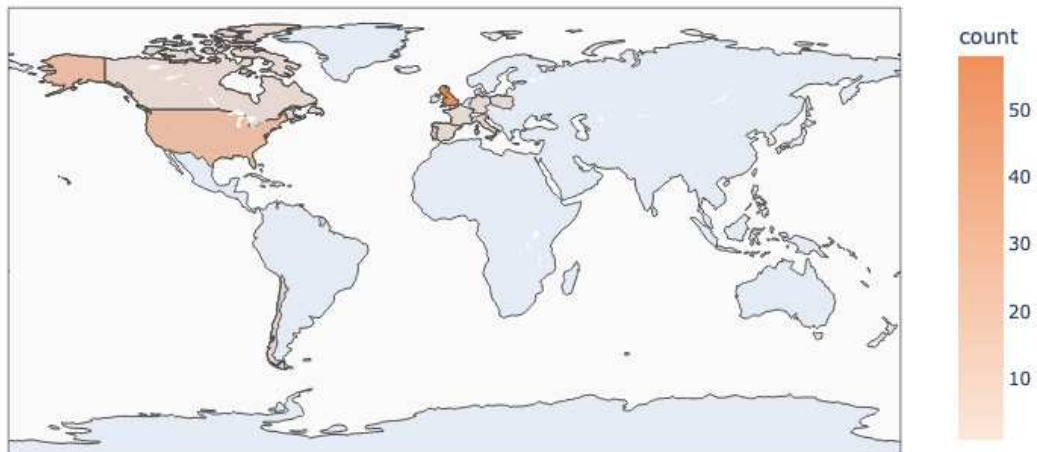
Section 6.4.1.1 includes remarks necessary for interpreting the results accurately. Following this, the quantitative analysis of the workers' answers is provided. Initially, Section 6.4.1.2 examines the number of previous experiences with longitudinal studies reported by the workers across each platform. Then, Section 6.4.1.3 discusses the results of the P1 part of the survey. Finally, Section 6.4.1.4 provides a detailed analysis of the P2 part of the survey.

6.4.1.1 Initial Remarks

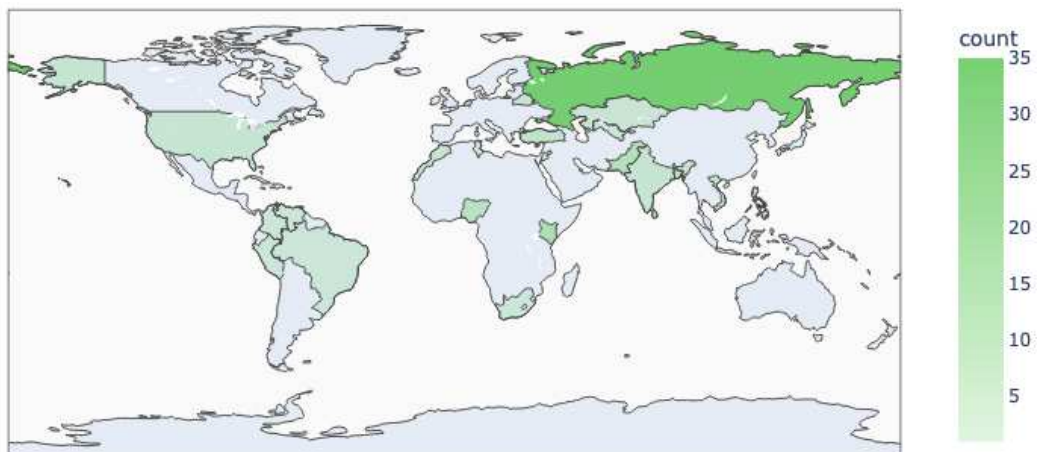
The worker provenance attribute (i.e., the platform used to recruit the worker who answered the survey) is used to break down the results across the three crowdsourcing platforms. It is important to note that the platform where the workers performed each previous experience reported is not considered even if they admitted having worked on multiple platforms, as described in Section 6.4.1.2. Consequently if a worker is recruited from the Amazon Mechanical Turk platform and refers to a longitudinal study they had participated on Prolific, their answers is included in the Amazon Mechanical Turk breakdown. The results are analyzed in depth in the following, while a summary is available in Table 6.3 and Table 6.4. For the sake of clarity, the paragraph dedicated to a given question is labelled with the question's index, as reported in Appendix D.



(a) Amazon Mechanical Turk.



(b) Prolific.



(c) Toloka.

Figure 6.2: Worker provenance distribution across the whole world, breakdown across each crowdsourcing platform.

As a second remark, several charts are reported in the following. The bar charts are used for the questions that required choosing one or more answers between a set of predefined options, using a radio button or a checkbox-based user interface control. Such kind of charts allows for highlighting the differences between the answer visually, across each crowdsourcing platform. The numbers right above the bars report the percentage, while the absolute number of workers that gave a particular answer is reported at the top of each chart.

6.4.1.2 Previous Experiences With Longitudinal Studies

To begin the investigation, the number of previous longitudinal studies in which each worker had participated is analyzed. The analysis reveals that 44.8% of workers reported participating in a single longitudinal study, while 27.8% and 27.4% reported participating in two and three longitudinal studies, respectively. However, this trend varied across the three platforms. For more detailed information on the experiences reported, refer to Table 6.2. Overall, the 300 workers recruited (i.e., 100 per each of the three platforms) report 547 previous experiences with longitudinal studies, with a grand mean of 1.82 previous experiences per worker. Among them, Amazon Mechanical Turk workers report a number of 187 previous experiences, while Prolific workers report 193, and Toloka workers report 167.

In general, it is slightly more likely to find workers with previous experience with longitudinal studies in Prolific (35.28%) rather than Amazon Mechanical Turk (34.19%), while Toloka workers are less used to such kind of studies (30.53%). Furthermore, 97 out of the 300 workers (32.3%) reported experiences that took place in a crowdsourcing platform different from the recruitment one (see also Figure 6.8).

Table 6.2: Previous experiences with longitudinal studies reported by the workers recruited.

Platform	Experiences	Percentage	Mean
Amazon Mechanical Turk	187	34.19%	1.85
Prolific	193	35.28%	1.89
Toloka	167	30.53%	1.67
Total	547	300 %	1.82

6.4.1.3 P1: Spreading Of Longitudinal Studies

The P1 part of the survey consists of 11 questions related to the workers' experiences with longitudinal studies. However, the results for 10 out of the 11 questions are presented in the following sections, as they were multiple-choice questions. The 11th question, which is text-based, requires thematic analysis and is therefore discussed in Section 6.4.2.

1.1: Amount Of Previous Experiences Figure 6.3 further details the previous experiences with longitudinal studies reported by the workers shown in Table 6.2. In more detail, 42% of the workers report a single experience with longitudinal studies when considering

the Amazon Mechanical Turk platform, while 29% report 2 experiences, and 29% report 3 experiences. The majority of workers (43%) report a single experience when considering Prolific, while 21% report 2 experiences, and 36% report 3. Finally, 50% of the workers report 1 experience when considering Toloka, while 33% report 2 experiences, and 17% report 3. This suggests that it is more likely to find workers able to report several previous experiences at once on the Prolific platform instead of the remaining two. This is also a further indication of the validity of the criterion described in Section 4.2.1 used to recruit the workers; the workers on Amazon Mechanical Turk and Toloka seem to be less used to longitudinal studies, thus to increase the odds of recruiting them a higher HIT completion threshold has to be set.

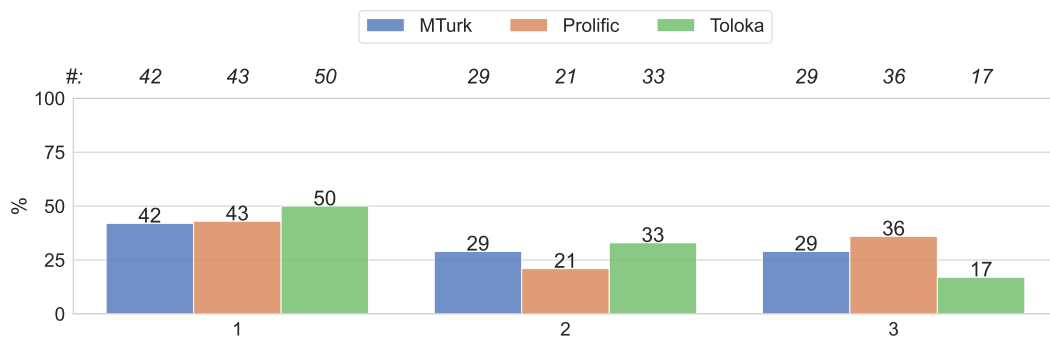


Figure 6.3: Number of previous experiences as reported by workers.

1.1.X.1: Timing Figure 6.4 describes the time elapsed since each experience reported, with a particular focus on participation up to 12 months earlier. The majority of the experiences reported (86.8%) indeed happened up to 12 months before participation in the survey, while the remaining 13.2% took place earlier. Anyhow, the distribution within the previous year is rather homogeneous, and roughly 13% of participation for each crowdsourcing platform happened more than 12 months earlier (Amazon Mechanical Turk vs. Toloka statistically significant, adjusted p-value < 0.05).

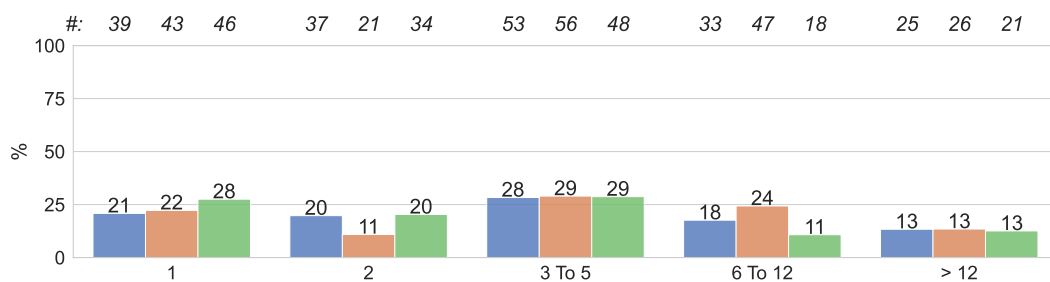


Figure 6.4: Months elapsed since each experience.

1.1.X.2: Sessions Figure 6.5 details how many sessions each experience reported was composed. The experiences reported by workers from Amazon Mechanical Turk and Toloka

have about 6 sessions on average, while for Prolific they are 7. In general, it appears that task requesters tend to publish slightly longer longitudinal studies on Prolific.

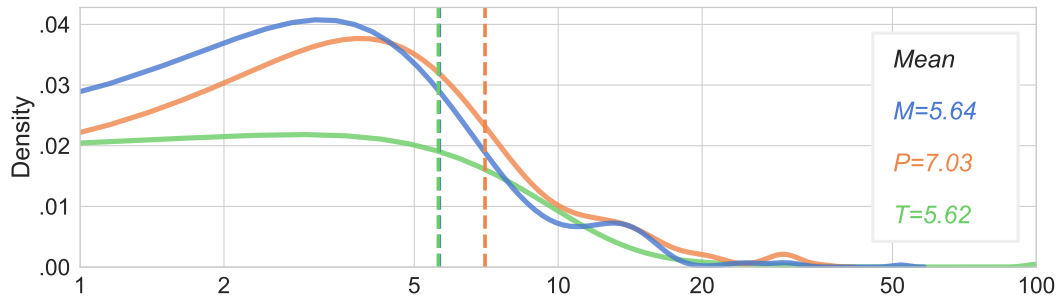


Figure 6.5: Number of sessions of each experience.

1.1.X.3: Interval Between Sessions Figure 6.6 details the time elapsed between each session of the experiences reported. The time interval ranges from 1 day to more than 30 days. The vast majority of time spans between subsequent sessions are of 30 days or less (89.8%). This indicates that a longitudinal study should not make workers wait more than a month to have them participate again (Amazon Mechanical Turk vs. Toloka statistically significant, adjusted p-value < 0.01). The question allowed workers to provide an additional free text for each experience reported; 6% of them provided it (18 out of 300), for 4.02% of the experiences (22 out of 547).

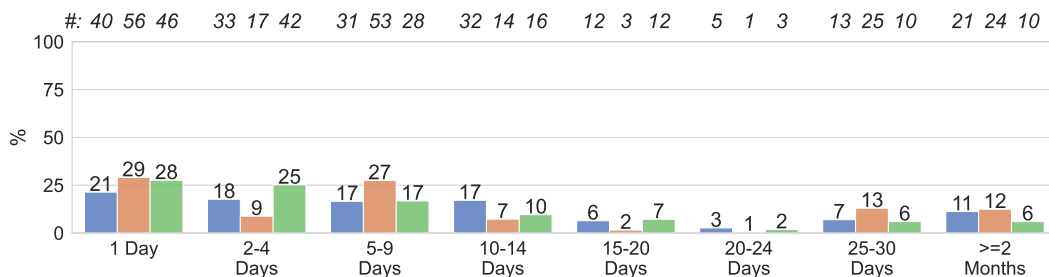


Figure 6.6: Time interval between sessions of each experience.

1.1.X.4: Session Duration Figure 6.7 details the duration of each session of the experiences reported. The vast majority of sessions (99.2%) last up to 2 hours, with the only exception of Amazon Mechanical Turk workers sessions which lasted for 3 hours or more. Moreover, only 8 Toloka workers, a single Amazon Mechanical Turk worker, and a single Prolific worker reported sessions of 2 hours; thus, the ideal session duration seems to be of 1 hour or less (Amazon Mechanical Turk vs. Prolific, Amazon Mechanical Turk vs. Toloka and Toloka vs. Prolific statistically significant; adjusted p-value < 0.01). The question allowed workers to provide an additional free text for each experience reported; 5.33% of them provided it (16 out of 300), for 4% of the experiences (22 out of 547).

1.1.X.5: Crowdsourcing Platform Figure 6.8 details on which platform the experiences

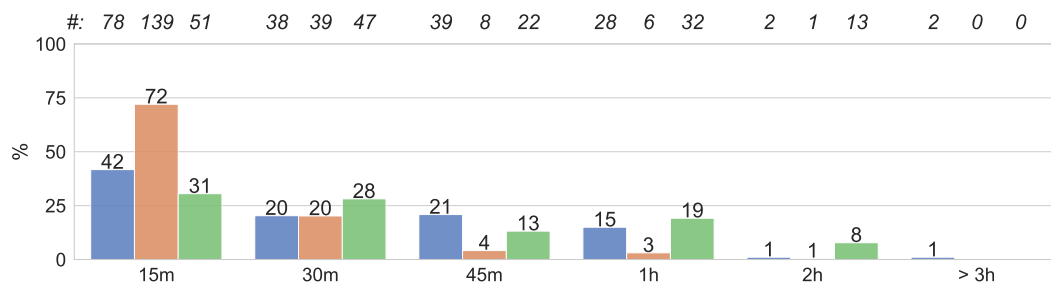


Figure 6.7: Average session duration of each experience (minutes and hours).

reported took place. The workers recruited to answer our survey from Amazon Mechanical Turk, Prolific, and Toloka tend to participate in longitudinal studies on the same platform, as expected. Nevertheless, it must be noted that while for Amazon Mechanical Turk and Prolific the percentages are rather high (i.e., 91% and 90% respectively) when considering Toloka the percentage drops to 63%. This shows that Toloka workers are those who tend more to work on multiple platforms, at least when considering longitudinal studies (Amazon Mechanical Turk vs. Prolific, Amazon Mechanical Turk vs. Toloka and Prolific vs. Toloka statistically significant; adjusted p-value < 0.01). The question allowed workers to provide an additional free text for each experience reported; 8.67% of them provided it (26 out of 300), and for 6.1% of the experiences (33 out of 547).

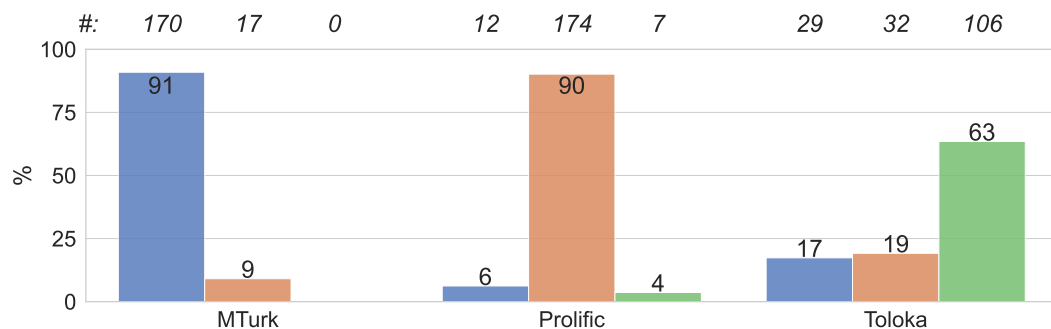


Figure 6.8: Crowdsourcing platforms where each experience took place.

1.1.X.6: Payment Model Figure 6.9 investigates the payment model adopted by the longitudinal studies in which the workers recruited participated. The majority of workers (72.5%) reported experiences with payment after each session. Only 9% of Amazon Mechanical Turk workers, 7% of Prolific workers and 7% of Toloka workers reported a combination of both approaches. This shows that even though it is more common to provide some kind of reward to a worker after each session of the study, employing a single and final reward can be a viable option (Amazon Mechanical Turk vs. Prolific, Amazon Mechanical Turk vs. Toloka and Prolific vs. Toloka statistically significant; adjusted p-value < 0.01). The question allowed workers to provide an additional free text for each experience reported; 10% of them provided it (30 out of 300), for 6.4% of the experiences (35 out of 547).

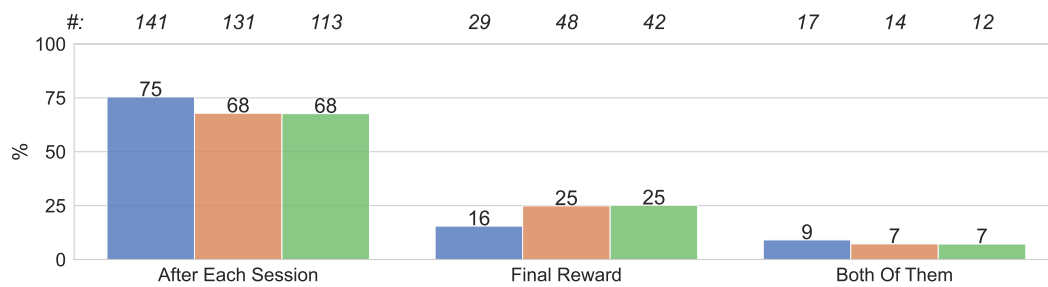


Figure 6.9: Model used to pay the workers during each experience (i.e., when the reward was provided).

1.1.X.7: Satisfaction Figure 6.10 investigates the satisfaction of each worker after the reported experiences. The majority of workers (91.6%) are keen to participate again in a longitudinal study. Such behavior is consistent for the Prolific and Toloka workers with a percentage of affirmative answers of, namely, 98% and 93%, while such a percentage is slightly lower for Amazon Mechanical Turk workers (83%). In general, workers indeed want to keep participating in longitudinal studies (Amazon Mechanical Turk vs. Prolific and Prolific vs. Toloka statistically significant; adjusted p-value < 0.01).

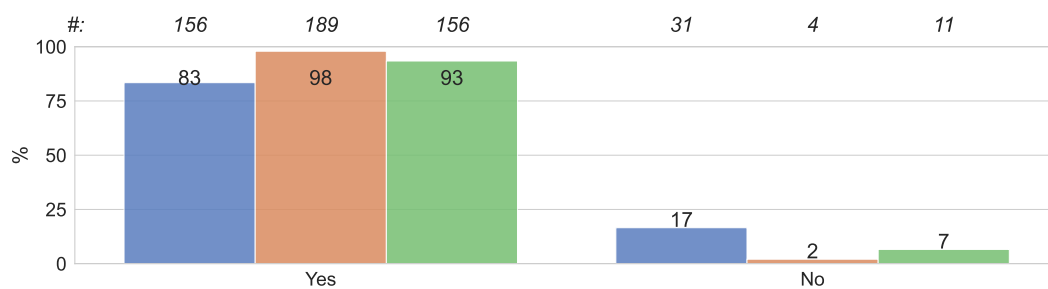


Figure 6.10: Workers willingness to participate again in each experience reported.

1.1.X.8: Driving Incentives Figure 6.11 addresses the underlying motivations that drive workers to participate in the experiences reported. The incentives related to money such as bonuses and rewards are the most important ones for the majority of workers (70.9%). However, when considering Prolific workers' personal interest and altruism (to help research) are reported to be important as well (34%); this may be due to the fact that Prolific is a platform mostly focused on helping academics on conducting their research projects. It should also be noted that 17% of Toloka workers found the reported longitudinal study educative. Even though the reward is the most important incentive for a worker, there are several other factors that should not be ignored when designing a longitudinal study (Amazon Mechanical Turk vs. Prolific, Amazon Mechanical Turk vs. Toloka, and Prolific vs. Toloka statistically significant with adjusted p-value < 0.01). The question allowed workers to provide an additional free text for each experience reported; 7.33% of them provided it (22 out of 300), for 4.94% of the experiences (27 out of 547).

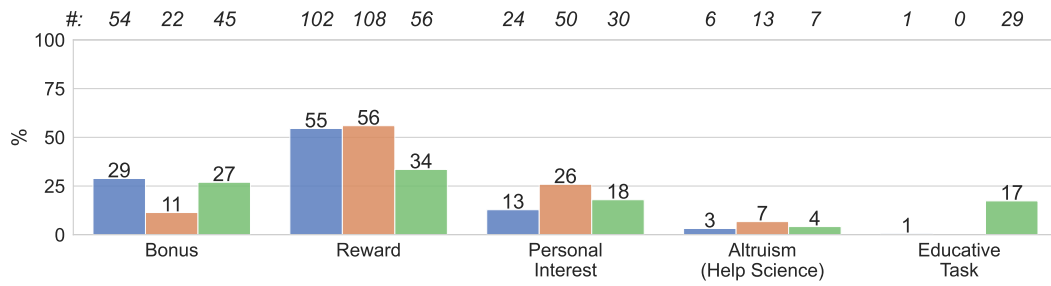


Figure 6.11: Incentives that drive workers to participate in each experience.

1.1.X.9: Termination Figure 6.12 investigates whether workers completed the longitudinal studies in which they participated. Almost every worker recruited on Prolific and Toloka that answered the survey completed the longitudinal study reported. The percentage is slightly lower (95%) for Amazon Mechanical Turk workers.

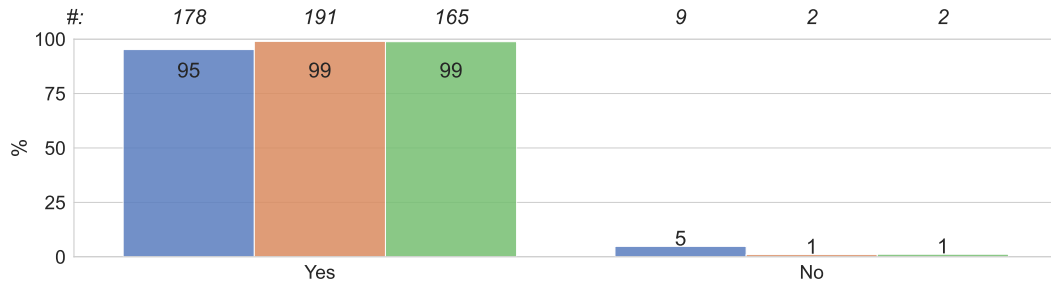


Figure 6.12: Completion of the longitudinal studies in which the worker participated.

1.1.X.9.1: Completion Incentives Figure 6.13 addresses the underlying motivations that drive workers to complete the experience reported and should be ideally compared with question 1.1.X.8. Indeed, the incentives related to money such as bonuses and rewards are still the most important ones for the majority of workers (70.2%), followed by the workers’ personal interest in the longitudinal study (6.2%) and altruism to help research (6.2%). The personal interest of Prolific workers highlighted when asked about what drove them to participate drops to 19%, becoming comparable with the other platforms. On the other hand, 15% of Toloka workers still find the study educative (Amazon Mechanical Turk vs. Prolific, Amazon Mechanical Turk vs. Toloka and Prolific vs. Toloka statistically significant with adjusted p-value < 0.01).

3: Popularity Figure 6.14 investigates the reasons that limit longitudinal studies’ popularity on crowdsourcing platforms according to workers’ opinions. The most prevalent reasons according to Amazon Mechanical Turk workers are that rewards and incentives are insufficient (27%, *Reward*) and that longitudinal studies are not optimally supported by current popular crowdsourcing platforms (23%, *Technical*). Toloka workers agree on the reward (20%) and the technical aspects (32%). The most important factor according to Prolific workers is the commitment required to perform the longitudinal study (39%,

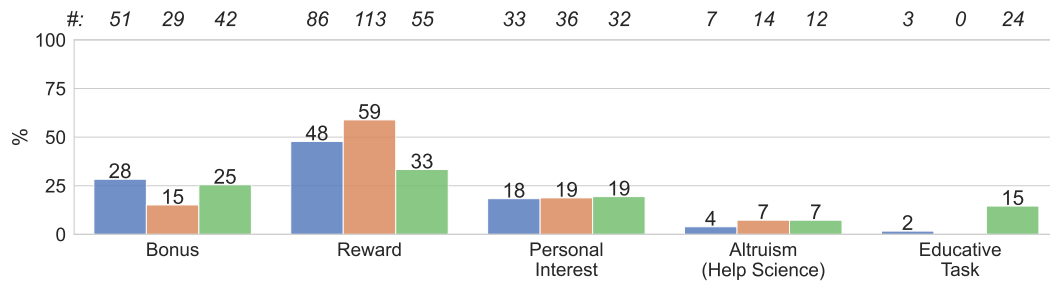


Figure 6.13: Incentives that drive workers to complete each experience.

Commitment) and the insufficient reward for participating required (20%). They also think that requesters do not need longitudinal participation since most of the tasks deal with static data to annotate (13%, *Data*) The reward is the most important reason that hampers the popularity of longitudinal studies across each platform, but other reasons should not be ignored while designing such a kind of studies (Amazon Mechanical Turk vs. Prolific, Amazon Mechanical Turk vs. Toloka, and Prolific vs. Toloka statistically significant; adjusted p-value < 0.01). The question allowed workers to provide an additional free text for each experience reported; 7.67% of them provided it (48 out of 300), for 8.78% of the experiences (48 out of 547).

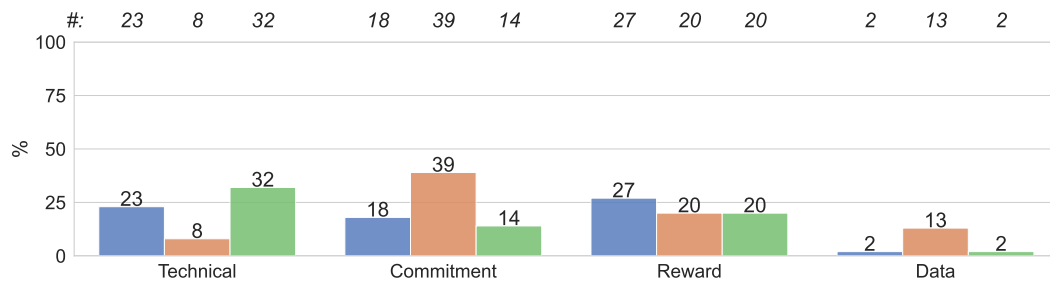


Figure 6.14: Reasons that limit the popularity of longitudinal studies in crowdsourcing according to workers.

6.4.1.4 P2: Design Of Future Longitudinal Studies

The P2 part of the survey is designed to address 10 different aspects that should be considered when designing future longitudinal studies, based on the insights gained from the responses of the 300 workers recruited.

1: Commitment Duration Figure 6.15 studies the amount of commitment that workers are willing to provide for a longitudinal study. The Amazon Mechanical Turk workers are keen to commit for about 19 days on average, while for Prolific workers such a value increases to 27 days. Lastly, Toloka workers would commit to a longitudinal study for a shorter duration, around 17 days on average. Thus, Prolific workers are generally keen to be committed to longitudinal studies for less time than workers from the other two platforms (Prolific vs

Toloka statistically significant with adjusted p-value < 0.05).

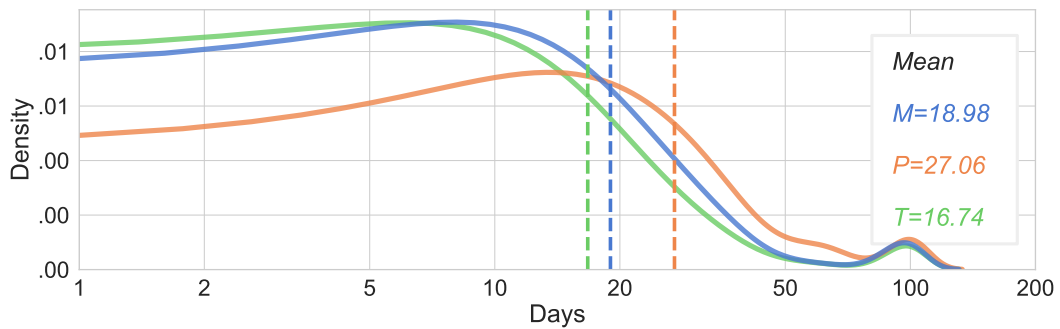


Figure 6.15: Ideal amount of commitment (days) for a longitudinal study according to workers.

2: Participation Decline Figure 6.16 investigates which are the reasons that drive workers to refuse participation in longitudinal studies. The majority of workers (71.1%) think that the longitudinal study’s length is the most important factor. This is the opinion of the vast majority of workers recruited from Prolific (85%), while for Toloka workers the amount is slightly lower (71%). Lastly, for Amazon Mechanical Turk workers the amount drops to 58%. Only 29% of the workers consider the frequency of the sessions of the longitudinal study as a factor that can lead to refusing participation. Thus, for Prolific and Toloka workers a study’s length is a major concern, while for Amazon Mechanical Turk workers its frequency should not be overlooked (Amazon Mechanical Turk vs. Prolific, Amazon Mechanical Turk vs. Toloka and Prolific vs. Toloka statistically significant; adjusted p-value < 0.01). The question allowed workers to provide an additional free text for each experience reported; 16.67% of them provided it (50 out of 300), for 9.14% of the experiences (50 out of 547).

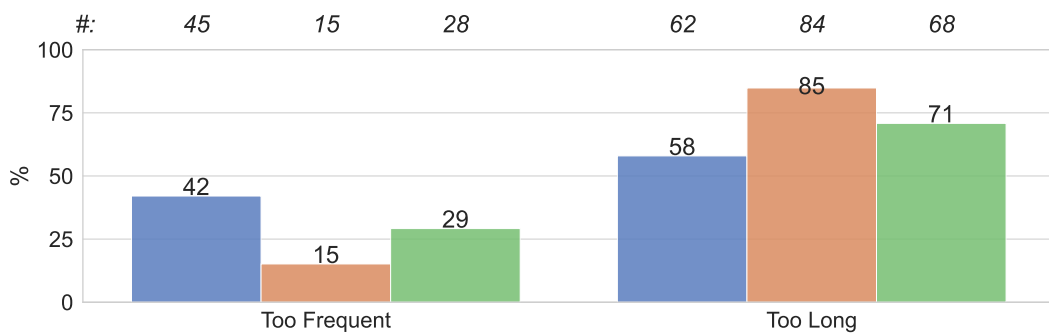


Figure 6.16: Reasons that drive workers to refuse participation in longitudinal studies.

3: Participation Frequency Figure 6.17 investigates the ideal participation frequency in longitudinal studies according to the workers. The vast majority of workers (92%) prefer short time spans across all platforms, as they prefer studies that have a daily to weekly participation commitment. This is particularly true for Toloka workers since 53% of them

find a daily frequency as ideal. This is also true, to some extent, for Amazon Mechanical Turk workers (39%). Prolific workers, on the other hand, are keen to participate also on a weekly basis (41%). Only a few of them across Amazon Mechanical Turk and Toloka (8% for both platforms) would prefer studies with longer time spans (Amazon Mechanical Turk vs. Prolific, Amazon Mechanical Turk vs. Toloka and Prolific vs. Toloka statistically significant; adjusted p-value < 0.01).

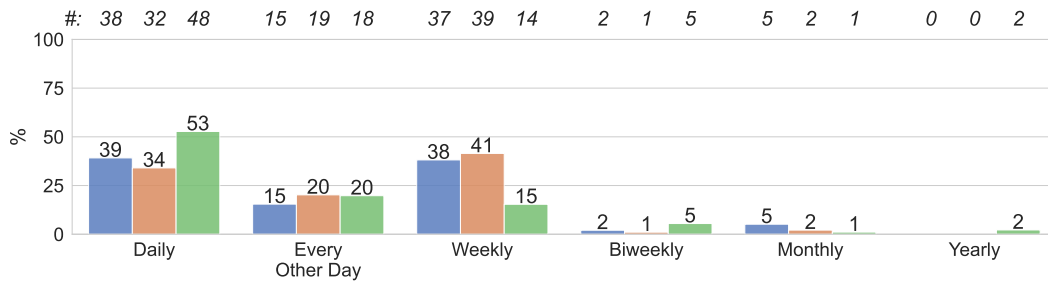


Figure 6.17: Ideal participation frequency according to workers.

4: Ideal Session Duration Figure 6.18 investigates the ideal session duration for longitudinal studies according to the workers. The Prolific workers prefer short sessions of less than 1 hour on average, while Amazon Mechanical Turk and Toloka workers share a more uniform preference with an average of about 2 hours. The figure does not show 3 outliers who provide non-reasonable duration (i.e., between 15 and 20 hours), thus being removed. In general, Amazon Mechanical Turk and Toloka workers are keen to work for a longer time within each session than Prolific workers (Amazon Mechanical Turk vs. Prolific and Prolific vs. Toloka statistically significant; adjusted p-value < 0.05).

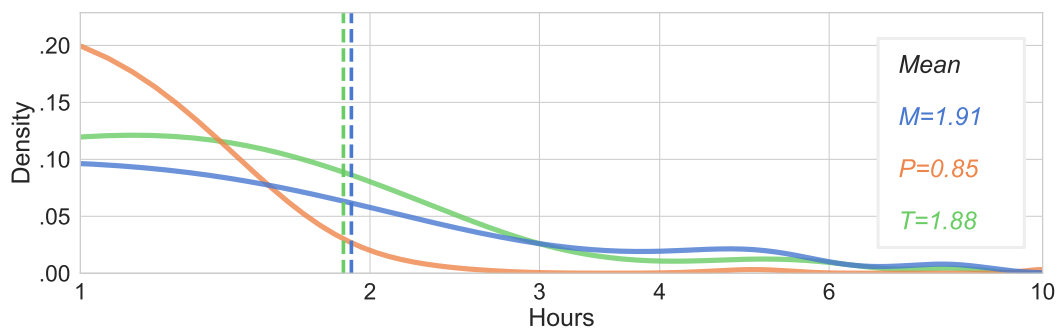


Figure 6.18: Ideal session duration for longitudinal studies according to workers.

5: Ideal Hourly Payment Figure 6.19 investigates the ideal hourly payment rate in USD\$ for the participation in a longitudinal study according to the workers. The Amazon Mechanical Turk workers aim to receive on average a higher hourly payment (about 14\$) than Prolific workers, while the Toloka ones are keen to expect a lower amount (about 8\$). The figure does not show 4 outliers who provide unreasonable amounts (i.e., amounts ranging between 80\$ and 100\$), thus being removed. In general, Toloka workers are those who expect to be

paid less. However, while on such a platform and Amazon Mechanical Turk the requester can propose an arbitrary reward amount for each HIT, on Prolific they have to propose an estimated completion time for the task and the platform enforces a minimum reward amount. Such a feature may thus have an impact on workers' ideal payment perception (Amazon Mechanical Turk vs. Toloka statistically significant; adjusted p-value < 0.05).

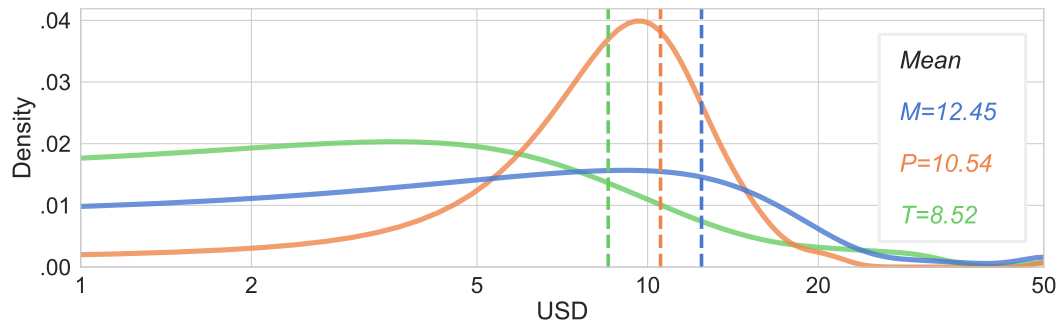


Figure 6.19: Ideal hourly payment (USD\$) for the participation in a longitudinal study according to the workers.

6: Ideal Daily Time Figure 6.20 investigates the ideal amount of time to allocate for participating in longitudinal studies according to the workers. The Amazon Mechanical Turk and Toloka workers are willing to allocate respectively on average up to 2.6 and 3.7 hours per day, while Prolific workers aim to work less, up to about 1.7 hours on average. The figure does not show 3 outliers who provide non-reasonable amounts of hours per day (i.e., between 20 and 25), thus being removed. In general, Toloka workers are those who are keen to perform more work on a daily basis in longitudinal studies, while Prolific workers expect to allocate a much lower amount of time (Amazon Mechanical Turk vs. Prolific statistically significant; adjusted p-value < 0.05).

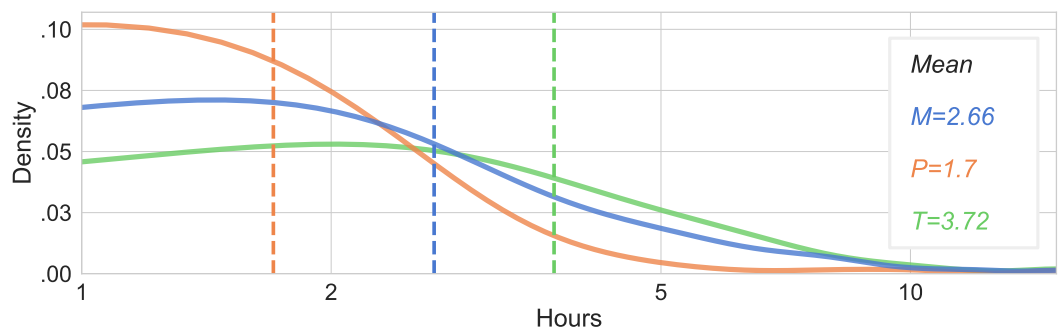


Figure 6.20: Ideal amount of daily time to allocate to participate in longitudinal studies according to the workers.

7: Participation Incentives Figure 6.21 investigates which are the most important incentives that drive workers into participating in longitudinal studies. In general, the type of

reward/payment mechanism has a major impact (79%). Overall, 24.7% of workers prefer a final bonus after the last contribution, 31.8% a partial payment after each session, and 20.5% an incremental payment after each contribution. Other aspects such as task diversity and variations of the same task to reduce repeatability play a minor but not negligible role since they motivate 18.7% of the workers. On the other hand, the presence of decrement in the payments after the initial session or eventual penalization for skipping one or more sessions have a negligible impact, since only the 2.2% of workers consider it (Amazon Mechanical Turk vs. Prolific, Amazon Mechanical Turk vs. Toloka, and Prolific vs. Toloka statistically significant; adjusted p-value < 0.01). The question allowed workers to provide an additional free text for each experience reported; 5.67% of them provided it (17 out of 300), for 3.11% of the experiences (17 out of 547).

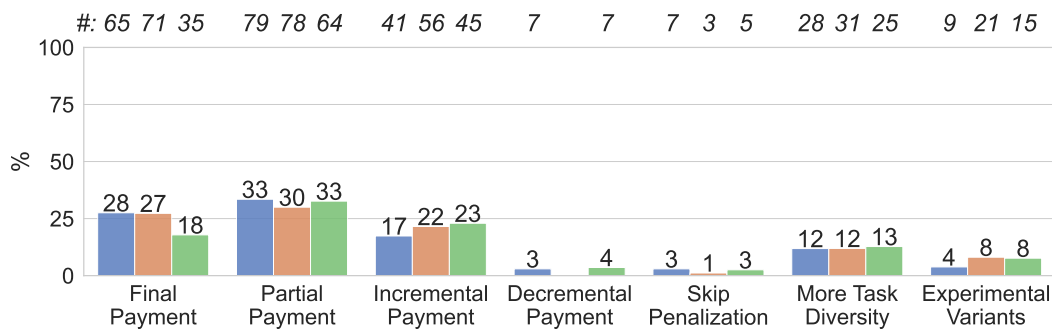


Figure 6.21: Most important incentives for participating in longitudinal studies according to the workers.

8: Ideal Task Type Figure 6.22 investigates which are ideal task types to perform in a longitudinal study fashion according to workers. There is no clear preference, with a rather homogeneous answer trend. The workers are willing to perform any kind of task across each platform, including content access, surveys, information finding and so on (Amazon Mechanical Turk vs. Prolific, Amazon Mechanical Turk vs. Toloka, and Prolific vs. Toloka statistically significant; adjusted p-value < 0.01). The question allowed workers to provide an additional free text for each experience reported; 4.67% of them provided it (14 out of 300), for 2.56% of the experiences (29 out of 547).

9: Involvement Benefits Figure 6.23 investigates which are the benefits of being involved in longitudinal studies according to workers. In general, workers are keen to think that the most important benefit characterizing longitudinal studies is that they allow them to be more productive since they are more operational (32.1%, *Produce*). The workers also think that these kinds of studies allow them to avoid spending regular time searching for new tasks (26.5%, *Search*). Furthermore, they also think that the intermediate payments increase trust in the requester (25.8%, *Trust*). Some of them also think that performing a longitudinal study allows them to avoid learning again how to perform the task (15.7%, *Learn*). The trends are homogeneous across all platforms, without any factor being considered more important than others (Amazon Mechanical Turk vs. Prolific, Amazon Mechanical Turk vs. Toloka, and Prolific vs. Toloka statistically significant; adjusted p-value < 0.01).

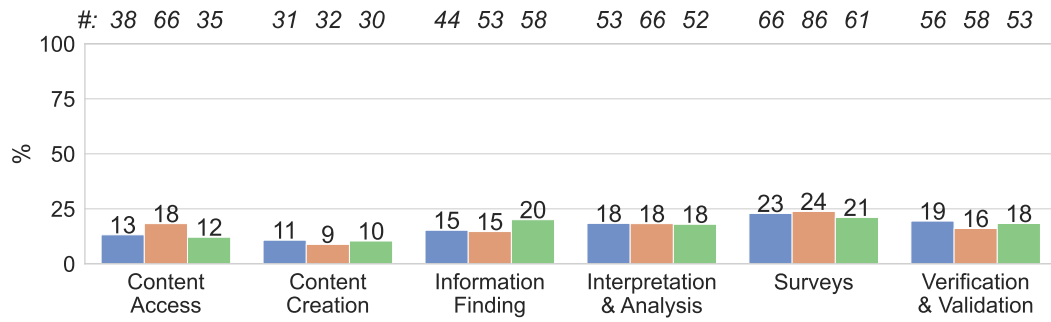


Figure 6.22: Ideal task type to be performed in longitudinal studies according to the workers.

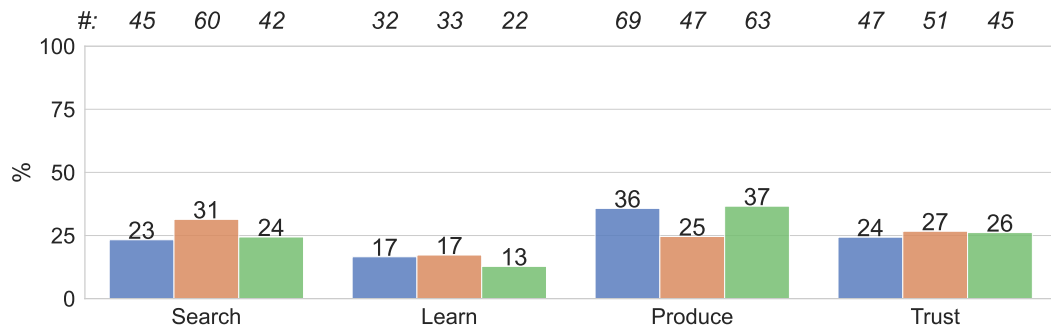


Figure 6.23: Benefits provided by involvement in longitudinal studies according to the workers.

10: Involvement Downsides Figure 6.24 investigates which are the downsides of being involved in longitudinal studies according to workers. In general, workers think that the long-term commitment required (27.9%, *Commitment*), the lack of flexibility (27.7%, *Flexibility*) and the rewards provided after completion only by task requesters (30.4%, *Reward*) are the most important downsides. The lack of diversity plays a minor role (14%, *Diversity*). However, it is interesting to notice how the lack of diversity has a not negligible impact according to Toloka workers (21%), while this is less evident for Amazon Mechanical Turk workers (Amazon Mechanical Turk vs. Prolific, Amazon Mechanical Turk vs. Toloka, and Prolific vs. Toloka statistically significant; adjusted p-value < 0.01). The question allowed workers to provide an additional free text for each experience reported; 7.67% of them provided it (23 out of 300), for 4.2% of the experiences (23 out of 547).

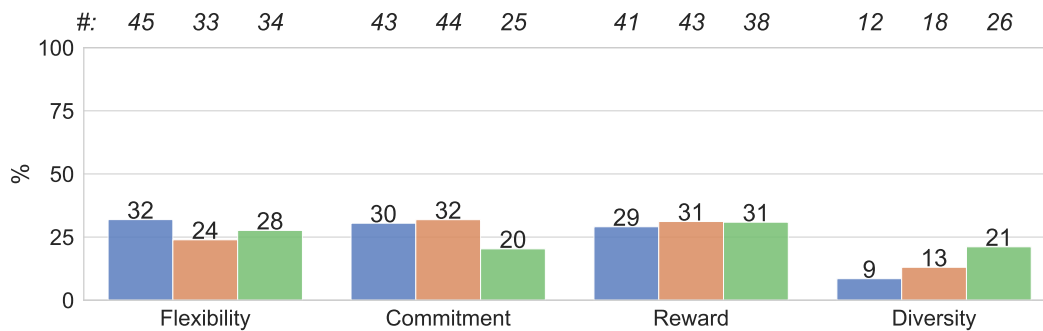


Figure 6.24: Downsides of being involved in longitudinal studies according to the workers.

6.4.1.5 Summary

The key findings extracted from the workers' responses, which could be analyzed quantitatively, are presented in Table 6.3 for the P1 part (Section 6.4.1.3) and in Table 6.4 for the P2 part (Section 6.4.1.4). Both tables provide a detailed summary of the answers, along with the code used to classify each question and a breakdown of responses across each crowdsourcing platform considered. Furthermore, Table 6.5 shows the outcome of statistical tests performed by comparing the answers provided across each platform. The name and answer type of each question are reported, and a checkmark (✓) indicates a statistically significant comparison with the adjusted p-value provided. Questions without any statistically significant comparisons are not reported.

The key findings from the quantitative analysis are finally summarized with the following list of take-home messages, starting from the analysis of longitudinal studies' prevalence according to workers' previous experiences (questions 1-11, P1 part), then moving to workers' opinions about future study design (questions 12-21, P2 part).

1. Workers with more experience with longitudinal studies can be found more easily on the Prolific platform.
2. Most of the experiences reported by the workers happened up to one year before participation in the survey.

3. It is easier to find longer longitudinal studies on the Prolific platform, in terms of sessions' number.
4. Most of the time intervals between different sessions of longitudinal studies range from 1 to 30 days.
5. Most of the sessions of longitudinal studies last up to 2 hours, and roughly half of them last for 15 minutes only.
6. The workers from the Toloka platform tend to participate in longitudinal studies available on other platforms more than others.
7. Most of the longitudinal studies provide some kind of reward to the workers partially, after each session.
8. Most of the workers want to keep participating in longitudinal studies.
9. The most important reason that drove workers into participating in the longitudinal studies reported is the reward.
10. Almost every worker completed the longitudinal studies reported.
11. The most important reason that drove workers into completing the longitudinal studies reported is the reward.
12. The most important reason that hampers the popularity of longitudinal studies is the reward.
13. The workers are keen to commit to longitudinal studies for about 20 days on average.
14. Most of the workers think that the length of a longitudinal study is the most important aspect that drives them to refuse to participate.
15. Most of the workers prefer a daily to weekly session frequency for longitudinal studies.
16. The workers think that the ideal duration for a session of longitudinal studies is of 1.5 hours on average.
17. The workers think that the ideal hourly payment for participating in longitudinal studies is about 10.5 USD\$ on average.
18. The workers think that the ideal amount of time to allocate for participating in longitudinal studies daily is about 2.7 hours on average.
19. The most important incentives that drive workers into participating in longitudinal studies are related to the reward provided.
20. The workers are willing to perform any kind of task in longitudinal studies and there is no clear preference.
21. The most important benefits of participating in longitudinal studies is that they allow workers to be more productive and allow them to avoid searching for new tasks.

22. The most important downsides of participating in longitudinal studies are the long-term commitment required, lack of flexibility and reward provided only at their completion.

Table 6.3: Summary of the key findings for the P1 part of the survey presented in the quantitative analysis.

Question	Amazon Turk	Mechanical Turk	Prolific	Toloka
<i>Amount of Previous Experiences</i>	42% 1 experience, 29% 2 experiences, 29% 3 experiences	43% 1 experience, 21% 2 experiences, 36% 3 experiences	50% 1 experience, 33% 2 experiences, 17% 3 experiences	
<i>Timing</i>	87% up to 1 year before, 13% later	87% up to 1 year before, 13% later	87% up to 1 year before, 13% later	
<i>Sessions</i>	~6 on average	~7 on average	~6 on average	
<i>Interval Between Sessions</i>	89% up to 1 month, 11% later	88% up to 1 month, 12% later	97% up to 1 month, 6% later	
<i>Session Duration</i>	98% up to 1 hour, 3% more	99% up to 1 hour, 1% more	91% up to 1 hour, 8% more	
<i>Crowdsourcing Platform</i>	91% MTurk, 9% Prolific, 0% Toloka	6% MTurk, 90% Prolific, 4% Toloka	17% MTurk, 19% Prolific, 63% Toloka	
<i>Payment Model</i>	75% after each session, 16% final reward, 9% both	68% after each session, 25% final reward, 7% both	68% after each session, 25% final reward, 7% both	
<i>Satisfaction</i>	83% yes, 17% no	98% yes, 2% no	93% yes, 7% no	
<i>Driving Incentives</i>	29% bonus, 55% reward, 13% personal interest, 3% altruism, 1% educative task	11% bonus, 56% reward, 26% personal interest, 7% altruism, 0% educative task	27% bonus, 34% reward, 18% personal interest, 4% altruism, 17% educative task	
<i>Termination</i>	95% yes, 5% no	99% yes, 1% no	99% yes, 1% no	
<i>Completion Incentives</i>	28% bonus, 48% reward, 18% personal interest, 4% altruism, 2% educative task	15% bonus, 59% reward, 19% personal interest, 7% altruism, 0% educative task	25% bonus, 33% reward, 19% personal interest, 7% altruism, 15% educative task	
<i>Popularity</i>	23% technical, 18% commitment, 27% reward, 2% data	8% technical, 39% commitment, 20% reward, 13% data	32% technical, 14% commitment, 20% reward, 2% data	

6.4.2 RQ13: Key Findings From Qualitative Analysis

Section 6.4.2.1 clarifies some remarks needed to correctly interpret the results. Then, the qualitative analysis of the workers' answers is provided. Initially, Section 6.4.2.2 addresses

Table 6.4: Summary of the key findings for the P2 part of the survey presented in the quantitative analysis.

Question	MTurk	Prolific	Toloka
<i>Commitment Duration</i>	19 days on average	27 days on average	17 days on average
<i>Participation Decline</i>	42% too frequent, 58% too long	15% too frequent, 85% too long	29% too frequent, 71% too long
<i>Participation Frequency</i>	92% up to 1 week, 3% bi-weekly, 5% monthly	95% up to 1 week, 1% bi-weekly, 2% monthly, 2% more than yearly	88% up to 1 week, 5% biweekly, 1% monthly, 2% yearly, 3% more than yearly
<i>Ideal Session Duration</i>	1.91 hours on average	0.85 hours on average	1.88 hours on average
<i>Ideal Hourly Payment</i>	12.45 USD\$ on average	10.54 USD\$ on average	8.52 USD\$ on average
<i>Ideal Daily Time</i>	2.66 hours on average	1.70 hours on average	3.72 hours on average
<i>Participation Incentives</i>	28% final payment, 33% partial pay, 17% incremental payment, 3% decremental payment, 3% skip penalization, 12% more task diversity, 4% experimental variants	27% final payment, 30% partial pay, 22% incremental payment, 1% skip penalization, 12% more task diversity, 8% experimental variants	18% final payment, 33% partial pay, 23% incremental payment, 4% decremental payment, 3% skip penalization, 13% more task diversity, 8% experimental variants
<i>Ideal Task Type</i>	13% content access, 11% content creation, 15% information finding, 18% interpretation and analysis, 23% surveys, 19% verification and validation	18% content access, 9% content creation, 15% information finding, 18% interpretation and analysis, 24% surveys, 16% verification and validation	12% content access, 9% content creation, 20% information finding, 18% interpretation and analysis, 24% surveys, 16% verification and validation
<i>Involvement Benefits</i>	23% search, 17% learn, 36% produce, 24% trust	31% search, 17% learn, 25% produce, 27% trust	24% search, 13% learn, 37% produce, 26% trust
<i>Involvement Downsides</i>	32% flexibility, 30% commitment, 29% reward, 9% diversity	24% flexibility, 32% commitment, 31% reward, 13% diversity	28% flexibility, 20% commitment, 31% reward, 21% diversity

Table 6.5: Summary of statistical tests comparing answer groups of each platform. Questions without statistically significant comparisons are not reported. Statistical significance is computed using adjusted p-values according to Section 6.2.2.

Question	Type	MTurk Vs. Prolific	MTurk Vs. Toloka	Prolific Vs. Toloka	Significance Level
<i>Timing</i>	mcq		✓		$p \leq 0.05$
<i>Interval Between Sessions</i>	mcq		✓		$p \leq 0.01$
<i>Session Duration</i>	mcq	✓	✓	✓	$p \leq 0.01$
<i>Crowdsourcing Platform</i>	mcq	✓	✓	✓	$p \leq 0.01$
<i>Payment Model</i>	list	✓	✓	✓	$p \leq 0.01$
<i>Satisfaction</i>	mcq	✓	✓	✓	$p \leq 0.01$
<i>Driving Incentives</i>	mcq	✓		✓	$p \leq 0.01$
<i>Completion Incentives</i>	mcq	✓	✓	✓	$p \leq 0.01$
<i>Popularity</i>	mcq	✓	✓	✓	$p \leq 0.01$
<i>Commitment Duration</i>	number			✓	$p \leq 0.05$
<i>Participation Decline</i>	list	✓	✓	✓	$p \leq 0.01$
<i>Ideal Session Duration</i>	number		✓		$p \leq 0.05$
<i>Ideal Hourly Payment</i>	number		✓		$p \leq 0.05$
<i>Ideal Daily Time</i>	number	✓			$p \leq 0.05$
<i>Participation Frequency</i>	mcq	✓	✓	✓	$p \leq 0.01$
<i>Participation Incentives</i>	list	✓	✓	✓	$p \leq 0.01$
<i>Task Type Incentives</i>	list	✓	✓	✓	$p \leq 0.01$
<i>Involvement Benefits</i>	list	✓	✓	✓	$p \leq 0.01$
<i>Involvement Downsides</i>	list	✓	✓	✓	$p \leq 0.01$

workers' loyalty and commitment to longitudinal studies, while Section 6.4.2.3 addresses the suitability of crowdsourcing platforms in supporting them. Lastly, Section 6.4.2.4 describes various suggestions about longitudinal studies design.

6.4.2.1 Initial Remarks

The workers are asked through open-ended questions about various aspects of longitudinal studies in both the P1 and P2 parts of the survey, as described in Section 6.2.1. Among the whole set of questions, 4 of them required a text-based answer and 11 questions allowed workers to provide a custom free text to detail their answers. In more detail, most of the questions in P1 (Appendix D.1) required an answer for each experience reported; thus, the maximum number of valid answers for such questions is 547. On the other hand, a single question of the P1 part and those in P2 part (Appendix D.2) required a single answer from each worker. For such questions, the maximum number of valid answers is thus 300.

In the following, the text-based answers are focused from a qualitative point of view. Each answer is read and coded according to the thematic analysis methodology described in Section 6.2.3. The answers are categorized into 7 overall themes, reported in Table 6.6 along with sample answers and the initial code assigned. Then, explicitly elicit the answers provided for three questions. For the remaining ones, the workers do not provide meaningful or useful answers.

6.4.2.2 1.1.X.7.2: Worker Loyalty And Commitment

The mandatory open question 1.1.X.7.2 (P1 part) asks workers about what drove them towards returning to a second session after completing the first one in the experience with a longitudinal study reported and why they would refuse to participate in the same study altogether. The workers provide 485 answers among the 547 previous experiences with longitudinal studies reported (88.66%). Table 6.7 shows the distribution of the answers across each theme emerged, while Table 6.8 shows a sample of such answers.

A majority of answers (272 out of 485, 56.08%) point out attributes of the task as influencing their decisions. Some workers mentioned that finding the task interesting (100 out of 272, 36.76%), easy to complete (54 out of 272, 19.85%), or well-paid (112 out of 272, 41.58%) encourages them to return. Others (15 out of 272, 5%) point out that the perceived reliability of securing the reward in the following session having succeeded in the first session is a driver to return. Several workers (58 out of 272, 41.58%) enjoy the agency provided by the tasks to express their views or opinions and get paid in return. On the other hand, low or unfair rewards, unavailability of workers during the follow-up sessions, or device-specific requirements, are common qualms that lead to abandonment after a session or refusal to participate in longitudinal studies.

Around 20.82% of answers (101 out of 485) are provided by workers who think that their preferences and attributes influence the decision to return to complete subsequent sessions in longitudinal studies. Some workers (4 out of 101, 3.96%) reflected on the sunk costs of having completed the first session as a driving factor to return [18]. Other workers (45 out of 101, 44.55%) express the satisfaction they felt towards completing the first session, the commitment required (12 out of 101, 11.88%), the overall involvement, or the opportunity

Table 6.6: Themes emerged while reading each text-based answer provided by the workers.

Theme	Description	Sample Answer	Initial Code
<i>task_features</i>	Aspects related to the task to be performed during a given session of the longitudinal study, such as its design, easiness, etc.	"It was easy to complete"	<i>task_easiness</i>
<i>worker_features</i>	Aspects related to workers' own beliefs and motivations, their satisfaction after participating in the longitudinal study, etc.	"It gave me the chance to be a part of change and real scientific study and know that my part contributed."	<i>worker_motivation</i>
<i>requester_features</i>	Aspects related to the requester who is publishing the longitudinal study, such as reliability, communicativeness, etc.	"Be reliable - offer a reasonable window during which the study can be completed and respond promptly to any messages from participants"	<i>requester_reliability</i>
<i>ls_features</i>	Aspects related to the longitudinal study as a whole, such as session scheduling, reward mechanism, etc.	"Performance rewards are a good way to maintain interest, as it feels like your time and effort are being rewarded"	<i>ls_progress</i>
<i>platform_features</i>	Aspects related to the crowdsourcing platform on which the longitudinal study is conducted such as its features, interface, general design, etc.	"Yes. I think there is a large enough pool to pull from and if set up properly and rewarded, people will respond"	<i>platform_adequacy</i>
<i>no_suggestion</i>	Answers provided by workers that acknowledge by explaining explicitly that they do not have any additional suggestions.	"Nothing comes to mind"	<i>no_suggestion</i>
<i>answer_useless</i>	Answers that do not convey anything related to the question proposed or that are made of random words and digits.	"Unique crowdsourcing business model"	<i>answer_useless</i>

Table 6.7: Distribution across each theme of the answers collected for question 1.1.X.7.2.

Theme	Answer Collected	Percentage
<i>task_features</i>	272	56.08%
<i>worker_features</i>	101	20.82%
<i>requester_features</i>	9	1.86%
<i>platform_features</i>	2	0.41%
<i>ls_features</i>	10	2.06%
<i>answer_useless</i>	91	18.76%
Total	485/547	100%

to gain further insights, learn, and develop their skills through the course of the studies (15 out of 101, 15%).

A few answers (9 out of 485, 1.86%) address aspects and characteristics of the task requester that influence loyalty and commitment to the longitudinal study reported. The most important aspects are communication with the requesters and their ability to remind participants of the subsequent study session. Lastly, 10 answers out of 485 (2%) are from workers that talk about aspects of the longitudinal study reported as a whole. They describe the kind of study they perform because they enjoy doing it and they explain how longitudinal studies are guaranteed work for which they do not have to fight for.

6.4.2.3 2: Crowdsourcing Platforms Suitability

The mandatory open-ended question 2 (P1 part) is used to ask workers about the adequacy of the crowdsourcing platform of provenance in the support they provide for longitudinal studies. The majority of them (273 out of 300, 91%) provide an answer that allows us to draw some kind of consideration. Table 6.9 shows the distribution of the answers across each theme emerged, while Table 6.10 shows a sample of such answers.

The vast majority of answers provided address, as one may expect given the question, aspects related to the crowdsourcing platform of provenance (244 out of 273, 55.86%). Given that the number of workers recruited from each platform is 300, breaking down those who answered the question across each platform allows finding 97 Amazon Mechanical Turk workers, 99 Prolific workers and 75 Toloka workers. Most of the Amazon Mechanical Turk workers (70 out of 97, 67.9%) think that the platform is adequate in general, without providing additional details, while 3 of them (2.91%) mention the easiness of sending reminders about upcoming longitudinal studies sessions. Only 7 of them (6.79%) find the platform not adequate in supporting the longitudinal study. In more detail, one of them states explicitly that it needs further improvements for longitudinal studies such as for scheduling, while another one argues that it is hard for a requester to ensure that the workers recruited are honest. Almost every Prolific worker (97 out of 99, 97.98%) finds the platform adequate in supporting longitudinal studies. When compared with the other platforms, they provide more detailed answers. Some of them mention that the platform

Table 6.8: Sample of answers provided by workers concerning loyalty to longitudinal studies.

Worker Answers
<i>It was a well-designed study and the requester was very specific about when the follow-up tasks would be posted, and they sent reminders as well.</i>
<i>I felt the study was interesting and the reward was excellent so happy to do it again</i>
<i>It was very well organized and efficient. I didn't have to wait much between sessions.</i>
<i>Because I find interesting seeing how differently sometimes my answers can be just after a few days due to changes in the circumstances.</i>
<i>I don't like that participating in some studies again because of im afraid of getting rejected</i>
<i>As long as the daily tasks are short and do not require an app download of any sort, I'll do them. I don't like downloading software or committing much time. I also don't like time windows. I like doing studies when I have free time, not during required blocks of time.</i>
<i>The individual studies were well-compensated and there was a generous bonus for completing all sessions of the study. Other than that, the study itself was quite unique and enjoyable to complete.</i>
<i>It's interesting to participate in longitudinal studies because it's pleasant to help with a research that monitors our learning/evolution over time in a given subject. This particular study was a monitored study that checked my performance on a repetitive memory task over the weeks. Also, the reward was excellent.</i>
<i>There would be random alerts on my phone (the study work took place within an app but was paid via Prolific) and I really struggled over the course of the fortnight duration - I was effectively a slave to my phone.</i>
<i>I don't find them any different to normal single part studies other than they can be more repetitive but so long as they meet the minimum payment reward on Prolific then I don't have any issue and I don't even care about bonuses for completing all parts because I complete all studies that I am invited to anyway and with Prolific I get instant alerts but you also get e-mail invitations when you aren't available so you can always complete them later on, it is really impossible to miss them and because each part is paid separately and approved individually it is more trustworthy for both participant and researcher.</i>

Table 6.9: Distribution across each theme of the answers collected for question 2.

Theme	Answer Collected	Percentage
<i>worker_features</i>	11	4.03%
<i>platform_features</i>	244	89.34%
<i>answer_useless</i>	17	6.22%
Total	273/300	100%

provides a detailed report of each task in which one participates, thus helping track the whole longitudinal study. Others (7 out of 99, 6.93%) report the availability of individuals with different backgrounds, skills, and so on. Other factors to consider are the easiness of contacting or sending reminders to the same workers using their identifier (16 out of 99, 16.16%). Also, according to two workers (2 out of 99, 2%), the workers' motivation and reliability in helping researchers must be considered. Interestingly, a worker mentions being recruited from the platform through a third-party application that relies on the platform's API. The vast majority of Toloka workers (68 out of 75, 60%) mention the platform's adequacy in general, again without providing particular details. Two of them (2 out of 75, 1.5%) detail their answers by reporting workers' availability and the easiness of contacting them using their identifier. A worker provides an interesting answer; they believe that the platform can not support adequately a longitudinal study because they live in a country with poor network infrastructure.

Some cross-platform factors further emerge when turning to the workers which are either unsure or deny the platform's adequacy. Workers tend to drop out of longitudinal studies. There is no easy way for them to assess requesters' honesty. The respondents also think that usually, workers do not search for longitudinal studies and that platforms should provide a way to separate these studies and standard crowdsourcing tasks.

6.4.2.4 11: Suggestions About Longitudinal Study Design

The last and optional question 11 (P2 part) asked workers to provide any suggestions to requesters that aim to design a longitudinal study. There are 199 out of 300 (66.33%) workers who provide some kind of suggestions. Table 6.11 shows the distribution of the answers across each theme emerged, while Table 6.12 shows a sample of such answers.

The majority of workers (139 out of 199, 69.85%) provide suggestions related to the features of the task to be performed within each session of the longitudinal study, such as its design, scheduling, participant filtering, etc. Six out of 134 workers (4.47%) suggest allowing a reasonable window for completion as this is rarely the only activity in someone's day/week. A worker argues that it is useful to allow skipping a given session if someone can't commit to it once or twice. A few workers (3 out of 134, 2.23%) stress the need of conducting pilot tests for the task to be performed. According to them, using adequate tests can help both requesters find participants that fit the needs of the study, as well as workers that are less likely to quit part-way through. A worker suggests offering different systems on which to take a given session of the study (i.e., desktop devices, smartphones, and so on), while another worker suggests avoiding requiring downloads of any kind. The workers stress the need to have clarity in the instructions and the user interface. They also suggest establishing an understandable sequence of events, identifying changes over time, and providing insight into cause-and-effect relationships. Some workers suggest that some variability may help retain an interest in the overall study. Turning to the suggestions concerning the whole structure of a longitudinal study (5 out of 134, 3.73%), the workers suggest planning the whole set of sessions from the beginning, while at the same time being flexible on the overall schedule, especially when recruiting workers from multiple geographic time zones. They suggest also establishing a sense of progression. This can be done for example by highlighting the differences in their previous answers at the end of each

Table 6.10: Sample of answers provided by workers concerning the adequacy of crowd-sourcing platforms in supporting longitudinal studies.

Worker Answers	Platform
<i>I think that this platform is good for longitudinal studies, especially when a Requester can send email reminders to the Workers about when the follow-up tasks are available to be completed.</i>	MTurk
<i>Yes, I have done tasks like that on this platform before and it went well for me.</i>	MTurk
<i>I don't think so because everything that gets released gets snatched up quickly. Also, the requesters on this platform don't respond much. Before, yes but not most likely not.</i>	MTurk
<i>Yes but it need further improvements for this specific type of tasks such as scheduling improvements etc.</i>	MTurk
<i>Yes, I think it is perfectly suitable given its nature. I do think coordinating longer studies can be more difficult on mturk compared to other platforms, as there are many other studies constantly on the platform and remembering longitudinal studies can be difficult while also keeping up with regular studies. To remedy this, requestors must often use e-mail reminders and other types of reminders, which I have no issues with at all.</i>	MTurk
<i>Yes, I believe it is. This platform is the host of many other studies all of which provide for a professional and safe environment (on both sides, for the requester and surveyee with full disclosure of all procedures. I've had previous experience with a longitudinal study on this platform and I have zero complaints.</i>	Prolific
<i>Yes. The messaging system on Prolific is very useful in this regard, the platform itself can easily be tailored to longitudinal studies, and both the researcher and the participant can rely on Prolific for any support required around the task.</i>	Prolific
<i>Yes I think Prolific works very well, I have Prolific Assistant so get the alerts if I'm on my PC so usually I start them just like any other study but even if you don't then you would be sent an e-mail invitation to remind you so you are very unlikely to ever miss any part of a study and I have completed all parts of any longitudinal studies that I have been part of. I think so long as all of the details are explained in the first part and the participant agrees to complete all of the following parts then they should have very high success rates and if anyone does drop out or has any reason to you can also communicate this via Prolific messaging.</i>	Prolific
<i>Not really, there should be an option to separate normal from longitudinal studies.</i>	Prolific
<i>Yes, but Prolific does not email you outside of itself. This can be a problem if the study requires out-of-band responses. With Mechanical Turk your requests hit email so I get message reminders when I am not at my desk.</i>	Prolific
<i>Yes, it's a nice platform to work, to earn rewards and to learn some new things so it would be a great platform for longitudinal studies too.</i>	Toloka
<i>Yes it fits. I think there is a large number of participants, which makes the study more accurate.</i>	Toloka
<i>I have had good experiences with tasks offered by Toloka. Proper instructions are provided.</i>	Toloka
<i>Yes, it has participants which login every or almost every day, they are interested in completing tasks they are already acquainted with.</i>	Toloka
<i>Yes, it is suitable because most people in this platform work more than five hours everyday</i>	Toloka

Table 6.11: Distribution across each theme of the answers collected for question 11.

Theme	Answer Collected	Percentage
<i>task_features</i>	139	69.85%
<i>worker_features</i>	7	3.52%
<i>requester_features</i>	9	4.52%
<i>platform_features</i>	2	1.01%
<i>ls_features</i>	5	2.51%
<i>answer_useless</i>	2	1.01%
Total	199/300	100%

session. A few workers (7 out of 134, 5.22%) provide suggestions about themselves and their beliefs. A worker stresses that many of them work from home and are self-employed, having thus to pay taxes on the earnings from crowdsourcing platforms; thus, the reward should consider also this aspect. Another worker describes that they like getting small payments and a bonus for completing all of the sessions at the end. Lastly, when considering aspects related to the task requesters (9 out of 134, 13.4%), the workers think that regular feedback from them is important. They also suggest that requesters should be as communicative and friendly as they can, also by leaving blank spaces for feedback in each study. Task requesters should also be keen to send reminders when needed and provide clear upfront information.

6.4.3 RQ14: Recommendations For Researchers And Practitioners

Although there is no standard approach for designing and conducting longitudinal studies on a crowdsourcing platform, our quantitative and qualitative analyses of workers' responses have allowed us to develop 9 recommendations that could serve as a framework. These recommendations should be considered by task requesters when designing longitudinal studies, as they provide useful guidelines and address workers' fears and needs that emerged during the survey.

R1: Be Communicative And Provide Feedback Communication is a critical factor to encourage workers' loyalty and decrease the abandonment rate, as emerges from their answers about what drove them towards returning to longitudinal studies and their suggestions to task requesters. Indeed, according to them, task requesters should inform workers about upcoming sessions, and the progress made throughout the study, and, eventually, they could contact the workers explicitly to invite them to participate in newly published studies. They should also provide information about the longitudinal study's overall progress and feedback concerning the quality of the work performed up to the current session. However, alerts, emails or notifications should be sent according to a regular schedule; sending them randomly could be detrimental to the worker experience. Also, they should keep in mind that platforms like Prolific provide only an internal notification system, without any way to send a standard email to the worker.

Table 6.12: Sample of suggestions provided by workers concerning longitudinal studies.

Worker Suggestions
<i>Establish the correct sequence of events, identify changes over time, and provide insight into cause-and-effect relationships.</i>
<i>Plan each session in a way that it makes the surveyee feel like they're making progress. Maybe at the end of each session highlight the differences in their previous answer to accentuate that feeling of progression.</i>
<i>Beside all of the aspects regarding time and money, fast communication between requester and worker and also regular feedbacks regarding workers task quality would be great to increase their (our :)) commitment.</i>
<i>Maybe offer different platforms on which to take the study (ie android, PC, mac, etc)</i>
<i>Just don't require downloads. Keep tasks short. No time frames.</i>
<i>A lot of us work from home and are self employed so we have to pay tax on these earnings. As long as it pays a decent amount for the time taken (at least £6 per hour), I would be more than happy to take part.</i>
<i>It is useful to allow one or two sessions to be skipped if the responder can't commit to absolutely every session.</i>
<i>Be reasonable with what you expect people to do. People who work full time and have caring responsibilities won't necessarily have the capacity/flexibility to do daily tasks that last an hour or more. If your study makes those demands then you're going to only be getting a certain kind of participant (e.g. unemployed).</i>
<i>Keep them to the point, don't give long, fatigued instructions, try not to ask the same question fifty different ways. Also, if you have a game, games are very attractive for me; I'd be interested in longitudinal studies where we have to play a game and collect something, like points, or something. And gives a good bonus! Good base pay, as well. At least 12 dollars an hour.</i>
<i>Ensure the timings are not onerous when considering participants from multiple geographic zones - they need adequate time to complete. A final bonus payment completion incentive helps reduce attrition - and on that note, keep the study shorter (say 2 weeks) to minimise participant drop-off.</i>
<i>I think you have to be as revealing as possible in the first part of the study so the participant knows in advance what they are signing up for, it would help if the participant gets a good idea or sampling of the task in full so there are no surprises if that is possible so it would be good to have them complete the worst part of it if there is one and if it is repetitive and hard to complete over a longer period then to explain that so they can make a judgement. So long as they know what is involved and what is expected of them in advance before they then agree to take part because then so long as they understand the commitment they are making and the schedule and timing they should be able to complete it.</i>
<i>Using good screeners can both help requesters find participants that fit the needs of the study, as well as participants that are less likely to quit part-way through. Also, compensation schemes that reward consistent participation are likely to increase the odds that participants complete all required sessions of the study.</i>

R2: Schedule Each Session Accordingly Crowd workers have free time to dedicate to participation in crowdsourcing tasks during different days of their working week. Scheduling the work required properly is particularly important for longitudinal studies since they can be composed of a potentially high number of sessions. Task requesters should be mindful and schedule carefully each of them. Determining a priori and stating explicitly such a number is useful since it would allow the worker to estimate the amount of commitment required. Communicating when the subsequent session is going to happen will provide some flexibility to the workers. Furthermore, task requesters should be careful when from multiple geographic time zones; for a worker, the session might start in the morning, while for another one during the night. It could be beneficial to split the work required in multiple batches spread across the whole 24-hour timespan of the day or, at least, provide a high enough time frame for workers to complete a session, with some of them suggesting 24 to 48 hours. Also, there should not be too much time between each session, since workers may become bored or not recall the overall study, and thus drop participation halfway through. The requester could also ponder about allowing workers to skip one or more sessions, to provide additional flexibility.

R3: Workers Fear Performance Measurement Crowdsourcing platforms measure worker performances and quality using various metrics and indicators, such as the time elapsed between accepting a given HIT and its successful submission and the overall completion rate. These indicators can be used by task requesters to filter the pool of available workers. In light of this, workers might avoid participating in longitudinal studies because they somehow believe that participating in a longitudinal study can increase the odds of being rejected at any time after a given session, once completed, thus impacting the completion rates and performance as measured by the platform. In other words, workers fear completion rates. A way to address such an issue is by disclosing and clarifying the whole study's workflow, having a particular focus on the rejection criteria. They should be described accurately along with the behaviors and causes that may trigger them.

R4: Longitudinal Studies Boost Reliability And Trustworthiness Even though longitudinal studies might increase the fear of performance indicators, task requesters should remember that workers find such kinds of studies more reliable than other studies. Such reliability refers to, first of all, the reward provided but also to avoiding having to learn how to perform the work required again or searching for new work, thus optimizing the time available for work. A successful longitudinal study demonstrates researcher honesty, increasing the overall trustworthiness. Hence, task requesters should employ a well-documented task design which is as consistent as possible across sessions, having a sound and understandable sequence of events.

R5: Worker Provenance Affects Their Availability Crowdsourcing platforms allow task requesters to recruit people from all over the world. This may include workers from countries characterized by not adequate network infrastructure. For instance, when considering the Toloka platform it is rather easy to find people from CIS countries [226] (Commonwealth of Independent States), as reported by a worker. Task requesters should carefully consider where to recruit each worker since their provenance can affect profoundly their availability, loyalty, and commitment.

R6: Design Cross-Device Layouts Workers may use various devices to perform crowdsourcing tasks. For instance, the Prolific platform offers task requesters a user interface control to explicitly allow the usage of certain device classes. Moreover, a worker could start working on a desktop device and then switch to a mobile device on the go. This can be particularly true for longitudinal studies since they are made of different sessions that can be performed over an arbitrary amount of days. The requester should thus design and build a layout as cross-platform as possible, thus offering the possibility of using different devices.

R7: Avoid Requiring Additional Software Workers may not agree with being required to download additional software to perform a crowdsourcing task. Task requesters should aim to provide a single (and possibly web-based) interface where the workers can perform the work required, whenever possible. This could be also useful to provide a consistent cross-device layout, as suggested by the previous recommendation. To such an end, the task requester could try to deploy the task's user interface using the tools and building blocks which are provided by some platforms such as Amazon Mechanical Turk and Toloka. However, it must be noted that other platforms (Prolific) do not allow that, so relying on external software or interfaces could not be an option.

R8: Provide Partial Payments The most important incentive to foster longitudinal studies' popularity and motivate workers in participating is the reward. Conducting a longitudinal study offers the possibility to provide a worker with a reward after each session, partially. Indeed, they feel more motivated when participating instead of providing a single final reward at the study's end. Task requesters should thus consider splitting the reward across each session, when applicable. Such a decision might help reduce workers' abandonment rate by further motivating them. Another approach could consist in having a payment that is initially low and then increases as the study progresses. Moreover, a final bonus reward can be provided to the workers once the study end, thus helping maintain a lower reward during the previous stages.

R9: Consider Deploying Pilot And Training Versions Piloting a task to be performed helps reduce worker attrition due to errors and unexpected scenarios within its business logic, and longitudinal studies do not make an exception. Furthermore, a longitudinal study may involve recruiting novice workers during subsequent sessions [362]. Task requesters may consider deploying a lightweight training version of the task to be performed. This will help first-timers and prepare them to perform the overall study as expected.

6.4.4 RQ15: Best Practices For Crowdsourcing Platforms

In the past, researchers have conducted longitudinal studies on crowdsourcing platforms to a certain extent. However, the support for such studies by commercial platforms is not as straightforward as it may seem. Through our analysis of workers' responses and our experience in deploying a crowdsourcing task, we have discovered that even simple goals, such as tracking the overall progress of the study for requesters and workers, are not easily achievable. As a result, we have synthesized a list of 5 best practices that we believe the designers of crowdsourcing platforms should adopt and prioritize to adequately support

longitudinal studies.

BP1: Allow Requesters Sending Reminders To Workers. One of the most pressing issues reported by the workers is the need of being reminded of an upcoming session when committing to a longitudinal study. Let us consider, for instance, Figure 6.15. The reward provided after each session is certainly critical for the workers, but also the commitment required is a prominent factor. This is particularly true for Prolific workers. Toloka workers also believe that longitudinal studies are not optimally supported, and this could be part of the problem. Furthermore, a worker answered by reporting that they enjoyed participating because he had been reminded daily. Hence, the crowdsourcing platform should allow task requesters to remind workers somehow. A solution could be allowing to schedule automatic reminders. They could be scheduled after each session, or after a fixed amount of time, and they should be attached with a customizable message if needed. The reminders could be sent as notifications on the platform's user interface, or as simple email messages.

BP2: Report To Workers The Overall Progress Similarly to reminding workers, allowing workers to understand their progress within a longitudinal study seem a reasonable requirement at a first glance, yet it is achievable to some extent only by Prolific workers today. Other platforms, indeed, provide feedback to the worker concerning a single task published, but this is seldom the case when conducting longitudinal studies. A first solution could be allowing requesters to display in advance the number of sessions of the whole study. This would provide first and prompt feedback to the worker. Another solution to enforce a sense of progress for the worker could be to compute and display performance metrics. They may measure various parameters such as the increment in response quality, the average time elapsed across each session, and so on. For instance, workers reported that they enjoy participating in longitudinal studies to monitor the changes in answers over time. This could thus be another interesting piece of information to be summarized and shown as a performance indicator.

BP3: Support More Advanced Worker Recruitment Strategies A task requester who designs a longitudinal study may need to recruit the same set of workers across a set of sessions. For instance, Roitero et al. [362] designed a longitudinal study about statements said by public figures related to the COVID-19 pandemic. They re-published the same set of HITs several times, where each time they contacted workers who previously participated. Thus, the crowdsourcing platform should provide a way to recruit those workers. One may argue by looking at Figure 6.15 that only a small subset of workers will participate again. After all, they are keen to commit to a longitudinal study for about 19 days on average. Indeed, Roitero et al. [362] measured the task abandonment using the definition provided by Han et al. [172], reporting a 50% abandonment ratio on average. In light of this, the crowdsourcing platform should also provide a way to compensate for the reduced amount of returning workers by explicitly asking the requester whether they want to recruit also novice workers.

BP4: Add Adequate User Interface Filters For The Workers When designing and publishing a study on a crowdsourcing platform, it is not possible to indicate that it will be conducted in a longitudinal fashion, by publishing additional sessions over time. Given that Figure 6.23 shows that several workers think, after all, that longitudinal studies allow

them to avoid spending regular time searching for a new task, we believe that platforms should provide workers with a user interface filter that allows them separating longitudinal studies from standard tasks and offer to requesters the possibility to indicate whether their studies are going to be longitudinal or not. This best practice might be considered rather obvious, yet the workers of every platform considered can only guess or rely on the study's description provided by the requester. This will make workers aware and help them to participate in such kinds of studies and, perhaps, task requesters will obtain the number of workers needed in lesser time.

BP5: Provide Support For Non-Desktop Devices Workers use a multitude of devices to participate in crowdsourcing tasks of any kind. Among the platforms considered, only Prolific allows task requesters to indicate the type of device class (i.e., mobile, desktop or tablet) required to perform a given task, and workers to filter the available tasks accordingly. Crowdsourcing platforms should thus provide task requesters with a way to design a layout adequate for each of these classes. This could be done by providing a set of predefined and responsive user interface components, as done to some extent by Amazon Mechanical Turk with its Crowd HTML elements (Section A.1.1). Otherwise, the platform could provide a way to design different layouts for the same task, one for each device class supported. Then, the workers should be allowed to choose studies compatible with a certain device class using an appropriate filter, similar to the choice between participating in a longitudinal or standard study. This best practice is general and not limited to longitudinal studies design, yet a worker reported that they dropped out of a longitudinal study because, indeed, it was available only for a single device class.

6.5 Summary

This chapter investigates the barriers to running longitudinal tasks on crowdsourcing platforms. A large-scale survey over three popular commercial crowdsourcing platforms is performed to investigate the popularity of longitudinal studies, the worker suggestions and motivational factors needed to successfully carry out this sort of study, and their strengths and weaknesses. Aspects pertaining to the design of longitudinal studies which are critical in shaping their success are identified. These include clear communication with participants, setting worker expectations around reward design and the required participation mode, and a correct strategy to manage or deal with participants abandoning a longitudinal study midway through it, among others.

A qualitative analysis is conducted apart from the quantitative analysis. It is based on workers' answers to the survey, carried out following an inductive thematic analysis approach, which highlighted the main codes and themes in workers' answers. As result, researchers, practitioners and crowdsourcing platforms are provided with a list of recommendations and best practices that should be implemented to successfully conduct and support longitudinal studies using crowdsourcing. The answers to the research questions can be summarized as follows.

RQ12 The detailed quantitative analysis of the workers' responses allows obtaining a breakdown across various crowdsourcing platforms concerning the spreading of longitu-

dinal studies. The analysis hints that workers with more experience with longitudinal studies can be found more easily on the Prolific platform, and the studies published tend to have a longer duration than those available on other platforms, in terms of session number. The time span between a session and the subsequent one ranges from a single day to an entire month. Each session lasts up to 2 hours, and workers from the Toloka platform tend to participate in studies available on other platforms more than the others. The reward is usually provided partially, after each session, and the workers want to keep participating in longitudinal studies. The most important reason that drives them into participating and completing a given study is the reward, and almost every one of them indeed completed the experiences reported. Unsurprisingly, the workers think that the reward provided can also hamper the popularity of longitudinal studies. The workers are keen to commit to such studies for about 20 days, and their duration is a key factor that may lead to refusing participation. They prefer a daily to weekly session frequency each each session should last for about 1.5 hours. The ideal hourly payment, according to the workers, should be 10.5 USD\$, given that they aim to allocate up to 2.7 hours of their daily time for participating. The workers can perform various kinds of tasks during longitudinal studies, without clear preferences. The most important benefit of participating in longitudinal studies is they allow being more productive and avoid searching for new work, while the most important downsides are the long-term commitment required and the lack of flexibility. The answers provided by the workers recruited from different crowdsourcing platforms lead to statistically significant comparisons.

RQ13 The qualitative analysis performed using an inductive thematic analysis approach to identify the main codes and themes in workers' responses, allows for obtaining useful insights. Such findings reveal several barriers to conducting longitudinal studies on crowdsourcing platforms. For instance, workers are more likely to participate in longitudinal studies that provide higher payments, better communication, and a clear sense of progress. Additionally, the lack of effective quality control mechanisms and a clear communication channel between the requester and the workers hinder the success of longitudinal studies on these platforms.

RQ14 The output of the quantitative and qualitative analyses is used to distil a list of 9 recommendations for researchers and practitioners that aim to conduct a longitudinal study. These recommendations include the use of incentives, effective communication channels, clear instructions and quality control mechanisms to improve the overall success of crowdsourcing based longitudinal studies.

RQ4 The output of the quantitative and qualitative analyses is used to synthesize a list of 5 best practices that crowdsourcing platforms should employ to improve their support for longitudinal studies. These best practices include supporting more advanced worker recruitment strategies, supporting non-desktop devices, adding adequate user interface filters for the workers, and implementing feedback mechanisms and communication channels.

The next chapter addresses the idea of having a multidimensional notion of truthfulness, where the crowd workers are asked to judge the truthfulness of information items using seven different dimensions found in the literature. The same statements used in Chapter 4 are employed.

The Multidimensionality Of Truthfulness

This chapter is based on the article published in the “Information Processing & Management” journal [395]. Section 2.1, Section 2.4, Section 2.5, Section 2.6, and Section 2.7 describe the relevant related work. Section 7.1 addresses the research questions, while Section 7.2 describes the experimental setting. Section 7.4 presents the results obtained. Finally, Section 7.5 summarizes the main findings and concludes the chapter.

7.1 Research Questions

Recent work has looked at the possibility to employ crowdsourcing methods to perform fact-checking at scale [232, 361, 363]. Truthfulness scales at different levels of granularity have been compared leading to the conclusion that coarse-grained (e.g., three levels) scales are to be preferred for crowdsourced truthfulness judgments [232]. However, a uni-dimensional truthfulness scale such as the ones described in Chapter 4 and Chapter 5 appear to be too simplistic to capture all the nuances of truthfulness. This chapter thus studies how crowdsourcing truthfulness judgments may be performed by taking a *multidimensional* labeling approach rather than asking annotators to label on a single scale between the “true” and “false” extremes. Specifically, a crowdsourcing task asking US-based crowd workers recruited from Amazon Mechanical Turk to judge the truthfulness of political statements is published. The task is not just based on a single multi-level scale (e.g. like done by Wang [436] with a 6-level scale), but rather using multiple dimensions of truthfulness. The participants are asked to judge a statement on a scale for each of the Correctness, Neutrality, Comprehensibility, Precision, Completeness, Speaker’s Trustworthiness, and Informativeness dimensions.

A large-scale crowdsourcing experiment asking crowd workers to judge political statements with the aim of identifying online misinformation is run. The same set of statements of the experimental setting described in Section 4.2 is used. The statements have been fact-checked by experts of PolitiFact (Section 3.1) and ABC Fact Check (Section 3.2). Differently

from the approaches described in Chapter 4 and Chapter 5, a multidimensional notion of truthfulness is employed. Independent judgments for each dimension are collected from each worker. The workers also judge the Overall Truthfulness of each statement and they had to justify their choice by providing a URL to the web page they used to verify the truthfulness of the statement. The following research questions are investigated:

RQ16 Are crowd workers able to reliably assess multiple dimensions of information truthfulness? How do their judgments correlate with expert judgments?

RQ17 Are all truthfulness dimensions independent from each other, and thus required? Can some dimensions be derived from (a combination of) the others? Is it possible to combine the individual dimensions in a way that improves agreement between crowd and expert judgments?

RQ18 What is the behavior of workers when choosing labels for truthfulness dimensions? Do their cognitive abilities have any influence?

RQ19 How meaningful and informative are the individual information quality dimensions?

RQ20 Can multidimensional judgments be used to accurately predict expert judgments and verdicts?

7.2 Experimental Setting

The experimental setting involves the same 180 political statements ranging from 2007 to 2019 described in Section 4.2 and sampled from the PolitiFact and ABC Fact Check datasets (Chapter 3). This allows directly seeing the impact of a multidimensional scale, as well as providing the research community with two sets of annotations referring to the same set of statements. Table 7.1 shows a sample of the statements used, similar to the one of Table 4.1

Table 7.1: Example of statements sampled from the PolitiFact and ABC Fact Check datasets.

Dataset	Label	Statement	Speaker	Year
PolitiFact	True	Washing your hands and covering your mouth when you cough makes a huge difference in reducing transmission of the flu.	Barack Obama	2009
ABC Fact Check	True	Under this government, the tax to GDP ratio has, in the period weve been in office, [been] an average of 22.7 per cent	Kevin Rudd	2013

7.2.1 The Seven Dimensions Of Truthfulness

The main difference of the experimental setting from the one described in Section 4.2 is that each worker is asked to assess seven different dimensions of truthfulness more than just the Overall Truthfulness of the statement. The following dimensions are used as presented to the workers, who are also shown an example for each dimension.

1. **Correctness:** the statement is expressed in an accurate way, as opposed to being incorrect and/or reporting mistaken information.
2. **Neutrality:** the statement is expressed in a neutral/objective way, as opposed to subjective/biased.
3. **Comprehensibility:** the statement is comprehensible/understandable/readable as opposed to difficult to understand.
4. **Precision:** the information provided in the statement is precise/specific, as opposed to vague.
5. **Completeness:** the information reported in the statement is complete as opposed to telling only a part of the story.
6. **Speaker's Trustworthiness:** The speaker is generally trustworthy/reliable as opposed to untrustworthy/unreliable/malicious.
7. **Informativeness:** the statement allows us to derive useful information as opposed to simply stating well known facts and/or tautologies.

A detailed description of each dimension and the examples provided to the workers can be found in Appendix E. The choice of dimensions is informed by previous work. In the information systems literature, information quality and user satisfaction are two major dimensions for evaluating the success of information systems [203]. These two facets can be further split along different characteristics. Given that we are mainly interested in news truthfulness, we focused on information quality characteristics, such as accuracy and precision. The ISO 25012 Model [193] derived these dimensions from various related works [203, 425, 435]. The dimensions of Correctness, Completeness, Precision, Comprehensibility, and Neutrality considered in our work are thus motivated by the ISO Model and are intended to describe information quality. In addition, we also considered two additional dimensions, Speaker's Trustworthiness and Informativeness, which find motivations in the literature. Jowett et al. [200] highlights the influence of the speaker's trustworthiness in relation to the judgment towards a statement the reliability of the source is one of the relevance dimensions catalogued by Barry et al. [33]. Ceolin et al. [60] and Maddalena et al. [266] use Informativeness among other dimensions to perform crowdsourcing tasks dealing with information quality assessment. It is important to note that these additions are necessary, since the ISO model focuses on data quality, while the experimental setting aims to assess the quality of the information represented by such data. Thus, the subset of dimensions from the ISO model that are relevant in this context is considered and extended with additional ones motivated by the literature.

In more detail, the same dimensions employed by Maddalena et al. [266] are considered in the experimental setting. Maddalena et al. perform a crowdsourcing experiment with the aim of understanding if the crowd is a valid alternative to the experts for the task of information quality assessment. They used almost the same dimensions previously detailed by Ceolin et al. [60], who present an experiment aimed to perform user studies considering web documents about the vaccination debate. Maddalena et al. slightly reformulate the description of some dimensions to adapt them and make them more adequate for the crowdsourcing context. Both studies found that using such a set of dimensions, crowd workers and experts perform well reaching a satisfactory level of external agreement when comparing the crowd and expert labels. Summarizing, those particular seven dimensions are considered because they find a theoretical grounding and are proven to lead to a good level of external agreement, allowing to capture information accuracy and appropriateness.

7.2.2 Crowdsourcing Task

The Amazon Mechanical Turk crowdsourcing platform is used to collect data. When a worker accepted a Human Intelligence Task (HIT), he/she was shown an input token and a URL to an external server which contained a deployment of our web application (i.e., the actual task). The worker carried out the assigned HIT on such an external application (Appendix A). If they successfully completed the HIT the worker is shown an output token, which has to be copied back to the platform's page to receive the payment upon approval.

The task itself is as follows. First, a (mandatory) questionnaire is shown to the worker, to collect background information such as age and political views. Then, the worker needs to provide answers to three Cognitive Reflection Test (CRT) questions. These questionnaires are those already described in Section 4.2.1. The workers are then asked to assess 11 statements selected from PolitiFact (6 statements) and ABC Fact Check (3 statements) dataset. Each HIT contains a statement for each truthfulness label of the PolitiFact and ABC Fact Check datasets, plus 2 special statements used for the purpose of quality checks. Each HIT is built using a randomization process to avoid all possible sources of bias. In more detail, each crowd worker is first asked to provide the Overall Truthfulness of the statement and a Confidence level of the knowledge of the topic. Then, the worker had to provide the URL that he/she used as a source of information to assess the Overall Truthfulness. Such a URL had to be found using the customized search engine (Section A.3.4) which allows to filter out PolitiFact and ABC websites from search results. Then, each worker is also asked to assess the seven different dimensions of truthfulness described in Section 7.2.1. Each judgment was expressed on the following Likert scale [250]: Completely Disagree (-2), Disagree (-1), Neither Agree Nor Disagree (0), Agree (+1), Completely Agree (+2). The quality checks concerning the selected URL, the gold questions, and the time spent on each statement described in Section 4.2.1 are implemented also for this task. The set of instructions shown to the workers and containing a detailed description of the assessment process is available in Appendix E.

Each HIT reward is 2 USD\$ including the set of 11 judgments, computed on the basis of the time needed to finish the task and the U.S. Minimum Salary Wage of 7.25 USD\$ per hour. Overall, 180 statements in total are used, as outlined in Section 4.2. Each statement is evaluated by 10 distinct crowd workers. Thus, 200 HITs are published on Amazon

Mechanical Turk and 2200 judgments in total are collected. The crowdsourcing task was launched on June 1st, 2020 and it finished on June 4th, 2020.

7.3 Descriptive Analysis

Overall, 200 crowd workers successfully complete the experiment. Amazon Mechanical Turk allows to select workers living within a certain country and each worker must provide some personal info when subscribing such as the home address. We request only U.S. citizens.

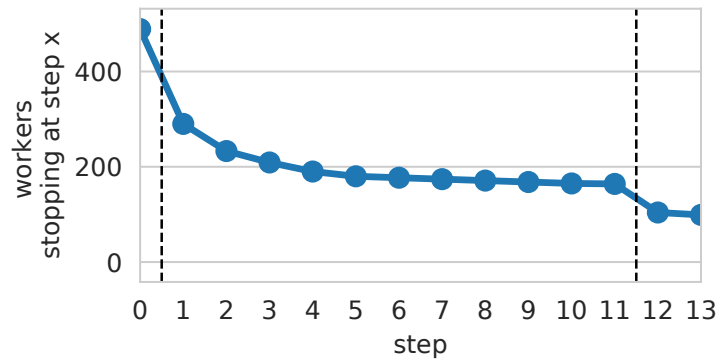
7.3.1 Worker Demographics

Nearly 49% of workers (95/200) are between 26 and 35 years old. The majority of workers (52%) have a college/bachelor's degree. As for total income before taxes, the 22% earn 50,000\$ to less than 75k\$, while the 18% earn 40k\$ to less than 50k\$. Turning to their political views, 33% identify their political views as Liberal, 22% as Moderate, and 16% as Conservative. The majority of workers (46%) consider themselves as Democrats, while the 28% as Republicans and the 23% as Independents. The majority of workers (53%) disagree with building a wall on U.S. southern border while the 40% agree. Finally, the vast majority of workers (85%) think that the government should increase environmental regulations to prevent climate change, while only 9% disagree. In general, the sample is well balanced, with the only exception of a few categories. Furthermore, it is aligned to those of other tasks (Section 4.3.2 and Section 5.3.2).

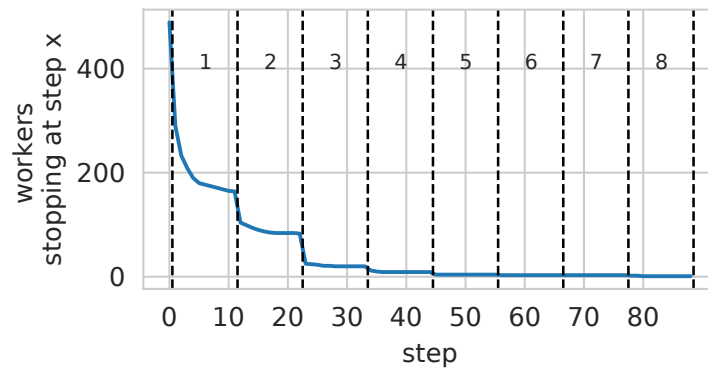
7.3.2 Task Abandonment

To quantify how many workers abandoned the task, the abandonment rate is measured using the definition provided by Han et al. [174]. Overall, 200/681 workers (about 29%) successfully complete the task while 355/681 workers (about 52%) abandon it (i.e., voluntarily terminate the task before completing it), and 126/681 (about 18%) fail (i.e., terminated the task due to failing the quality checks too many times). Furthermore, 184/651 workers (about 27%) abandon the task before really starting it; in other words, right after the completion of the initial questionnaire.

Figure 7.1a shows the abandonment rate breakdown across task steps. A worker reaches the next step when he/she completed the judgment of a single statement. Therefore, a task is completed if the worker judges each statement within the current attempt. This definition does not make any assumption on task success. Step 0 is the questionnaire, and each submission attempt occurs every 11 steps (since each HIT is composed of 11 statements). The abandonment rate monotonically decreases when the step number increases. There are two consistent drops of such amount that occur (highlighted by the dashed vertical lines in the figure). Many workers abandoned the task when they completed only the questionnaires, i.e., at step 0. The second drop occurs between step 11 and step 12, i.e., when they complete and fail the first attempt thus becoming bored or frustrated. Some workers performed up to 8 attempts before abandoning the task. These abandonment



(a) Questionnaire and first attempt.



(b) All attempts.

Figure 7.1: Abandonment rate shown as number of workers that reached a certain number of steps in the task. The abandonment monotonically decreases as the step number increases.

distributions are aligned with those described in previous work [174] and in Section 4.3.2 and Section 5.3, thus providing a first confirmation of the quality of the data.

7.4 Results

Section 7.4.1 discusses the reliability of multidimensional judgment. Section 7.4.1 studies the relationships and independence of each dimension of truthfulness. Section 7.4.3 addresses the usage of worker behavior as a proxy for quality. Section 7.4.4 evaluates the informativeness of the multidimensional judgments (not to be confused with the truthfulness dimension called *Informativeness*). Lastly, Section 7.4.5 studies the usage of a machine learning-based approach to analyze the usefulness of the multidimensional assessments and worker behavior in predicting expert judgments.

7.4.1 RQ16: Reliability Of Multidimensional Judgment

The reliability of multidimensional judgment is assessed by analyzing firstly the distributions of the individual and aggregated judgments provided by crowd workers (Section 7.4.1.1). Then, Section 7.4.1.2 studies the external agreement with experts, while Section 7.4.1.3 studies the internal agreement among workers. Section 7.4.1.4 addresses worker behavior while judging each truthfulness dimension. Lastly, Section 7.4.1.5 summarizes the main findings about the reliability of judgments.

7.4.1.1 Distributions Of Judgments

Figure 7.2 analyzes the correlation between the different dimensions. The heatmaps in the lower triangular matrix show the individual judgments collected for each dimension. There is a total of 28 heatmaps, one for each pair of dimensions. For each heatmap, each cell shows how many times the judgments are equal for the considered pair of dimensions. The histograms on the diagonal show the distributions of the individual judgments for both PolitiFact (blue) and ABC Fact Check (orange), for each dimension. Note that half of ABC Fact Check judgments are collected compared to PolitiFact. Each distribution is skewed to the right (i.e., towards higher truthfulness values) showing that workers tend to agree with statements more than disagree, or at least do not have a strong opinion. Since the subset of statements is balanced, as described in Section 7.2.2, this means that workers tend to agree also with false statements. However, this may be due to the scale used, which is different with respect to the original (Section 4.2.1). The individual judgments are then aggregated using the arithmetic mean since previous work [356, 232] and Section 4.4.2 show that it allows obtaining a better result. The scatterplots in the upper triangular matrix show how the aggregated judgments of each pair of dimensions correlate, for both PolitiFact (blue) and ABC Fact Check (orange). Each point within a plot represents a statement. The histograms on the bottom row show the distributions of the aggregated judgments for both PolitiFact (blue) and ABC Fact Check (orange). The distributions become roughly bell-shaped and lightly skewed to the right for each dimension.

Overall, the correlation values shown in Figure 7.2 for both individual (heatmaps in the lower triangular matrix) and aggregated judgments (scatterplots in the upper triangular

matrix) is always positive, as it would be expected since all the seven dimensions share the same positive connotation. Correlations are sometimes even quite high (e.g., $\rho = 0.86$ between aggregated Correctness and Overall Truthfulness for PolitiFact statements), thus demonstrating some relations between different dimensions. However, some correlations are lower (e.g., $\tau = 0.24$ and 0.2 for Neutrality and Comprehensibility), thus highlighting somehow higher independence between those dimensions.

7.4.1.2 External Agreement

Figure 7.3 shows a chart of the workers scores aggregated with respect to the corresponding expert scores. Three dimensions are reported: Overall Truthfulness, Correctness, and Precision. Before commenting on these charts, some remarks are needed. First, it must be noted that the set of expert judgments is available only for the dimension named Overall Truthfulness, thus the remaining dimensions show the perceived value of the statements on each dimension with a breakdown on the PolitiFact and ABC Fact Check categories. So, while Overall Truthfulness is meant to be correlated with experts' judgment, Precision captures orthogonal and independent information. This is reflected by the different trends of the workers' median scores reported in the figure. Second, the judgment scales used by the workers and by the experts on the Overall Truthfulness are slightly different: The crowd workers provided their judgments on a five-level Likert scale, while experts provided their judgment on either a six (for PolitiFact) or a three-level (for ABC Fact Check) ordinal scale. These scales are different both on the number of levels (six or three versus five) and also on the psychological interpretation of such scale.

That being said, it is now possible to analyze Figure 7.3. The ground truth is on the horizontal axis (PolitiFact on the left and ABC Fact Check on the right) and the aggregated crowd judgments on the vertical axis. Each dot is a statement, the boxplots show the breakdown of the distributions (quantiles and median) for each ground truth level. Focusing on Figure 7.3a, it can be seen that the median values are clearly increasing for Overall Truthfulness (directly corresponding to the ground truth), but not necessarily so for the other dimensions (not directly related to the ground truth) when going towards increasing truthfulness according to the ground truth (in other words, moving towards the right-hand side of the charts). Considering Overall Truthfulness, it is an indication that crowd workers provided judgments which are in agreement with the experts, despite the two sets of judgments being on two different scales, both theoretically and psychologically. The correlation between Overall Truthfulness and the ground truth can be compared with shown with the similar three plots shown in Figure 4.2 (one for each scale used to collect truthfulness judgments). There is no noticeable qualitative difference, despite the judgments being again of different scales. Overall, we can say that the crowd workers recruited for participating in the task described in Section 7.2.2 provided judgments of comparable quality to the one described in Section 4.2.1.

Furthermore, the charts on the centre and on the right of Figure 7.3 show that the specific dimensions of Correctness and Precision have a different appearance, and it can thus be considered somehow orthogonal to Overall Truthfulness. The other dimensions (not shown) show similar behavior to either Precision or Overall Truthfulness. It must be remarked that expert judgments for the dimensions with the exception of Overall

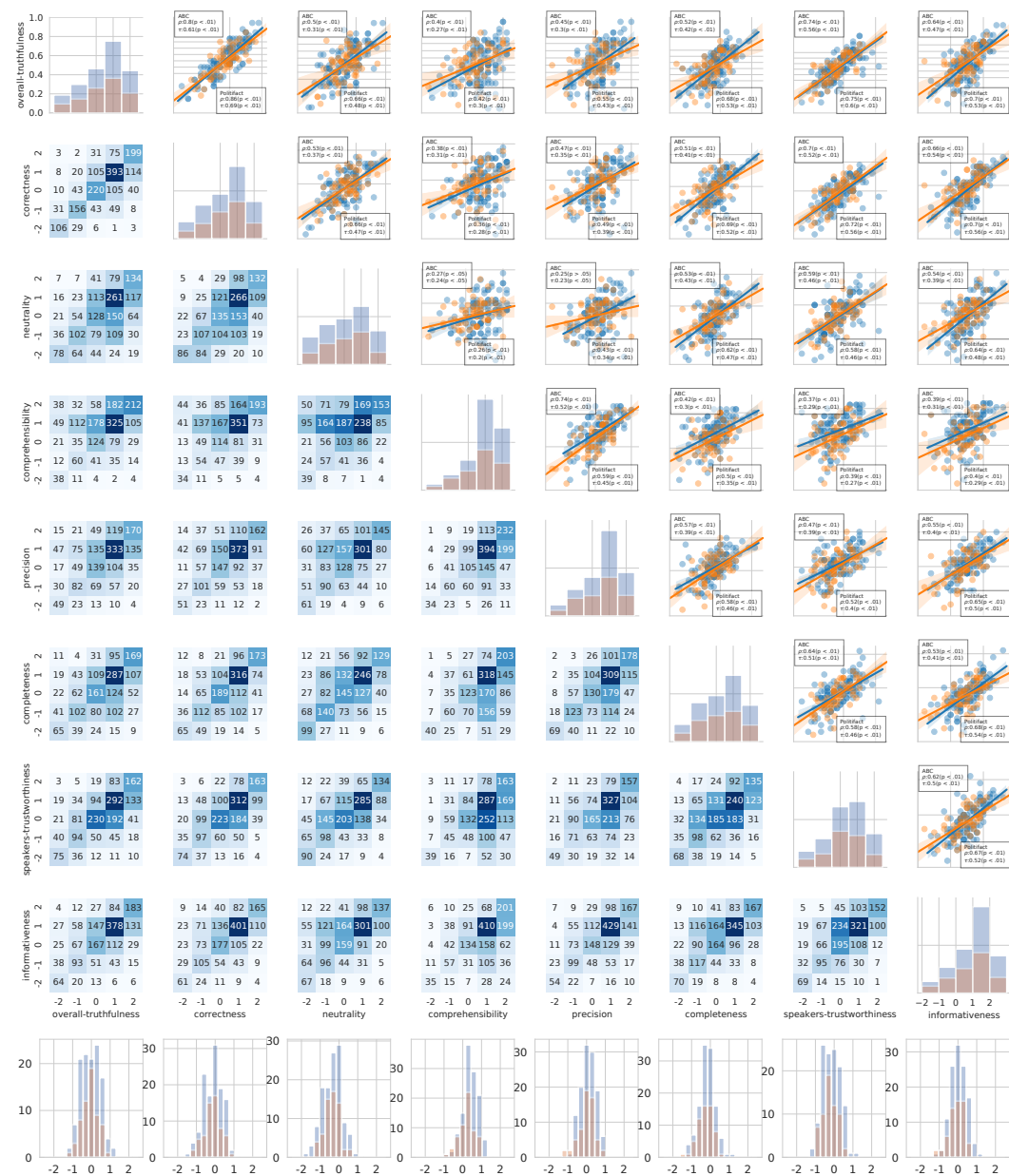
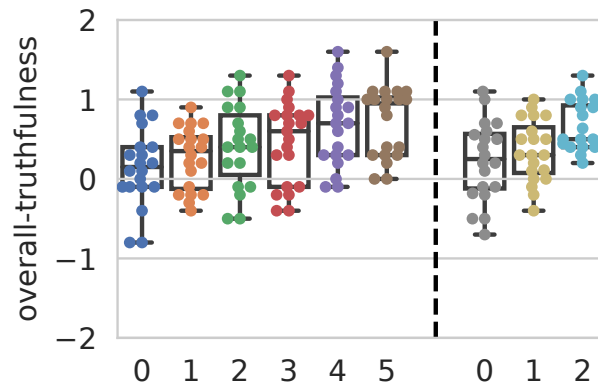
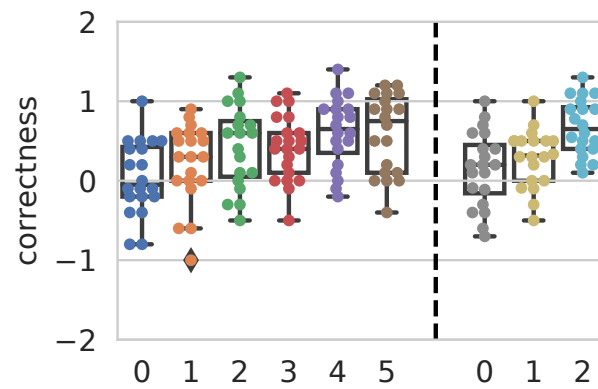


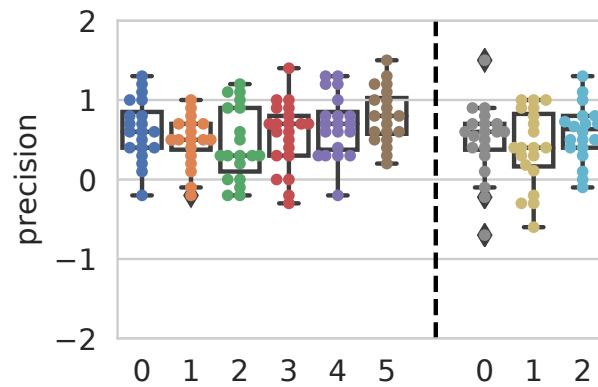
Figure 7.2: Correlation between dimensions: individual in the lower triangle and diagonal, aggregated in the upper triangle, aggregated distribution in the last row; breakdown on PolitiFact (in blue) and ABC Fact Check (in orange) categories (better on screen and using the zoom feature). Workers values are skewed towards positive values, i.e., Agree (+1) and Completely Agree (+2) (diagonal and bottom plots), and different dimensions have different correlation values (upper and lower triangle).



(a) Overall Truthfulness



(b) Correctness



(c) Precision

Figure 7.3: Correlation with the ground truth of the Overall Truthfulness and behavior of the Correctness and Precision dimensions with a breakdown on PolitiFact and ABC Fact Check labels. Mean as aggregation function.

Truthfulness are not available. Thus it does not make sense to directly correlate the other dimensions with the expert judgments, as each dimension can measure different aspects from the ground truth (e.g., the Precision of a statement is not necessarily related to its truthfulness). However, it might make sense to combine different dimensions to obtain a better measure of truthfulness.

The number of times in which the aggregated values shown in Figure 7.3a correspond to a value which is at the same distance between two values of the judgment scale used (i.e., the average is $x.5$, for x in the scale, $0 \leq x \leq 4$) is computed. This allows investigating the perceived disagreement between the expert and crowd judgments on Overall Truthfulness, given that the two sets of judgments collected are on different scales. This happens for about 20.5% of statements. The result is compared to each judgment scale used in Section 4.2.2, since the set of statements is the same. Thus, the percentages of statements are very close, respectively of 19.4%, 18.3% and 23.9% when considering the three-, six-, and one-hundred-level scales. This is an indication that the perceived disagreement between experts and crowd workers is not dependent on the scale used to collect the judgments, but it is attributable to other factors.

In order to check if the agreement between experts and crowd workers can increase when considering a coarse-grained scale, the ground truth levels are grouped together, as done in Section 4.4.2.2. Figure 7.4 shows the correlation values between Overall Truthfulness and expert ground truth obtained by binning PolitiFact ground truth categories into 3 bins using mean as aggregation function. With respect to Figure 7.3, this binning allows to slightly improve some correlation values and to obtain a clearer trend of increasing median values for Overall Truthfulness. This result holds across each truthfulness dimension (the charts show five of them) and is consistent with Section 4.4.2.2 findings.

7.4.1.3 Internal Agreement

The Krippendorff's α [224] metric is used to measure the internal agreement both on the different ground truth level and at the unit level, as done in Section 4.4.1.2 and Section 5.4.1.2. The choice of using Krippendorff's α is motivated not only by previous work [232], but also by theoretical reasons since other agreement metrics are not suitable for this setting. Cohen's κ [81] is used to compute agreement in the case of two assessors. Fleiss' κ [147], which generalizes Cohen's κ to multiple assessors, can be only used when they assign categorical ratings, i.e. when they classify items. None of these can be applied to the current setting, where there are several assessors (i.e., 10) and an ordinal classification problem (i.e., the categories are ranked). For these reasons, Krippendorff's α is chosen to compute the agreement with multiple assessors on non-nominal scales. For further analysis on the agreement, metrics see [68, 147, 231].

Results show that, overall, the agreement level is rather low. The α values for all the dimensions are in the $[-0.02, 0.08]$ range when computed for the statements altogether, in the $[-0.02, 0.1]$ range when computed on the three ABC Fact Check categories, and in the $[-0.02, 0.1]$ range (with a mean value of 0.03) for the PolitiFact categories, with the exception of the Mostly-False and Half-True categories which are in the $[-0.02, 0.14]$ range (with a mean value of respectively 0.09 and 0.05). It is known that α values are dependent on the amount of data and the evaluation scale considered [68]. Since both factors are fixed in the

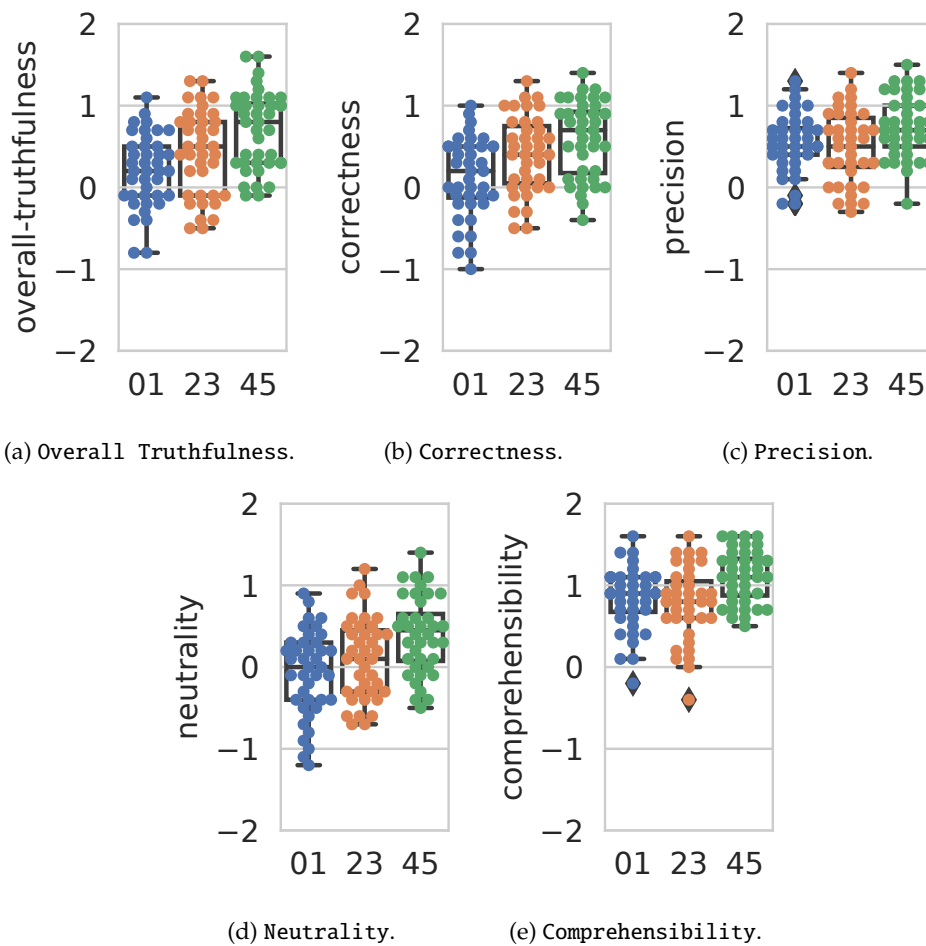


Figure 7.4: Correlation with the ground truth of Overall Truthfulness and a sample of the other dimensions. PolitiFact has been grouped into 3 bins. Mean as aggregation function. The binning allows seeing more clearly the increasing median trends. Compare to Figure 7.3.

experiments, this might be an indication that workers agree more when judging statements on the middle of the truthfulness scale.

7.4.1.4 Behavioral Data

Concerning the analysis of workers' behavior while judging each truthfulness dimension, Figure 7.5 shows the average time spent by each worker to select a value for the Overall Truthfulness for each statement position. There is a clear indication of a learning effect since the average time spent to select a value decreases while the statement position increases. To support this finding, the statistical significance of the time values between each statement position is measured. The differences are statistically significant with a $p < .01$ level when considering positions 1 and 2 compared to any other position. When considering positions 3 and 4, there are statistically significant differences with a $p < .05$ level only with respect to the first two and the last two positions. These findings confirm that there is a learning effect within the first two positions which can last up to the fourth position, and after the fourth statement, the workers evaluate the subsequent statements in the same amount of time.

Workers spend most of the time assessing Overall Truthfulness because they are required to provide also a URL as justification for their choice. When considering other dimensions workers spend much less time selecting a value and there are no clear trends visible. This is probably due to the fact that workers ponder about the value to assign to other dimensions while assessing Overall Truthfulness. In more detail, the average time spent to select a value for other dimensions corresponds to 1.7 seconds for Confidence, 3 for Correctness, 4.1 for Neutrality, 5 for Comprehensibility, 6.2 for Precision, 7.1 for Completeness, 8.3 for Speaker's Trustworthiness and 9.4 for Informativeness, much lower than the average time spent to assess the Overall Truthfulness, which is 85 seconds.

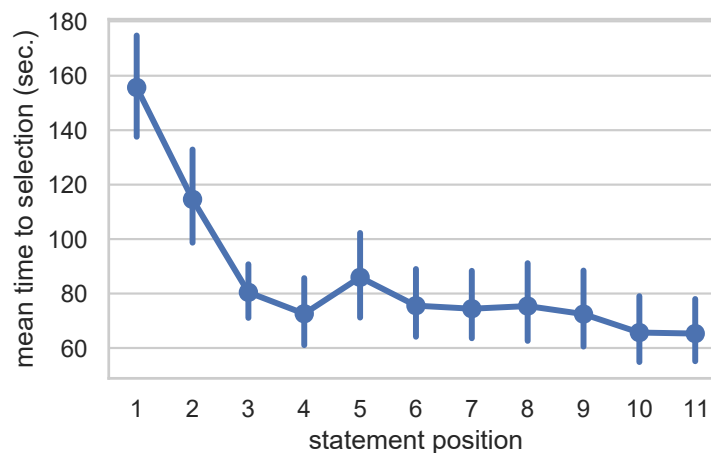


Figure 7.5: Average time (in seconds) spent by workers to judge the Overall Truthfulness for each statement position.

7.4.1.5 Summary

Overall, from the analyses in Section 7.4.1 several remarks can be drawn.

1. Workers tend to agree with statements more than disagree, and since the dataset is balanced this holds also for false statements (Figure 7.2).
2. Workers have on average a similar level of agreement on the set of statements they judge and an increasing ground truth level corresponds to increasing judgments by them, for Overall Truthfulness (Figure 7.3 and Figure 7.4)
3. Workers tend to agree more when judging statements on the middle of the truthfulness scale.
4. Workers learn how to judge the Overall Truthfulness (Figure 7.5).

7.4.2 RQ17: Independence of the Dimensions

The results reported so far show that the various dimensions, as well as Overall Truthfulness, are correlated to some extent. It must be understood whether they anyway measure different aspects, or if some of them could indeed be derived from the other ones. Figure 7.2 show higher and lower correlations. The charts on the bottom left, concerning non-aggregated judgments, show higher correlations for Correctness with both Overall Truthfulness and Speaker's Trustworthiness. The same is confirmed for aggregated judgments, shown on the top right, for which also Pearson's ρ and Kendall's τ correlation values are included. Focusing on the correlation of Overall Truthfulness with the seven dimensions (first row/first column) it appears clear that Neutrality, Comprehensibility, and Precision (0.48, 0.30, 0.43 τ respectively for aggregated judgments over PolitiFact statements and 0.31, 0.27, 0.30 τ for ABC Fact Check statements) do not correlate well with Overall Truthfulness. Completeness, Speaker's Trustworthiness, and Informativeness are slightly higher (0.53, 0.60, and 0.53 τ respectively for aggregated judgments over PolitiFact statements and 0.42, 0.56, 0.4 τ for ABC Fact Check statements) but not as high as Correctness. To summarise, given a statement each of the various dimensions measures a different aspect of truthfulness and is different from the Overall Truthfulness as well. This is true both when we look at individual worker assessments as well as at assessments aggregated over all workers who judged the same statement.

Reconsidering Figure 7.3 and Figure 7.4 allows finding further confirmation of the independence of the dimensions since it is true that all trends are similar, but there are also clear differences. Seeking for further evidence, the ANOVA analysis is employed to correlate the Overall Truthfulness as a function of the other dimensions. The ω^2 index [314] to measure the size of the effect of each dimension in estimating the Overall Truthfulness after fitting the ANOVA model. The Overall Truthfulness score is mainly influenced (by one order of magnitude) by the Correctness ($\omega^2 = 0.228$), followed by Speaker's Trustworthiness ($\omega^2 = 0.019$) and Informativeness ($\omega^2 = 0.017$). Comprehensibility ($\omega^2 = 0.008$), Completeness ($\omega^2 = 0.001$), Precision ($\omega^2 = 0$), and Confidence ($\omega^2 = 0$) have almost no effect. Another ANOVA model is fitted to investigate the interactions between

dimensions. Results show that all interactions are weak ($\omega^2 \leq 0.04$) suggesting that indeed all the dimensions are somehow orthogonal and measure different aspects of the truthfulness of the statements. Nevertheless, the analysis of the interaction between dimensions also shows that all dimensions are used by the workers when judging the statements, and thus all dimensions are necessary (i.e., there is no redundant one). Investigating if other dimensions can be added to the existing ones in order to capture even more aspects when evaluating a statement is left for future work.

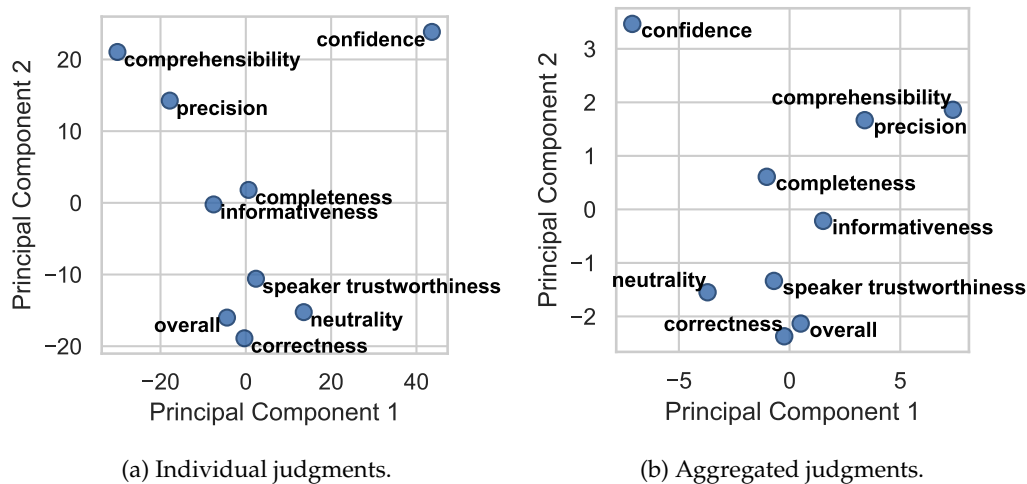


Figure 7.6: Principal components for the statements \times judgments matrix.

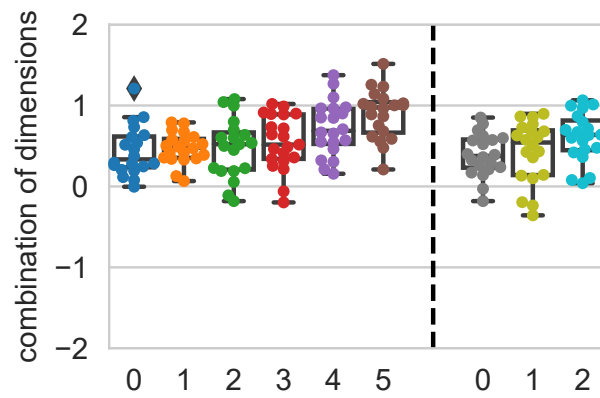
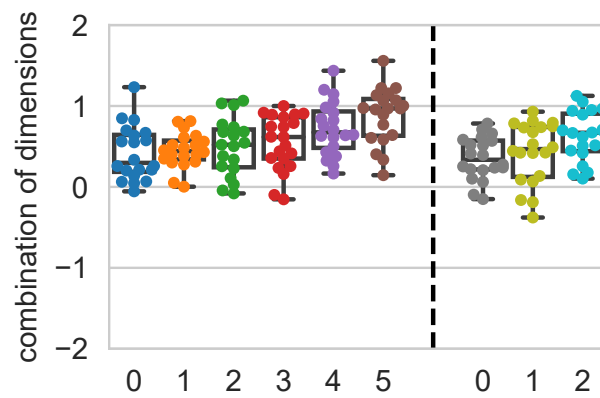
The individual and aggregated judgments to build a statement \times judgments matrix, To further study the relationships and independence of dimensions. The Principal Components Analysis (PCA) of such a matrix is computed to find the orthogonal bases which explain the maximal variance of data. In the computed space with the new coordinate system, the two components (i.e., dimensions) which explain the majority of the variance are considered. Figure 7.6 shows the result of the PCA analysis on the individual (Figure 7.6a) and aggregated (Figure 7.6b) judgments. The most similar dimensions to Overall Truthfulness are Correctness, Speaker's Trustworthiness, and to a lesser extent, Neutrality. This can be seen especially by focusing on the position of the other dimensions with respect to Overall Truthfulness. This behavior holds for both the individual and aggregated judgments. It makes sense that when a worker provides a judgment for the Overall Truthfulness of a statement, the dimensions which are more correlated with its judgments are the ones identified by the PCA analysis. On the contrary, Figure 7.6 shows that other dimensions, such as Confidence, Comprehensibility, and Precision, are not related to any other dimension and are the most distant from Overall Truthfulness as well. Again, this behavior makes sense thinking about the process of judging the truthfulness of a statement. In future work a study must be conducted to investigate if the same behavior is present in the expert judges. Summarizing, the PCA analysis confirms that all dimensions are needed and different, and allows drawing meaningful information on the relationships and similarities between those dimensions.

The following addresses whether it is possible to combine the individual dimensions in a way that improves agreement between the crowd and expert judgments. Since the individual dimensions measure different aspects, a hypothesis can be that a combination of the judgments on certain individual dimensions could lead to a better approximation of the ground truth than using the Overall Truthfulness only. The judgments collected for each truthfulness dimension can be combined together and used to predict the ground truth categories for both PolitiFact and ABC Fact Check. To do so, the ANOVA analysis is employed using the ω^2 index to estimate the size of the effect of each dimension when used to estimate the ground truth. It must be noted that the ground truth values for the statements are not available in the real setting, thus the combination of dimensions is being estimated in a sort of ideal scenario. After computing the ω^2 index of each dimension, the 10 judgments collected for each statement are aggregated using the weighted mean function, where the weights are the ω^2 values. Figure 7.7a shows the correlation values between the label obtained by combining each dimension and ground truth categories. Overall, combining the dimension still allows to obtain increasing median values when moving towards higher truthfulness values, but it does not seem an improvement of Figure 7.3a.

Another approach involves using the CRT answers. First, all the judgments are aggregated using weighted mean where the weights are the ratio of correct answers given by each worker to the CRT questions normalized in $[0.5, 1]$ interval (i.e., we weight more the judgments from high-quality workers). Then, all the dimensions are combined using a weighted mean function where the weights are the ω^2 scores computed above. Figure 7.7b shows the result. There is no significant difference with respect to the aggregation shown in Figure 7.7a.

To better understand this somehow negative results in the combination of dimensions, the ANOVA analysis is used again. In more detail, two ANOVA models are fitted. The former correlates the ground truth values to all the dimensions. The latter correlates the ground truth values to the Overall Truthfulness dimension alone. Results show that the residual in both cases is very similar, indicating that there is no major difference when trying to predict the ground truth label using the Overall Truthfulness alone or a combination of all the dimensions. Similar analyses have been proposed to understand the contributions of each component to the quality of a system [141, 142, 143, 355, 466]. The ω^2 index for the latter model is rather low (i.e., 0.02), indicating that indeed the Overall Truthfulness dimension alone is not sufficient to predict the ground truth label, neither are naive combinations of the dimensions.

It seems that an effective combination of dimensions cannot be achieved by simple models. More complex approaches are left for future work. These approaches include hierarchical models (that might require a modification in the experimental analysis), the combination of dimensions by means of complex (e.g., non-linear) functions, or even the usage of other data as the URL provided. Requesting additional information from the worker such as a confidence value and a textual justification for each dimension might help. This, however, will probably require a slightly different experimental design to avoid overloading the worker.

(a) Combination using ω^2 values.

(b) Combination using CRT scores.

Figure 7.7: Truthfulness dimensions first aggregated using the mean function then combined. Compare with Figure 7.3.

7.4.3 RQ18: Worker Behavior

An attempt to consider worker behavior as a proxy for worker quality is made, considering the still inconclusive results from the combination of dimensions. The aim is to boost the correlation values between the collected judgments and the ground truth. The main idea is to give more weight to the workers with higher quality and to use the CRT answers to estimate worker quality.

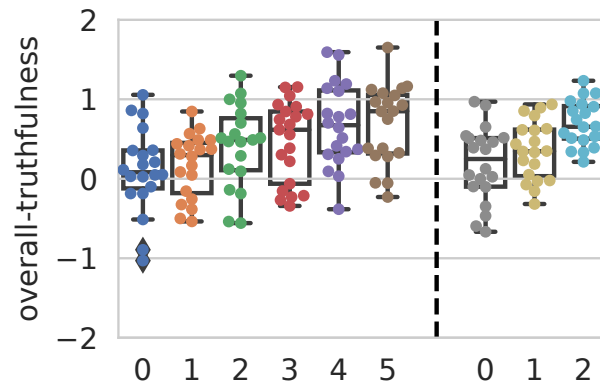
The individual judgments are aggregated with the weighted mean function, using as weights the normalized CRT scores. For each worker, the amount of correct answers (out of 3) is considered for the CRT questionnaire and the score is normalized in the [0.5, 1] range. Figure 7.8a shows the correlation of the Overall Truthfulness values obtained by such a weighted mean with PolitiFact and ABC Fact Check ground truth. Figure 7.8b shows the result when grouping the categories into 3 bins. The resulting plots are very similar to the top left plots in Figure 7.3 and Figure 7.4, thus it seems that this approach does not improve the correlation with the ground truth. More complex worker behaviors and their relations with aggregation functions will be investigated in future work.

It must be also remarked that when considering the individual (i.e., not aggregated) judgments for each statement without gold questions, the majority of workers tend to use distinct labels to provide the judgment. Each worker provides 8 judgments by choosing labels from a set of five possible values, without considering self-reported confidence. Only 12% of workers used the same label for all dimensions, whereas 29% used two distinct labels, 39% used 3 distinct labels, 18% used 4 distinct labels, and 2% used all 5 distinct values. The majority of workers tend to use most of the judgment scale to provide their judgments. This is another confirmation of dimensions independence (RQ17) and Section 7.4.2) and shows how different dimensions cover different aspects of truthfulness.

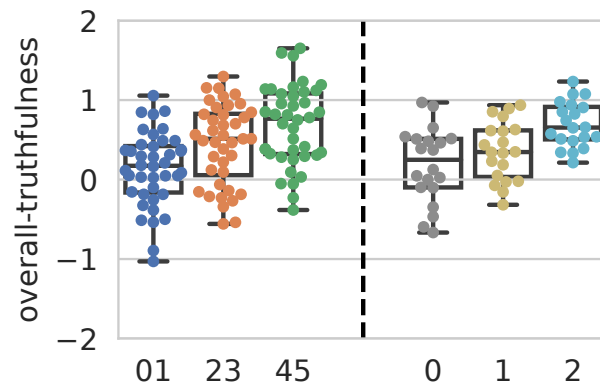
7.4.4 RQ19: Dimension Informativeness

The informativeness of the multidimensional assessments is evaluated. It must not be confused with the truthfulness dimension called Informativeness used to evaluate the statements. First, the possibility of synthesizing these judgments computationally is tested. The two dimensions for which we found computational counterparts are Comprehensibility and Correctness. Readability measures determine the understandability of text which might affect Comprehensibility. The readability of all the statements is computed for 10 measures:

- Flesh Kincaid Reading Ease;
- Flesh-Kincaid Grade Level;
- Automated Readability Index [389];
- Gunning Fog Index [218];
- Dale-Chall [93];
- Simple Measure of Gobbledygook (SMOG, [280]);
- Coleman-Liau Index [82];
- Forcast [57];
- Lesbarhets Index and Rate Index (LIX, RIX, [235]).



(a) CRT scores.



(b) PolitiFact categories are grouped together.

Figure 7.8: Overall Truthfulness judgments aggregated with the weighted mean function. Compare with Figure 7.3 and Figure 7.4.

All of them show a low correlation with the *Comprehensibility* scores (with a maximum $\rho = 0.19$ for RIX). Thus, the information provided by the workers with the *Comprehensibility* scores is hardly captured by automated readability scores, and thus it is a significant measure to be crowdsourced. Also, the *Correctness* scores are compared with the statement polarity computed using the *Textblob* Python library.¹ However, polarity measures the statement emphasis, while *Correctness* focuses on the content level. As a result, their correlation is weak ($\rho = 0.13$).

Then, the contribution of each judgment dimension to understanding the motivations behind the overall judgment [236] is investigated as follows. For ABC Fact Check statements, the ground truth provides also an assessment rationale (e.g., “cherry picking”). The *Word Mover’s Distance* (*wmd*) [228] is computed between each rationale and the name of each dimension, and we check whether it correlates with the scores of that dimension. Consider the case where there are two statements, *statement_i* and *statement_j*, their *Precision* scores are 2 and 1 respectively, and their ground truth rationales are “exaggeration” and “wrong”. In such a situation, the correlation is computed between the two scores (i.e., 2 and 1) and the semantic similarity of the word pairs (rationale, dimension name):

$$\text{corr}((1, 2), (\text{wmd}(\text{exaggeration}, \text{precision}), \text{wmd}(\text{wrong}, \text{precision}))). \quad (7.1)$$

The scores show a weak correlation with the semantic distance between the labels and the corresponding dimension name (with a peak at 0.3 Pearson’s ρ correlation for *Informativeness*). However, combinations of similarity scores and metrics scores show a higher correlation (e.g., *Overall Truthfulness* values vs. *Informativeness* similarity 0.38, *Speaker’s Trustworthiness* vs. *Completeness* similarity 0.3). These preliminary insights indicate that the dimensions scores can help identify the motivation behind the overall assessment of a statement. The combinations of similarities and scores will be further investigated in the future.

7.4.5 RQ20: Learning Truthfulness from Multidimensional Judgment

This section describes a machine learning-based approach to analyze the usefulness of the multidimensional assessments and of the worker behavior in supporting the prediction of expert judgments, both for *PolitiFact* and *ABC Fact Check*. Two approaches are followed. First, a number of supervised approaches in being able to predict the exact truthfulness verdicts provided by experts is evaluated (Section 7.4.5.1). Second, unsupervised and hybrid approaches are used to estimate truthfulness scores that are semantically close to the ground truth (Section 7.4.5.2).

7.4.5.1 Supervised Approach

The aim is predicting *PolitiFact* and *ABC Fact Check* judgments, considering for *ABC Fact Check* both the three-level scale and the original verdicts, with 30 different labels in the

¹<https://textblob.readthedocs.io/en/dev/>.

sample used (Section 7.2.2). The latter is the scale initially used by experts when judging truthfulness and it is semantically more informative than the simplified one.

The following features are considered, and computed for each judgment. The one-hot-encoding of the worker identifiers in order to identify which worker provided the judgments, followed by the worker judgments on all the dimensions, and the 300-dimensional embedding of the string obtained from the concatenation of the query issued by the worker, and the title, snippet, and domain of the URL selected. The SISTER (SIMple SenTence EmbeddeR) implementation is used² to such an end. The rationale behind this set of embeddings is trying to capture the semantic relationship between the expert classification and the piece of information used by the worker to justify its judgment. After computing the features, the dataset is divided into training and test sets. To avoid any possible bias or overfitting the effectiveness metrics are computed over 3 folds obtained using stratified sampling. The following baselines are considered. The first (i.e., “Most Frequent”) predicts always the most frequent class present in the training set. The second (i.e., “Weighted Sampling”) predicts, for each instance in the test set, a weighted random choice among the classes present in the training set, where the weights are the frequencies of each class. The process for the second baseline is repeated 1000 times for each fold. Finally, the third baseline (i.e., “Random Choice”) simply returns a random class. Apart from the three baselines, the following supervised classification algorithms are used:

- Random Forest;
- Logistic Regression,
- AdaBoost;
- Naive Bayes;
- Support Vector Machine (SVM).

The `sklearn` implementation of the algorithms is used.³ The parameters used to train the algorithms, reported to allow reproducibility, can be found in the repository containing the dataset released (Chapter 3).

Table 7.2 reports the effectiveness scores obtained when predicting the PolitiFact and ABC Fact Check verdicts. To deal with class imbalance, the weighted-averaged version of the Precision, Recall, and F1 metrics is reported. In other words, the effectiveness scores of all classes weighted by their frequency are aggregated. The Random Forest algorithm is able to predict the expert verdict better than both the random baselines and the other algorithms, for all the datasets considered. To investigate the reason behind the differences in effectiveness between Random Forest and the other algorithms, the importance of the features used by the algorithm is addressed.⁴ Random Forest considers equally all the features in the embedding vector, which are the most important for such an algorithm. The rest of the features (i.e., the one-hot encoding of the worker ids and the worker judgments) have an importance which is lower than the embedding vector, but still a presence. As evidence of that, if either the workers’ identifiers vector or the judgments are removed, the effectiveness metrics decrease. Thus, it seems that Random Forest is able to use all the input features to correctly classify the training instances and to effectively generalize to novel

²<https://github.com/tofunlp/sister>

³https://scikit-learn.org/stable/supervised_learning.html.

⁴https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html

ones. This is an important result, as it indicates that multiple signals from the workers, namely their search sessions, can be leveraged to successfully predict the expert verdicts. It is important to notice that this is also true for the 30-class scenario of the original—and more semantically meaningful—ABC Fact Check verdicts.

Table 7.2: Effectiveness metrics when predicting the expert judgment. Baselines above the dashed line.

Algorithm	Accuracy	Precision	Recall	F1
PolitiFact 6 Levels				
Random Choice	.167	.167	.167	.167
Random Forest	.556	.561	.556	.554
Random Forest (bootstrap CI)	[.477, .569]	[.482, .574]	[.477, .569]	[.476, .568]
Logistic Regression	.391	.417	.392	.392
AdaBoost	.327	.340	.327	.327
Naive Bayes	.165	.185	.165	.064
SVM	.225	.213	.226	.207
ABC Fact Check 3 Levels (Simplified)				
Random Choice	.333	.333	.333	.333
Random Forest	.667	.670	.667	.665
Random Forest (bootstrap CI)	[.594, .716]	[.595, .720]	[.594, .716]	[.592, .715]
Logistic Regression	.557	.563	.557	.555
AdaBoost	.560	.562	.560	.559
Naive Bayes	.579	.584	.579	.576
SVM	.392	.391	.392	.379
ABC Fact Check 30 Levels (Original)				
Random Choice	.033	.033	.033	.033
Most Frequent	.134	.018	.134	.032
Weighted Sampling	.067	.067	.067	.066
Random Forest	.518	.562	.518	.491
Random Forest (bootstrap CI)	[.426, .538]	[.460, .605]	[.426, .538]	[.398, .514]
Logistic Regression	.195	.151	.195	.143
AdaBoost	.148	.088	.148	.073
Naive Bayes	.203	.221	.203	.181
SVM	.154	.052	.154	.075

The statistical significance of the metric scores when comparing them against the best baseline is also investigated. To such an end, the Wilcoxon signed-rank test is used (paired data, non-parametric test). The results for multiple comparisons are corrected using the Bonferroni correction. None of the comparisons is statistically significant, and all have $p > 0.05$. This is most likely due to the low number of data points considered in the test (i.e., 3 since the data is split using 3 folds). As a further analysis, the scores for the different folds for each effectiveness metric are plotted, and the best baseline is highlighted with a dashed line (note that the baseline always obtains the same effectiveness score for all the folds). Figure 7.9 shows the results. It is reasonable to assume that the best-performing algorithms

are significantly better than the best baseline even though the statistical significance does not hold. The same behavior holds for the PolitiFact 6 levels and ABC Fact Check 3 levels case (not shown).

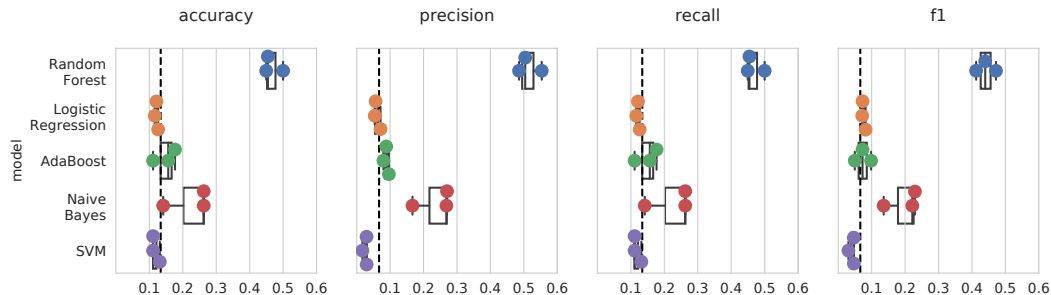


Figure 7.9: Effectiveness over the 3 folds for the ABC Fact Check 30 levels case. The dashed line represents the best baseline.

As a final analysis, the bootstrap technique is employed to compute the 95% confidence interval for the most effective algorithm (i.e., Random Forest). 100 000 stratified samples are employed and the 2.5-th and 97.5-th percentiles are computed [42, 254], in order to compute the 95% likelihood that the computed range covers the true statistic mean. The results are shown in Table 7.2. Even considering the 2.5-th percentile, Random Forest is significantly more effective than the best baseline.

Given that the purpose of this chapter is to study the impact of using a multidimensional scale, the performances of the machine learning techniques when different sets of dimensions are used are investigated. The aim is predicting the PolitiFact and ABC Fact Check judgments. In more detail, the same algorithms considered in Table 7.2 are trained by using three groups of features:

1. all the dimensions apart from Overall Truthfulness;
2. only the Overall Truthfulness dimension;
3. all the dimensions and Overall Truthfulness.

Results (not shown) are almost indistinguishable from the ones of Table 7.2, with very small fluctuations. Nevertheless, it is almost always the case that the effectiveness metrics obtained when training the algorithms with (1) all the dimensions apart from Overall Truthfulness are a little higher than the ones obtained when training considering (3) all the dimensions and Overall Truthfulness. Both approaches lead to obtaining higher effectiveness metrics than the ones obtained considering (2) only the Overall Truthfulness dimension. As before, the statistical significance between all the pairs of approaches is investigated using the Wilcoxon signed-rank test and corrected for multiple comparisons. All differences are not statistically significant.

Summarizing, the results indicate that using all the dimensions to train a supervised approach leads to obtaining the best (even though not significant) effectiveness metrics. Furthermore, Overall Truthfulness does not provide a significant improvement when used as a feature and is outperformed when all the other dimensions are used.

7.4.5.2 Unsupervised Approach

The use of unsupervised approaches for truthfulness prediction is evaluated in addition to using a supervised approach as above described in Section 7.4.5.1. Considering both supervised and unsupervised approaches gives a complete overview of the expected effectiveness of the methods that can be used to predict a given verdict. The aim is predicting a verdict that is semantically close to, and which polarity agrees with the ground truth. However, predicting the exact label used in the ground truth is not a goal. In particular, the focus is on the ABC Fact Check verdicts, which are semantically rich. This analysis helps in understanding the links and relationships between experts' and workers' judgments. In particular, the aim is to understand if the weighted embeddings derived from the workers' judgments only are aligned with the judgments produced by the experts.

To such ends, the predictions are evaluated by checking Word Mover's semantic distance and sentiment difference. The sentiment scores are computed using Flair⁵. Sentiment scores range between -1 and +1, and while semantic similarity tells us whether the rationale for the judgments is similar, sentiment difference tells us whether the polarities agree (e.g., comprehensible and accurate have a higher semantic distance than comprehensible and incomprehensible, but the sentiment difference is higher in the second case). The results are compared with the worst, best, and average combinations obtainable by picking judgments in our ground truth. Picking a random verdict from the ground truth for each statement would lead to an average semantic distance of 2.48 in the best scenario, and 4.41 in the worst. The average distance from random judgments is 3.40. Also, the worst possible sentiment difference is 1.97 and the best (excluding the case when we pick the exact right judgment) is 0.02. The average sentiment difference is 1.00. Here the focus is on the statements, considering the average value of the judgments given by the workers. The strategies are described in the following.

Weighted Average Word Embeddings The starting assumption is that the quality dimensions are positively connoted: when a worker assigns a +2 score to comprehensibility, the overall verdict is assumed to imply that the statement is comprehensible. So, the word embedding of each dimension name is computed weighted on the basis of the corresponding score. Then, the resulting embeddings are averaged to obtain an expected representation of the verdict's embedding. The term having the closest embedding to this average embedding is searched in the embedding dictionary. The resulting labels have an average semantic distance from the ground truth of 4.14 and an average sentiment difference of 1.31: the performance does not improve the random selection of judgments from the ground truth. This is also because such a method searches the whole embedding dictionary, while the ground truth judgments belong to the same semantic area of quality assessment.

Averaging the embeddings introduces some information loss, but this loss is quite limited because the embeddings belong to the same semantic space. To investigate this aspect further, Figure 7.10 shows the plots of the embeddings of each dimension and compares them to the average embedding. These plots are obtained by using t-Distributed Stochastic Neighbor Embedding (t-SNE) [424] to produce a meaningful bi-dimensional representation of the embeddings. Each plot includes an ellipsis representing the 95%

⁵<https://github.com/flairNLP/flair>

confidence interval for each of the sets of embeddings. Each set of embeddings can be thought of as a sample of the population of judgments that we can collect about the quality of the statements analyzed, weighted on the embedding representing the quality dimension's name. The significant overlap between the distribution of each set of embeddings and their average shows that the information loss is limited.

Linear Regression Lastly, Linear Regression is tested as a supervised approach based on weighted average word embeddings. For each statement, an average word embedding of the judgments is built as mentioned in the previous approach. The word embedding of the corresponding ground truths is computed, and the linear regression model that links the two is built. A 3-fold cross-fold evaluation is employed. Every time the linear regression model predicts a verdict, the closest term is searched in the embedding dictionary. The resulting average distance between the predicted judgment and the ground truth is 3.38 and the average sentiment difference is 0.41.

This method improves the performance of the random selection baseline. This indicates that the link between worker assessments and expert judgment is not straightforward as the previous approach hypothesized, but a linear model is already capable of capturing it to some extent. In the future, more sophisticated models will be tested. Workers' profiles should be also considered.

7.5 Summary

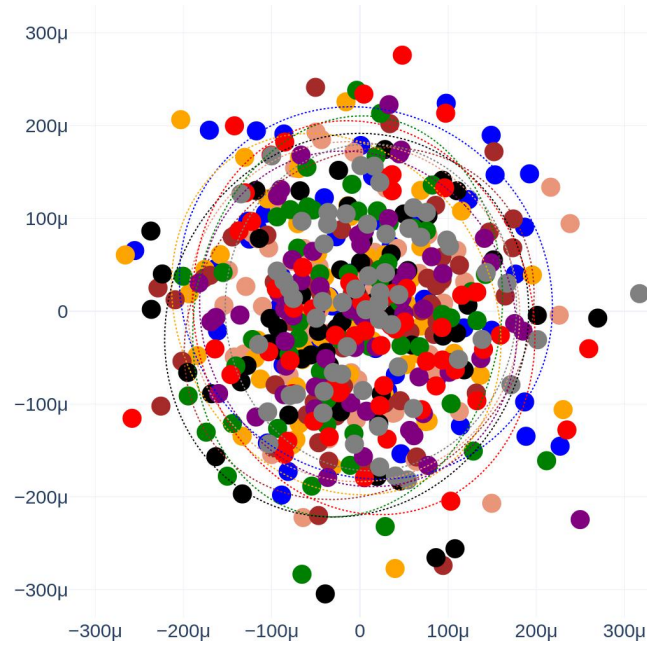
This chapter presents a study of the impact of crowdsourcing truthfulness judgments using multiple dimensions rather than just one. This allows for increased explainability of the collected labels as well as additional opportunities for quality control as crowd workers are asked to provide more input which can be cross-correlated. The answers to the research questions are summarized in the following. The answers to the research questions can be summarized as follows.

RQ16 Extensive evidence that the truthfulness judgments provided by crowd workers over the seven dimensions of truthfulness are sound and reliable is provided. The analyses of the internal agreement among workers do not show any issue with any of the dimensions. The agreement with the ground truth provided by experts is good when the same notion (i.e., Overall Truthfulness) is measured, and reasonable for the individual dimensions, with differences that can be justified by the meaning of each dimension.

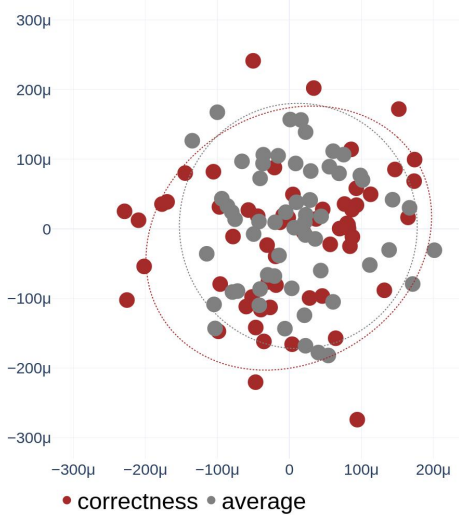
RQ17 Several analyses show that the seven dimensions are independent, not redundant, and measure different aspects. It has been not possible to use this independence to combine the assessments on the single dimensions to obtain a higher agreement with the ground truth.

RQ18 Different crowd workers behave differently. Nevertheless, it has been not possible to leverage such behavioral signals to improve the correlation between the aggregated crowd judgments and the ground truth of expert judgments.

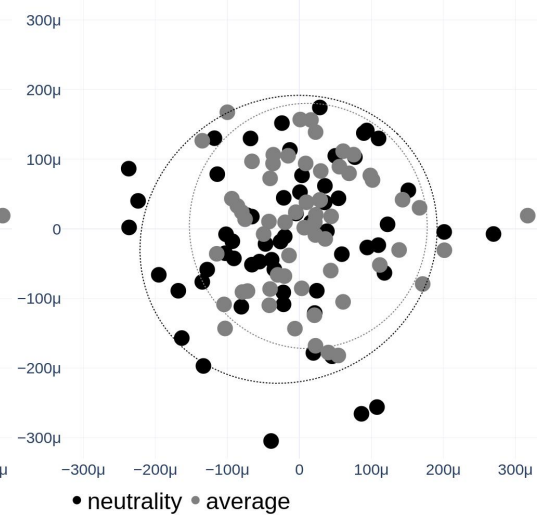
RQ19 The analyses on the informativeness of the different dimensions show that the crowd data are not easy to be generated automatically and that the different dimensions can be



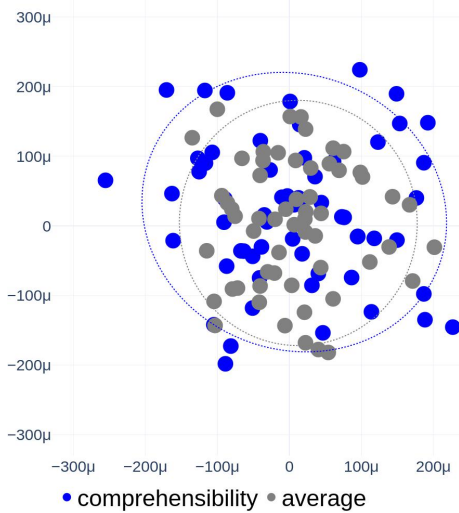
(a) The whole set of dimensions.



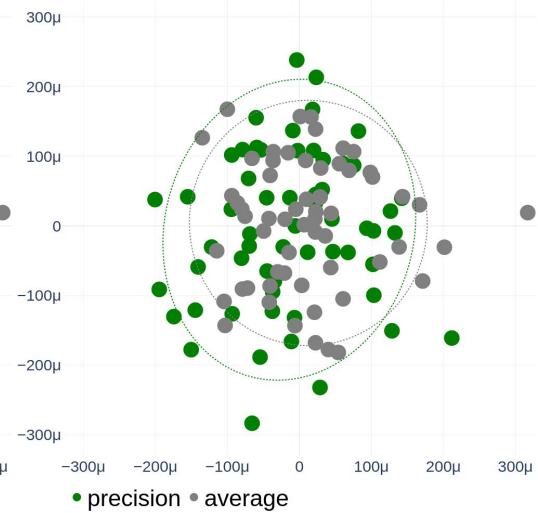
(b) Correctness.



(c) Neutrality.



(d) Comprehensibility.



(e) Precision

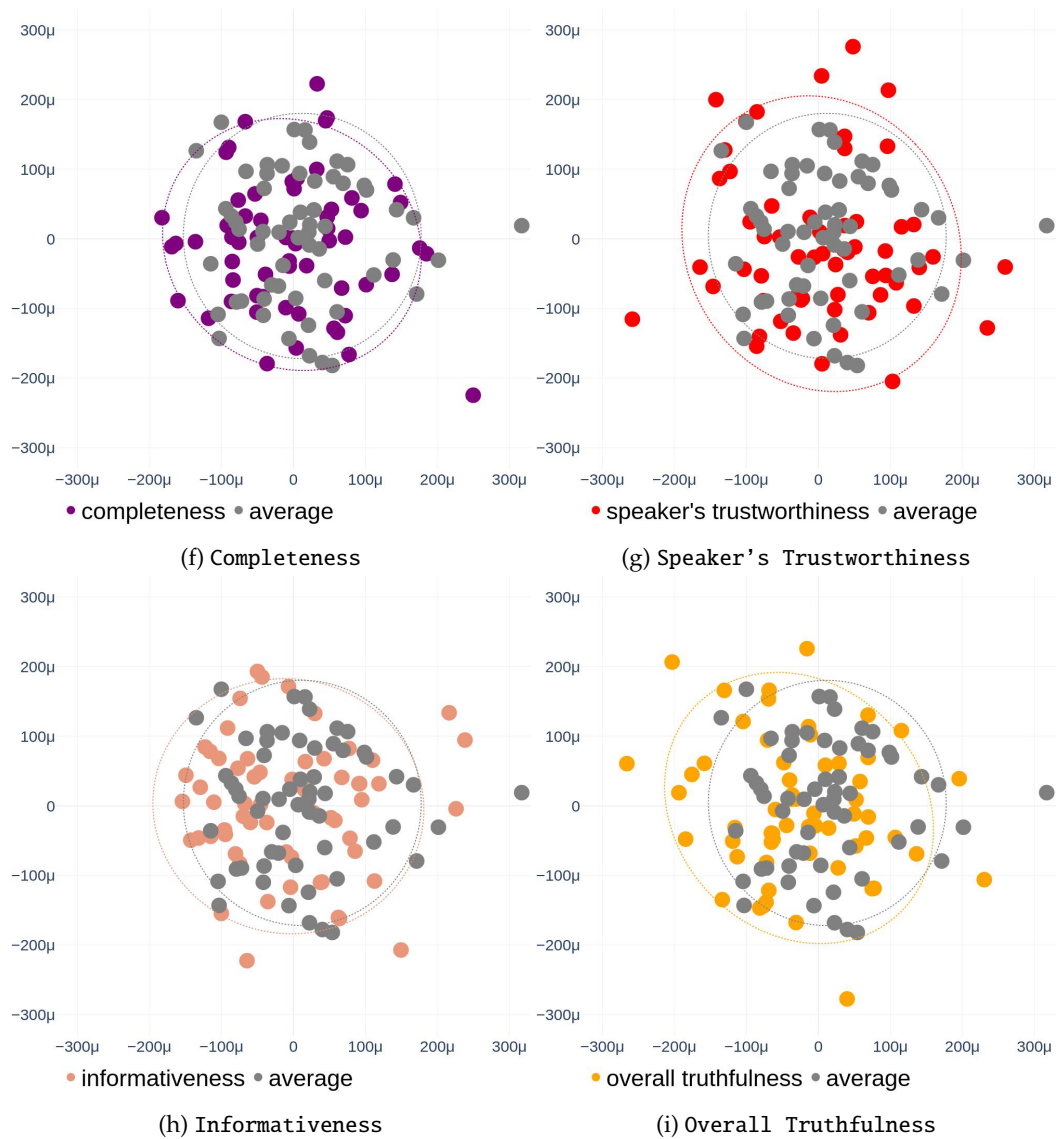


Figure 7.10: Visualization of the embeddings space. All dimensions are compared to average. The coloring in the first visualization follows the legend used in the other plots.

useful to understand the reasons behind the crowd worker's judgment.

RQ20 Signals derived from workers, and in particular their judgments and search sessions, can be leveraged to effectively predict the expert verdicts, both for PolitiFact and ABC Fact Check.

The next chapter proposes a comprehensive and systematic investigation of the cognitive biases that may manifest during the fact-checking process. The PRISMA methodology is used to such an end. A list of countermeasures and a bias-aware pipeline for fact-checking are proposed.

Characterizing Cognitive Biases In Fact-Checking

This chapter is based on the article under review in the “Information Processing & Management” journal [360]. Section 2.1 and Section 2.5 describe the relevant related work. Section 8.1 details the research questions. The PRISMA methodology used is described in Section 8.2, while the results are described in Section 8.3. Finally, Section 8.4 summarizes the main findings and concludes the chapter.

8.1 Research Questions

This chapter aims to provide a comprehensive and systematic investigation of the cognitive biases that may manifest during the fact-checking process, compromising its effectiveness in a real-world scenario. The purpose of this review is thus threefold:

1. to systematically identify and categorize cognitive biases that are relevant to the fact-checking process,
2. to provide a detailed analysis of these biases and real-world examples to illustrate their impact on fact-checking, and
3. to propose potential countermeasures that can help mitigate the influence of these biases on the fact-checking process.

Achieving these objectives allow for offering valuable insights and practical guidance for researchers, practitioners, and policymakers working in the field of fact-checking and information assessment. In more detail, in this work, the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) methodology [289, 317] is adapted to systematically collect and report the review. The methodology involves several phases and the most updated version has been released in 2020. Thus, such version of the “PRISMA Abstract Checklist” (Appendix F.1) and the “PRISMA Checklist” (Appendix F.2) are used.

Such an approach aims to characterize cognitive biases in fact-checking and, as far as it can be understood, this is the first attempt to do so. However, it must be acknowledged that this proposal is not conclusive and should be taken as a starting point for further investigation in this area. The following research questions are investigated:

RQ21 Which are the cognitive biases that might manifest while performing the fact-checking process?

RQ22 Can the cognitive biases that manifest while fact-checking information items be categorized according to some classification scheme?

RQ23 Which countermeasures can be employed to prevent the manifestation of cognitive biases in a fact-checking context?

RQ24 Can a fact-checking pipeline that minimizes the risk of cognitive biases manifesting be proposed?

8.2 Methodology

The PRISMA methodology employed for the systematic investigation of cognitive biases is introduced in Section 8.2.1, while Section 8.2.2 details the eligibility criteria, information sources, and search strategy used. Lastly, Section 8.2.3 describes the data collection and selection processes.

8.2.1 Preferred Reporting Items For Systematic Reviews and Meta-Analyses

The choice to use the PRISMA methodology as a reference over other methods is primarily motivated by its well-established reputation and proven effectiveness in conducting systematic literature reviews. PRISMA is an evidence-based, transparent approach that has been widely adopted in various research fields to conduct high-quality systematic reviews and meta-analyses [289, 317]. Its clear and structured framework facilitates the identification, assessment, and synthesis of relevant data, ensuring that the review process is rigorous, replicable, and unbiased.

By following the PRISMA methodology as a reference for our study, its robust framework is adapted to the specific context of fact-checking and cognitive biases, providing a comprehensive and systematic investigation of the relevant biases. This tailored adaptation allows following PRISMA's structured approach – which involves predefined eligibility criteria, search strategies, and data extraction – that helps minimize the risk of bias in the review process. This is particularly important in the context of this thesis, which aims to examine those cognitive biases that could potentially influence our own assessment and analysis of the relevant literature.

8.2.2 Eligibility Criteria, Information Sources, And Search Strategy

The list of the whole set of cognitive biases that may affect humans is derived by exploring the related literature. The starting point is the list of 220 cognitive biases obtained

by joining together the set of biases identified by Caverni et al. [56], Haselton et al. [180], Hilbert [186], and Kahneman et al. [205]. The full list of biases considered is reported in Appendix F.3. Note that since a standard conceptualization or classification of biases is a debated issue [159, 186], the systematic review relies on the literature to identify biases and maximize recall, thus including two biases even if their difference is subtle.

To ensure the comprehensiveness of the data sources and the reliability of our findings, a systematic approach to identify the 220 cognitive biases from the literature is employed. The process involved the following steps. An extensive search in multiple academic databases and online resources is performed, focusing on the fields of cognitive psychology and decision-making. The search terms included different combinations of keywords such as “cognitive bias”, “heuristics”, “systematic error”, “judgment”, and “decision-making”. Then, manual searches in the reference lists of relevant articles and reviews are performed to identify any additional sources for an exhaustive list of cognitive biases. Each of the 220 cognitive biases found in the literature is analyzed one at a time by focusing on its definition, causes, and domains of application. Biases are included in the list if they are described in peer-reviewed literature, have a clear definition, and are potentially applicable to the fact-checking context. Biases are excluded if they are not well-established, lacked a clear definition, or are not relevant to fact-checking.

8.2.3 Data Collection And Selection Process

There is evidence in the literature that some of the 220 cognitive biases might manifest during the fact-checking process. Thus, each of these biases is considered. The biases with some kind of evidence in the literature include, for example, the Backfire Effect [451] and the Belief Bias [239], which were studied in the context of misinformation by Lewandowsky et al. [243]. Then, a fact-checking scenario for each of the inspected biases is sketched. The aim is to investigate if and how they can manifest. If such a scenario can be defined, the cognitive bias analyzed is included in the systematic review. A given bias is not included if such a scenario can not be defined or if the literature provides explicit evidence that it can not manifest in a fact-checking context. An exhaustive list of fact-checking-related biases is thus compiled, even though this process is subjective to some extent.

As for the selection process, the methodology illustrated in Figure 8.1 is followed. Two authors of the systematic review (denoted as *Assessor 1* and *Assessor 2*) individually and independently selected in the full set of 220 candidate biases the ones they think can manifest in the fact-checking process. To this aim, each assessor inspects the full list of cognitive biases found while exploring the literature and checks each bias definition as well as a set of practical examples for such a bias. Furthermore, each assessor provides a motivation for the inclusion/exclusion of a particular bias with an example of manifestation in the fact-checking scenario, thus following the process described in Section 8.2.2. Then, *Assessor 1* and *Assessor 2* cross-check the list produced by the other author and resolve the conflicts that may emerge by discussing and reaching a final agreement. Note that also in this case assessors maximize recall, thus they decided to include a bias in the list even if the chances for the bias to manifest are rather low. After that step, a third author of the systematic review (denoted as *Assessor 3*) checks the conflict-free list of biases. As in this stage, *Assessor 3* found no inconsistencies, and the bias selection process ends. Note that

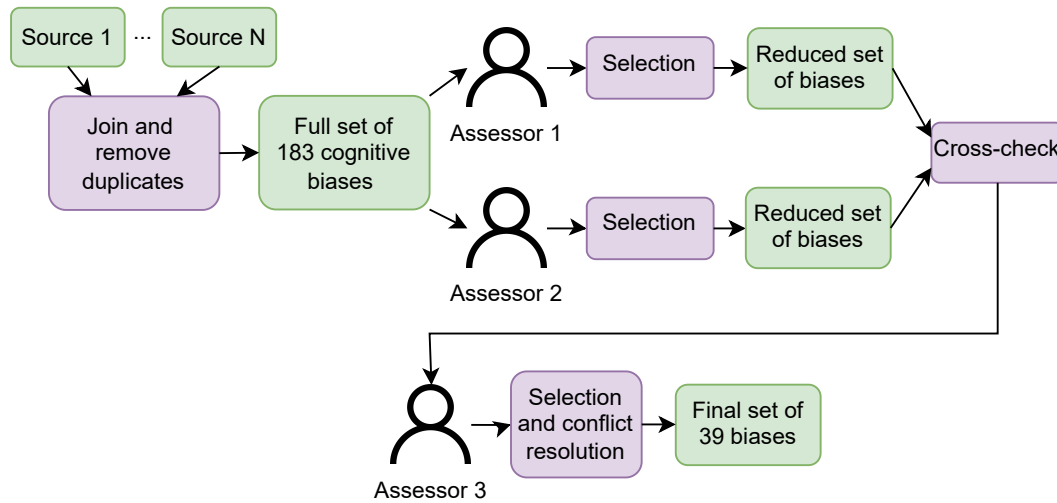


Figure 8.1: Data collection and selection process of the PRISMA-inspired approach.

while this selection process is somehow subjective, we believe we implemented a sound process by implementing discussion points, redundancy, and cross-checks.

8.3 Results

Section 8.3.1 reports the list of cognitive biases selected and investigated. Section 8.3.2 describes their categorization. Section 8.3.3 proposes a list of countermeasures, while Section 8.3.4 presents the bias-aware assessment pipeline to be applied in the context of fact-checking.

8.3.1 RQ21: List Of Cognitive Biases

The 39 out of 220 cognitive biases that can manifest in a fact-checking context are listed in alphabetical order. For each bias, a reference to literature is provided, along with a short description. Also, a situation where it can manifest is framed, when possible. The authors involved in the selection process described in Section 8.2.3 find 39 out of 220 cognitive biases relevant for this systematic review. It must be remarked that while compiling the list outlined below the authors maximize recall, thus risking having some false positives. In other words, a bias is included in the list even if the chances for it to manifest are rather low. The list of biases is presented in the following.

- B1. *Affect Heuristic* [388]: to often rely on emotions, rather than concrete information, when making decisions. This allows one to conclude quickly and easily, but can also distort the reasoning and lead to making suboptimal choices. This bias can manifest when the assessor likes, for example, the speaker of an information item.
- B2. *Anchoring Bias* [307]: to rely too much on an information item (typically the first one acquired) when making a decision. This effect can occur when the assessor inspects

more than one source of information when judging the truthfulness of an information item.

- B3. *Attentional Bias* [27]: the effect for which perception might be affected by recurring thoughts. This effect may occur due to the overwhelming amount of certain topics on news media over time, for example for an assessor who is asked to evaluate the truthfulness of a COVID-19-related claim.
- B4. *Authority Bias* (also called *Halo Effect*) [350]: to attribute higher accuracy to the opinion of an authority figure (unrelated to its content) and be more influenced by that opinion. This bias can manifest when the assessor is shown the speaker/organization making the claim.
- B5. *Automation Bias* [91]: to rely on automated systems which might override correct decisions made by a human assessor. This bias can occur when the assessor is presented with the outcome of an automated system that is designed to help him/her make an informed decision on a given information item.
- B6. *Availability cascade* [227]: process for which a collective belief appears to be more plausible. This might occur when the information item presented to the assessor contains popular beliefs or popular facts.
- B7. *Availability Heuristic* [166]: to overestimate the likelihood of events that are recent in the memory. This bias can occur when the assessors are evaluating recent information items.
- B8. *Backfire Effect* [451]: the reaction of humans to increase their original belief when presented with opposed evidence. This bias can in principle always occur in fact-checking.
- B9. *Bandwagon Effect* [219]: to do (or believe) things because many other people do (or believe) the same. This effect manifests for example when an assessor is asked to evaluate an information item related to recent or debated topics, for which the media coverage is high.
- B10. *Barnum Effect* (also called *Forer Effect*) [144]: to fill the gaps in vague statements by including personal experiences or information. This bias can in principle always occur in fact-checking.
- B11. *Base Rate Fallacy* [445]: to focus on specific parts of information which support a statement and ignore the general information. This bias is related to the fact that the assessors are asked to report the piece of text or sources of information motivating their assessment.
- B12. *Belief Bias* [239]: the process for which the logical strength of someone's argument is biased by the validity of the conclusion. This bias is most likely to occur when the assessors are asked to evaluate factual statements.

- B13. *Choice-Supportive bias* [202]: to remember one's own choices as better than they were. This bias might occur when an assessor is asked to perform a task more than one time when they is asked to revise the judgment and it might prevent assessors from revising their initially submitted score.
- B14. *Compassion Fade* [239]: to act more compassionately towards a small group of victims. This bias can occur for example when the information item to be evaluated is related to minorities.
- B15. *Confirmation Bias* [308]: the tendency to focus or search the information item which confirms prior beliefs. This bias can in principle always occur in fact-checking if the assessor is asked to provide supporting evidence for its score.
- B16. *Conjunction Fallacy* [421]: to assume that specific conditions are more probable than other conditions (e.g., being a woman and a nurse is perceived more probable than being a woman and a bank teller). This bias might occur when the assessor is presented with the name of the speaker and/or when the information item to assess contains references to identifiable subjects.
- B17. *Conservatism Bias* [264]: to revise one's belief insufficiently when presented with new evidence. Note that this bias is different from the *Confirmation Bias*: the *Conservatism Bias* deals with the revision of a belief, while the *Confirmation Bias* deals with new information. As for the *Confirmation Bias*, this bias can occur when the assessor is asked to provide supporting evidence for its score.
- B18. *Consistency Bias* [78]: to attribute past events as resembling present behavior. This bias is present when the assessor has evaluated an information item in the past referring to the same speaker / party and is asked to provide another assessment on an information item coming from the same subject.
- B19. *Courtesy Bias* [199]: to give an answer which is socially accepted to avoid offending anyone. This bias is related to the personal experience/background of the assessor and also to the context in which they is experiencing it.
- B20. *Declinism* [135]: the predisposition to see the past with a positive connotation and the future with a negative one. This bias is related to the temporal part of the pieces of information that the assessor is asked to evaluate.
- B21. *Dunning-Kruger Effect* [121]: the tendency of non-expert individuals to overestimate their abilities. This bias can occur when an assessor is not trained and is under/over-confident about a given subject. Is less likely to happen with expert assessors and more likely to happen with crowd workers or non-expert assessors in general.
- B22. *Framing Effect* [270]: to draw different conclusions from the same information item based on the context, the alternatives, and the delivery method.
- B23. *Fundamental Attribution Error* [178]: not to judge the actor in a given situation but to picture themselves in the same situation.

- B24. *Google Effect* [46]: to forget information that can be found readily online by using search engines. This bias can manifest when a worker is required to use a search engine to find evidence, and/or when they are asked to assess an information item at different periods. An example of this is an assessor forgetting part of the statement right after reading it because they know that it is easily retrievable again if needed by consulting a search engine.
- B25. *Hindsight Bias* (also called “*I-knew-it-all-along*” effect) [354]: to see past events as being predictable at the time those events happened. Since it may cause distortions of memories of what was known before an event, this bias may manifest when an assessor is required to evaluate an event after some time or when is asked to evaluate the same information item multiple times at different periods.
- B26. *Hostile Attribution Bias* [338]: to interpret someone’s behavior as hostile even if it is not.
- B27. *Illusion of Validity* [128]: to overestimate someone’s judgment when the available information is consistent. This bias can happen for example when an assessor works with a set of positive records and previous true pieces of information from a specific person and predict they will have the same outcome for the subsequent set of misinformation items.
- B28. *Illusory Correlation* [169]: to perceive the correlation between non correlated events. This bias can manifest when an assessor works on multiple misinformation items in a single task and they might perceive non-existing patterns between the items.
- B29. *Illusory Truth Effect* [304]: to perceive a statement as true if it is easier to process or it has been stated multiple times. This bias can manifest for example when using straightforward or naive gold questions in a task to check for malicious assessors.
- B30. *Ingroup Bias* [296]: to favor people belonging to the same own group. This bias can manifest for example when the assessors are required to work on misinformation items related to their political party, city, etc.
- B31. *Just-World Hypothesis* [241]: to believe that the world is just. This bias can happen for example when the assessor is working with statements related to major political institutions, as Rubin et al. [365] showed that people tend to assign to them higher scores in belief.
- B32. *Optimism Bias* [381]: to be over-optimistic, underestimating the probability of undesirable outcomes and overestimating favorable and pleasing outcomes. This bias can occur when the statement is vague and does not present factual information.
- B33. *Ostrich Effect* (considered as a sub case of the *Optimism Bias*) [209]: to ignore an obvious (negative) situation. As for the B32. *Optimism Bias*, this effect can manifest when the statement is vague and does not present factual information.
- B34. *Outcome Bias* [29]: to judge a decision by its eventual outcome instead based on the quality of the decision at the time it was made. This bias can manifest when the information item under consideration is related to a past event.

- B35. *Overconfidence Effect* [120]: excessive confidence in one's own answers to questions. This effect can manifest when the assessor is an expert in the field, for example, an expert journalist who performs fact-checking related to his writing or a medical specialist who assesses information items related to their area of expertise.
- B36. *Proportionality Bias* [240]: the innate human tendency to assume that big events have big causes; it is also used to explain the human tendency of some individuals to accept conspiracy theories. This bias can occur when the factual information being assessed deals with the causes and effects of a particular event.
- B37. *Saliency Bias* [296]: to focus on items that are more prominent or emotionally striking and ignore those that are unremarkable, even though this difference is often irrelevant by objective standards. For example, an information item detailing the numerous death of infants will receive different attention from an information item detailing a less emotionally striking fact. If those two facts are presented in the same information item, the score assessed for the prominent fact might drive the overall assessment of the whole information item.
- B38. *Stereotypical (e.g., Gender) Bias* [183]: to discriminate against a personal trait (e.g., gender). Like B30. *Ingroup Bias*, this can happen when the assessor, especially a crowd worker, identifies him/herself with the group related to the misinformation piece they are assessing.
- B39. *Telescoping Effect* [410]: to displace recent events backwards in time and remote events forward in time, so that recent events appear more remote, and remote events, more recent. This effect might occur when the information item presented contains temporal references.

8.3.2 RQ22: Categorization Of Cognitive Biases

The 39 biases described in Section 8.3.1 are categorized using the classification scheme proposed by Dimara et al. [108]. The result of this categorization is shown in Table 8.1. B18. *Consistency Bias*, B19. *Courtesy Bias*, and B36. *Proportionality Bias* are not classified. The first column of the table identifies the task category as defined by Dimara et al. In detail, they propose the following tasks: *estimation*, *decision*, *hypothesis assessment*, *causal attribution*, *recall*, *opinion reporting*, and *other* (i.e., tasks that could not be assigned to any other grouping). Then, the biases are associated with a set of 5 sub-categories that Dimara et al. [108, Section 3.4] define as *flavors*. The flavors are the following:

- *Association*: cognition is biased by associative connections between information items;
- *Baseline*: cognition is biased by comparison with (what is perceived as) a baseline;
- *Inertia*: cognition is biased by the prospect of changing the current state;
- *Outcome*: where cognition is biased by how well something fits an expected or desired outcome;
- *Self-perspective*: cognition is biased by a self-oriented viewpoint.

Table 8.1 shows the result of the classification process and in particular the categorization of the 39 biases that can manifest in fact-checking across the 7 tasks and 5 task subcategories.

Table 8.1: Taxonomy of biases, adapted from Dimara et al. [108]. The biases are classified according to their task (rows) and each task’s subcategory, called “flavor” (columns).

	Association	Baseline	Inertia	Outcome	Self Perspective
Causal Attribution	–	–	–	<i>B26. Hostile Attribution Bias</i> <i>B31. Just-World Hypothesis</i>	<i>B1. Affect Heuristic</i> <i>B23. Fundamental Attribution Error</i> <i>B30. Ingroup Bias</i>
Decision	<i>B4. Authority Bias</i> <i>B5. Automation Bias</i> <i>B22. Framing Effect</i>	–	–	–	–
Estimation	<i>B7. Availability Heuristic</i> <i>B16. Conjunction Fallacy</i>	<i>B2. Anchoring Bias</i> <i>B11. Base Rate Fallacy</i> <i>B14. Compassion Fade</i> <i>B21. Dunning-Kruger Effect</i> <i>B35. Overconfidence Effect</i>	<i>B17. Conservatism Bias</i>	<i>B27. Illusion of Validity</i> <i>B34. Outcome Bias</i>	<i>B32. Optimism Bias</i> <i>B37. Salience Bias</i>
Hypothesis Assessment	<i>B6. Availability Cascade</i> <i>B29. Illusory Truth Effect</i>	–	–	<i>B10. Barnum Effect</i> <i>B12. Belief Bias</i> <i>B15. Confirmation Bias</i> <i>B28. Illusory Correlation</i>	–
Opinion Reporting	–	–	<i>B8. Backfire Effect</i>	<i>B9. Bandwagon Effect</i> <i>B38. Stereotypical Bias</i>	–
Recall	<i>B24. Google Effect</i> <i>B39. Telescoping Effect</i>	–	–	<i>B13. Choice-Supportive Bias</i> <i>B20. Declinism</i> <i>B25. Hindsight Bias</i>	–
Other	<i>B3. Attentional Bias</i>	–	–	<i>B33. Ostrich Effect</i>	–

8.3.3 RQ23: List Of Countermeasures

The literature allows specifying 10 countermeasures that can be employed in a fact-checking context to help prevent manifesting the cognitive biases outlined in Table 8.1. We detail each countermeasure in the following (C1–C10). The countermeasures are selected as follows. First, the literature is inspected to identify works dealing with specific biases. Then, the works that address the setting of assessing the truthfulness of an information item are considered. Note that this approach only details how to remove specific biases. Researchers and practitioners should be aware that the removal of one bias as a result of the application of a countermeasure might result in a manifestation of another one, and that there is no systematic way to safely remove all the possible sources of bias from a fact-checking scenario. Indeed, the aim should be finding a good compromise between the possibility of bias manifestation and the specific experimental setting. The list of countermeasures is presented in the following.

- C1. *Custom search engine.* Researchers and practitioners should be extremely careful with the system supplied to the assessors to help them retrieve some kind of supporting evidence as the system itself can be biased [104, 294, 316, 448]. Researchers should employ a custom and controllable search engine when asking the assessors to evaluate an information item. The assessors might be influenced by the score assigned to the news by a news agency or an online website for the very same information item. Thus, the researcher may tune the search engine parameters to limit the bias that each assessor encounters during a fact-checking activity due to the result source. Moreover, the assessors should be always asked for confirmation or rejection if there is any kind of automatic system designed to provide support during the assessment activity. Such a practice limits B5. *Automation Bias*.
- C2. *Discussion.* Researchers should allow a synchronous discussion among assessors when possible. In fact, when evaluating the truthfulness of an information item each individual is more prone to accept statements that are consistent with their set of beliefs [232, 243, 361]. Reimer et al. [345] and Szpara et al. [404] proved the effectiveness of inter-assessor synchronous discussion to reduce intra-assessor biases. Pitts et al. [333] and Zheng et al. [471] show how discussion among assessors improves the overall assessment quality.
- C3. *Engagement.* It is important to keep the assessors engaged when performing a fact-checking task. Furnham et al. [150] show that if assessors are engaged they are less likely to experience B2. *Anchoring Bias* and B33. *Ostrich Effect*, while Cheng et al. [72] show that engaged assessors are less likely to experience both B22. *Framing Effect* and B28. *Illusory Correlation*.
- C4. *Instructions.* Another important aspect to consider consists of formulating an adequate set of instructions. Gillier et al. [160] have shown that a set of instructions helps assessors in coming up with new ideas when performing a crowdsourcing task. Gadiraju et al. [152] explains that assessors can perform a task even if they have a sub-optimal understanding of the work requested. Thus, task instructions clarity should be carefully taken into account. Furthermore, the assessors should be encouraged to

- be skeptical about the information that they are evaluating [243]. Ecker et al. [123] and Schul [372] prove that pre-exposure warning (i.e., telling explicitly a person that they could be exposed to something reduces the overall impact on the person itself. Thus, showing a set of assessment instructions can be seen as a pre-exposure warning against the impact of misinformation on the assessor.
- C5. *Provide evidence.* Requiring the assessors to provide supporting evidence for their judgments is another effective countermeasure which provides several advantages. It ensures that the assessor focuses on verifiable facts. Lewandowsky et al. [243] explain that such a countermeasure increases the perceived familiarity with the piece of information, reinforcing the assessor's perceived trustworthiness of the claim. They also show that reporting a small set of facts as evidence has the effect of discouraging possible critiques by other assessors, thus reinforcing the assessment provided. Jerit [195] observes such a phenomenon in public debates. Furthermore, asking the assessors to come up with arguments to support their assessment has proven to reduce the: B2. *Anchoring Bias* (as shown by Mussweiler et al. [297]), B11. *Base Rate Fallacy* (as shown by Kahneman et al. [206]), B22. *Framing Effect* (as shown by Cheng et al. [72] and Kim et al. [217]), B27. *Illusion of Validity* (as shown by Kahneman et al. [206]), and B28. *Illusory Correlation* (as shown by Matute et al. [274]). However, requesting evidence may be a source of bias itself. Luo [264] and Wood et al. [451] show that such a request can lead to the manifestation of, respectively, B17. *Conservatism Bias* and B8. *Backfire Effect*. Thus, the requester of the fact-checking activity should address this matter carefully.
- C6. *Randomized or constrained experimental design.* Using a randomized or constrained experimental design is helpful in reducing biases. Different assessors should evaluate different information items, and the same information item should be evaluated by different assessors. Moreover, each set of items should be evaluated according to a different order, and the assignment of an information item to a given assessor should be such that the item overlap among every two assessors is minimum. If such a constraint can not be satisfied, a randomization process should minimize the chances of overlap between items and assessors [64, 184].
- C7. *Redundancy.* Redundancy should be employed when asking more than one assessor to fact-check a set of information items. Each item can thus be characterized by a final score, that should be computed by aggregating the individual scores provided by each assessor. In this way, the individual bias of each assessor is mitigated by the remaining assessors. If the population of assessors is diverse enough, one can ideally expect a less biased fact-check as a result. The population of assessors should thus be as variegated as possible, in terms of both background and experience [106].
- C8. *Revision.* Asking the assessors to revise and/or double-check their fact-check or even provide them with alternative labels is a useful countermeasure to reduce many biases. In more detail, Cheng et al. [72], Effectiviology [124], Kahneman [204], Kahneman et al. [206], Kim et al. [217], and Mussweiler et al. [297] show that assessment revision helps reducing: B2. *Anchoring Bias*, B7. *Availability Heuristic*, B9. *Bandwagon Effect*, B11. *Base*

Rate Fallacy, and *B22. Framing Effect*. Furthermore, Bollinger et al. [43], Cooper et al. [86], Hettiachchi et al. [185], and Mussweiler et al. [297] show that providing feedback to assessors while performing a given task is useful to reduce biases such as: *B2. Anchoring Bias*, *B3. Attentional Bias*, and *B37. Salience Bias*.

- C9. Time*. Researchers should be careful when setting the time available for each assessor to fact-check a given information item. An adequate amount of time should be left to the assessor. There are pros and cons to granting the assessor a small or huge amount of time. For instance, one may guess that giving the assessor more time they will carefully ponder the decision and thus avoid *B2. Anchoring Bias*. However, Furnham et al. [150] show that overthinking might actually increase such a bias. Moreover, Effectiviology [124] show that assessors left with an adequate amount of time experienced a reduction of the *B9. Bandwagon Effect*.
- C10. Training*. Dugan [115], Kazdin [210], Lievens [249], Pell et al. [324], and Szpara et al. [404] show that training an assessor increases accuracy and reduces the chances for bias to manifest. Thus, assessors' training is a useful countermeasure against biases within any context.

8.3.4 RQ24: Towards A Bias-Aware Judgment Pipeline

Section 8.3.3 presents different countermeasures to reduce the risk of cognitive biases manifest in a fact-checking context. It can be thus proposed a fact-checking pipeline that minimizes such a risk. Table 8.2 outlines such pipeline. The table shows for each part of the overall fact-checking activity the corresponding countermeasures along with a brief recap and the biases that may manifest. More specifically, the first column of the table details the task phase where the specific countermeasure can be applied: before the task happens (i.e., pre-task), when the assessor is performing the task (i.e., during the task), or after the task, when the assessment has been made (i.e., post-task); furthermore, a set of countermeasures that are not bound to a specific task purpose (i.e., general purpose) is listed. The other columns detail the countermeasures that can be adopted and, for each of them, its brief description along with the set of biases that the specific countermeasure reduces.

To provide an example, the first line of the table deals with the adoption of the countermeasure *C6. Randomized or constrained experimental design*, i.e., to randomize the process that assigns assessors and information items to enforce diversity and randomness in the pairing. This is a general-purpose countermeasure since it can be applied in different task phases (e.g. when designing the task offline or dynamically when a new assessor is assigned to a new information item). It allows to mitigate or remove: *B2. Anchoring Bias*, as the assessor is less likely to rely on a specific information item given that they inspects more than one with different characteristics, and *B9. Bandwagon Effect*, as the assessor is less likely to be presented with a set of items all related to the personal belief or to debated topics with high coverage.

Given that the body of literature investigating cognitive biases in the context of misinformation is not exhaustive, the assessment pipeline might include some bias for which there are not many literature studies or were the literature piece had artifacts (e.g., small

Table 8.2: Constituting elements of a bias-aware assessment pipeline.

Task Phase	Countermeasure Adopted	Brief Description	Biases Involved
General Purpose	<i>C6. Randomized or constrained experimental design</i>	Employ a randomization process when pairing assessors and information items	<i>B2. Anchoring Bias</i> <i>B9. Bandwagon Effect</i>
	<i>C7. Redundancy</i>	Use more than one assessor for each information item, and a variegated pool of assessors	Bias in General
	<i>C9. Time</i>	Allocate an adequate amount of time for the assessors to perform the task	Bias in General
Pre-Task	<i>C3. Engagement</i>	Put the assessors in a good mood and keep them engaged	<i>B2. Anchoring Bias</i> <i>B22. Framing Effect</i> <i>B28. Illusory Correlation</i> <i>B33. Ostrich Effect</i>
	<i>C4. Instructions</i>	Prepare a clear set of instructions to the assessors before the task	<i>B21. Dunning-Kruger Effect</i> <i>B35. Overconfidence Effect</i>
	<i>C10. Training</i>	Train the assessors before the task	Bias in General
During the Task	<i>C1. Custom Search Engine</i>	Deploy a custom search engine	Bias in General
	<i>C5. Provide evidence</i>	Ask the assessors to provide supporting evidence	<i>B2. Anchoring Bias</i> <i>B5. Automation Bias</i> <i>B8. Backfire Effect</i> <i>B11. Base Rate Fallacy</i> <i>B22. Framing Effect</i> <i>B27. Illusion of Validity</i> <i>B28. Illusory Correlation</i>
	<i>C8. Revision</i>	Ask the assessors to revise the assessments	<i>B2. Anchoring Bias</i> <i>B7. Availability Heuristic</i> <i>B9. Bandwagon Effect</i> <i>B11. Base Rate Fallacy</i> <i>B22. Framing Effect</i>
	<i>C2. Discussion</i>	Synchronous discussion between assessors	Bias in General
Post-Task	<i>C7. Redundancy</i>	Aggregate the final scores	Bias in General

sample sizes). Thus, the proposed pipeline should not be taken as final, but rather should be updated as new evidence of effects of specific cognitive biases get published in the literature.

8.4 Summary

The characterization discussed in this chapter addresses the problem of cognitive biases that may manifest while performing fact-checking tasks. The study summarizes the misinformation literature with the aim of creating a comprehensive list of cognitive biases that may affect the fact-checking process. This is the first attempt to comprehensively study – and handle – the cognitive biases that may be present at the different stages of the fact-checking workflow. The answers to the research questions can be summarized as follows.

RQ21 There is a subset of 39 out of 220 cognitive biases that are likely to manifest while performing the fact-checking process.

RQ22 It is possible to build a taxonomy of the 39 cognitive biases using a classification scheme proposed in the literature.

RQ23 There are ten countermeasures that can be employed to limit the impact of the 39 cognitive on the fact-checking process.

RQ24 The constituting blocks of a pipeline that contrast cognitive biases are described. Each block is assigned with one or more countermeasures.

The next chapter describes an exploratory analysis of the data collected while performing the experiments described in Chapter 7.1. The aim is to identify a set of potential cognitive biases that may occur when crowd workers perform fact-checking tasks. A novel set of crowdsourced truthfulness judgments is collected to validate the hypotheses derived from the analysis.

The Effect Of Cognitive Biases In Fact-Checking Tasks

This chapter is based on the article published at the 2022 ACM Conference on Fairness, Accountability, and Transparency [111]. Section 2.1 and Section 2.5 describe the relevant related work. Section 9.1 addresses the research questions. Section 9.2 describes the exploratory study conducted, while Section 9.3 describes the experimental setting of a novel crowdsourcing study. Section 9.5 presents the results obtained. Finally, Section 9.6 summarizes the main findings and concludes the chapter.

9.1 Research Questions

This chapter investigates which systematic biases may decrease data quality for crowd-sourced truthfulness judgments. Initially, an exploratory study is conducted on an earlier collected data set containing crowdsourced truthfulness judgments for political statements (Section 7.2). These data also contain information on the political leaning of statements as well as individual worker characteristics (e.g., workers' level of education and political leaning). The findings from these exploratory analyses are used to formulate specific hypotheses concerning which individual characteristics of statements or workers and what cognitive worker biases (Section 8.3.1) may affect the accuracy of crowd workers' truthfulness judgments. To test these hypotheses, a new, preregistered crowdsourcing study is conducted. Supplementary materials related to the work described in this chapter (e.g., task design, preregistration, data sets, and analysis code) are openly available.¹ The following research questions are investigated:

RQ25 What individual characteristics of crowd workers and statements may lead to systematic biases in crowd workers' truthfulness judgments?

RQ26 What cognitive biases can affect crowd workers' truthfulness judgments?

¹<https://osf.io/8yu5z/>

RQ27 Are different truthfulness dimensions affected by different biases?

9.2 Exploratory Study

The exploratory study is conducted on the dataset collected to evaluate the multiple dimensions of truthfulness. The experimental setup, the dataset sample, and the crowd-sourcing task design used to collect the data are those described in Section 7.2. Thus, this section details the exploratory study and describes the hypotheses formulated as a result. Section 9.2.1 describes the preprocessing steps performed on the previously collected data. Section 9.2.2 describes the exploratory analyses. Finally, Section 9.2.3 presents the seven different hypotheses derived.

9.2.1 Data Preprocessing

Several preprocessing steps are performed on the data described in Section 7.2 so that they fit the purposes of the analysis. Specifically, the judgment scales are transformed and mapped into a common set of labels (Section 9.2.1.1) and three worker-related judgment bias metrics are computed (Section 9.2.1.2).

9.2.1.1 Scale Transformations

Each statement in the data set described in Section 7.2 contains a truthfulness judgment from either PolitiFact (Section 3.1) or ABC Fact Check (Section 3.2), as well as truthfulness judgments from crowd workers. However, these different types of judgments all adhere to different (ordinal) scales. PolitiFact judgments are made on a six-level scale. ABC Fact Check judgments are made on a three-level scale. The worker judgments are made on a five-level (Likert) scale. Comparing the different judgments require that an alignment of those scales. Assuming that all the PolitiFact, ABC Fact Check, and Likert judgment scales are linear equally spaced scales (the same assumption has been made in previous the previous experiments and discussed in more detail in Section 4.2.2) the PolitiFact and Likert scales are converted to the three-level scale used by ABC Fact Check. This means transforming each judgment to one of three labels: **Negative** (-1), **Neutral** (0), and **Positive** (1). This might lead to some confusion, since two out of three labels are verdicts used by ABC Fact Check. However, they must be interpreted as a new and shared set of labels used to map the three sets of judgments. Such a decision is needed given the intra-dataset inconsistencies that occur for both PolitiFact (Section 3.1) and ABC Fact Check (Section 3.2) described in the respective sections. The procedure is detailed in the following:

- PolitiFact: **Pants-On-Fire** and **False** into **Negative** (-1), **Mostly-False** and **Half-True** into **Neutral** (0), and **Mostly-True** and **True** into **Positive** (1).
- ABC Fact Check: **False** and **True** maintain the same semantic meaning, while **In-Between** is mapped into **Neutral** (0).
- Likert scale: -2 and -1 are mapped into **Negative** (-1), 0 into **Neutral** (0), and +1 and +2 into **Positive** (1).

9.2.1.2 Judgment Bias Metrics

Three different metrics are computed to quantify and evaluate judgment bias. Both *external* errors (i.e., when comparing crowd judgments with the ground truth) and *internal* errors (i.e., when comparing crowd judgments with other crowd judgments for the same set of items) are considered. The metrics are the following:

- *External Error (eE)*: the difference between a worker's Overall Truthfulness judgment (Section 7.2.1) and the respective item's ground truth judgment as assessed by the expert. This metric assesses the degree to which a crowd worker overestimates or underestimates the Overall Truthfulness of a particular statement. Its values range in $[-2, 2]$: for example, if the ground truth label (i.e., from PolitiFact or ABC Fact Check) for an item is positive (1) but the crowd worker's judgment is negative (-1), eE for this particular annotation is equal to -2.
- *External Absolute Error (eAE)*: the *absolute* difference between a crowd worker's Overall Truthfulness judgment and the respective item's ground truth label. Its values range in $[0, 2]$. In contrast to eE, this metric quantifies the *magnitude* of bias. It is the absolute value of eE. The mean squared error is not used here to avoid penalizing larger errors (e.g., an error of 2 should not be more than double the error of 1).
- *Internal Error (iE)*: the difference between a worker's judgment and the average judgment of other crowd workers for the same statement. Its values range in $[-2, 2]$. Such metrics are computed nine times in total, i.e., one for Overall Truthfulness, one for workers' Confidence, and one for each of the seven truthfulness dimensions (Section 7.2.1). These nine metrics quantify the degree to which a specific judgment was above or below other crowd workers' judgments on a particular dimension.

9.2.1.3 Worker Bias Metrics

Aggregate bias metrics that evaluate each worker's individual degree of bias based on the annotation bias metrics described in Section 9.2.1.2 are computed. Specifically, each worker's mean eE (eME), mean eAE (eMAE), and – for Overall Truthfulness, Confidence, and each of the seven dimensions – mean iE (iME) are computed. These 11 worker-specific metrics are used as dependent variables for the exploratory study.

9.2.2 Exploratory Analyses

A series of exploratory analyses on the dataset described in Section 7.2 is performed to identify potential systematic biases in crowd workers' truthfulness judgments. Specifically, different worker-related attributes (e.g., political views and average time per judgment) are used as independent variables and the aggregate worker bias metrics described in Section 9.2.1.3 as dependent variables. The workers in the dataset are quite balanced in terms of demographics (e.g., age group and income) and political views (e.g., conservative versus liberal orientation). Note that the results reported in this subsection (e.g., *p*-values from hypothesis tests) are exploratory. These analyses are only conducted to identify concrete hypotheses to be tested on novel data (Section 9.2.3).

9.2.2.1 Exploring Worker's eME

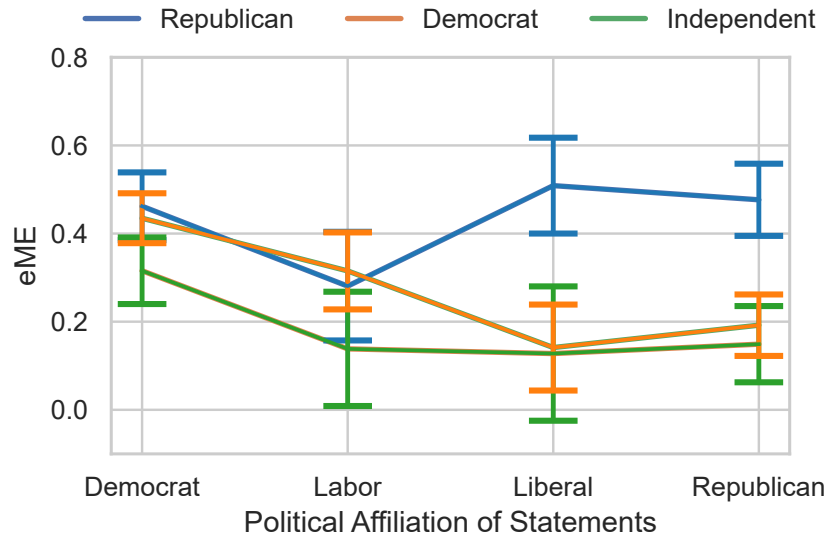
The exploratory analysis begins by computing workers' eME, corresponding to the average difference between a crowd worker's judgment and the respective item's ground truth label. Workers overall tend to overestimate truthfulness (mean eME = 0.32, $\sigma = 0.42$, $t = 10.93$, $p < 0.001$; result from a one sample t -test; test value = 0). Looking at specific worker characteristics using linear regression and ANOVA models (incl. post-hoc tests), shows that workers who identified as Very Conservative and/or Republican tended to overestimate truthfulness more than other worker groups (i.e., Tukey-adjusted $p = [0.006, 0.050]$ compared to other political views for *very conservative* workers; Tukey-adjusted $p = [0.012, 0.089]$ compared to other party affiliations for Republican workers). The results further show that workers who agree to the southern border question (the full questionnaire is reported in Appendix G) overestimate truthfulness more than workers who disagree (Tukey-adjusted $p = 0.004$); although this effect seemed to be explained by workers' political affiliation, as 78% of those workers also identified as Republican.

When looking for explanations for the aforementioned systematic biases, a slight trend that workers (especially those who identified as Republican) particularly overestimate the truthfulness of those statements that confirmed their political views can be found (Figure 9.1a). Ironically, this led the average worker to judge the truthfulness of statements affiliated with other parties more accurately than their own, due to the general trend toward overestimating truthfulness. This phenomenon could be explained by different cognitive biases (Section 8.3.1), i.e., the *B1. Affect Heuristic* (crowd workers may overestimate truthfulness when they like the statement speaker) or the *B15. Confirmation Bias* (crowd workers may overestimate truthfulness when they support the underlying political message).

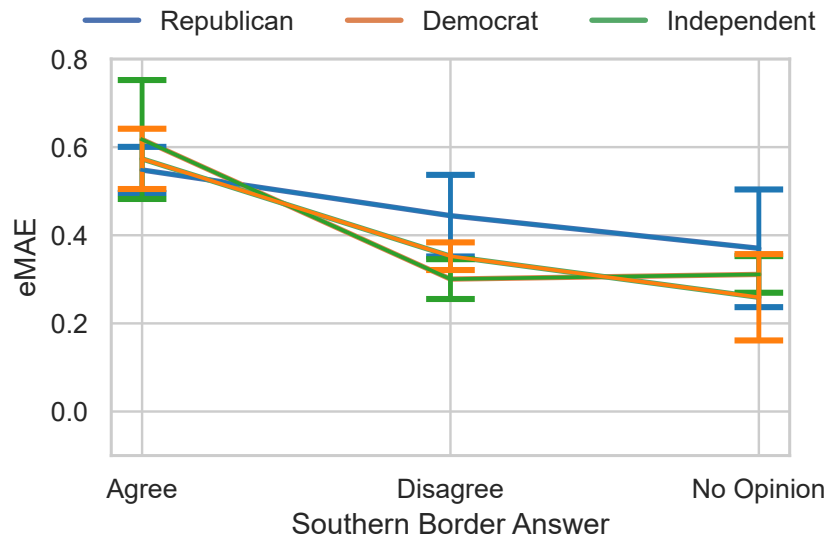
9.2.2.2 Exploring Worker's eMAE

The eMAE, which corresponds to the mean absolute difference between a crowd worker's judgment and the respective item's ground truth label, is also considered. The mean eMAE in the data is 0.42 ($\sigma = 0.31$), reiterating that the average worker was somewhat biased in their annotations (i.e., eAE ranged from 0 to 1.11). Moreover, in line with the findings above, we found that workers who identified as Very Conservative (Tukey-adjusted $p = [0.012, 0.200]$), Republican (Tukey-adjusted $p = [0.031, 0.129]$), or agreed on the southern border question (Tukey-adjusted $p < 0.001$) were more biased than others (i.e., had a higher eMAE, as shown in Figure 9.1).

Furthermore, the more biased worker groups mentioned above generally took less time for their judgments compared to other workers. Although not any effect of cognitive reasoning on eMAE has been found when considering all independent variables at the same time, workers with lower cognitive reasoning also tend to do the task quicker. It could thus be that cognitive reasoning abilities explain some of the variances between worker groups but that the effect was too small to be detected in this exploratory study. Another explanation could be workers' *belief in science* (Appendix G.2). Indeed, 78% of the Disagree (-1) answers regarding additional environmental regulations came from Very Conservative workers. Given the clear scientific stance regarding the environment, some workers may simply not trust scientific results and therefore distrust statements in which scientific results



(a) Mean eME per political affiliations of statements and workers.



(b) Mean eMAE per southern border answer and political affiliations of workers.

Figure 9.1: Mean eME in the dataset (Section 7.2). Four workers who considered themselves something else other than Democrat, Independent, or Republican are excluded.

are brought up as evidence. Although there are too few of these Disagree (-1) answers overall to detect a direct effect here, *belief in science* may be an underlying variable that influences the accuracy of crowd workers' truthfulness judgments.

Interestingly, the analyses also reveal a positive relationship between workers' average confidence in their judgments and eMAE ($\beta = 0.14, p < 0.001$), which might be an indication of the *B35. Overconfidence Effect* cognitive bias (Section 8.3.1).

9.2.2.3 Exploring Worker's iME

Finally, the iME is investigated, which corresponds to the mean difference between crowd workers' judgments and other crowd workers' judgments on the same statements. Workers with some postgraduate or professional schooling (no postgraduate degree) have higher confidence in their abilities to judge truthfulness compared to most workers with lower or higher education status (Tukey-adjusted $p = [< 0.001, 0.018]$). The analyses also reveal that the more a worker identifies as being Conservative, the higher their self-reported Confidence compared to other workers who annotated the same items. In general, Confidence was higher in worker groups with greater bias, which further pointed to a potential *B35. Overconfidence Effect* bias in some workers. This could also indicate that the Confidence dimension acts as a proxy for explaining the political skewness of the results.

By far the strongest predictor of eME among the iME measures is the Correctness dimension ($\beta = 0.51, p < 0.001$). This suggests that workers might see the Correctness dimension as commensurable to Overall Truthfulness (as previously identified in Section 7.4.2), and indicates that workers who judge Correctness higher than others are likely also overestimating Overall Truthfulness.

Furthermore, workers who identify as Democrat or Republican judge truthfulness higher on most dimensions than workers who identify as Independent or something else, which usually leads to more accurate judgments for the latter group due to the general tendency toward overestimation of truthfulness. Even though these differences were small, this might be an indication that workers with higher *trust in politics* (as here represented by Republicans and Democrats) exhibit more overall bias because they overestimate truthfulness to a greater degree than workers with lower *trust in politics*, as here represented by other workers). This suspicion is underlined by the finding that workers who answered with "no opinion" to the southern border question tended to judge the Speaker's Trustworthiness lower than other workers (Figure 9.1).

Lastly, the analyses also reveal that iME for Speaker's Trustworthiness is the strongest predictor among the iME measures for eMAE ($\beta = 0.16, p = 0.040$). This again could point to a potential *B1. Affect Heuristic* (see Section 8.3.1 and Section 9.2.2.1).

9.2.3 Hypotheses For The Novel Data Collection

From our exploratory study (Section 9.2), seven different hypotheses are derived. Such hypotheses are tested on novel data. The hypotheses are differentiated based on whether they refer to general worker traits (e.g., their *trust in politics*) or task-related cognitive biases (e.g., the *B1. Affect Heuristic*).

9.2.3.1 RQ25: General Worker Traits

These hypotheses refer to expectations about which worker groups may be more prone to biased judgments compared to others.

Hypothesis 1a (H1a): Workers with stronger *trust in politics* are less accurate in judging the Overall Truthfulness of statements compared to other workers.

- *Rationale:* Workers who consider themselves Democrat or Republican (i.e., the most “traditional” political parties) are less accurate in their truthfulness judgments than other workers in the exploratory study. Overly high *trust in politics* (i.e., the conviction that politicians and governmental bodies are trustworthy and aim to do the right thing) may lead some workers to strongly identify with political parties and could be the underlying reason for this bias. Such workers may not be skeptical enough when considering politicians’ statements and therefore overestimate the likelihood of statements being true.

Hypothesis 1b (H1b): workers with stronger *belief in science* are more accurate in judging the Overall Truthfulness of statements compared to other workers.

- *Rationale:* workers who answer with Disagree (-1) to the environmental regulations question tend to be more biased than others in the exploratory study. An hypothesis is that the underlying responsible variable could be workers’ *belief in science* (i.e., the conviction that scientific results are trustworthy and important for societal development). Workers with low belief in science may automatically doubt the truthfulness of statements that refer to scientific findings, e.g., related to climate change. This may undermine workers’ ability to give accurate truthfulness judgments.

Hypothesis 1c (H1c): Workers with better *cognitive reasoning abilities* are more accurate in judging the Overall Truthfulness of statements compared to other workers.

- *Rationale:* In the exploratory study, workers with lower cognitive reasoning abilities tend to perform the task quicker, which was generally associated with greater bias. An hypothesis is that such a relationship could exist but that it might be hard to detect, although a direct association of workers’ cognitive reasoning abilities with their bias has not been found; especially given that many study participants have been exposed to the CRT before [168].

9.2.3.2 RQ26: Cognitive Biases

These hypotheses are predictions about cognitive biases that may affect crowd workers.

Hypothesis 2a (H2a): workers generally overestimate truthfulness.

- *Rationale:* workers overestimate truthfulness in the exploratory study, finding the same in novel data is expected.

Hypothesis 2b (H2b): workers’ tendency to over- or underestimate the Overall Truthfulness of a statement is related to the degree to which they like the statement speaker.

- *Rationale*: the exploratory study reveal several relationships that hint at a potential the *B1. Affect Heuristic* bias (Section 8.3.1).

Hypothesis 2c (H2c): workers’ tendency to overestimate or underestimate the Overall Truthfulness of a statement is related to the degree to which they personally support the goal of the statement.

- *Rationale*: some relationships part of the exploratory study hint at a potential *B15. Confirmation Bias* (Section 8.3.1).

Hypothesis 2d (H2d): workers with higher *mean confidence* in their ability to correctly judge the truthfulness of items exhibit more bias compared to other workers.

- *Rationale*: workers’ confidence in their judgments are directly related to their degree of bias in the exploratory study. Thus, finding similar *B35. Overconfidence Effect* in novel data collected.

9.3 Experimental Setting

A novel crowdsourcing experiment is conducted to test the hypotheses detailed in Section 9.2.3. The hypotheses, research design, and data analysis plan have been preregistered before data collection.²

9.3.1 Crowdsourcing Task

The experimental design is the same as the one used to evaluate the multiple dimensions of truthfulness, described in Section 7.2. Specifically, the same interface and set of HITs are used, to keep the new task as similar as possible. However, three additional variables are individuated to study the hypotheses described in Section 9.2.3 (i.e., *trust in politics*, *belief in science*, and *affect for statement speaker*). Such additional variables require modifications to the original task. To such an end, the generalized version of the “Citizen Trust in Government Organizations” questionnaire [165] (CTGO, reported in Appendix G.1) is used to measure workers’ trust in politics. The “Belief in Science Scale” [92] (BISS, reported in Appendix G.2) is used to record workers’ belief in science. These two questionnaires are placed in the task right after the original initial questionnaire (reported in Appendix B.1). Finally, a single, five-point Likert scale is added to capture the degree to which the workers like the speaker of the statement. This scale also includes an additional answer option that allowed the worker to state that they do not know the speaker.

The crowdsourcing task aims to collect data from at least 255 crowd workers. Such a required sample size is computed in a power analysis for a Between-Subjects ANOVA (Fixed effects, special, main effects, and interactions; see Section 9.5.1) using the software *G*Power* [133]. Here, a small effect size of $f = 0.10$ is specified based on the findings of the exploratory study. Also, a significance threshold $\alpha = 0.05/7 = 0.007$ (due to testing multiple hypotheses and statistical power of $(1 - \beta) = 0.8$ are set. Then, the presence of

²The preregistration is available at <https://osf.io/5jyu4>.

three between-subjects groups (i.e., Republican, Democrat, and Independent/else) and four within-subjects groups (i.e., Republican, Democrat, Liberal, and Labor). The required sample size is computed for each of the hypotheses using their respective degrees of freedom.

The task publishes 200 HITs on the Amazon Mechanical Turk platform to evaluate the set of 180 statements outlined in Section 7.2 and described in detail in Section 4.2. A total of 2200 judgments is thus collected. Crowd workers who are based in the United States are recruited. Each crowd worker is rewarded 2 \$USD for completing the task. This amount was based on the minimum time required to complete the task and the United States minimum wage of 7.25 \$USD per hour.

9.3.2 Variables

The crowdsourcing task allows recording 9 descriptive and exploratory variables, 7 independent variables, and 3 dependent variables. The descriptive and exploratory variables are those collected using the questionnaires reported in Appendix B.1 and not any conclusive hypothesis tests are performed using their values. They are *iE*, *iME* (Section 9.2.2.3), age group, gender, level of education, income, political views, and the opinions on US southern border and about US environmental regulation.

The dependent variables considered are *eE*, *eME* (Section 9.2.2.1), and *eMAE* (Section 9.2.2.2). Lastly, the independent variables recorded and addressed are the following:

- *trust in politics* (continuous; $[-2, 2]$): the degree to which workers trust in media and politics as measured by the CTGO questionnaire (i.e., averaging all responses). Higher scores mean greater trust in politics.
- *belief in science* (continuous; $[-2, 2]$): the degree to which workers believe in science as measured by the BISS questionnaire (i.e., averaging all responses). Higher scores mean greater belief in science.
- *Cognitive reasoning* (ordinal; $[0, 4]$): worker's cognitive reasoning abilities as measured by the CRT. The time spent on CRT is also measured and considered as a proxy for cognitive effort. Higher scores mean greater cognitive reflection.
- *Political party affiliation* (categorical): whether workers consider themselves as Republican, Democrat, Independent or something else (i.e., not represented by any of the three previous political parties). Question Q5 of the initial questionnaire.
- *Affect for the statement speaker* (ordinal; $[-3, 3]$): each worker rated on a five-point Likert scale the degree to which they like each statement claimant. The option "I don't know the speaker" is also added.
- *Mean confidence* (ordinal; $[-2, 2]$): workers' average self-reported confidence regarding the accuracy across their truthfulness judgments (on a five-point Likert scale).
- *Statement support* (categorical): the degree to which workers support the cause of the statement (whether true or false) is approximated with their personal political orientation.

9.4 Descriptive Statistics

Overall, 302 workers completed the crowdsourcing task.

9.4.1 Worker Demographics

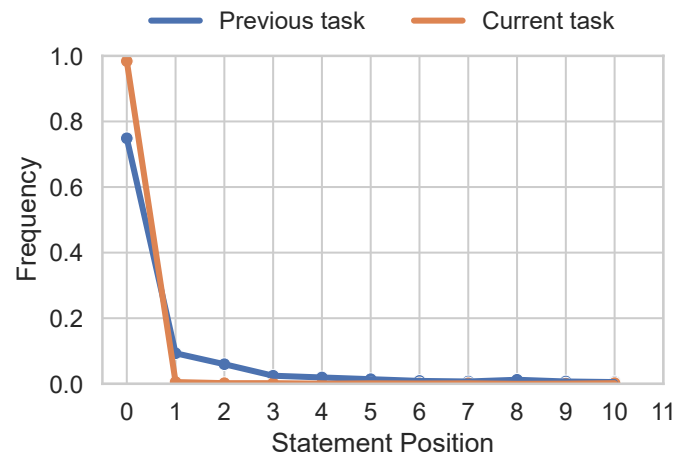
Nearly 36% of workers are between 26 and 35 years old, while the 34% are between 35 and 50 years old. The majority of workers (52%) have a college/bachelor's degree. Concerning the total income before taxes, 25% of workers earn \$50k to less than \$75k, while 19% earn \$75k to less than \$100k. When considering workers' political views, 27% identify as moderate, 27% as conservative, and 26% as liberal. The majority of workers (53%) consider themselves Democrats, while the 27% as Republicans and the 17% as Independents. The majority of workers (53%) agree with building a wall at US southern border, with 25% of them disagreeing. Finally, the vast majority of workers (84%) think that the government should increase environmental regulations to prevent climate change, while only 9% disagree. In general, the sample is well balanced apart from a few categories and similar to the one described in Section 7.3.2, except that most workers in that setting disagree with building a wall at the US southern border.

9.4.2 Task Abandonment

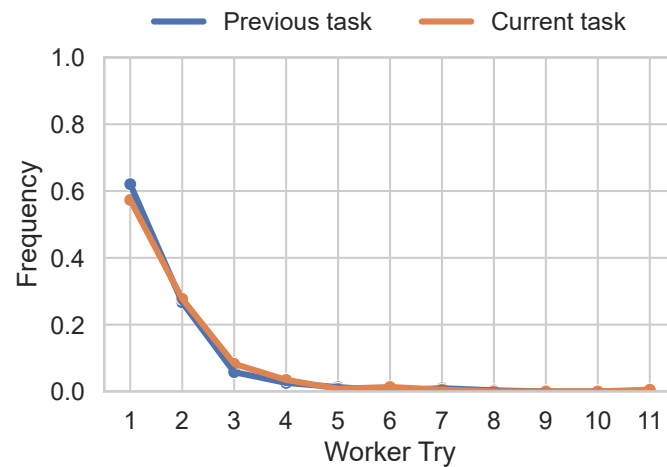
The abandonment rate of the crowdsourcing task is measured using the definition provided by Han et al. [174] (i.e., how many workers voluntarily terminated the task before completion). Overall, 2742 workers participated. Figure 9.2 shows that about 302 (11%) workers completed the task, while 2065 workers (75%) voluntarily abandoned it. Furthermore, 375 workers (14%) failed at least one quality check at the end of the task. Each worker had up to 10 tries to complete the task.

Figure 9.2a shows how many workers abandon the task per number of statements annotated. The vast majority of workers (98%) abandon the task when reaching the first statement. The number of workers who abandon the task after the first statement is negligible. There is an 18% increase in abandonment rate when comparing the values with those described in Section 7.2.2 (i.e., the previous version of the task), compared to which two additional questionnaires and an evaluation dimension are added. Thus, the task described in Section 9.3.1 requires somewhat more effort from workers. A higher number of workers may have become bored or frustrated sooner. Indeed, when considering the previous task, it can be seen that a fraction of workers abandons the task even after reaching the fourth statement. Despite this difference, the general trend is that workers abandon the task when reaching the first statement.

Figure 9.2b shows how many workers fail at least one quality check after submitting their work within their current try. The majority of workers who fail the task perform it only once (216, 58%), with 103 (27%) workers doing it a second time. The remaining 15% of workers who fail the task perform it from three up to 10 times. The failure rate drops from 18% to 14% compared to the task in Section 7.2.2, meaning that those who submitted their work were less likely to fail. However, the failure distribution of our task is in line with one of the previous tasks.



(a) Abandonment distribution.



(b) Failure distribution.

Figure 9.2: Comparison of workers' abandonment and failure distribution. The orange lines represent the task described in Section 9.3.1. The blue lines represent the task described in Section 7.2.2

9.4.3 Agreement

The internal agreement among workers is computed using Krippendorff's α [224] on the unit level. The use of this metric is motivated in Chapter 4 and Chapter 7. There is a low level of agreement overall between the workers for each considered truthfulness dimension, which is in line with the other tasks (Section 4.4.1 and Section 7.4.1).

The external agreement between workers' aggregated scores for the Overall Truthfulness and corresponding experts' values is also measured. It must be recalled that the judgment scales used by the experts and the workers are different. Whereas the experts used six- (PolitiFact) or three-level scales (ABC Fact Check), the workers evaluated the statements using a five-level scale. Figure 9.3 thus shows a box plot for each ground truth label. PolitiFact is shown to the left of the dotted line, and its labels range from 0 (Pants-On-Fire) to 5 (True). ABC Fact Check is shown to the right of the dotted line its labels range from 0 (False) to 2 (True). The figure can be directly compared with Figure 7.3. It shows that workers tend to provide judgments with higher mean value when moving from left to right (i.e., when considering ground truth values with a higher value), in agreement with the experts. Although Figure 9.3 shows only the Overall Truthfulness, Precision and Correctness dimensions as examples, the patterns for the remaining dimensions are similar. Although Overall Truthfulness directly correlates with the ground truth, all the other dimensions capture orthogonal and independent information not directly measured by the experts. Moreover, the inter-quartile range is lower for ABC Fact Check statements when compared to the analysis described in Section 7.4.1.2.

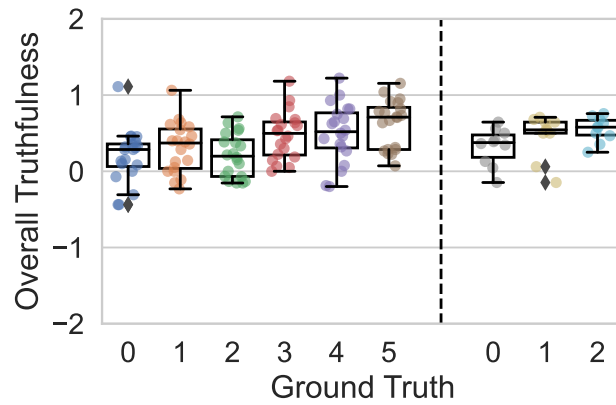
9.5 Results

Section 9.5.1 discusses the testing of the hypotheses outlined in Section 9.2.3. Section 9.5.2 addresses the research questions.

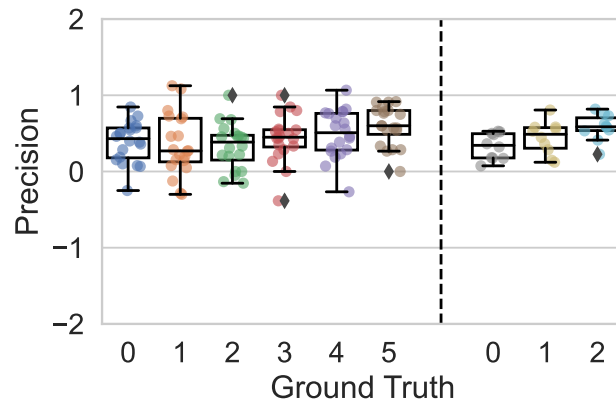
9.5.1 Hypothesis Tests

The hypotheses testing involves conducting statistical analyses. A multiple linear regression is performed to predict eMAE from *trust in politics* (H1a), *belief in science* (H1b), *cognitive reasoning abilities* (H1c), and *mean confidence* (H2d). A one-sample *t*-test is performed to assess H2a (i.e., comparing eME to a test value of 0). Furthermore, a Spearman correlation analysis allows testing H2b (i.e., computing a correlation between affect for the statement claimant and eE). Finally, H2c is tested by conducting a factorial mixed ANOVA with eE as the dependent variable, workers' political party affiliation as the between-subjects factor, and statement's political affiliation as the within-subjects factor (i.e., H2c describes an interaction effect between these two variables).

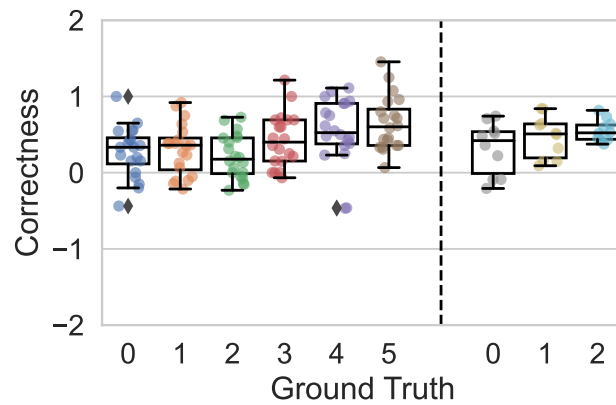
The multiple linear regression analysis reveals no evidence for a relationship between eMAE and *trust in politics* (H1a; $\beta = -0.04, p = 0.020$) or *cognitive reasoning abilities* (H1c; $\beta = 0.02, p = 0.152$). However, *belief in science* (H1b; $\beta = 0.07, p = 0.003$) and *mean confidence* (H2d; $\beta = 0.06, p = < 0.001$) are both significant predictors of eMAE. Workers with stronger *belief in science* and those with greater *mean confidence* are biased in their truthfulness judgments compared to others, partly unexpectedly. Furthermore, workers generally



(a) Overall Truthfulness



(b) Precision



(c) Correctness

Figure 9.3: Correlation with the ground truth of three dimensions with a breakdown on Politifact and ABC Fact Check labels. Compare with Figure 7.3

overestimated truthfulness, as their mean eME (i.e., 0.33, $\sigma = 0.46$) lay significantly above 0 in the one-sample t -test we performed (**H2a**; $t = 12.18, p < 0.001$).

The Pearson correlation analysis reveals a significant positive relationship between *affect for statement speaker* and eE (**H2b**; $r = 0.25, p < 0.001$). Thus, the more the workers like the statement speaker, the more they overestimate truthfulness. Also, the more workers dislike the statement speaker, the more they underestimate truthfulness.

The final analysis is the ANOVA with the statement's affiliated party and worker's affiliated party as independent variables and eE as the dependent variable. Its analysis reveals no evidence in favor of an interaction effect between the two independent variables (**H2c**; $F = 1.59, p = 0.112$), which means no conclusion can be drawn about whether workers have different degrees of over- or underestimating truthfulness depending on whether the statement party match their personally favored party or political direction.

Summarizing, there is evidence in favor of some of the hypotheses suggesting that:

- workers with greater confidence are more biased in their truthfulness judgments (**H2a**);
- workers generally overestimate truthfulness (**H2b**);
- workers' truthfulness judgments are affected by the degree to which they like the statement speaker (**H2d**).

There is also evidence for a relationship between *belief in science* and bias in truthfulness judgments. However, the results show that workers with a stronger *belief in science* are more biased than others, in contrast to **H1b**.

9.5.2 Exploratory Analyses

Next to the descriptive analyses (Section 9.4) and hypothesis tests (Section 9.5.1), exploratory analyses have been performed on the data collected. Section 9.5.3 studies the worker characteristics that lead to systematic biases while crowdsourcing truthfulness judgments. Section 9.5.4 addresses the manifestation of cognitive biases. Lastly, Section 9.5.5 addresses the effect of biases on the individual truthfulness dimensions.

The exploratory analyses aim to further detail the outcomes of the hypothesis tests and to identify interesting trends that have not been covered by the planned ones. It must be noted that the results reported in this subsection are indeed of exploratory nature, as the exploratory analyses have not been preregistered. In light of this, the following results must be considered by keeping in mind the initial exploratory study (Section 9.2 and the experimental variables (Section 9.3.2), together with the outcome of the hypothesis tests (Section 9.5.1).

9.5.3 RQ25: The Role Of Workers' And Statements' Political Affiliations

The ANOVA analysis conducted shows no evidence for an interaction effect between workers' and statements' political affiliations in predicting eE (**H2c**). This suggests that workers may not overestimate or underestimate truthfulness systematically based on whether they support the political party that the statement is affiliated with. The same model also shows no evidence for the main effect of workers' political affiliation on eE ($F = 1.43, p =$

0.232), thus suggesting that workers' political affiliation may not matter at all here. However, there is a significant main effect for statements' political affiliation ($F = 10.55, p < 0.001$).

Comparing the different statement affiliations shows that workers overestimate the truthfulness of statements relevant to the Australian Labor Party significantly more than those relevant to other parties (mean eE = 0.51, Tukey-adjusted $p = [< 0.001, 0.018]$). Workers also judge the truthfulness of statements affiliated with the Australian Liberal party significantly lower than those affiliated with other parties (mean eE = 0.08, Tukey-adjusted $p = [< 0.001, 0.014]$). Republican and Democrat statements were rated roughly equally on average. This suggests that the political parties connected to the statements may matter for predicting bias in crowd workers' truthfulness judgments, even –or perhaps especially– when those parties are not well-known among the crowd worker population (i.e., the crowd workers in the crowdsourcing task described in Section 9.3.1 are all US-based).

9.5.4 RQ26: Predicting eMAE

The multiple linear regression identifies workers' *belief in science* and *mean confidence* as significant predictors of eMAE. Interestingly, the individual Pearson correlation analyses show that only *mean confidence* correlates considerably with eMAE ($r = 0.20, p < 0.001$), whereas *belief in science* does not (see also Figure 9.4). This suggests that *belief in science* only becomes a relevant predictor of eMAE when also taking *trust in politics* and/or *cognitive reasoning abilities* into account, as in the multiple linear regression analysis. These two variables might thus still play an important role in predicting workers' eMAE, even though such evidence has not been found.

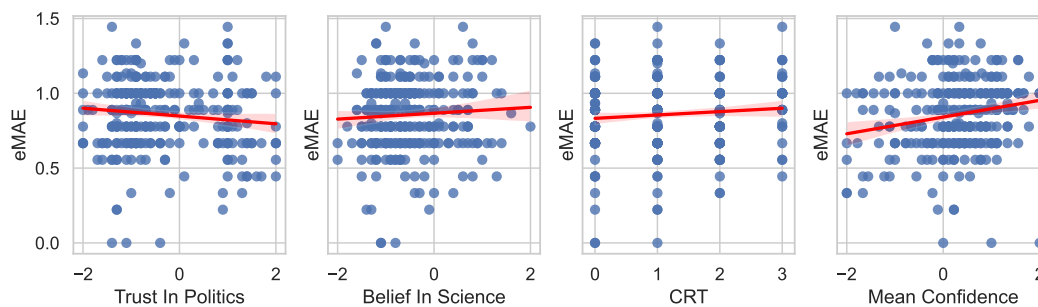


Figure 9.4: Scatter plots showing the relationships between workers' eMAE and their *trust in politics* (H1a, left-hand plot), *belief in science* (H1b, center-left plot), *cognitive reasoning abilities* (H1c, center-right plot), and *mean confidence* (H2d, right-hand plot).

9.5.5 RQ27: Looking At Individual Truthfulness Dimensions

The research question concerns whether different truthfulness dimensions are affected by different biases. Next to an overall tendency towards overestimation of truthfulness, our hypothesis tests (Section 9.5.1) reveal that workers' *belief in science*, *mean confidence*, and the degree to which they like the statement speaker (*affect for statement speaker*) may be related to bias in their truthfulness judgments. Thus, the truthfulness dimensions have been analyzed

to look at which specific dimension is particularly affected by these biases to get some more insight into the nature of these biases.

In more detail, the best iME predictors of eMAE are *Neutrality* and *Comprehensibility*. Workers thus exhibit more bias when they judge *Neutrality* higher ($\beta = 0.10, p = 0.001$) or *Comprehensibility* lower than others ($\beta = -0.08, p = 0.013$). Moreover, workers' *belief in science* affects no other truthfulness dimensions except *Overall Truthfulness*, while the *mean confidence* of a worker is a significant predictor for all iME measures. There are also other interesting relationships, i.e., between workers' *trust in politics* and lower scores on *Neutrality* ($\beta = -0.09, p = 0.028$), and between *cognitive reasoning abilities* and higher scores on *Comprehensibility* ($\beta = 0.08, p = 0.027$). Finally, *affect for statement speaker* is positively related to all considered truthfulness dimensions.

9.6 Summary

The work described in this chapter addresses the impact of worker biases in crowd-sourced fact-checking, addressing three research questions. To perform the analyses, an exploratory study is conducted to derive several hypotheses, using the dataset described in Section 7.2. These hypotheses are then in a novel crowdsourcing experiment. Below, we summarize our findings. The answers to the research questions can be summarized as follows.

RQ25 The first research question concerns what individual characteristics of crowd workers may lead to systematic biases in crowd workers' truthfulness judgments. In this context, no evidence has been found for any influence of workers' *trust in politics* (**H1a**) or *cognitive reasoning abilities* (**H1c**). The results do indicate a relationship between workers' degree of *belief in science* (**H1b**). However, in contrast to what was expected, workers who report a stronger belief in science are less accurate in their truthfulness judgments.

RQ26 The second research question concerns what cognitive biases can affect crowd workers' truthfulness judgments. The results indicate that several cognitive biases can affect crowd workers' truthfulness judgments. Although no evidence for a *B15. Confirmation Bias* has been found in this context (i.e., there was no interaction effect between workers' and statement's party affiliation on truthfulness judgments; **H2c**), workers generally overestimate truthfulness (**H2a**). The findings also suggest an influence of the *B1. Affect Heuristic*: the more workers like the speaker of a statement, the more they overestimate the statement's truthfulness (and vice versa; **H2b**). Finally, there is evidence for *B35. Overconfidence Effect* in crowd workers. The higher workers' self-report confidence in their ability to judge the truthfulness of statements, the less accurate their judgments generally are (**H2d**).

RQ27 The third research question concern whether different truthfulness dimensions are affected by different biases. The study returns exploratory evidence that more biased workers judge the *Neutrality* of statements higher, and the *Comprehensibility* of statements lower than others. Moreover, workers' *trust in politics* is negatively correlated with their *Neutrality* judgments.

The next chapter describes a machine learning-based architecture that can predict the

truthfulness of information items and jointly generate a human-readable explanation for it. The underlying models are validated and calibrated, and an extensive human evaluation of the impact of generated explanations is performed.

Chapter 10

A Neural Model To Jointly Predict and Explain Truthfulness

This chapter is based on the article published in the “Journal of Data and Information Quality” [48]. It is an extension of the one published at the 2021 Truth and Trust Online Conference [47]. Section 2.1 and Section 2.7 describe the relevant related work. Section 10.1 addresses the research questions, while Section 10.3 describes the experimental setting.

10.1 Research Questions

This chapter proposes and experimentally evaluates a system that jointly makes a truthfulness prediction and provides an explanation within the same model. This is novel as compared to classic post-hoc explainability methods that are built on top of existing machine learning models. As such, the generated explanations more closely reflect the decisions made by the veracity prediction model. In addition to this, it shows that large transformer models are flexible enough to multitask, and are thus able to explain their actions without detriment to the original task. This allows human end users to better interface with transformer models, fostering a more trustworthy relationship between humans and deep learning models. It is hoped that automated fact-checking systems will become more widely adopted by creating an automated system that is capable of both evaluating the truthfulness of a statement and simultaneously generating a human-interpretable explanation for this decision. The following research questions are investigated:

RQ28 How a deep learning model can be designed to classify information truthfulness and, at the same time, generate a natural language explanation supporting its classification decision?

RQ29 Can such a model result in both accurate classification decisions and high-quality natural language explanations?

RQ30 Are machine-generated explanations useful for humans to better judge information truthfulness?

RQ31 Can the deep learning model be calibrated in a way that outputs reliable confidence scores for the truthfulness predictions?

10.2 RQ28: E-BART Definition

Many of the systems in the reviewed literature use separate Transformer models for veracity prediction and explanation generation. On the other hand, the architecture proposed in this chapter called E-BART jointly outputs a truthfulness prediction, as well as a human-readable, abstractive explanation addressing.

Adapting the BART-Large encoder-decoder model to this downstream task involved developing a Joint Prediction Head, shown in Figure 10.1. This head sits atop the BART [244] model, and manipulates the transformer hidden states into the form of the desired output. Both the BART base model and the Joint Prediction Head can be fine-tuned as a single unit to customise pre-trained BART weights to the joint prediction task.

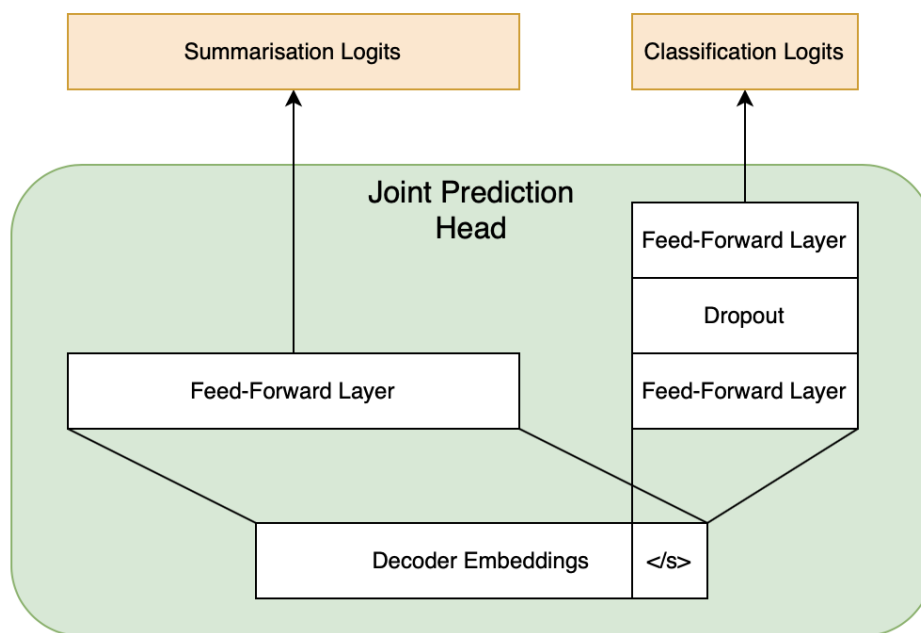


Figure 10.1: The Joint Prediction Head of the E-BART architecture.

The Joint Prediction Head is also shown using the green color in Figure 10.2. The head takes as input the final decoder hidden state embeddings. It then passes all embeddings to a single feed-forward layer to produce a series of logits which form the basis of the predicted explanation. To facilitate classification, the hidden state embeddings corresponding to the final sequence separator token (</s> in BART) are extracted and passed to a

small feed-forward network to shape the output to the desired number of classes. The logits obtained from this are then passed to a final soft-max layer to produce probabilities for each class. Unlike BERT [103], which uses embeddings corresponding to the [CLS] token which is pre-pended to the input to perform classification, in BART the final sequence separator token is used instead as the decoder can only attend to the left of the current token. This conditions the classification of the entire input sequence. The summarization component of the Joint Prediction Head consists of a single feed-forward layer with an input dimension equal to the decoder embedding dimension of 768, while the output dimension is equal to the vocabulary size of the model. The argmax of the raw logits is used by the head during the greedy generation process. It is instructive to consider the training and inference processes separately, as they differ slightly due to the auto-regressive nature of the BART decoder.

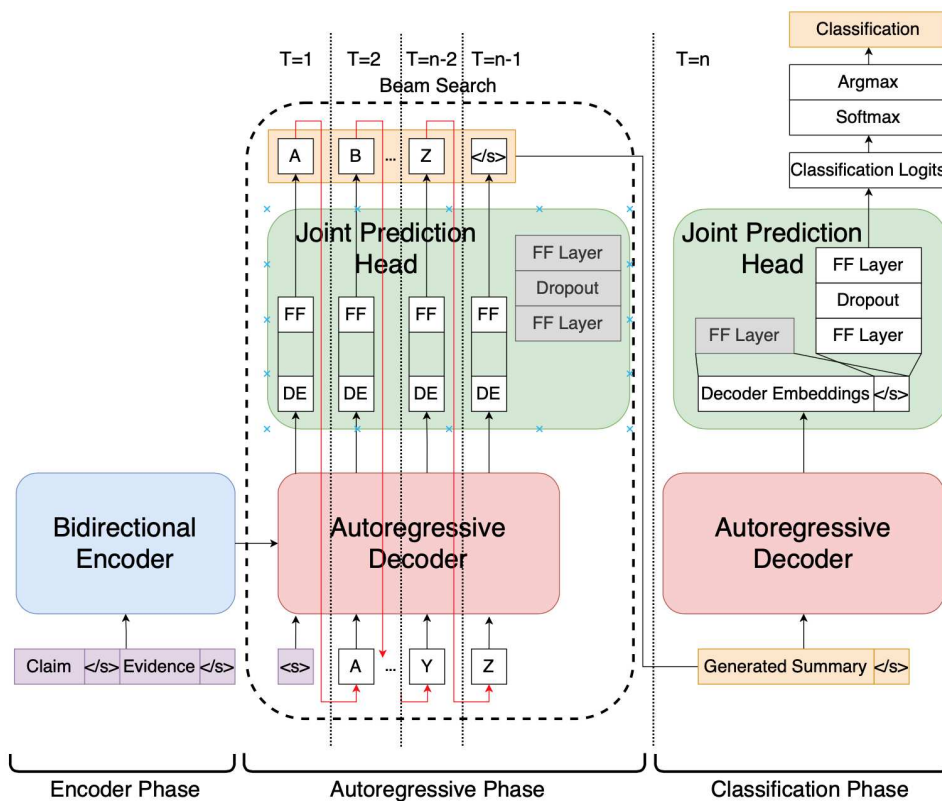


Figure 10.2: The inference process of the E-BART architecture.

Figure 10.2 shows the inference process. Running inference on the model begins by running the encoder with the tokenised input to generate the encoder hidden states, as before. The decoder is presented with the start sequence token (<s> in BART), in contrast to the training process. It generates logits auto-regressively, guided by a beam search. The final phase of inference runs the decoder with the entire generated sequence presented at its input. At this point, the Joint Prediction Head extracts the embeddings corresponding to the token immediately before the final sequence separator token from the generated

sequence. This is done to mirror the training process. These embeddings are passed to the classification component of the Joint Prediction Head, and then to a soft-max layer to produce the final classification.

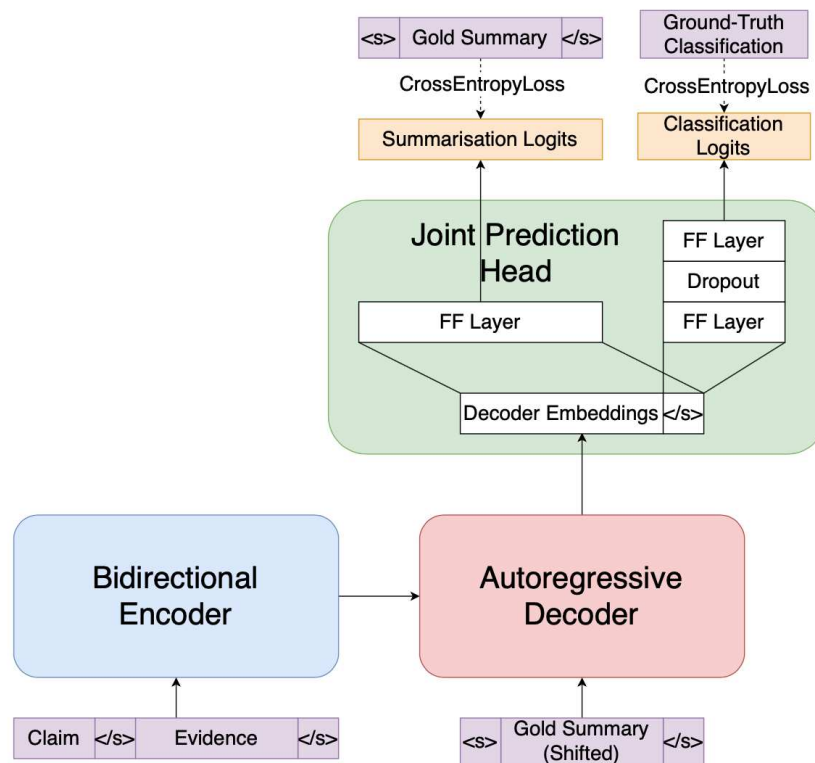


Figure 10.3: The training configuration of the E-BART architecture.

Figure 10.3 shows the training process. The encoder generates hidden states from the tokenised input that are then injected into the decoder, during the training phase. The tokenised gold summary is presented to both the input and summarization output of the decoder, with the input shifted right by one token. This conditions the decoder to predict the next token given the current token. Concurrently, the classification labels are presented to the classification output of the Joint Prediction Head. Note that the model weights for each of the two main “pathways” in the Joint Prediction Head (i.e., generation and classification path) are not shared but up until the head (like in the base BART model), the weights are shared for both classification and generation. The loss is calculated as the weighted sum (with parameters α and $(1 - \alpha)$) of the Cross Entropy Loss computed between the summarisation logits and the gold summary, and the Cross Entropy Loss between the classification logits and the ground truth classification. Thus, the first objective function is to corrupt the text of the summary with a noising function and train the model to learn to reconstruct the original summary text. The second objective function is to adjust the weights of the network such that the logits from the last layer are as close as possible to the class representation of the model (i.e., a classical classification task). This way, the loss

is optimized for both the generation task and the classification task jointly.

The code, training method, and regularization parameters used in the original paper [244] and repository¹ detailing the BART model are used. Lewis et al. [244, Section 5] provides further details. In more detail, the BART pre-trained weights are initially used and then further training is performed on the BART and the developed Joint Prediction Head jointly. It must be noted that the E-BART architecture while being heavily based on BART, has some important differences. Indeed, E-BART does not simply apply a classification layer on the top of the original BART model but rather uses the Joint Prediction Head (Figure 10.2 and Figure 10.3) that allows to perform a joint modelling and thus be able to both classify the truthfulness of a statement and generate an explanation for it, at the same time.

10.3 Experimental Setting

The experimental setting involves three datasets: FEVER (Section 3.3), e-FEVER (Section 3.4) and e-SNLI (Section 3.5). Section 10.3.1 describes the training methodology of the E-BART architecture, while Section 10.3.2 presents its evaluation.

10.3.1 Training Methodology

Two different versions of the model are trained to evaluate the performance of the proposed model on the FEVER and e-FEVER datasets. The first model, E-BARTSmall, is trained on the subset of the e-FEVER training set that did not include null explanations (Section 3.4). This results in 40.702 examples. To process the data, the + character used to separate page titles from the pieces of evidence is removed. The model inputs are tokenised and formatted as `<s>claim</s>evidence</s>`. The truthfulness labels are made numerical and explanations are tokenised in a similar manner. The processed dataset is used to fine-tune the BART-Large model with Joint Prediction Head for 3 epochs. The second model, E-BARTFull, is trained in exactly the same way as E-BARTSmall. However, its training includes the entire e-FEVER training set, including examples with null explanations (50.000 examples).

The models are trained on two platforms: Google Colab, using an NVIDIA T4 GPU with 16 GB of memory, and Microsoft Azure, using a 12 GB NVIDIA Tesla K80 GPU. The models based on BART have roughly 375.000 000 parameters and take approximately 5 hours to train them. In other words, to fine-tune BART and the Joint Prediction Head for 3 epochs on the NVIDIA T4 GPU using the e-FEVER dataset.

10.3.2 Evaluation Methodology

The development split of the e-FEVER dataset is prepared identically to the training split, producing E-FEVERFull and e-FEVERSmall which, respectively, include and do not include examples null explanations. It has been noted when evaluating the truthfulness prediction accuracy of the models that including the NOT ENOUGH INFO class could under-represent the actual classification performance. Table 10.1 shows an example whose ground truth label

¹<https://github.com/pytorch/fairseq/blob/main/examples/bart/README.md>.

is NOT ENOUGH INFO. Manual inspection shows that the explanation and evidence indicate that the statement is indeed refuted, which is correctly predicted by E-BART. Hence, two sets of results are reported, one including and one not including examples that have a e-FEVER label of NOT ENOUGH INFO. Section 10.4 presents a set of experiments performed to evaluate the proposed approach.

Table 10.1: Statement where the ground truth label is NOT ENOUGH INFO and the one predicted by E-BART is REFUTES.

Statement	Evidence	Explanation Generated
Marnie was directed by someone who was “The Master of Nothing”.	Alfred Hitchcock Sir Alfred Joseph Hitchcock (13 August 1899-29 April 1980) was an English film director and producer, at times referred to as “The Master of Suspense”. Marnie (film) Marnie is a 1964 American psychological thriller film directed by Alfred Hitchcock.	Marnie was directed by Alfred Hitchcock, who was “The Master of Suspense”.

10.4 Results

Section 10.4.1 reports the results of a set of experiments performed to evaluate E-BART and validate the usage of the joint models. Section 10.4.2 tests the impact of the explanations generated by relying on crowd workers. Finally, in Section 10.4.3 the network is calibrated to derive meaningful confidence scores.

10.4.1 RQ29: E-BART Evaluation And Validation

The effectiveness of the E-BART model is evaluated on the original FEVER dataset in Section 10.4.1.1. Section 10.4.1.2 address the e-FEVER dataset, while Section 10.4.1.3 the e-SNLI one. Section 10.4.1.4 and Section 10.4.1.5 describe the two experiments performed to validate the usage of the joint models.

10.4.1.1 Evaluation: Original FEVER

The classification performance of E-BART on the original FEVER development set is reported to compare with existing models. The DOMLIN system [397] was used for evidence retrieval (discarding its truthfulness predictions) to provide evidence for 17.000 out of the 20.000 examples in the development set. The E-BART model is used to generate truthfulness predictions for the 17k examples, and then label the remaining with NOT ENOUGH INFO, as specified by Stambach et al. [397]. Results are reported for the development set rather than the test set, as ground-truth labels are not published for the latter.

E-BART_{Small} and E-BART_{Full} achieve on the FEVER dataset label accuracies of respectively 75.0 and 75.1, outperforming state-of-the-art methods. For comparison, other published model accuracies on this dataset include: UKP-Athene (68.5) [175], UNC (69.6) [310],

BERT-BASED (74.6) [391], DOMLIN (72.1) [397], and UCL MR (69.7) [462]. E-BART compares favorably to the existing literature despite the e-FEVER training set having 95,000 fewer examples compared to FEVER, which the other models are trained on. It is hypothesized that the performance improvements are derived from using BART as a base model, and from requiring the model to further attend to the most relevant evidence in forming an explanation. The most noteworthy comparison is between E-BART and DOMLIN, which use identical evidence retrieval mechanisms, thus isolating the contribution of E-BART over standard truthfulness predictors.

10.4.1.2 Evaluation: e-FEVER

Table 10.2 shows the results obtained on the development e-FEVER dataset. Since there have been no other results reported for the dataset, a comprehensive snapshot of E-BART’s performance is presented. Perhaps unsurprisingly, both models perform better on e-FEVERSmall, which contains less inconclusive examples. More surprising is the consistency of E-BART’s performances regardless of whether it is trained on e-FEVERSmall or E-FEVERFull. This indicates that E-BART is robust to situations where evidence is sparse. The ROUGE [253] metrics evaluate the consistency between the generated and e-FEVER dataset explanations, but are not necessarily representative of explanation quality. For instance, the explanation generated by GPT-3 may include some additional information compared to E-BART. Whether this additional information results in a better explanation compared to something more succinct is largely subjective and dependent on the system’s use case.

Table 10.2: Effectiveness of E-BART on the e-FEVER dataset.

Model	Dataset	Accuracy No N.E.I	Accuracy Full	ROUGE 1	ROUGE 2	ROUGE L	ROUGE Sum
E-BARTSmall	e-FEVERSmall	87.2	78.2	73.581	64.365	71.434	71.585
E-BARTSmall	E-FEVERFull	85.4	77.1	59.447	50.177	57.697	57.782
E-BARTFull	e-FEVERSmall	87.1	78.1	64.530	55.283	62.691	62.820
E-BARTFull	E-FEVERFull	85.2	77.2	65.511	57.598	64.071	64.144

Table 10.1, Table 10.3, and Table 10.4, present examples from the development set. Table 10.3 shows qualitatively that the model can even express the fact that it was not able to find relevant evidence. In the generated explanation, the phrase “however does not say whether...” indicates in a human-readable way that the evidence retrieval mechanism was not able to provide sufficient evidence for the model to produce a confident prediction. Table 10.4 reports an example that shows that the model has identified a contradiction between the statement that Janet Leigh was incapable of writing, and the evidence which suggested that she had written four books. The model has generated a coherent explanation that details this finding, extracting relevant facts from the evidence to support the justification. Finally, it has generated the correct label for the statement, which, importantly, is also consistent with the generated explanation.

Table 10.3: Statement where the ground truth label is SUPPORTS and the one predicted by E-BART is SUPPORTS.

Statement	Evidence	Explanation Generated
CBS is the network that aired The Millers	The Millers The multi-camera series aired from October 3, 2013 to July 18, 2015 and ran 34 episodes over two seasons on CBS. CBS announced the cancellation of The Millers on November 14, 2014, four episodes into the show’s second season.	The Millers aired on CBS, however it does not say whether it was the network that aired it.

Table 10.4: Statement where the ground truth label is REFUTES and the one predicted by E-BART is REFUTES (E-BARTSmall).

Statement	Evidence	Explanation Gold	Explanation Generated
Janet Leigh was incapable of writing.	Janet Leigh Janet Leigh (born Jeanette Helen Morrison; July 6, 1927-October 3, 2004) was an American actress, singer, dancer and author. Janet Leigh She also wrote four books between 1984 and 2002, including two novels.	Janet Leigh was capable of writing.	Janet Leigh wrote four books between 1984 and 2002, including two novels.

10.4.1.3 Evaluation: e-SNLI

The e-SNLI task presents a challenge similar to the one of e-FEVER, whereby the entailment between two sentences is predicted and an explanation is generated. Such an idea is similar to predicting the truthfulness of a statement with respect to evidence. A different version of the E-BART architecture is trained specifically on the e-SNLI dataset. The data is prepared by enumerating the labels, removing noisy data, and tokenizing the summaries. The first and second sentences were concatenated and tokenised in the same way as the statement and explanation for the e-FEVER evaluation.

E-BART achieves a label accuracy of **90.1** and a BLEU [320] score of **32.70**. The model proposed together with the e-SNLI dataset, e-INFERSENT, achieves an accuracy of 84.0 and BLEU score of 22.4 [54]. The first two gold explanations are used as references to calculate the BLEU metric for the explainable models as done by Camburu et al. The following are the best-performing models published in the literature² which, however, do not provide explanations. They are reported for further comparison: CA-MTL 92.1 [331], D-CRCo-AN 90.1 [216], MT-DNN 91.6 [260], LMtransformer 89.9 [343], SemBERT 91.9 [470], and SJRC 91.3 [469].

The performance of E-BART compares favorably with the state-of-the-art for this different task, strengthening the result from the FEVER task, and further confirming that E-BART is flexible enough to provide explanations without compromising its performance on the main task. Table 10.5 and Table 10.6 provide some examples to complement the BLEU metric by

²<https://nlp.stanford.edu/projects/snli/>

manually verifying the explanation quality. The generated explanation generated shown in Table 10.5 provides insight into E-BART’s capability to focus the important details of the input. Sentence 1 also contains superfluous information about an additional person in a wool hat, whereas Sentence 2 makes no mention of this. The generated explanation reflects this aspect, as it only discusses the differences between the two sentences in terms of the tan girl, rather than the person in the wool hat. Similarly, Table 10.6 provides an example of where the model has to extract a small, difficult detail in the input sentences that suggests the two sentences may not perfectly entail each other. The fact that Sentence 2 makes a stronger claim about the family going to the beach means that the text in Sentence 1 is not wholly representative of the situation. The E-BART architecture can identify this detail and thus label the sentences as `Neutral`. The generated explanation reflects this, providing insight into the fact that the model has identified that Sentence 2 contains important details omitted in Sentence 1.

Table 10.5: Example where the ground truth label is `Entailment` and the one predicted by E-BART is `Entailment`.

Sentence 1	Sentence 2	Explanation Generated
One tan girl with a wool hat is running and leaning over an object, while another person in a wool hat is sitting on the ground.	A tan girl runs leans over an object	One tan girl with a wool hat is running and leaning over an object is a rephrasing of a tan girl runs leans over an object.

Table 10.6: Example where the ground truth label is `Neutral` and the one predicted by E-BART is `Neutral`.

Sentence 1	Sentence 2	Explanation Generated
A young family enjoys feeling ocean waves lap at their feet.	A young man and woman take their child to the beach for the first time.	Just because a young family enjoys feeling ocean waves lap at their feet does not mean that they take their child to the beach for the first time.

10.4.1.4 Validation: Experiment 1

The performance of E-BART is compared to a pipeline that produces a truthfulness prediction and generates an explanation using two independent models, to test the abilities of the underlying E-BART two models and further investigate RQ29. The architecture and training methodology were kept as consistent as possible, to ensure that the results between the two methods are comparable. The separate pipeline, named `Separate-Bart`, uses a BART-based sequence classifier, and a BART-based model for language generation. Both E-BART and `Separate-Bart` are initialized with the same pre-trained weights, trained and

evaluated on e-FEVERSmall. The BART classifier is trained with the statement and evidence as input to both the encoder and decoder, in contrast to E-BART which uses the much shorter gold summary as input to the decoder. This meant that, due to memory constraints, the inputs are truncated to a maximum length of 256 tokens (which only truncate 4.56% of examples). In addition to this, a virtual batch size of 32 is used (batch size four, with eight gradient accumulation steps) to overcome convergence issues. A batch size of two with two gradient accumulation steps is used when training the sequence generator model, also due to memory restrictions on the hardware available for this experiment. In comparison, the joint model is trained with a batch size of four and no additional gradient accumulation.

Before detailing the results, it is worth remarking on the trade-off between effectiveness and resources needed to train a single model instead of two separate models. Training two separate models would require either the double amount of resources if they are trained in parallel (since each model needs to be fitted into a GPU device independently), or the double amount of time if they are trained sequentially. On the contrary, a single model has to cope with more parameters, but it can be trained on two tasks at the same time. Table 10.7 indicates that the prediction performance of E-BART and Separate-Bart is almost identical, with the latter being slightly more effective. Manual inspection of the generated explanations revealed that both were of similar quality in terms of expressiveness and cohesiveness. This experimental result reinforces what was seen in the practical evaluations on e-FEVER and e-SNLI: that E-BART is able to jointly provide an explanation without diminishing the performance on its main task.

Table 10.7: Effectiveness of E-BART and Separate-Bart on the e-FEVERSmall dataset.

Model	Accuracy No N.E.I	Accuracy Full	ROUGE 1	ROUGE 2	ROUGE L	ROUGE Sum
E-BART	87.2	78.2	73.581	64.365	71.434	71.585
Separate-Bart	88.1	78.9	73.070	63.634	71.005	71.136

10.4.1.5 Validation: Experiment 2

This experiment aims to investigate whether the internal consistency between the predicted truthfulness and predicted explanation differs between E-BART and Separate-Bart. Thus, the same models from Experiment 1 are used (Section 10.4.1.4) but an additional “judge” model is trained to predict the truthfulness of a statement, given an explanation. The ground truth truthfulness labels and dataset explanations from e-FEVERSmall are used to train the BART-based sequence classifier. As such, its weights are not conditioned on those of E-BART or Separate-Bart, meaning that it is independent of both models.

The experiment is run by taking the statements from the development set and the predicted explanations from E-BART. The statements and explanations are then provided to the “judge” model to produce a truthfulness prediction. This “judge” truthfulness prediction is then compared against the truthfulness prediction from E-BART, and the accuracy is computed. The process is repeated for Separate-Bart, and the results are presented in Table 10.8.

The results show a higher accuracy for E-BART as determined by the “judge” model. This indicates that the truthfulness prediction and explanation generated by E-BART are more consistent with each other than those generated by Separate-Bart. Ultimately, this means that joint models are one step closer to being truly interpretable compared to models that generate explanations separately in a post-hoc manner. While this is not conclusive proof, it does provide some evidence that there are consistency gains to be made when using joint prediction and explanation models.

Table 10.8: Internal consistency of E-BART and Separate-Bart on the e-FEVERSmall dataset.

Model	Accuracy No N.E.I	Accuracy Full
E-BART	91.8	86.8
Separate-Bart	90.4	85.8

10.4.2 RQ30: Testing The Impact Of The Explanations Generated

The benefit of explanations generated by the E-BART model is validated experimentally by relying on crowd workers. The human annotators perform a set of crowdsourcing experiments as detailed in the following.

10.4.2.1 Crowdsourcing Task

The crowdsourcing task is published on the Amazon Mechanical Turk platform. Four versions of the same task are published to test the impact of machine-generated explanations of truthfulness. In each version of the task, the workers are provided with a statement from the FEVER dataset. The workers are asked to provide both truthfulness judgments using the two-levels judgment scale along with a sentence justifying their judgments, as this has been shown to improve assessment quality [230]. The labels of the judgment scale are False and True. Each worker is asked to judge the truthfulness of four statements, two labelled in the ground truth as REFUTES, and two labelled as SUPPORTS. Each statement is judged by ten distinct crowd workers. To avoid bias, a randomization process is performed while generating the statement-worker assignments (i.e., in the HITs published using the platform). The same assignments (i.e., same HITs) are kept for consistency within the whole set of task versions (apart from one case, as explained later). The crowd workers were only allowed to complete one version of the task. To ensure the high quality of the collected data and to avoid adversarial behaviour, the workers were required to spend at least 2 seconds judging each statement. The four versions of the task designed, implemented and published under the settings above are the following:

1. Task 1: the workers are provided with the statement from the FEVER dataset. They are asked to provide a truthfulness judgment and a justification.
2. Task 2: the workers are provided with the statement from the FEVER dataset and the explanation generated by the E-BART architecture. They are asked to provide a truthfulness judgment and a justification.

3. Task 3: the workers are provided with the statement from the FEVER dataset and the ground truth explanation. They are asked to provide a truthfulness judgment and a justification
4. Task 4: the workers are provided with the statement from the FEVER dataset, the ground truth explanation and the explanation generated by the E-BART architecture. They are asked to provide a truthfulness judgment and a justification. They are also required to indicate which explanation they find more informative.

A manual inspection of the dataset allowed understanding that for some HITs of the task, the ground truth and the explanation generated by E-BART were the same. Thus, the statements have been re-sampled by requiring the two explanations to be different by at least 1 character. The experimental setup detailed above allows making multiple comparisons, both implicit and explicit. By comparing Task 1 (no explanation) with Task 2 (E-BART explanation) and with Task 3 (ground truth explanation) the effect of showing the explanation to the worker can be tested. Also, an implicit comparison between the two explanations (E-BART and ground truth) can be made. In addition, Task 4 (explanation preferred) allows to explicitly judge which explanation is the one preferred by the workers.

10.4.2.2 External Agreement

Figure 10.4 and Figure 10.5 show the external agreement between the ground truth and the crowd when considering both the individual worker judgments and the judgments aggregated over the ten workers who judge the same statement, using majority vote as aggregation function. The accuracy scores can be computed by inspecting the pilots for the different versions of the task, as follows:

- Task 1: 0.70 for raw and 0.83 for aggregated judgments;
- Task 2: 0.73 for raw and 0.90 for aggregated judgments;
- Task 3: 0.64 for raw and 0.65 for aggregated judgments;
- Task 4: 0.64 for raw and 0.71 for aggregated judgments.

A non-parametric ANOVA with post-hoc test³ is run to account for statistical significance. A non-parametric Kruskal-Wallis H test (one-way non-parametric ANOVA) is used to test the probability that samples came from the same distribution, given that normality assumptions are violated. The null hypothesis that the population medians of all of the considered tasks are equal can be rejected since a p-value < 0.05 is obtained. A post-hoc test must be run to identify which tasks differ in their medians. Thus, the Conover's [83, 84] test is employed. Results show that the task pairs for which the null hypothesis can be rejected (i.e., those which are statistically significantly different) are:

- Task 1 – Task 2 ($p < 0.05$);
- Task 2 – Task 3 ($p < 0.01$);
- Task 2 – Task 4 ($p < 0.01$), for both raw and aggregated judgments.

³<https://scikit-posthocs.readthedocs.io/en/latest/>

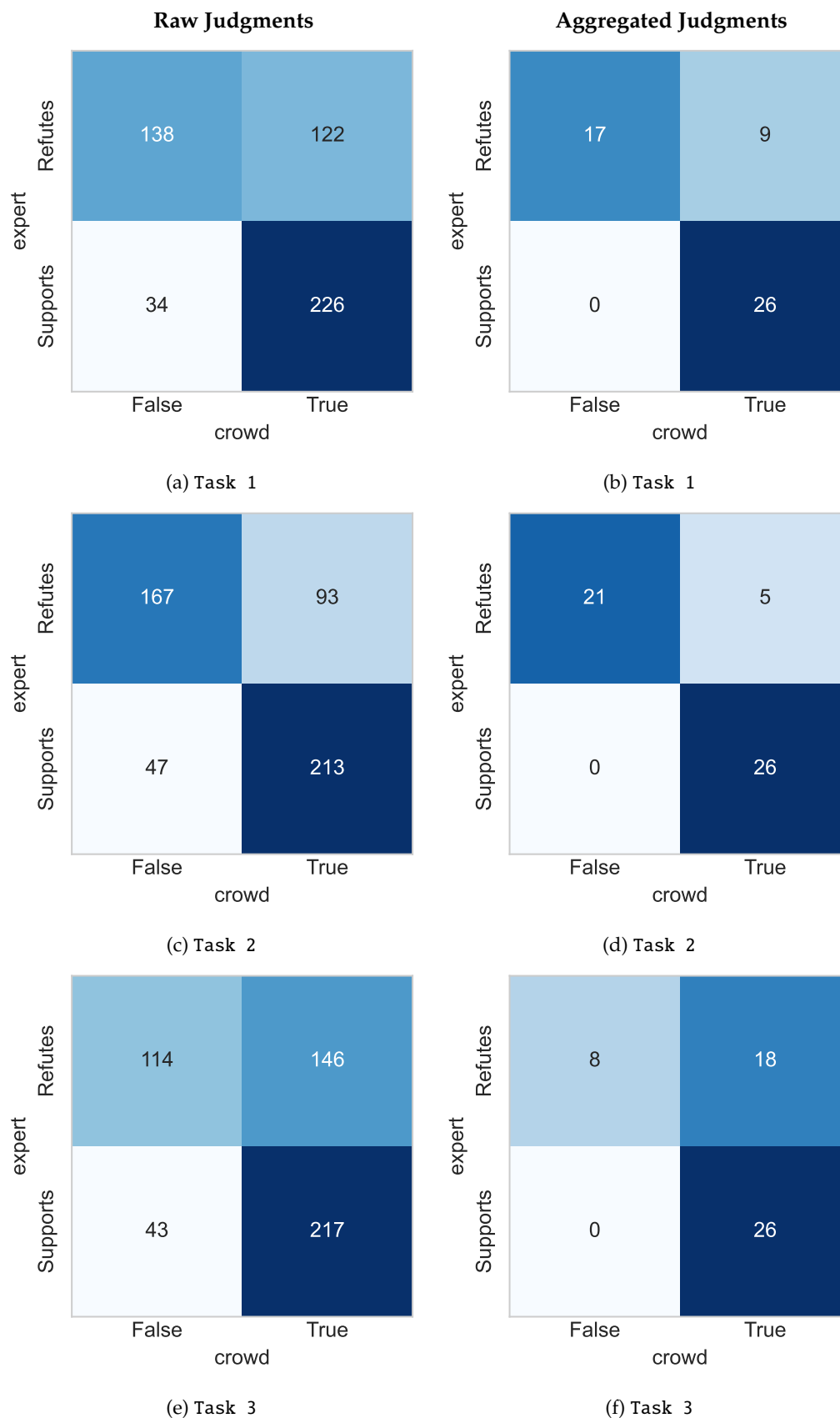


Figure 10.4: External agreement between ground truth and crowd for raw (first column) and aggregated (second column) truthfulness judgments. Each cell represents either the count of judgments (first column), or statements (second column). Correctly classified statements lay on the main diagonal.

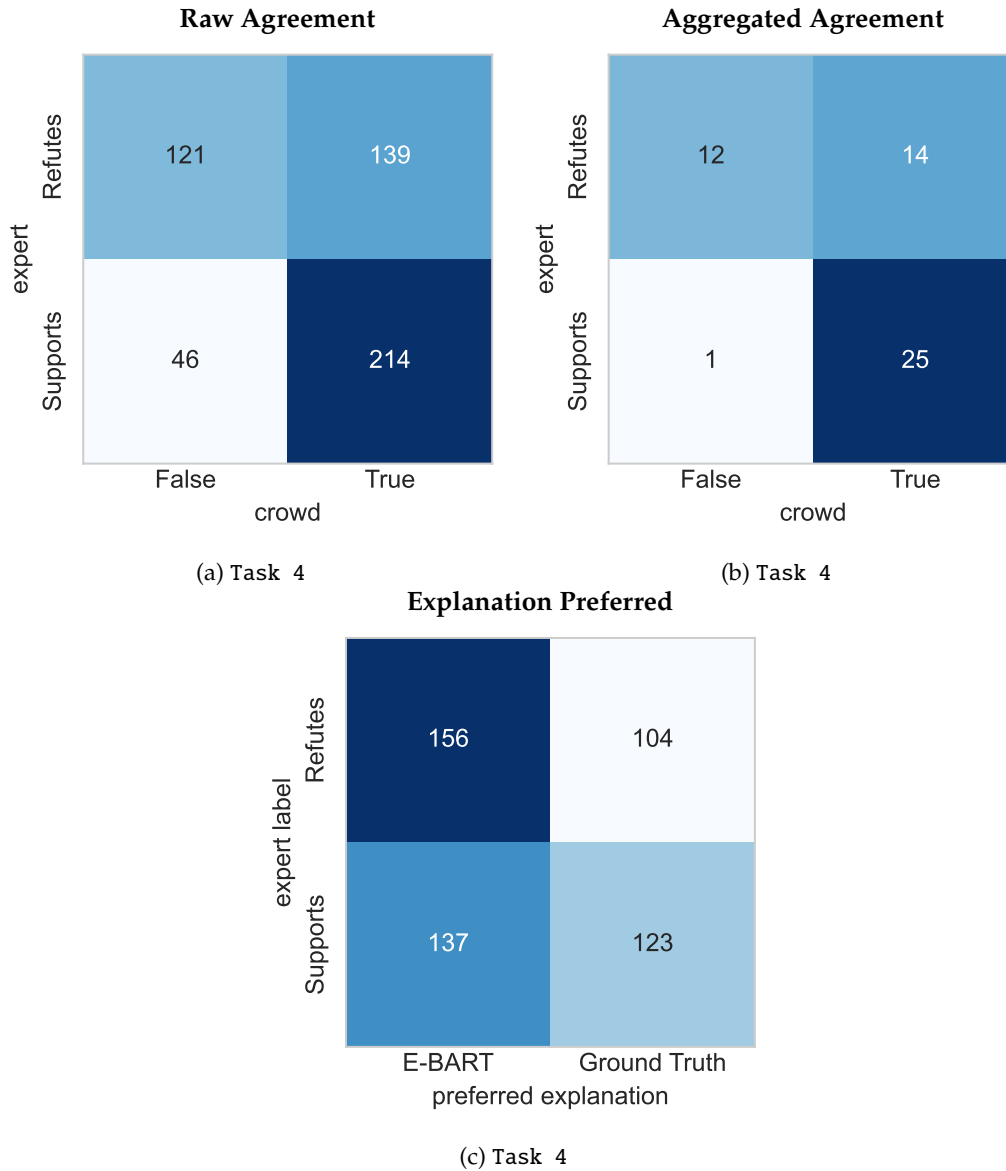


Figure 10.5: External agreement between ground truth and crowd for raw (left chart) and aggregated (center chart) truthfulness judgments. The worker preferences are shown in the right chart. Each cell represents either the count of judgments (left and right charts), or statements (center chart).

10.4.2.3 Internal Agreement

The internal agreement among workers is measured using Krippendorff's α [224, 467], as done in Section 4.4.1.2, Section 5.4.1.2, and Section 7.4.1.3. Thus, α is computed on individual judgments and the agreement scores for the four tasks are respectively of:

- 0.19 for Task 1;
- 0.24 for Task 2;
- 0.10 for Task 3;
- 0.10 for Task 4.

These values indicate a similar low level of agreement between the tasks.

10.4.2.4 Summary

The results obtained allow drawing the following remarks. Showing the E-BART explanation (Task 2) is better than showing no explanation (Task 1), while this is not true when considering the ground truth explanation (comparison between Task 1 and Task 3). The implicit comparison between Task 2 and Task 3 shows that crowd workers are more accurate when considering the E-BART explanation in place of the ground truth explanation. Such result is also confirmed when making the explicit comparison (Figure 10.5c).

Additionally, displaying the E-BART explanation (Task 2) reduces the number of *false positives* (i.e., statements that are false but are erroneously perceived as being true by the crowd workers) from 122 to 93. Such a result is not true when considering the ground truth explanation. Thus, the explanations automatically generated by E-BART have the effect of making people more skeptical about statements (see also Table 10.3 for an example). This behavior does not hold for *false negatives* (i.e., statements that are true but are erroneously perceived as being false by the crowd workers). Note that in misinformation settings *false positives* are potentially more dangerous than *false negatives*, and it is better to be erroneously skeptical than not recognizing false statements.

Looking back at accuracy scores and their implications, performing a simple aggregation of crowd judgments under conditions of Task 2 allows achieving 90% non-expert label accuracy, which is a promising step towards crowdsourced truthfulness judgments (Section 4.5).

10.4.3 RQ31: Network Calibration And Generation Of Confidence Scores

Producing confidence scores along with the truthfulness classification would further enhance the transparency and interpretability of the E-BART predictions. These confidence scores provide insight into how confident the model is in making its prediction. The confidence could be impacted by a number of factors, including the quality and quantity of the provided evidence, or the similarity of the statement to the training data. E-BART makes its predictions using a soft-max layer over a series of logits with dimensions equal to the number of target classes. The output of the soft-max layer produces a "probability" score for each class. However, it is unlikely that this probability is well-calibrated. That is, the output from the soft-max layer is not likely to be representative of the true probability of correctness [167]. It is not clear if such an issue is present in all transformer-based models,

even though this calibration error is present in most modern deep neural networks. Thus, it works investigating this issue for the E-BART model.

A number of post-processing techniques exist to correct the calibration error, allowing the output of the final soft-max to be correctly interpreted as the confidence of correctness. Some techniques include Bayesian Binning into Quantiles [298], Platt Scaling [334], Histogram Binning [465]. Huang et al. [191] propose a tutorial on calibration measures. Temperature Scaling [167] is a technique that has demonstrated high efficacy on a range of neural networks including multi-class classifiers, and is relatively easy to implement. Temperature Scaling introduces a single parameter, the temperature ($T > 0$), and uses it to produce a confidence prediction given a series of logits z_i . Equation 10.4.3 show the resulting computation, where σ_{SM} is the soft-max function. Applying temperature scaling to a model does not change its classification prediction, and therefore does not alter the accuracy of the model [167].

$$\hat{q}_i = \max_k \sigma_{SM}(z_i/T)^k \quad (10.1)$$

The temperature parameter must be tuned on the validation set rather than the original training set [167]. The first 9999 items from the e-FEVER development set are used to train the temperature parameter, while the remaining examples are used for validation. This is needed since there is no test split for the e-FEVER dataset. The temperature parameter is inserted after the final fully-connected layer of the original model, and the original model parameters are frozen. The temperature parameter is then trained using the LBFGS optimizer [258] for 10,000 iterations. It is important to note that the E-BART model is run in “auto-regressive (inference) mode” to produce the logits for input to the temperature parameter. Finally, the model is tested using the held-out validation data. The testing proceeds by running E-BART in inference mode, and applying the temperature parameter prior to the final soft-max function.

A reliability diagram is produced for the model before and after calibration, to evaluate the calibration itself. In addition, the Expected Calibration Error [298] (ECE) and Maximum Calibration Error [439, 167] (MCE) were calculated using ten bins. Figure 10.6 shows the E-BARTFull model before and after calibration. The dotted 45-degree line on the reliability diagrams specifies the perfect calibration. Deviation from this line indicates that the predicted confidence scores differ from the actual accuracy. Visual inspection of the reliability diagrams indicates that the original E-BART model was in fact not well calibrated, with ECE of 11.44% and MCE of 26.35%. However, following temperature scaling, the model performs more closely to the ideal calibration, with ECE and MCE reduced to 1.61% and 6.92%, respectively. This increase in calibration means that the output of the final soft-max layer can be more accurately interpreted as a confidence score.

Table 10.9 and Table 10.10 are provided to demonstrate how the change in calibration impacts the model output. The former shows a situation in which E-BART predicts the correct truthfulness and indicates that it is confident in its prediction. The latter demonstrates a situation where the model is not confident in its prediction. Here, the ground truth truthfulness indicates that the statement is refuted, and even the ground truth explanation does a poor job of communicating the evidence. In this case, E-BART predicts that the statement is supported, however, it indicates that it has low confidence in this prediction.

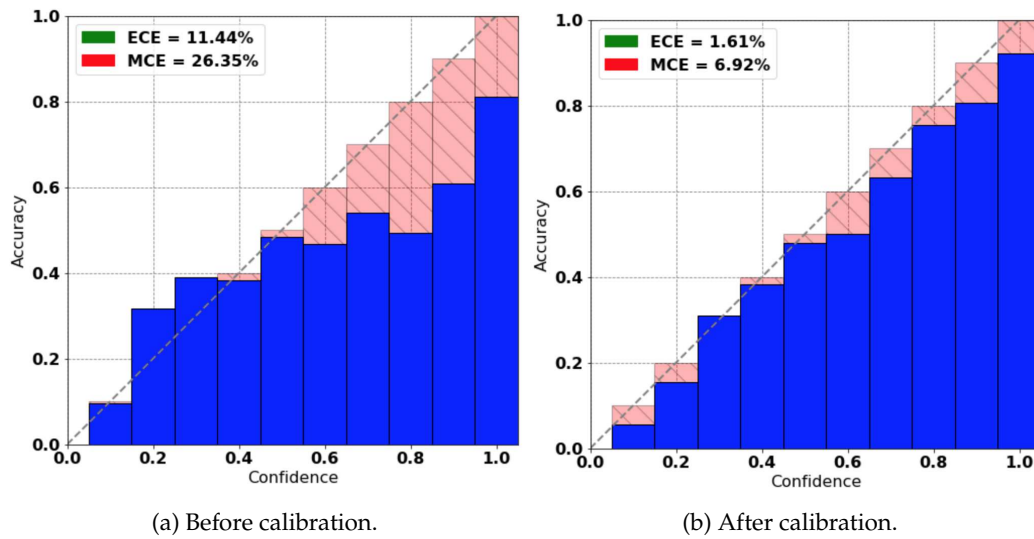


Figure 10.6: Confidence as a function of accuracy for the E-BARTFull model. The calibration error is corrected using Temperature Scaling [167].

The only indication that perhaps the model is not accurate without providing confidence in this instance is the low quality of the explanation. Now, however, there is a quantitative representation of confidence. This would demonstrate to the end user that in this instance, the model should not be trusted.

10.4.4 Summary

The work described in this chapter explores the potential of Automated Fact-Checking (AFC) models that can jointly make a truthfulness prediction and provide a human-readable explanation for that prediction. The E-BART architecture is proposed to such an end and its performances are evaluated on the e-FEVER and e-SNLI tasks. Experimentation revealed that E-BART could achieve results comparable to the state-of-the-art and simultaneously generate coherent and relevant explanations. The hypothesis is that predicting the truthfulness and providing an explanation makes AFC systems more transparent, and fosters greater trust in the system. The answers to the research questions can be summarized as follows.

RQ28 The E-BART architecture is designed and proposed by developing a Joint Prediction Head that sits atop the BART model. The Joint Prediction Head manipulates the transformer's hidden states. Such modelling allows both classifying the truthfulness of a statement and generating an explanation for it, at the same time.

RQ29 The evaluation of E-BART show that the architecture is competitive with the state-of-the-art on the e-FEVER and e-SNLI tasks. Additionally, generating explanations does not significantly impede the model from performing well in its main task of truthfulness prediction. The truthfulness predicted and the explanations are more internally coherent when generated jointly than separately.

Table 10.9: Statement where the ground truth label is SUPPORTS and the one predicted by E-BART is SUPPORTS. Confidence score is 0.851, original confidence was 0.987 (E-BARTFull).

Statement	Evidence	Explanation Gold	Explanation Generated
Ekta Kapoor worked on an Indian soap opera that premiered in 2000.	Kyunki Saas Bhi Kabhi Bahu Thi (Because a mother-in-law was once a daughter-in-law , too) is an Indian soap opera that premiered on 3 July 2000 on Star Plus. Ekta Kapoor Some soap operas she had worked on include [...]. Pavitra Rishta (Sacred Ties) is a 2009 Indian soap opera produced by Ekta Kapoor of Balaji Telefilms, that aired on Zee TV. Kasautii Zindagii Kay (The criterion of life), often abbreviated as KZK, is an Indian soap opera created by Ekta Kapoor 's Balaji Tele-films for the channel STAR Plus.	Ekta Kapoor worked on the Indian soap opera Kyunki Saas Bhi Kabhi Bahu Thi, which premiered in 2000.	Ekta Kapoor worked on the soap opera Kyunki Saas Bhi Kabhi Bahu Thi, which premiered in 2000.

RQ30 The extensive human evaluation of the impact of the explanations generated by E-BART using a crowdsourcing task reveals that the explanations generated by E-BART generally increase the human ability to spot misinformation and make people more skeptical about claims. The explanations are also competitive with the ground truth ones.

RQ31 The E-BART architecture is calibrated using an approach based on the Temperature Scaling technique in order to make it produce reliable confidence scores for the truthfulness predictions. Indeed, the results show that the model was not well calibrated before.

The next chapter concludes this thesis. Initially, it summarizes each contribution. Then, for each meta-research question, provides a list of practical implications, sketches the future work and addresses the current limitations of the approaches proposed.

Table 10.10: Statement where the ground truth label is REFUTES and the one predicted by E-BART is SUPPORTS. Confidence score is 0.432, original confidence was 0.550 (E-BARTFull).

Statement	Evidence	Explanation Gold	Explanation Generated
Henry III assumed the throne when he was 2 years old.	Henry III of France Henry III (19 September 1551 - 2 August 1589; born Alexandre Édouard de France, Henryk Walezy, Henrikas Valua) was a monarch of the House of Valois who was elected the monarch of the Polish-Lithuanian Commonwealth from 1573 to 1575 and ruled as King of France from 1574 until his death.	Henry III was elected the monarch of the Polish-Lithuanian Commonwealth when he was 2 years old.	Henry III was born on 19 September 1551 and died on 2 August 1589.

Discussion

Section 11.1 summarizes the main contributions provided by the experiments performed in this thesis. Section 11.2 lists their practical implications, while Section 11.3 describes the limitations. Section 11.4 sketches future research directions. Each section addresses separately the meta-research questions detailed in Section 1.6. Section 11.5 concludes this thesis. Finally, Section 11.6 presents the acknowledgements.

11.1 Contributions

The contributions that this thesis allows obtaining are summarized in the following. Initially, Section 11.1.1 reports those related to the assessment of online (mis)information. Then, Section 11.1.2 addresses the characterization of cognitive biases and the study of their impact on the fact-checking activity. Lastly, Section 11.1.3 outlines the contributions related to predicting the truthfulness of information items and generating explanations.

11.1.1 MRQ1: Information Truthfulness Judgment

The results obtained show that the judgment scale used to collect the truthfulness judgments does not affect their quality. The crowdsourced data correlates well with expert fact-checker judgments when properly aggregating workers' answers and merging truthfulness levels, even if the agreement among workers is low. The background of the crowd workers has an impact on the judgments they provide (RQ1–RQ4).

The crowd workers are able to detect and objectively categorize online recent information related to the COVID-19 pandemic. Both the crowdsourced and expert judgments can be transformed and aggregated to improve quality. The background of crowd workers, together with other signals (e.g., source of information, behavior), impacts the quality of the data. The longitudinal study demonstrates that the time span has a major effect on the quality of the judgments, for both novice and experienced workers. Extensive failure analysis of the statements misjudged by the crowd-workers is provided (RQ5–RQ12).

The first systematic and large-scale survey study characterizing longitudinal studies from the perspective of crowd workers is provided. It is conducted over three popular commercial crowdsourcing platforms and a detailed quantitative and qualitative analysis of the answers collected is performed. A set of 9 recommendations for researchers and practitioners who wish to conduct longitudinal studies over crowdsourcing platforms and a set of 5 best practices for commercial crowdsourcing platforms to support such studies are provided, informed by the outcome of the analysis of the cross-platform study (RQ12–RQ15).

The truthfulness judgments provided by crowd workers over the seven dimensions of truthfulness are sound and reliable. The agreement with the ground truth provided by experts is good when the same dimension is measured, and reasonable for the individual dimensions, with differences that can be justified by the meaning of each dimension. Several analyses show that the seven dimensions are independent, not redundant, and measure different aspects. The analyses on the informativeness of the different dimensions show that the different dimensions can be useful to understand the reasons behind the crowd worker's judgment. The signals derived from workers can be leveraged to effectively predict the expert verdicts, in particular, their judgments and search sessions (RQ16–RQ20).

11.1.2 MRQ2: Cognitive Biases

The characterization of cognitive biases summarizes the misinformation literature to create a comprehensive list of those biases that may affect the fact-checking process. The subset of 39 out of 220 cognitive biases that are likely to manifest is detailed using the PRISMA methodology. Furthermore, a classification is proposed together with a list of countermeasures to limit their impact. A bias-aware assessment pipeline for fact-checking is thus proposed, and each countermeasure is mapped on a constituting block of the pipeline itself (RQ21–RQ24).

The experiments performed to study the impact of worker biases in crowdsourced fact-checking show that there is no evidence for any influence of workers' trust in politics or cognitive reasoning abilities. On the other hand, there is a relationship between workers' degree of belief in science. In contrast to the expectations, workers who reported a stronger belief in science were accurate in their truthfulness judgments (RQ25–RQ27).

11.1.3 MRQ3: Predict And Explain Truthfulness

The potential of automated fact-checking models that make a prediction and jointly provide a human-readable explanation for it is explored. The novel E-BART architecture is proposed and its performance is evaluated within classification tasks. The experiments reveal that E-BART could achieve results comparable to the state-of-the-art and simultaneously generate coherent and relevant explanations.

The human evaluation of the impact of explanations generated reveals that they generally make people more accurate in detecting misinformation and more skeptical of a statement they encounter online. The E-BART architecture is calibrated to make it produce reliable confidence scores for the truthfulness predictions (RQ28–RQ31).

11.2 Practical Implications

The analyses described in this thesis allow various remarks which can be helpful in practice for researchers and practitioners. Section 11.2.1 describes the outcomes of the experiments to the assessment of (mis)information spread online, while Section 11.2.2 discusses the characterization of cognitive biases and their impact on the fact-checking activity. Then, Section 11.2.3 briefly addresses a use case of the model proposed to predict the truthfulness of information items and generate explanations for it. Finally, Section 11.2.4 describes a publisher work where the multiple dimensions of truthfulness described in Chapter 7 are used in the context of product reviews quality judgment.

11.2.1 MRQ1: Information Truthfulness Judgment

The main practical implications of experiments performed to understand the ability of human assessors to address misinformation, such as the study of the effect of judgment scales and workers' background described in Chapter 4 and the multidimensional notion of truthfulness proposed in Chapter 7, are summarized by the following list:

- Crowd workers can detect and objectively categorize online (mis)information related to political statements.
- Crowd workers can detect and objectively categorize online (mis)information related to recent information items such as those related to the COVID-19 pandemic.
- Crowd workers can assess the truthfulness of political statements using a multidimensional scale.
- Researchers should not rely on questionnaire answers, which are not a proxy for worker quality. In particular, a worker's background/bias is not helpful to increase the quality of the aggregated judgments.
- The agreement among workers usually does not provide a strong signal. This is further confirmed also when leveraging the multiple dimensions of truthfulness.
- There is a major effect on the quality of the judgments if they are collected for the same documents at multiple time-spans. Batches which are closer in time to each other are more similar in terms of workers' quality, and experienced/returning workers have generally higher quality than novice workers. Thus, a researcher that aims maximizing the agreement with expert labels should rely on experienced workers.
- Researchers should expect the labelling quality to vary depending on statement features and peculiarities. There are statements which are objectively difficult to evaluate. Statements for which there is little or no information will be of lower quality. Workers might focus only on part of the statement/source of information to give a particular truthfulness label, so asking for a specific textual justification might help in increasing to quality of the judgments.
- The truthfulness categories at the lower end of the scale (such as Pants-On-Fire and False) are evaluated very similarly from crowd workers.
- Researchers should expect that the workers' values will tend to be skewed towards the positive values of the Likert scale (i.e., Agree (+1) and Completely Agree (+2)). Such behavior is present but less evident when the individual judgments are aggregated using the mean function.

- The truthfulness labels can be transformed and aggregated to improve judgment quality since it provides an increase of the agreement with expert labels. The arithmetic mean should be used as an aggregation function, as it provides a high level of agreement with the expert labels.
- The seven dimensions of truthfulness used are independent. Thus, researchers can re-use the same set of dimensions proposed in this thesis to collect truthfulness judgments using crowdsourcing.
- Researchers should expect no correlation between the judgments gathered on every single dimension and the corresponding set of computational measures that can be automatically computed for the same dimension; crowd workers and computational measures provide a different signal.
- Researchers should avoid the usage of naive techniques to combine the multiple dimensions as they do not improve the external agreement with expert labels.
- The usage of a custom search engine stimulates workers to use multiple sources of information and report what they think are good sources to explain their label.
- The quality checks implemented in the crowdsourcing tasks are helpful to obtain high-quality data.
- Researchers should expect faster response times as the worker proceeds into the task since he/she will learn while doing it. For this reason, the same statement should be presented in different positions in the task to avoid any possible source of bias.

11.2.2 MRQ2: Cognitive Biases

The characterization of cognitive biases that can be applied in the setting of fact-checking described in Chapter 8, the proposed bias-aware assessment pipeline for fact-checking and the study about their effect described in Chapter 9 detailed have various practical implications for crowdsourcing truthfulness judgments as well as adjacent domains such as the collection of document viewpoint annotations [110, 112, 288]. The following list summarizes the main practical implications.

- Researchers and practitioners are allowed to fully understand which cognitive biases can manifest in a fact-checking task.
- The findings presented may help human expert fact-checkers in revising their processes [58].
- Artificial intelligence researchers and developers can be more informed while building models that are robust against biased data and result in fair decisions.
- Although crowd workers are generally reliable when judging the truthfulness of statements, individual characteristics (e.g., their *belief in science*) or cognitive biases (e.g., the *B1. Affect Heuristic* or *B35. Overconfidence Effect*) can negatively affect the accuracy of their judgments. The recommendation is to assess, document, and –where possible– mitigate these biases [112, 126, 192] either by adapting the task design or corrective post-processing of the collected data.
- Table 8.2 presents possible actions towards the mitigation of the identified risks in producing biased fact-checking decisions.
- Filtering the information available to human assessors to avoid them being biased by

the evidence they should not consider during their judgment process (see *C1. Custom Search Engine*)

- Allowing for an open discussion leads to highlighting possible extreme individual views (see *C2. Discussion*).
- Asking the assessors to revise their judgment (i.e., adopting *C8. Revision*) might help in mitigating *B2. Anchoring Bias*.
- Instructions given to assessors are another possible source of bias and thus making sure they are presented as intended (see *C4. Instructions*) is a way to avoid wrong priming and bias.
- Requesters should hide unnecessary information (e.g., statement speaker identities or political affiliations) where possible to mitigate the influence of biases such as the *B1. Affect Heuristic*.
- Task requesters should measure relevant concepts such as workers *belief in science* [92], where applicable, to enable effective assessment of systematic biases.
- Requesters could consider prioritizing workers with moderate political affiliation, *belief in science*, and confidence in their judgment abilities, as this thesis suggests that overly strong convictions in these contexts can lead to worse quality in truthfulness judgments.
- The usage of unnecessary instruments should be avoided. For instance, the outcome of the cognitive reasoning test (CRT) shows no relationship to the quality of truthfulness judgments has been found. Requesters should be aware that each such test may reduce the *cognitive reasoning abilities* of crowd workers to eventually perform the actual task. Thus, although assessment and mitigation of systematic biases are recommended, it must be noted that requesters should also not overdo it in this respect.
- Judgments coming from workers with high self-reported Confidence in their ability to identify misinformation should be carefully adjusted, as such workers tend to be more biased than others.

11.2.3 MRQ3: Predict And Explain Truthfulness

Turning to leverage the truthfulness judgment collected using machine learning-based approaches, researchers can use the crowd judgments gathered to predict the expert judgments in a supervised learning scenario. In such a case, researchers should use the Random Forest algorithm as it provides the best effectiveness metrics, which are higher than baselines.

11.2.4 Multidimensional Reviews Quality Judgment

This section is based on the article published in the “Information Systems” journal [61]. It is an extension of the one published at the 20th International Conference on Web Engineering [62]. Part of the related work described in Section 2.6 is of interest. This section sketches a work that is an implication of the work described in Chapter 7.

Online reviews can be a valuable source of information, as they allow users to gain from the experience of others who have expressed their opinion about the next product to buy, room to book, etc. Opinions provided by Web users are useful if those of higher quality can

be identified and those that can be characterized as low quality can be dismissed (e.g., for bias, incompleteness, irrelevance, and so on). Over the past years, research has characterized reviews' trustworthiness in several ways: user reputation and quality assessment are among them. However, while reviews are about specific products or services, they often express multifaceted views on the target object. To judge the quality and trustworthiness of a review, it is important to understand which arguments it provides, its strengths and which aspects of a target product provide positive or negative evidence. In other words, reviews are a means for users to express their opinions on a given product or service. Reviews can be seen in the form of ratings-descriptions pairs. Such a form indicates a rating (often on a 1–5 Likert scale) for the quality of a given target product, enriched with textual descriptions motivating the score. The textual description of a review can provide one or more arguments to support the corresponding Likert scale rating given. Thus, argumentation reasoning is used to analyse these textual descriptions. Research on the assessment of the quality and credibility of product reviews has focused mostly on linguistic aspects, e.g., based on readability and linguistic errors [156, 222, 311, 455]. Wathen et al. [441], on the other hand, proposes an approach that looks into credibility factors. Lastly, Wyner et al. [456] looks for a junction between natural language processing and argumentation reasoning. While it classifies more thoroughly the diverse tokens as different kinds of arguments, it does so in a semi-supervised fashion.

Formal Argumentation implements argumentation theory, which is the interdisciplinary study of how conclusions can be reached from premises through logical reasoning. In such a formal setting, arguments are the atomic unit of analysis and the scope of the theory is that of analysing the complex graph resulting from relations between them, considering whether an argument attacks or supports another argument, and identifying which argument survives. Hence, descriptions within reviews are analyzed to identify arguments that support the corresponding scores. A formal semantics of value-based argumentation that extends the model of Baroni et al. [30] can be proposed. Such a semantic allows for describing the conflict and support dynamics between tokens within a set of reviews of a given product. In more detail, the approach proposed falls within the growing family of weighted argumentation frameworks extending standard Dung's setting. It also relies on an ordering of weighted attacks, with some differences from previously proposed frameworks. The conflict and support dynamics between tokens within a set of reviews of a given product are formulated within a graph structure where nodes represent arguments and edges are attacks among them. The nodes of the graph are interpreted as reviews, requiring that reviews occurring in the same graph refer to at least one common feature of the product under evaluation. Edges of the graph express the attack relation between two reviews assigning different scores to the feature in common. The direction of the attack is given by the relevance of the tokens and the values of the reviews. The semantics of the graph is defined by a standard formal argumentation theory labelling function on vertices: (i) a review is labelled *in* when all its attackers are *out*; (ii) a review is labelled *out* when at least one of its attackers is *in*; (iii) a review is labelled *undec* if not all its attackers are *out* and no attacker is *in*.

Figure 11.1 illustrates this semantics through an example. R1 and R4 are labelled as *in* because either all their attackers are *out* (R1) or they do not have any attackers (R4). R2 and R3 are labelled as *out* because their attacker is *in*. R5 and R6 have only attackers

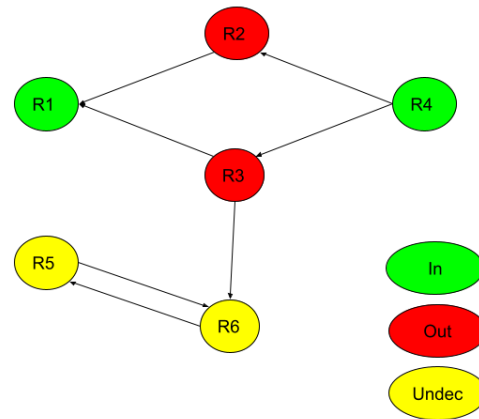


Figure 11.1: Example of labeling of reviews following the standard argumentation theory adopted.

undec or out and are thus labelled as undec. This semantics uses a scoring system for tokens within reviews generated from natural language processing which allows translating reviews and their relations in a graph construction algorithm. A detailed description of the overall pipeline and the formal details of the semantics is not provided in this section. The implementation of the pipeline is freely available to the research community.¹

The model obtained is evaluated on the Amazon Review Dataset [276], in particular on the Amazon Fashion 5-core dataset, which consists of 3,176 unique reviews provided by 406 users about 31 products. One of the goals of analysing via crowdsourcing is a stratified sample with the same number of reviews per number of stars assigned. However, the 5-core dataset does not contain a sufficient number of reviews for this aim. In more detail, the 5-core dataset is a subset of the complete Amazon Review Dataset where each of the users and items has 5 reviews. However, the dataset contains just a fraction of the subset of 5-core user/item pairs. To evaluate the approach proposed the whole 5-core dataset is first rebuilt. The Amazon Review Dataset consists of 883,636 reviews (871,502 after duplicate removal) provided by 746,352 users about 185,241 products. To rebuild the whole 5-core dataset, 5 reviews for each product ASIN code provided by 5 different authors are sampled. Each user/product pair must be unique across the whole dataset. In this way, there are 5 different authors of reviews for every product and 5 reviews provided by the same author for 5 different products, thus complying with the 5-core assumption. Hence, the final 5-core dataset is made of 148,588 reviews about 29,958 products, which corresponds to 16.81% of

¹<https://github.com/davideceolin/FAReviews>

the original Amazon Review Dataset. The sample used in the first version of the approach proposed [62] is extended by sampling reviews from the whole 5-core dataset to obtain the same number of reviews per each number of stars/upvotes. The final augmented sample balanced along the number of stars/upvotes is composed of 670 reviews which correspond to 0.45% of the original Amazon Review Dataset.

A crowdsourcing task is deployed to collect 670 reviews by randomly selecting one of the products reviewed at a time and then drawing one of its reviews until a balanced number of reviews per review score value is obtained. This leads to an amount of 134 reviews for each score. Each worker is asked to evaluate the quality of 10 reviews and each review is evaluated by 5 workers. Workers are located in the US, and the tasks (which are rewarded 0.9\$) are performed through the Amazon Mechanical Turk platform. Each worker is presented with a product description as provided in the Amazon dataset. Then, the review is shown to the worker who is required to judge the review on a 5-level Likert scale (from *Completely Disagree* (-2) to *Completely Agree* (+2)), across 7 quality dimensions (Appendix H) based on previous work on multidimensional quality assessment described in Chapter 7. However, with reviews, it is very hard for the workers to determine the truthfulness of information because they need to assess the authenticity of the review itself, which is often subjective. So, The dimensions are slightly adapted to represent more subjective aspects like reliability. The data obtained allow concluding that the classification performed by the argumentation framework is correlated with the quality of the reviews, mostly with their *Comprehensibility* and *Completeness*. ARI [389] is the readability measure that determines argumentation-based judgments with the highest correlation to quality dimensions. The readability scores alone would not be able to point out the reviews having higher *Overall Truthfulness*, as all readability scores show a very weak correlation with *Comprehensibility* assessments. This result allows supporting the argumentation-based approach and the need for logical reasoning to be performed on top of the ranked arguments to obtain labeling that correlates with *Overall Truthfulness* and *Comprehensibility*.

To summarize, the crowdsourcing task and the multidimensional scale originally used to provide truthfulness assessment allow confirming the ability of argumentation graphs of providing useful explanations about product reviews. The argumentation-based framework proposes represents a first step towards a reliable and transparent assessment of the quality of online opinions.

11.3 Limitations

There are limitations that can be pointed out concerning the experiments and the approaches employed in this thesis. Initially, Section 11.3.1 addresses those limitations that are related to the assessment of online (mis)information. Then, Section 11.3.2 outlines the issues related to the characterization of cognitive biases proposed and the study of their impact on the fact-checking activity that should be acknowledged. Lastly, there are no particular limitations concerning the model proposed for predicting the truthfulness of information items and generating explanations that should be pointed out.

11.3.1 MRQ1: Information Truthfulness Judgment

According to publicly available information about the PolitiFact assessment process [251, 114, 313] (Section 1.2), only the final label assigned by PolitiFact experts to a statement is considered. Each statement is judged by three editors and a reporter, who come up with a consensus for the final judgment. Having such information would allow, for example, a more detailed comparison of the disagreement between workers and the ground truth. However, such information is not publicly released, at least for the PolitiFact dataset. Part of the studies performed (e.g. those described in Chapter 5, Chapter 7 and Chapter 9) employ only statements sampled from the PolitiFact dataset. To generalize the findings, a comparison with multiple datasets is needed. Furthermore, having a single statement evaluated by 5 or 10 distinct workers only does not guarantee a strong statistical power, and thus an experiment with a larger worker sample might be required to draw definitive conclusions and reduce the standard deviation of the findings. The longitudinal study performed to study the effect of information recency shows a relatively low amount of returning workers. Using a different platform such as Prolific might help provide a better worker retention rate.

No ground truth for each of the seven dimensions of truthfulness (Section 7.2.1) exists. Having such information, a more direct comparison between the expert and worker annotations could be performed. However, the such matter would constitute a different research project, complementary to the work described in this thesis, and not free from obstacles as information quality dimensions are more (e.g., Comprehensibility) or less (e.g., Precision) subjective. Furthermore, comparisons with an expert-provided ground truth might even be misleading since differences in the single dimensions may be due to subjectivity. As reported in Section 7.2.2, workers are asked to provide answers using the five-levels Likert scale Completely Disagree (-2), Disagree (-1), Neither Agree Nor Disagree (0), Agree (+1), Completely Agree (+2). The adopted scale is not nominal, since the considered categories are ordered. Furthermore, the considered scale does not represent a mere ranking, since the categories have a clear semantic meaning. Also, the perceived distances between the considered categories might be not consistent for all the workers (e.g., the perceived distance between Completely Disagree (-2) and Neither Agree Nor Disagree (0) might not be double as the perceived distance between Completely Disagree (-2) and Disagree (-1)). On these bases, using the mean as an aggregation function might indeed be incorrect, since it assumes equidistant categories, and that might not be true for all the workers. Nevertheless, it is assumed that the adjacent categories are perceived as equidistant, also due to the labels used which also include a numeric value. If this assumption is correct, using the mean as an aggregation function should not introduce any error. Furthermore, the alternatives are not free from limitations. Both the median and the majority vote aggregation functions would discard possibly useful and significant signals and information. For a further discussion on label aggregation using the mean function when crowdsourcing truthfulness, see Section 4.2.2.

A further limitation of the work described in Chapter 7 concerns the combination of multiple dimensions to improve agreement between workers and expert labels. As stated in Section 7.4.2, an attempt to combine the individual dimensions in a way that improves agreement between the crowd and expert judgments have been made. Further analyses

have been performed, without finding any increase in either the internal agreement among assessors or the external agreement between the crowd workers and experts. Furthermore, the behavior of malicious workers could be also considered a limitation. A set of quality checks to ensure the high quality of the collected data has been implemented in various tasks, as discussed for example in Section 7.2.2. The workers which do not pass the quality checks (i.e., those who are malicious or non-diligent) are not allowed to submit the task, while it can be ensured that workers which passed the quality checks performed the task in good faith. To verify this, analyses of the collected data (distribution of answers, time spent, behavioral analysis, etc.) are usually running and they do not lead to any worker with suspicious or outlier behavior. Thus, it can be assumed that the workers which submitted the tasks are of high quality. Furthermore, it must be remarked that the abandonment rate monotonically decreases as the worker continues to go through a given task. Most workers abandon their task right after reading the instructions, followed by those performing one judgment, etc. This is evidence showing that a worker prefers to abandon the task if they find it not appealing, rather than after having attempted to do it maliciously. Such results are aligned in various crowdsourcing tasks conducted (Section 4.3.2, Section 5.3.2, and Section 7.3.2). Summarizing, there is no reason to suspect any anomalous pattern in the collected data, but this aspect should be further considered and analyzed. Also, a possible limitation could consist of the number and sample of statements chosen for performing the experiments. However, the statements sampled have been inspected without leading to any visible bias or difference with respect to the language level, terminology, length, etc.

11.3.2 MRQ2: Cognitive Biases

The approach proposed in Chapter 8 provides valuable insights into the set of cognitive biases affecting human assessors during the fact-checking process, yet there are several limitations that should be acknowledged. First, bias identification and classification are subjective. The identification and classification of the 220 cognitive biases are based on the assessors' interpretation and understanding of the single biases, which introduces subjectivity. Different researchers and practitioners might identify and categorize these biases differently, which could lead to alternative interpretations of the results and thus to an alternative set of selected biases. Furthermore, although the PRISMA-inspired methodology described in Section 8.2 aims to provide a comprehensive list of cognitive biases affecting fact-checking, it is possible that some biases were overlooked or not considered due to the vast amount of literature on cognitive biases. Thus, the list proposed may not be exhaustive or complete. Then, it should be noted that the research focuses on the cognitive biases affecting human assessors during the fact-checking process. It is important to note that the findings may not be generalizable to all possible fact-checking contexts or human populations, as cognitive biases may manifest differently depending on the specific context and individual differences.

Regarding the set of countermeasures presented, it should be noted that while a list of countermeasures to mitigate the effects of cognitive biases in the fact-checking process is provided, it is difficult to ascertain the extent to which these countermeasures are effective and general. The effectiveness of these countermeasures could be influenced by various factors, such as the specific context, individual differences, and the nature of the misinfor-

mation. It should be also remarked that the approach proposed treats each cognitive bias as an independent factor. However, it is important to acknowledge that biases may interact with each other in complex ways, potentially magnifying or attenuating their effects on the fact-checking process. It should be also noted that Chapter 8 is primarily a literature review, and as such, empirical tests to validate the potential impact of the identified cognitive biases on the fact-checking process involving human subjects have not been conducted. Finally, it should be noted that the cognitive biases identified and discussed are based on the current state of knowledge. As the field of cognitive psychology continues to evolve, new biases may be discovered or existing ones may be refined or redefined. Thus, the findings and analysis reported in Section 8.3 should be periodically updated to reflect the most current understanding of cognitive biases and their potential impact on the fact-checking process.

11.4 Future Directions

The work described in this thesis is a further step in the direction of targeting misinformation in real-time and it is possible to sketch several interesting future research directions. A more complex approach, combining automatic machine learning classifiers, the crowd, and a limited number of experts can lead to better solutions. Section 11.4.1 describes various research lines that can be leveraged in future work to foster online (mis)information assessment, while Section 11.4.2 proposes other experiments concerning the impact of cognitive biases. Then, Section 11.4.3 addresses future work related to predicting the truthfulness of information items and generating explanations. Finally, Section 11.4.4 introduces an ongoing work that will converge in the future with the topics of this thesis.

11.4.1 MRQ1: Information Truthfulness Judgment

The crowdsourcing methodology described in this thesis should be combined with machine learning to assist fact-checking experts in a human-in-the-loop process [99]. This could be done by extending information access tools such as FactCatch [306] or Watch 'n' Check [63]. Eventually, a rating or flagging mechanism to be used in social media could be implemented, in such a way that it allows users to judge the truthfulness of statements. This is a complex task which will require a discussion about ethical aspects such as possible abuses from opposing groups of people as well as dealing with under-represented minorities and non-genuine behaviors derived from outnumbering. Moreover, a thorough study of the perceived distance between multiple truthfulness scales would inform more sophisticated ways for aggregating and merging crowdsourced judgments. The resource created could be also used to better understand how the agreement obtained in crowdsourced judgments can assist experts in evaluating the checkworthiness of statements.

The experiments described by Figure 4.1 involve different pointwise scales to collect information truthfulness judgments (Section 4.2.2). There exists another scale worth studying, which is magnitude estimation [293]. Such a mechanism is a psychophysical scaling technique for the measurement of sensation, where observers assign numbers to stimuli in response to their perceived intensity. It has many applications and it has been used also in other fields such as Information Retrieval [268]. A crowdsourcing experiment con-

cerning information truthfulness where crowd workers are required to judge statements using magnitude estimation could be performed to understand if such a technique provides improvements in worker quality and behavior.

The longitudinal study described in Chapter 5 should be reproduced using statements verified by other fact-checking organizations, e.g., statements indexed by Google Fact Check Explorer² to allow generalizing the findings. The longitudinal study should be also reproduced on other crowdsourcing platforms. Qarout et al. [341] show, for instance, that experiments replicated across different platforms show significantly different data quality levels. Furthermore, 1:1 interviews with the crowd workers should be conducted to understand more profoundly the benefits and the downsides of participating in longitudinal studies. Also, intervention studies should be employed to help test features and experimental settings which will improve worker retention as well as increase both participant and requester satisfaction. Overall, this would lead to a more sound and robust process which will allow both workers and requesters to do fruitful work and conduct longitudinal studies. Another possibility is to perform a longitudinal study using the task design described in Section 7.2.2. Its outcome could be compared with the dataset collected.

The systematic survey about longitudinal studies described in Chapter 8 can be expanded on the findings proposed by conducting 1:1 interviews with crowdsourcing workers and task requesters, as well as intervention studies aimed at testing new features and experimental settings to improve worker retention and increase satisfaction among both participants and requesters. Through these interviews, a deeper understanding of the factors that motivate workers to participate in longitudinal studies on crowdsourcing platforms could be gained. Also, the challenges faced by requesters in designing and implementing effective longitudinal studies should be investigated, thus finding potential solutions to address them. Additionally, the intervention studies will involve testing new features and experimental settings on crowdsourcing platforms to determine their effectiveness in improving worker retention and satisfaction. This will enable the development of a more robust process that supports the successful conduct of longitudinal studies on crowdsourcing platforms, benefiting both workers and requesters.

The seven dimensions of truthfulness introduced in Chapter 7 do not show correlation, with the only exception of Overall Truthfulness and Correctness. This means that such a set of dimensions can be re-used to collect truthfulness labels using crowdsourcing. Conversely, this also means that such a set may not be the optimal one. Therefore, the relationships and correlations between dimensions should be leveraged in the future to characterize and find an optimal and definitive subset of dimensions to be used. It must be also remarked that the approach described favors explainability when compared to systems or data collection approaches based on a single quality dimension. Indeed, the judgment over multiple dimensions could allow understanding of which facet(s) of the statement causes uncertainty and/or disagreement, and thus help to make an informed decision about the final truthfulness label of the statement. Furthermore, expert annotated data should be gathered in the future for all the multiple dimensions of truthfulness, since a ground truth is available for the Overall Truthfulness only. Lastly, more sophisticated techniques should be used in the future to improve the results of combining the dimensions

²<https://toolbox.google.com/factcheck/explorer>

to improve the agreement.

Another interesting future work consists in taking advantage of the geographical data of the crowd workers. Access to the tasks described in this thesis is often (but not always) restricted to US workers. However, tracking or asking for such data poses additional issues to be addressed as this policy may go against the workers' will to not be tracked, even though these data could be leveraged to correlate the workers' quality with the geopolitical situation of their country. Furthermore, the experimental settings allowed to gather of a large amount of data that will likely undergo additional analyses and applications (e.g., confidence values, selected URLs, complex combinations of the dimensions, text justifications, etc.) by the research community. A possible example of these future analyses could be the exploitation of the URLs collected as evidence for each assessment and of the content of these web pages, to build a corpus of documents for each truthfulness level.

Conversational information seeking [94] involves interaction sequences between one or more users and an information system. The possible interactions are primarily based on natural language dialogue, while they may include other types of interactions, such as click, touch, and body gestures. The usage of many conversational agents such as smart speakers, assistants, and chatbots is increasing over time [41]. The COVID-19 pandemic pushed this trend even further. For example, people who are working from home are more likely to ask for updates on news and information [330]. Allowing people to address information truthfulness using such conversational agents could provide several benefits. The possibility to perform such a task by using a smart speaker in a voice-only fashion could keep people more engaged and interested. This is reasonable also when considering chatbots since people would perform the task using an interface to which they are already used. A crowdsourcing experiment concerning information truthfulness could be conducted to understand whether it provides improvements in worker quality, behavior, satisfaction, and engagement. Tools that enable such kinds of crowdsourcing experiments already exist [275, 342] and may be fruitfully used.

People are rarely exposed to one information element from a single source of information at a time [10]. Also, people are able to evaluate diverse information in a short amount of time. The whole set of crowdsourcing experiments performed and described in this thesis is performed in a pointwise fashion, where each worker addresses one information item at a time, sequentially. One may think about allowing workers to address multiple elements at the same time, in a pairwise fashion [67]. Therefore, the crowd workers could be exposed to (at least) two statements to study the impact on worker quality and behavior. This approach could be further extended by asking crowd workers to produce a ranked list of statements. Indeed, machine learning systems usually use pairwise or listwise approaches to learn how to rank a set of items [246]. This means that there is a discrepancy between how crowdsourcing-based approaches and machine learning techniques collect and aggregate labels. Such crowd-powered data could be leveraged by using machine learning-based approaches to learn how to rank information elements when their truthfulness is considered.

Limiting the time available to a worker to express a judgment can increase the quality of the collected data [265]. It is interesting to study whether the same effect occurs while judging the truthfulness of information items. A crowdsourcing experiment where the workers have a constrained time to provide their judgment could be performed. A time-bound assessment, if effective, would allow optimizing the cost of a crowdsourcing task by finding

the sweet spot where the worker is not under pressure but also not allowed to multitask, since they are paid with respect to an estimate of the time required to complete the task on commercial crowdsourcing platforms, as happens for instance on Prolific (Appendix A.1.3).

11.4.2 MRQ2: Cognitive Biases

Modern computational propaganda and social media platforms configurations are characterized by communication techniques that not only misinforms but also exhaust critical thinking, degrading the public's ability to share a system of interpretation of the social reality [434]. It is thus important to characterize the socio-technical features, platform metrics, and algorithmic configurations that affect the content production pipeline, to improve communities' resilience to the degradation of the public sphere. Further (cross-disciplinary) work is needed to better understand how theories studied in social, psycholinguistic, and cognitive science may explain the findings described in this thesis. For instance, the methodology detailed by Sethi et al. [377] could be used to study the emotional aspects of misinformation as perceived by the crowd workers.

The characterization of cognitive biases that can manifest during the fact-checking process described in Chapter 8 sets the basis for future work. Researchers and practitioners can use the set of identified cognitive biases and design ad hoc experiments to further investigate how to manage and mitigate the effects of these biases. For instance, Section 8.3.1 and Table 8.2 show that asking the assessors to revise their judgment (i.e., adopting *C8. Revision*) might help in mitigating *B2. Anchoring Bias*. Thus, researchers and practitioners can set up a between-subject experiment for truthfulness classification where assessors are divided into two disjoint sets. While both sets of assessors are presented with an initial piece of information before each assessment (i.e., the anchor point), the former set of assessors is asked to revise their truthfulness judgment before submitting it, while the latter is not. Researchers can then measure the data gathered for the two sets of annotators and empirically verify the amount of *B2. Anchoring Bias* in the annotations. Practitioners might decide to use the less biased experimental setting to collect the final set of judgments, depending on the outcome of the analysis. Furthermore, future research should consider experimental designs to investigate the actual effects of these biases on assessors' performance and the efficacy of the proposed countermeasures. Also, the possible interactions between cognitive biases and their cumulative impact on the ability of human assessors to accurately evaluate information should be investigated.

Additional experiments should be performed using the multidimensional truthfulness judgments collected. For example, there is evidence that some dimensions are prone to the effect of cognitive biases. Thus, test strategies to correct and de-bias the different truthfulness dimensions should be tested. Implementing such strategies will allow producing a set of non-biased datasets. Such resources can then be used as training data for state-of-the-art deep learning algorithms that target the automatic assessment of misinformation. Then, confidence scores from those algorithms can be derived and compared with the self-reported workers' confidence scores. These de-biasing approaches could be further enhanced by analyzing the different ways disinformation targets subgroups [290].

11.4.3 MRQ3: Predict And Explain Truthfulness

Several researchers propose approaches to study news attributes with the aim of determining whether such news is fake or not [188, 344, 384]. These approaches should be investigated to further automate the fact-checking process using machine learning techniques.

The machine-learning-based approaches described in Chapter 10 include an unsupervised approach to predict truthfulness judgments by computing static word embeddings. Such embeddings are leveraged to identify the semantic similarity between labels. Since such a method might suffer from information loss due to averaging, in the future different and more sophisticated approaches to compute word embeddings should be used. Query terms and justification texts provided by workers of high quality can potentially be leveraged to train a machine-learning model and build a set of fact-checking query terms. The E-BART architecture described in this thesis opens to future work which is made possible because of the unique joint modelling technique used. Future work should, for instance, exploit saliency maps to investigate the effect of evidence on classification performance and vice-versa investigate the relationships among the two models used in our network, the discriminative and the generative one.

11.4.4 Statistical Power In Crowdsourcing

This section is based on the article under review in the “ACM Transactions on Information Systems” journal [357]. Test collections are a mechanism which can be used to reliably measure the effectiveness of Information Retrieval (IR) systems. The most expensive (both in terms of human time and money) step when building a new test collection is the process of collecting the judgments that can be used to perform the measurement. Since this approach does not scale up, researchers proposed to use crowd workers as a valid alternative to classical human assessors and demonstrated the effectiveness of such an approach [11, 229, 267, 279, 358, 359, 369, 459].

A formal study of statistical power and significance in the setting of crowdsourcing experiments received little attention [127], as opposed to what happened for other disciplines like information retrieval. However, some researchers pay attention to statistical power and significance in crowdsourcing experiments, even though a formal study detailing the effects and consequences of the experimental design is missing. Kittur et al. [220] point out the relationship between a good experimental formulation and obtaining adequate results from the crowd. Ribeiro et al. [348] propose a tool to conduct Mean Opinion Score [399] tests to evaluate signal processing methods using crowdsourcing and considered the statistical significance of the crowd sample employed. Behrend et al. [35] compares the viability of using crowdsourcing platforms to recruit participants as opposed to university students when conducting survey data for behavioral research and considers statistical significance. Eickhoff et al. [127] consider statistical significance to increase the robustness of crowdsourcing tasks by identifying malicious workers. Landy et al. [234] compare the research outcome of fifteen research groups on a common subject. They also study how the design choices influence the significance of the results.

The design of the crowdsourcing tasks aimed at collecting relevance judgments to build

a so-called test collection is left to researchers and practitioners, who select and design the annotation task with a focus on the number of documents to be judged. Conversely, the number of workers assigned to judge each document is often set in a heuristic way using a rule of thumb. As a result, the annotated dataset produced by those crowdsourcing tasks does not have any guarantee to satisfy a set of predefined statistical requirements. These requirements include, for example, the guarantee that it is possible to have enough statistical power to distinguish each pair of documents in a statistically significant way. Achieving this would mean that observing that the relevance of a document is higher than the relevance of another one allows concluding that such a difference is real.

A methodology to estimate the number of crowd workers required to produce a test collection that achieves a given statistical power can be proposed. Such a methodology extends prior work based on the t-test and one-way ANOVA and allows the researchers and practitioners to estimate beforehand the number of workers to employ in a crowdsourcing task to obtain as a result an annotated dataset with a minimum statistical power that is guaranteed by design. The methodology is being experimentally evaluated using multiple publicly available datasets. The results show that it can provide a reliable estimation of the number of workers required to distinguish in a statistically significant way document, and of the number of documents required to distinguish in statistically significant way workers.

As it is, the methodology is being studied in the field of information retrieval systems evaluation. However, it is general and can be adopted in other domains. In the misinformation assessment setting performed using crowdsourcing-based approaches, each statement is evaluated by multiple human workers to filter out noise, other malicious workers and cognitive biases that might affect their judgments. These judgments are aggregated to improve their quality, usually by computing a value using some kind of aggregation function. A lot of comparisons among each statement are made. For this reason, a methodology that can compute the required number of judges (in this case crowd workers) and able to ensure that statement differences are statistically significant would be extremely useful. Such a methodology, in other words, would allow concluding that when saying that “statement X is perceived as more truthful than statement Y”, something statically sound is being stated.

11.5 Conclusions

This thesis addressed the challenging problem of the ever-increasing amount of (mis)information which is spreading online along three main research directions. It demonstrates that non-expert human judges can objectively identify and categorize misinformation using crowdsourcing-based approaches. Then, it proposes a characterization of the cognitive biases that may manifest during the fact-checking process, also investigating their effect by conducting an additional crowdsourcing experiment. Lastly, it proposes a machine-learning-based model to predict the truthfulness of an information item and generate an explanation to explain such a prediction. The whole set of data collected and analyzed is publicly released to the research community at: <https://doi.org/10.17605/OSF.IO/JR6VC> [392].

A collaborative process between non-expert human judges, expert fact-checkers and automatic fact-checking models would provide a scalable and decentralized hybrid mecha-

nism to cope with the increasing volume of online misinformation, while the characterization and study of cognitive biases that might manifest while performing the fact-checking activity can serve as a reference to build a more sound, robust, more aware, and bias-free pipeline to effectively crowdsource reliable truthfulness judgments at scale. Also, predicting truthfulness judgments and generating explanations at the same time allow systems built atop automatic fact-checking model to be more transparent, and fosters greater trust in them.

There is still much work to do before achieving the long-term goal of building a system to directly judge the truthfulness of statements as they appear on some social media using crowdsourcing, as proposed by Demartini et al. [99]. However, the work described in this thesis represents indeed a step towards the design and development of systems to overcome the spreading of online misinformation that is robust, trustworthy, explainable, and transparent. In other words, systems which are aligned with the key principles that fact-checking organizations must follow.³

11.6 Acknowledgments

The work described in Chapter 4, Chapter 5, and Chapter 7 has been supported by a Facebook Research award⁴ and by two Australian Research Council Discovery Project DP190102141⁵ and DE200100064⁶). Also, the MISTI “MIT International Science and Technology Initiatives - Seed Fund” (MIT-FVG Seed Fund Project) and the project HEaD “Higher Education and Development - 1619942002 / 1420AFPLO1” (Region Friuli-Venezia Giulia) provided support. I also thank Devi Mallal from RMIT ABC Fact Check for facilitating access to their dataset.

Damiano Spina is the recipient of an Australian Research Council (ARC) DECRA Research Fellowship, and Associate Investigator of the ARC Centre of Excellence for Automated Decision-Making and Society CE200100005,⁷ and a research collaborator of RMIT FactLab. Gianluca Demartini is a Chief Investigator of the ARC Training Centre for Information Resilience IC200100022.⁸ They both contributed to the characterization described in Chapter 8. The work described in Chapter 9 and Section 11.2.4 is partially supported by The Credibility Coalition.⁹ Lastly, the work described in Section 11.2.4 is partially supported also by the project “Departments of Excellence 2018-2022” awarded to the Department of Philosophy “Piero Martinetti” of the University of Milan and by the PRIN2020 Grant 2020SSKZ7R of the Italian Ministry of University and Research (MIUR). Any opinions, findings, and conclusions expressed in the chapters of this thesis are from the researchers that participated in each experiment and do not necessarily reflect those of the sponsors.

³<https://www.ifcncodeofprinciples.poynter.org/>

⁴<https://research.facebook.com/research-awards/the-online-safety-benchmark-request-for-proposals/>

⁵<http://purl.org/au-research/grants/arc/DP190102141>

⁶<http://purl.org/au-research/grants/arc/DE200100064>

⁷<https://purl.org/au-research/grants/arc/CE200100005>

⁸<https://purl.org/au-research/grants/arc/IC200100022>

⁹<https://credibilitycoalition.org/>

Crowd_Frame: Design and Deploy Crowdsourcing Tasks

This appendix on the article published at the fifteenth ACM International Conference on Web Search and Data Mining [393]. It describes in detail Crowd_Frame [393], a software system implemented to support the whole crowdsourcing activity workflow. Section A.1 describes the design and deployment workflow of a crowdsourcing task on three different platforms. Section A.2 motivates the need for such an implementation. Section A.3 analyzes in detail the overall architecture of each software component. Section A.3.2.3 presents a case study that uses Crowd_Frame. Section A.7 discusses the state of the implementation. Section A.8 sketches possible development directions.

A.1 Crowdsourcing Platforms

Understanding in more detail how the crowdsourcing platforms support the task design and deploy workflow, which may be challenging, is useful to grasp the improvements provided by the usage of Crowd_Frame. Three crowdsourcing platforms are described in the following, namely Amazon Mechanical Turk (Section A.1.1), Toloka (Section A.1.2) and Prolific (Section A.1.3).

A.1.1 Amazon Mechanical Turk

The task design workflow for a requester who uses Amazon Mechanical involves three phases. Initially, the requester must choose the type of task to publish (Figure A.1). The platform provides several templates ready to be customized. The requester can also choose using a blank template.

Then, the requester starts designing the overall task. Initially, they inputs name and description and sets five configuration parameters (Figure A.2). The parameters are the number of workers to recruit, the time allowed to perform the task, the amount of the reward in USD\$, the expiration date of the task, and the auto-approve threshold in days for

Select a customizable template to start a new project

The screenshot shows the Amazon Mechanical Turk task creation interface. On the left, there is a sidebar with a list of task types under the 'Survey' and 'Vision' categories. The main area displays a preview of a task. At the top, there is a 'View instructions' button. Below it, the task instructions are: 'What is your favorite color for a bird?'. There is a text input field with the example text 'example: pink'. Below that, there is a checkbox labeled 'Check this box if you like birds' which is checked. Below the checkbox, there is a scale input labeled 'On a scale of 1-10, how much do you like birds?'. Below the scale, there is a text area labeled 'Write a short essay describing your favorite bird' with the placeholder text 'Lorem ipsum...'. At the bottom right, there is a 'Create Project' button.

Figure A.1: Type selection for a task on Amazon Mechanical Turk.

the workers' submissions. The requester can also set various criteria to filter the workers to recruit. Amazon Mechanical Turk uses the word "Qualification Type" to indicate such criterion. The qualification types include the country of provenance of the worker, the age or household income, etc. It is also possible to define custom qualification types.¹ The requester designs the interface of the task after setting the parameters and criteria. They must use a set of custom markup tags defined as Crowd Elements.² Each task must contain the tag `<crowd-form>`, which is thus the most important element of the codebase. Listing A.1 shows the code of a sample task designed using the Crowd Elements.

The requester can preview each HIT of the task when the design is finalized. The task can thus be published an arbitrary amount of times. Amazon Mechanical Turk uses the word "batch" to indicate a single set of workers recruited within a given crowdsourcing task. The requester can recruit multiple batches of workers, at the same time. Recruiting a batch of workers involves providing a special CSV file to set the input and output data. The platform considers each column of such a CSV file as data. Each row of the file is assigned to a single worker. The values of the row assigned are used to initialize each input or output. The requester must thus provide a file with n rows to recruit n workers. Each worker accepts the HIT initialized using such a mechanism and completes the task. The requester approves or denies the payment. The final data are provided by the platform through a second CSV file when each worker of the batch completed the HIT.

¹<https://docs.aws.amazon.com/AWSMechTurk/latest/AWSMechanicalTurkRequester/WorkWithCustomQualType.html>

²https://docs.aws.amazon.com/AWSMechTurk/latest/AWSMturkAPI/ApiReference_HTMLCustomElementsArticle.html

Edit Project

1 Enter Properties 2 Design Layout 3 Preview and Finish

Project Name: This name is not displayed to Workers.

Describe your survey to Workers

Title:
Describe the survey to Workers. Be as specific as possible, e.g. "answer a survey about movies", instead of "short survey", so Workers know what to expect.

Description:
Give more detail about this survey. This gives Workers a bit more information before they decide to view your survey.

Keywords:
Provide keywords that will help Workers search for your tasks.

Setting up your survey

Reward per response: \$
This is how much a Worker will be paid for completing your survey. Consider how long it will take a Worker to complete your survey.

Number of respondents:
How many unique Workers do you want to complete your survey?

Time allotted per Worker:
Maximum time a Worker has to complete the survey. Be generous so that Workers are not rushed.

Survey expires in:
Maximum time your survey will be available to Workers on Mechanical Turk.

Auto-approve and pay Workers in:
This is the amount of time you have to reject a Worker's assignment after they submit the assignment.

Figure A.2: Parameters configuration for a task on Amazon Mechanical Turk.

```

1 <script src="https://assets.crowd.aws/crowd-html-elements.js"></script>
2 <crowd-form answer-format="flatten-objects">
3   <crowd-instructions link-text="View instructions" link-type="button">
4     <short-summary>
5       <p>Provide a brief instruction here</p>
6     </short-summary>
7     <detailed-instructions>
8       <h3>Provide more detailed instructions here</h3>
9       <p>Include additional information</p>
10    </detailed-instructions>
11    <positive-example>
12      <p>Provide an example of a good answer here</p>
13      <p>Explain why it's a good answer</p>
14    </positive-example>
15    <negative-example>
16      <p>Provide an example of a bad answer here</p>
17      <p>Explain why it's a bad answer</p>
18    </negative-example>
19  </crowd-instructions>
20  <div>
21    <p>What is your favorite color for a bird?</p>
22    <crowd-input name="favoriteColor" placeholder="example: pink"
23      → required>
24    </crowd-input>
25  </div>
26  <div>
27    <p>Check this box if you like birds</p>
28    <crowd-checkbox name="likeBirds" checked="true" required>
29    </crowd-checkbox>
30  </div>
31  <div>
32    <p>On a scale of 1-10, how much do you like birds?</p>
33    <crowd-slider name="howMuch" min="1" max="10" step="1" required>
34    </crowd-slider>
35  </div>
</crowd-form>

```

Listing A.1: Interface of a sample task on Amazon Mechanical Turk built using Crowd Elements.

A.1.2 Toloka

The task design workflow for a requester who uses Toloka involves three phases. Initially, the requester chooses the type of task to be performed (Figure A.3) and inputs name and description. Toloka uses the word “Project” to indicate a crowdsourcing task.

Then, the requester designs the task interface. They can either use standard HTML

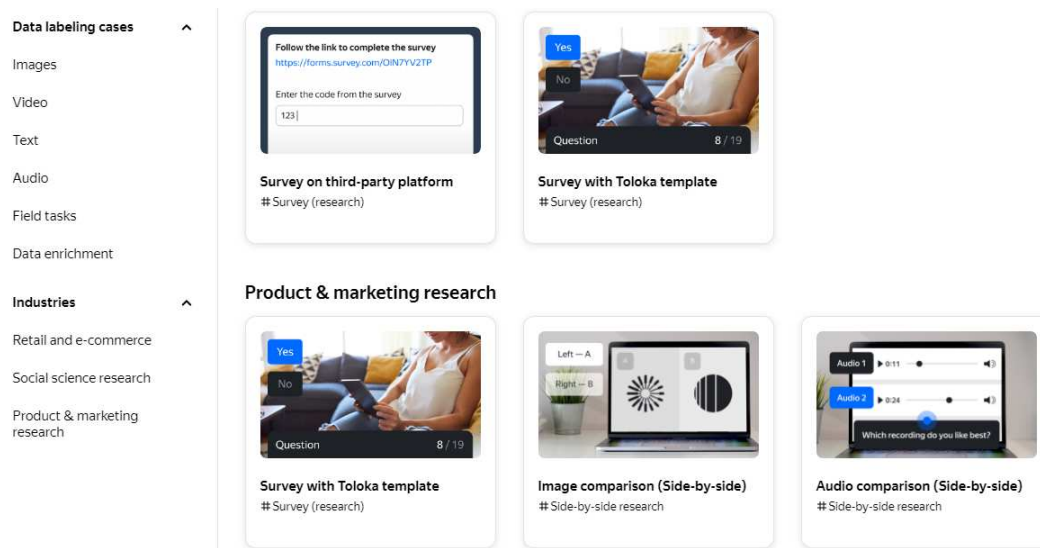


Figure A.3: Task type selection for a project on Toloka .

markup tags together with JavaScript code and CSS styling or Toloka 's template builder³ (Figure A.4). The template builder uses a set of predefined JSON objects used to initialize the overall interface. The input and output data specification⁴ for the task must be declared in a separate section of the user interface (Figure A.5). The data can be specified either manually using special JSON objects or using the user interface itself. The input and output data specified are then referenced while building the task interface. Listing A.2 show the JSON specification of the sample image classification task interface shown in Figure A.4. Listing A.3 and Listing A.4 show the JSON specification of the corresponding input and output data. The requester then writes the instructions that the workers will read when accepting a HIT. Toloka uses either the words "Toloker" or "User" to identify workers.

The requester can define multiple pools of workers to recruit for a given task when its design is finalized. Toloka uses the word "Pool" to indicate groups of workers who share a predefined set of attributes. In more detail, the requester specifies for each pool its name and description along with the set of attributes. These attributes may include the language spoken, the world region of provenance, the operating system used, and many others. Then, the requester specifies a speed/quality balance percentage. In other words, the percentage of top-rated workers who can access the pool. Requiring more quality workers means a slower pool completion time. The requester can also specify a set of quality control mechanisms and rules. They can also require overlap for each HIT published within the pool. Lastly, the requester sets the reward for completing each HIT. Figure A.6 shows part of the Toloka's pool configuration interface. Each pool can be started, paused and stopped independently. Multiple pools can be active at the same time.

The requester must provide values for the input and output data defined during the design phase for at least one pool to start publishing the task designed on Toloka and

³<https://toloka.ai/docs/template-builder/index.html>

⁴<https://toloka.ai/docs/guide/concepts/incoming.html>

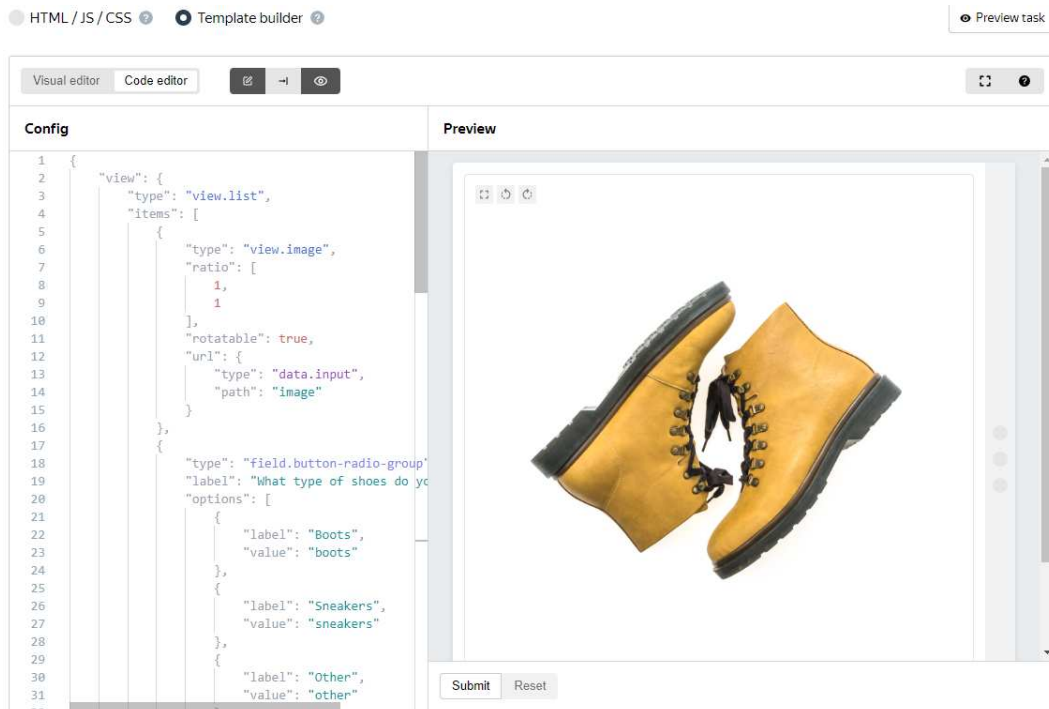


Figure A.4: Interface design for a project on Toloka using its template builder.

The screenshot shows the Toloka input and output data specification interface. It is divided into two main sections: "Input data" on the left and "Output data" on the right. The "Input data" section has a "Title" field with the value "image", a "Type" dropdown menu set to "URL", a checked "Required" checkbox, an unchecked "Array" checkbox, and a "Delete" button. A "Save" button is located at the bottom right of this section. The "Output data" section has a "Title" field with the value "result", a "Type" dropdown menu set to "string", an "Allowed values" section with two input fields containing "boots" and "other", a "Not required" button, a "Min length" field with "Not required", a "Max length" field with "Not required", a checked "Required" checkbox, an unchecked "Array" checkbox, and an "Add field" button. An "Add field" button is also present at the bottom left of the "Output data" section.

Figure A.5: Input and output data specification for a project on Toloka.

```
1  {
2    "view": {
3      "type": "view.list",
4      "items": [
5        {
6          "type": "view.image",
7          "ratio": [1,1],
8          "rotatable": true,
9          "url": { "type": "data.input", "path": "image" }
10       },
11       {
12         "type": "field.button-radio-group",
13         "label": "What type of shoes do you see?",
14         "options": [
15           { "label": "Boots", "value": "boots"},
16           ...
17         ],
18         "data": { "type": "data.output", "path": "result"},
19         "validation": { "type": "condition.required" }
20       }
21     ]
22   },
23   "plugins": [
24     {
25       "type": "plugin.toloka",
26       "layout": { "kind": "scroll", "taskWidth": 600 }
27     },
28     {
29       "1": {
30         "type": "action.set",
31         "data": { "type": "data.output", "path": "result" },
32         "payload": "boots"
33       },
34       "2": {
35         ...
36       },
37       "type": "plugin.hotkeys"
38     }
39   ], "vars": {}
40 }
```

Listing A.2: Image classification JSON configuration for the interface of a project on Toloka built using the template builder.

thus recruiting workers. The mechanism is similar to the one of Amazon Mechanical Turk, described in Section A.1.1. A special file containing the values for the input and output data must be provided. The file can comply with the XLSX, TSV or JSON formats. Each column

```

1 | {
2 |   "image": {
3 |     "type": "url",
4 |     "required": true
5 |   }
6 | }

```

Listing A.3: Input data specification for a project on Toloka.

```

1 | {
2 |   "result": {
3 |     "type": "string",
4 |     "allowed_values": [ "boots", "other" ],
5 |     "required": true
6 |   }
7 | }

```

Listing A.4: Output data specification for a project on Toloka.

(or attribute, if the file is in JSON format) is labelled with the prefixes `INPUT:` or `OUTPUT:` and then matched with the corresponding input or output data defined by the requester. Each row of the file is assigned to a worker and its values are used to initialize each data. In other words, the requester must provide a file with n rows to recruit n workers. Listing A.5 show the content of a sample XLSX file used to recruit two workers within a pool for a task having a single input data named `image`.

Each worker accepts the HIT assigned and completes the task. The requester can either approve or deny the reward for each HIT submitted. The final data are provided by Toloka through a second TSV file at the end of each pool of tasks.

```

1 | INPUT:image
2 | https://labs-images-testing.s3.yandex.net/presets/for%20tb%20and%20datas_
   | ↪ et/leather-boots.jpg
3 | https://labs-images-testing.s3.yandex.net/presets/for%20tb%20and%20datas_
   | ↪ et/pair-trainers.jpg

```

Listing A.5: Input data initialization for two HITs within a pool of a project on Toloka.

A.1.3 Prolific

The task deployment workflow for a requester that uses the Prolific platform involves four phases. Initially, the requester sets five parameters concerning the task (Figure A.7). Prolific uses the word “Study” to indicate a task. The preliminary parameters are the study

Audience

We added default quality control rules to this pool to enhance annotation quality. You can change them in quality control settings

General Information

Use filters to select performers who will get access to your tasks. Specify the languages or region by phone number. Otherwise, the tasks will be available to performers from all over the world. You can [copy](#) audience and quality control filter settings from another pool. [Learn more](#)

My tasks may contain shocking or pornographic content. [Learn more](#)

Languages = Value

+ Add filter + Add skill

Speed/quality balance
Note that fewer users means slower pool completion
[Learn more](#)

Top % Online

Specify the percentage of top-rated Tolokers who can access tasks in the pool

Speed 100% 90% 80% 70% 60% 50% 40% 30% 20% 10% Quality

90% of the best Tolokers were selected
The task is available to 0 active Tolokers

Price
for Image classification

Price per task suite, \$* 0.02
Set at least \$0.02 for simple tasks
Set at least \$0.05 for complex tasks

Toloker interest at this price Medium

Recommended number of tasks per suite 10
For Image classification

Overlap* 3
For simple tasks recommended overlap is 3

Price per 1 task \$0.008
Including 30% fee

Create pool

Figure A.6: Worker attributes configuration for a pool of a project on Toloka.

name and description and the devices that crowd workers can use for participating in the task. Prolific uses the word “Participant” to indicate a worker. The requester can also indicate whether the study needs to use the workers’ microphone, camera, or audio and if the workers must download additional software to perform the task.

The second phase of the task deployment workflow on Prolific requires the requester to specify the data collection modality (Figure A.8). The most evident difference when comparing Prolific with Amazon Mechanical Turk (Section A.1.1) and Toloka (Section A.1.2) is that Prolific does not offer any modality to design the task in-house. The task requester must rely on external tools to design and deploy the task and provide its URL to Prolific. There are two approaches to provide an anonymous identifier to the task deployed externally to detect each worker. The former consists of simply asking workers explicitly for their Prolific identifiers. The latter involves sending automatically the identifier of each worker by appending it to the external URL provided. The requester can send along also identifiers for the current task and session using the latter modality. Prolific uses the word “Session” to indicate the current batch of workers recruited. Furthermore, the requester can choose between two approaches to allow workers to confirm task completion from the software deployed externally. The former consists of embedding an URL in the software’s user interface (e.g., in a button) to redirect the worker to Prolific. The latter consists of providing a completion code that the worker copies and pastes manually on Prolific’s interface.

The third phase involves configuring the audience required for the task. Prolific uses the word “audience” to indicate the workers to recruit for the current task. The requester can indicate how many workers must be recruited and from which country. The requester

STUDY DETAILS

What is the title of your study?

Give your study an internal name (only visible to you)

Describe what participants will be doing in this study. [Read our tips](#)

H₁ H₂ B / U

In this study I will ask you to tell me your favourite ice cream and then ask you how you are feeling.

Which devices can participants use to take your study?
 Mobile Tablet Desktop

Does your study require any of the following?
 Audio Camera Microphone Download software

i The devices and tool options will be displayed to participants on their study preview. These options don't screen participants. To screen participants use the "Prescreen participants" option in the [Audience](#).
[Read about device compatibility](#)

Figure A.7: General parameters set up for a study on Prolific.

can choose between three sampling modalities to recruit the workers. Furthermore, various criteria can be used to further filter the audience of workers.


The fourth phase of the task deployment workflow on Prolific requires the requester to specify the time needed to complete the task and the final reward in pounds. The platform thus computes the hourly pay rate and acknowledges the requester whether the amount is sufficient or not. The platform does not prevent the requester from publishing the task if the amount is insufficient. However, the platform could stop the task from continuing if the true median completion time exceeds the completion time estimated and suggest the requester provide additional payments. The requester can publish the task and wait for completion once its deploy is finalized. The status of each HIT can be monitored in real-time and the payment can be approved or denied.


A.1.4 Discussion

The design and deployment workflow of a crowdsourcing task is often cumbersome and counter-intuitive, as described in Section 1.5. Let us consider Amazon Mechanical Turk. The task interface code must be written in a single box and is a combination of a custom subset of HTML tags and CSS and JavaScript statements. The business and presentation logic are mixed. Hidden form fields with JSON objects are used as values to store data. They must build a file for each batch of the task. La Barbera et al. [232] deploy a misinformation assessment task directly on such a platform. They crowdsource judgments for 120 political-related statements judged by 10 distinct crowd workers. They deploy 400 different HITs. The task is created using a paginated structure using the custom markup tags provided by Amazon Mechanical Turk. They had to implement custom Javascript code to show/hide elements of the user interface since there must be only a single `<crowd-form>`.

DATA COLLECTION

New How do you want to collect your data?

 **External study link**
Provide your own URL

 **Survey builder (beta)**
Create surveys with up to five questions on Prolific

How to record Prolific IDs

To link answers in your survey tool to participants in Prolific, you'll need to set up your survey tool to record our participants' unique Prolific IDs. This enables you to match our participant [demographic data](#) with their answers. If you receive a poor quality submission, you can also [reject it in our platform](#).

What is the URL of your study?

How do you want to record Prolific IDs? *(Select an option below for instructions)*

I'll add a question in my study
 I'll use URL parameters
 I don't need to record these

To link answers in your survey tool to participants in Prolific, **you'll need to set up your survey tool** to record our participants' unique Prolific IDs. Check out our [integration guide](#) instructions for the most commonly used survey tools.

Prolific ID
 Study ID
 Session ID
[Configure parameters](#)

How to confirm participants have completed your study

When participants start your study they will leave the Prolific app. When they return, we need to capture a unique Completion Code to prove they completed your study.

[Read more about study completion](#)

How do you want to confirm participants have completed your study? *(Select an option below for instructions)*

I'll redirect them using a URL
 I'll give them the Completion Code to copy & paste

Completion code
Add Completion URL as a last step in the survey tool to let Prolific know the study was completed by the participants.

Submissions will be:

URL to redirect: [Copy](#)

Figure A.8: Data collection set up for a study on Prolific.

AUDIENCE

Recruit participants

How many participants are you looking to recruit?

Location

Where should your participants be located?

All countries available

USA

UK

More

Study distribution

How do you want to distribute your sample?

New Representative sample

Distribute your study based on UK or USA census data.

Balanced sample

Distribute your study evenly to male and female participants.

Standard sample

Distribute your study to available participants on Prolific.

Selected

Prescreen participants

YOUR CRITERIA

[+ Add screener](#)

We've found **121,868** matching participants who have been active in the past 90 days

Figure A.9: Audience configuration for a study on Prolific.

STUDY COST

How long will your study take to complete? Max. time: 13 mins

Participants are paid according to your estimated study completion time. If the median completion time exceeds your estimate we will ask you to make additional payments. [Read more about study completion time](#)

How much do you want to pay them?

9.00/hr

Hourly rate

£6.00 £9.00 Good £12.00+

Figure A.10: Study cost configuration for a study on Prolific.

Toloka is affected by most of the difficulties that a requester faces when using Amazon Mechanical Turk. The platform further complicates the overall picture. The quality control rules and criteria used to initialize pools are non-trivial to set up. The allocation mechanisms of the HITs can be confusing because the platform groups them in overlapping task suites. On the other hand, the pool-based mechanism to recruit workers allows more fine-grained control over the workers to be recruited. Prolific is the platform that offers the most streamlined and easy workflow, but the absence of any way to deploy a task without relying on external tools can be a daunting challenge for many task requesters.

A.2 Aims

Several tools exist to help task requesters while employing crowdsourcing based approaches (Section 2.8). However, none of these tools allows for coping with the difficulties of the workflow detailed in Section A.1. A solution to these problems is to rely on crowdsourcing platforms only to recruit the workforce needed. The recruited workers reach the task deployed using external software in an external platform to perform it. Then, they return to the crowdsourcing platform of provenance to receive the reward. The task design and deployment processes can be handled by external software.

Soprano et al. [393] developed `Crowd_Frame`. It is a software system that allows to easily design and deploy diverse types of crowdsourcing tasks regardless of the chosen platform. The tasks can be composed of different sets of HITs. Each task is deployed in a customizable and controllable environment. The software is freely available and downloadable⁵ by the research community and it has already been used by researchers and practitioners to deploy several tasks [47, 48, 61, 111, 361, 362, 394, 395].

A.3 System Design

Requesters can use `Crowd_Frame` to instantiate and configure tasks using a simple user interface. They deploy the task into a controllable and customizable environment when the configuration is finalized. A wrapper is used by the requester as a bridge between the crowd workers recruited and the task deployed.

A.3.1 General Architecture

`Crowd_Frame` is a client-side application developed using Angular,⁶ an open source framework for web development. Figure A.11 shows its architecture. The software is composed of four main components namely `Generator`, `Skeleton`, `Search Engine`, and `Logger`. It relies on Amazon S3⁷ to store tasks configuration and source files. It is an object storage service built to store and retrieve any type and amount of data. The software also uses Amazon DynamoDB⁸ to store the data produced by each worker while performing the

⁵https://github.com/Miccighe1/Crowd_Frame

⁶<https://angular.io/>

⁷<https://aws.amazon.com/it/s3/>

⁸<https://aws.amazon.com/it/dynamodb/>

task. It is a fully managed NoSQL key-value and document database service. It is serverless and can scale autonomously according to the workload.

A requester uses the Generator to configure the task. The configuration is uploaded to a private S3 bucket (i.e., a storage resource). Then, the requester can publish a set of HITs on the chosen crowdsourcing platform. Each crowd worker interacts with the wrapper. The wrapper redirects the worker to the application deployed on a public S3 bucket. The worker interacts with the Skeleton to perform the task. The Skeleton may embed a custom and configurable Search Engine. The Logger analyses the worker's behavior during the task. The data produced is stored on a DynamoDB table. Each component uses also external services to function.

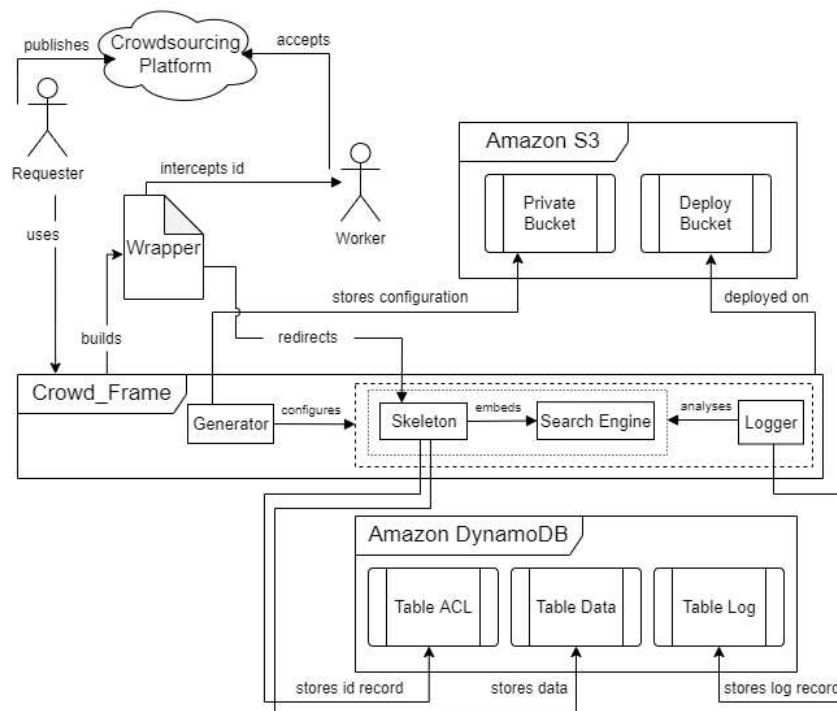


Figure A.11: General architecture of Crowd_Frame.

A.3.2 Generator

The Generator component allows requesters to design and customize the configuration of a crowdsourcing task, as shown by Figure A.11.

A.3.2.1 Use Cases

The diagram shown in Figure A.12 provides a high-level description of the interaction between a requester and the Generator component of Crowd_Frame. The requester authenticates him/herself to unlock the usage of the component. Then, they designs or customizes

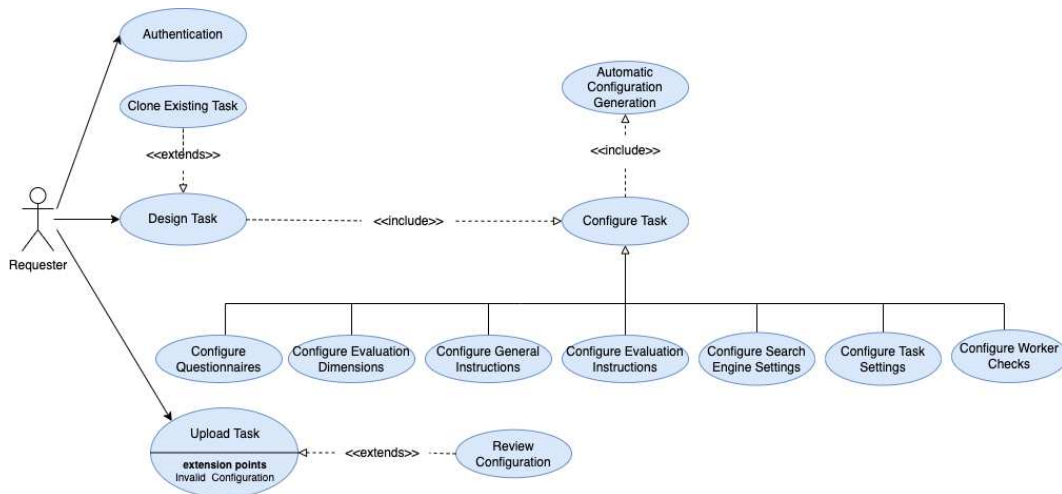


Figure A.12: Use case diagram for a requester that designs and deploy a task using Crowd_Frame.

the crowdsourcing task. The component provides an option to clone the configuration of a task previously deployed. The configuration phase involves seven different steps, described in detail in Section A.3.2.2. The overall configuration is automatically generated and updated during each step. The requester finalizes the upload of the configuration when satisfied. They may need to review errors that invalidate the configuration to be able to upload it.

A.3.2.2 Architecture

The implementation of the Generator component is a form composed of eight separate steps. A requester configures a certain aspect of the task within each step. Table A.1 provides a brief description of the role of each configuration step. The first step concerns questionnaires. It allows the creation of one or more questionnaires that workers will fill out before or after task execution, depending on the requester's needs. Three different types of questionnaires are available. The Standard option allows implementing simple questionnaires where the worker responds to questions that may require either a predefined and limited number of alternative answers, a text-based answer or a numerical value. The CRT option allows implementing Cognitive Reflection Tests [148], often used to estimate workers' cognitive abilities. The Likert option allows implementing questionnaires where workers choose from a range of possible responses to a specific question [250]. It is possible to easily extend the Generator to add additional types of questionnaires.

The second step concerns the evaluation dimensions. It allows configuring, in other words, the aspects the worker evaluates for each element of the HIT assigned. A dimension may require expressing a judgment using some sort of scale. Every possible type of rating scale [5] is available in Crowd_Frame. A requester can thus configure categorical, interval, and ratio scales such as magnitude estimation [292]. Other features can be activated within a given evaluation dimension. For instance, the requester may require the worker to provide

a textual justification or to search for an URL using the custom search engine integrated.

The third step concerns task instructions. The requester writes the set of instructions using a rich textual editor. The worker is shown such instructions before the task. Similarly, the fourth step concerns evaluation instructions. The worker is shown with such instructions within the task's body while evaluating each element of the HIT assigned. Such instructions explain how they should assess each element along each dimension. The fifth step concerns the Search Engine. It allows choosing the search provider wanted and adding a list of domains to filter from search results.

The sixth step concerns general task settings. The requester can set the maximum amount of tries for each worker. This allows workers to revise their work if they do not pass the quality checks at the end of the task and helps in reducing task abandonment [172, 174]. The requester can also configure the minimum time that each worker must spend on each HIT element. Such a check is useful to prevent automated bots from completing a task. The requester can set a countdown to limit the time available for the worker to complete the assessment. Maddalena et al. [265] show that such a constraint leads to an improvement in worker quality. It also allows for optimizing the cost of a crowdsourcing task. The requester can enable an annotation interface to require the worker to annotate and label texts, such as for social media conversation annotation [475]. The requester may choose previous batches of workers to block for the current task. For example, this can be useful when performing a longitudinal study, where the same task is repeated multiple times with new workers only [362]. Lastly, the requester is required to upload the HITs for the task.

The seventh step concerns additional checks on workers. The requester can manually specify a set of worker identifiers allowed or blocked for the current task. It allows finetuning the admitted workers without blocking an entire batch. The eighth and last step shows a summary of the configuration created. The requester can upload the configuration to the private S3 bucket if satisfied.

A.3.2.3 Case Study

Roitero et al. [362] use Crowd_Frame to deploy a misinformation assessment task to understand whether the crowd workers can identify and correctly classify online misinformation concerning the COVID-19 pandemic, as described in Chapter 5. Their task uses two questionnaires to collect workers' backgrounds and to estimate workers' cognitive abilities. Each worker recruited is shown a set of statements. A six-level assessment scale is used to provide truthfulness labels. Each worker must provide a URL using a custom search engine and a textual justification to support the rationale behind each assessment. Figure A.13 shows the resulting interface.

A requester can easily replicate Roitero et al. [362] setup using Crowd_Frame. They use a single standard questionnaire and three CRT questionnaires. They set a single evaluation dimension which is the overall truthfulness. The dimension uses a six-level categorical scale and they implement quality control on the values provided. In more detail, they require evaluating a statement that is obviously true with a higher truthfulness value than a statement that is obviously false. They also filter out search results originating from three fact-checking-related websites. A sample configuration that can be used to replicate Roitero

Table A.1: Summary of each configuration step of the Generator component.

Step #	Name	Description
1	Questionnaires	Allows creating one or more questionnaires that workers will fill before or after task execution.
2	Evaluation Dimensions	Allows configuring what the worker will assess for each element of the HIT assigned.
2	Evaluation Dimensions	Allows configuring what the worker will assess for each element of the HIT assigned.
3	Task Instructions	Allows configuring the instructions shown to each worker before starting the task.
4	Evaluation Instructions	Allows configuring the instructions shown to each worker while assessing each element of the HIT assigned.
5	Search Engine	Allows configuring the search provider wanted. Furthermore, it is possible to add a list of domains to filter from search results.
6	Task Settings	Allows configuring the overall task settings, such as the maximum amount of tries for each worker, the usage of an annotation interface, and more.
7	Worker Checks	Allows configuring additional filters and checks on the workers recruited.
8	Summary	Allows reviewing and uploading the final configuration.

et al. [362] is available on the repository of Crowd_Frame.⁹

A.3.3 Skeleton

The Skeleton component allows workers to perform the task after being recruited, as shown by Figure A.11.

A.3.3.1 Use Cases

The diagram shown in Figure A.14 provides a high-level description of the interaction between a worker and the Skeleton component of Crowd_Frame. The worker accesses the task deployed after being recruited. They are initially forced to read the general instructions of the task before starting to work on the HIT assigned. Then, they finally start working. The task may require filling up one or more questionnaires, at the start or the end. Three types of questionnaires can be configured, as described in Section A.3.2.2. The workers thus evaluate every dimension concerning every element of the HIT. A dimension may require

⁹https://github.com/Miccighel/Crowd_Frame/tree/master/examples/misinformation_assessment

SHOW INSTRUCTIONS

QUESTIONNAIRES
STATEMENTS
END

12345678

Statement: "Republicans DO NOT want to throw doctors" in jail.

Speaker: National Republican Senatorial Committee

Date: 2022-05-03

B - Use the search engine below to search for evidence for the statement. Then, select the most relevant result.

SEARCH

42 results found. Select

[National Republican Senatorial Committee- NRSC](#)
<https://www.nrsc.org/>
 NRSC Chairman Rick Scott: Republican Civil War Has Been Cancelled The Biden Administration's School Reopening Circus #OpenTheSchools When They Win, You Lose Latest Videos Check out the latest video update from NRSC Chairman Rick Scott on the importance of taking back the Senate.

[Republicans Announce Take Back Our Country Tour - NRSC](#)
<https://www.nrsc.org/press-releases/republicans-announce-take-back-our-country-tour-2022-10-17/>
 WASHINGTON, D.C. – Today, Republican National Committee (RNC) Chairwoman Ronna McDaniel, National Republican Senatorial Committee (NRSC) Chairman Senator Rick Scott, and National Republican Congressional Committee (NRCC) Chairman Congressman Tom Emmer announced the Take Back Our Country Tour.

[New Polls Show Republicans Are on the Right Path - NRSC](#)
<https://www.nrsc.org/press-releases/new-polls-show-republicans-are-on-the-right-path-2022-10-24/>
 Washington, D.C. – Two weeks out from Election Day and new polls show Republicans are in a great place to have big wins across the country. Republicans are united and enthusiasm is high this midterm because they know what is at stake at the ballot box. The latest NBC poll showed 78% of Republicans showing a high interest, and only 69% of Democrats.

[National Republican Senatorial Committee - Wikipedia](#)
https://en.wikipedia.org/wiki/National_Republican_Senatorial_Committee
 The National Republican Senate Committee (NRSC) is the Republican Hill committee for the United States Senate, working to elect Republicans to that body. The NRSC was founded in 1916 as the Republican Senatorial Campaign Committee. It was reorganized in 1948 and renamed the National Republican Senatorial Committee. [1]

[Republicans anxious about cash-strapped NRSC amid Scott's feud with...](#)
<https://www.cnn.com/2022/09/06/politics/rick-scott-mitch-mccconnell-republican-senate-fundraising/index.html>
 Sen. Thom Tillis, a Republican of North Carolina, said the spat over the NRSC is "a distraction from what voters are going to be motivated by," maintaining he does have confidence in Scott to...

Items per page: 5 | 1 - 5 of 42

C - Evaluate seven quality dimensions of the statement considering the evidence found.

Truthfulness

Assess the overall truthfulness of the statement.

Lie (The statement is not accurate and makes a ridiculous claim)
 False (The statement is not accurate)
 Mostly False (The statement contains an element of truth but ignores critical facts that would give a different impression.)
 Half True (The statement contains an element of truth but ignores critical facts that would give a different impression.)
 Mostly True (The statement contains an element of truth but ignores critical facts that would give a different impression.)
 True (The statement is accurate and there's nothing significant missing.)

Write your justification here *

This is a justification written by the current worker

BACK
NEXT

Figure A.13: Worker interface of a misinformation assessment task deployed by Roitero et al. [362] using Crowd_Frame.

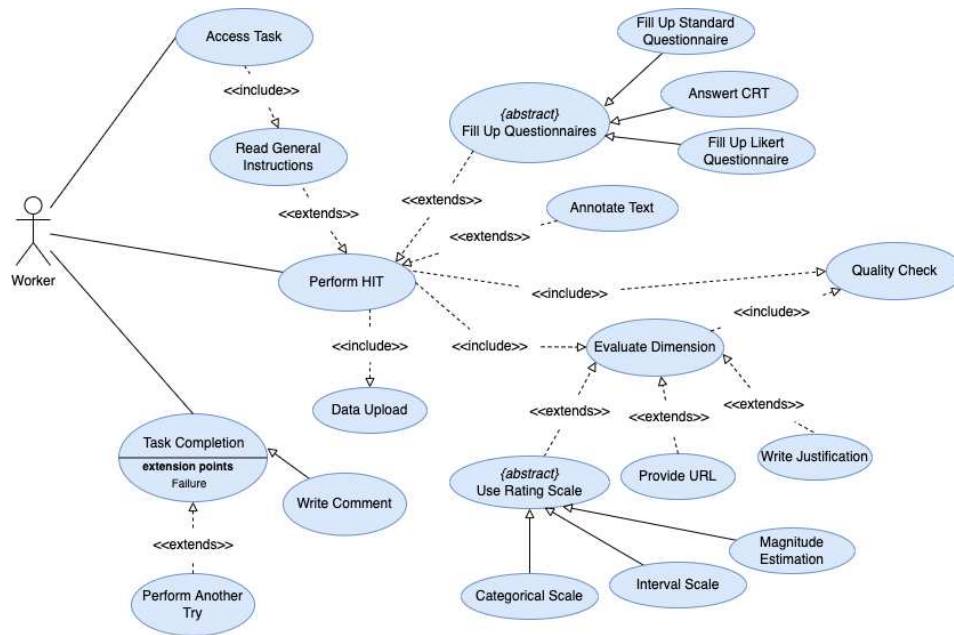


Figure A.14: Use case diagram for a worker that performs a task deployed using Crowd_-Frame.

providing a judgment using a rating scale, a URL using the search engine or writing a justification. Every type of rating scale is implemented, as described in Section A.3.2.2.

Data is uploaded throughout the whole task. Quality checks are performed on user behavior during the performing of the HIT assigned or on the values provided for each dimension. The task may thus terminate allowing the worker to write a final comment to the requester. The worker can perform another try if the quality checks fail due to some reason.

A.3.3.2 Architecture

Figure A.15 details the interaction between a crowd worker and the Skeleton component. The worker reaches the application deployed on the public bucket through the wrapper. The identifier is sent as a URL parameter. When the worker reaches the current deployment the identifier is stored in an access control list (i.e., the DynamoDB table shown in Figure A.11). This allows matching the worker with the data produced while performing the task. Furthermore, it allows tracking of how much time the worker still has to complete the work assigned. The requester can prevent the workers from accessing multiple times to the task deployed if needed.

The worker is initially shown the general instructions. The task unlocks after the initial check on the identifier. The Skeleton fetches the task configuration from the private bucket and instantiates the layout required. Then, the worker performs the task configured by the requester. The Skeleton creates a page for each of the $[0, M]$ questionnaires shown at the task's start. Then, the Skeleton initializes a page for each of the $[M + 1, P]$ elements to be

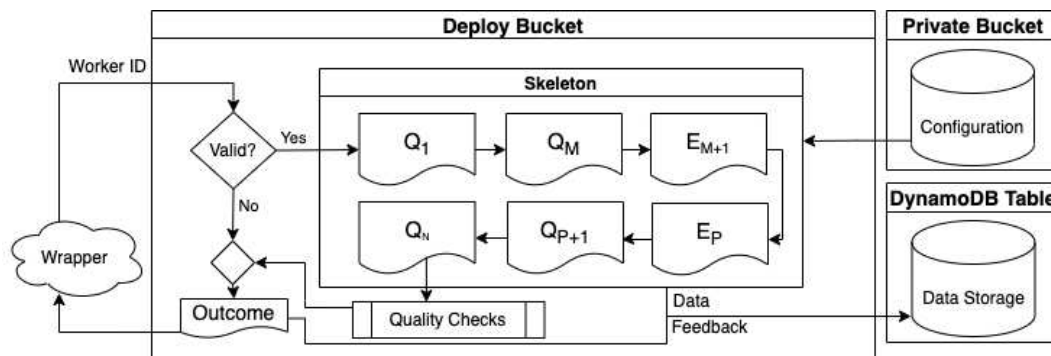


Figure A.15: Crowd_Frame Skeleton execution flow.

evaluated. Finally, the Skeleton initializes a page for each of the $[P + 1, N]$ questionnaires shown at the task's end. The appearance of each page depends on the type of questionnaire of the dimension's parameters. The evaluation instructions are shown on every page. The worker can go back and forth on each element and perform the work required. The quality checks trigger after the last element or questionnaire. The outcome page is shown to the worker if the quality checks are satisfied. The worker may provide the final comment on the same page. Finally, the worker can reach the crowdsourcing platform to receive the reward. During the whole execution flow, the Skeleton stores the data produced in the DynamoDB table.

A.3.3.3 Wrapper

Figure A.11 shows the presence of a wrapper between the workers recruited on a crowdsourcing platform and a task deployed using Crowd_Frame. The mechanism is used to identify a worker and assign a HIT to him/her. It relies on the possibility to provide input variables to the platform. The requester thus provides the crowdsourcing platform with a file containing input and output tokens. The input token is an alphanumeric string used to match the worker with a certain HIT. The output token is used by the worker to confirm the successful completion of the HIT. The wrapper fetches the platform-specific identifier of the worker. Figure A.16 shows the wrapper's interface used to redirect a worker recruited on Amazon Mechanical Turk. The wrapper's interface for a task deployed on Toloka is similar to the one shown by Figure A.16. On the other hand, Prolific does not require this at all the wrapper, since the requester must provide a direct URL to the task deployed externally.

The worker recruited on the crowdsourcing platform clicks the link shown by the wrapper to reach the task and is redirected to the deployment on the public S3 bucket. They is then automatically and implicitly matched with an input token which is not already allocated to someone else. Crowd_Frame implements a custom allocation scheme that ensures consistency in HITs assignment. Then, they is shown an output token when completing the task. Such a token must be copied back into the wrapper, depending on the platform used. If the output token provided matches with the input token, the worker can be paid. Worker identifiers and input/output tokens are fundamental to ensure correct matching between workers and HITs.

Steps Needed To Complete The Task

1. Make sure you have the latest version of your browser
2. Follow the link below to view the task page
3. Copy the output token, which will be shown at the end
4. Insert the output token below, click Submit, and get paid
5. Beware that:
 1. We will perform cross checks so random input/output tokens will not be accepted
 2. If you close the task's tab you will not be able to perform it again

<https://crowdsourcing-deploy-us.s3.us-east-2.amazonaws.com/Sample/Test-1/index.html>

(*) If you paste the output token and the "Submit" button is not enabled, please perform a mouse click.

Figure A.16: Wrapper interface for to a task configured Crowd_Frame that relies on Amazon Mechanical Turk for recruiting workers.

A.3.3.4 Data Format

The data produced throughout the task is stored on DynamoDB tables, as shown in Figure A.11. DynamoDB is a NoSQL database service, thus the records inserted in tables may contain structured data. The Skeleton component uses two different tables. One of these tables contains the access control list for the task deployed. In more detail, the access control list contains a single record for each worker. The attributes of the table vary depending on the platform of the provenance of the workers.

Listing A.6 shows the CSV dump of a single record present in the access control list table of a task. The attributes include timestamps useful to understand when the worker participated, completed, or abandoned the task. There are also the identifiers and the tokens of the HIT assigned. It is possible to understand whether the worker should be paid or not and how many tries they performed, and so on.

```

1  "identifier","access_counter","batch_name","folder","generated","in_prog
   ↪  res", "ip_address","ip_source","paid","platform","status_code","task
   ↪  _name","time_arrival","time_completion","time_expiration","time_expi
   ↪  ration_nearest","time_expired","time_removal","token_input","token_o
   ↪  utput","try_current","try_left","unit_id","user_agent","user_agent_s
   ↪  ource"
2  "<<anonymized>>","1","Batch-9-Toloka",<<anonymized>>","true","false","<
   ↪  <anonymized>>","cf","true","toloka","200",<<anonymized>>","Fri, 17
   ↪  Mar 2023 09:25:36 GMT","Fri, 17 Mar 2023 09:30:04 GMT","Fri, 17 Mar
   ↪  2023 10:25:36 GMT","Fri, 17 Mar 2023 10:25:54 GMT","false","", "PKPBSI
   ↪  NUKHS","BHPCIBOKBDL","1","10","unit_44","Mozilla/5.0 (Linux; Android
   ↪  11; TECNO KG7h) AppleWebKit/537.36 (KHTML, like Gecko)
   ↪  Chrome/108.0.0.0 Mobile Safari/537.36","cf"

```

Listing A.6: Format for a DynamoDB table that contains the access control list of a task deployed using Crowd_Frame.

The remaining table contains the data produced by the worker. It contains several records for each worker, depending on how the worker performs the task. The attributes of the table are always the same. The sequence attribute is a combination of worker and HIT identifiers, an index of the current element evaluated, and a sequence number. Such a combination is needed to identify each piece of data stored. The access attribute describes how the number of accesses to a HIT's element. The action attribute describes whether the worker is progressing through the HIT assigned or they is revising a previously evaluated element. The element attributes describe which kind of data is stored. The index element refers to the current element of the HIT. The sequence_number attribute is used to identify the sequence of data stored for the current worker. The time attribute is the data upload timestamp. The try attribute identifies the try performed by the worker. The data attribute contains the actual data produced during the task that needs to be stored. Listing A.7 shows the CSV dump of a single record present in the data table of a task.

Each piece of data is a JSON object. A piece of data can consist of the values provided for a set of evaluation dimensions, the outcome of quality checks, the answers provided

```

1 | "identifier","sequence","access","action","data","element","index","sequ_
   | ↪ ence_number","time","try"
2 | "<<anonymized>>","<<anonymized>>-unit_118-1-10","1","Next",{...},"docume_
   | ↪ nt","6","10","Tue, 23 Aug 2022
   | ↪ 11:49

```

Listing A.7: Format for a DynamoDB table that contains the data produced by the worker during a task deployed using Crowd_Frame.

for a questionnaire and so on. Listing A.8 shows a sample JSON payload stored in the DynamoDB table for a single element of a HIT assigned to the worker. The worker evaluates nine different dimensions, and one of them requires providing a URL. The payload thus shows the presence of queries and responses retrieved by the search engine.

A.3.3.5 Cost Estimation

The architecture implemented using Amazon Web Services is pay-per-use, with some fluctuation depending on the request size, number and other factors. The AWS pricing calculator¹⁰ allows making a rather precise cost prediction. The prices considered in the following refer to the AWS Region US-East-2 (Ohio). The most recent variant of the task deployed by Soprano et al. [395] (described in Chapter 7) and Draws et al. [111] (described in Chapter 9) is used to provide a sample estimation of the impact of the Skeleton component on the overall cost. The variant of the task deployed by Soprano et al. [395] and Draws et al. [111] considered involves 200 workers recruited from Prolific in a single batch.

The first service to consider is S3. The pricing¹¹ depends on various cost components, all determined based on the storage class chosen. Crowd_Frame uses the Standard class, usually recommended for general-purpose storage without particular requirements. The service applies a fee of \$0.023 for the first 50 TB/Month of data stored region-wide. Then, it applies a fee of \$0.005 for every 1000 HTTP requests of type PUT, COPY, POST and LIST, lowered to \$0.004 for all the remaining requests. Lastly, a fee is applied for bandwidth usage. Each AWS account receives free 100 GB of data transferred out of S3 to the rest of the internet. A fee of \$0.09 is applied beyond such a quantity for the first 10 TB/Month. The transfer of data from the internet to S3 is always free. The service applies additional fees for the usage of other features not required by Crowd_Frame. The amount of data stored in the S3 bucket is the size of the configuration and source code file for each batch of the task. A worker transfers a total of $7 + 3 = 10$ files each time they access the task. The access control list contains the number of accesses for each worker.

The size of the configuration of the task deployed is roughly 0.98 MB. Its source code weighs roughly 5.38 MB. The service thus stores $(5.38 + 0.98) = 0.00636$ GB/Month. The workers access the task 214 times and thus send a total $214 * (7 + 3) = 2140$ HTTP requests to perform it. The amount of data transferred given by is $214 * 0.00636 = 1.36$ GB/Month. Equation A.1 shows the detailed computation of the storage cost component. Equation A.2

¹⁰<https://calculator.aws/#/>

¹¹<https://aws.amazon.com/s3/pricing/>

```
1  {
2    "info": {
3      "action": "Next",
4      "access": 1,
5      "try": 1,
6      "index": 6,
7      "sequence": 10,
8      "element": "document"
9    },
10   "answers": { ... },
11   "notes": [],
12   "dimensions_selected": {
13     "data": [ ... ],
14     "amount": 9
15   },
16   "queries": {
17     "data": [ ... ],
18     "amount": 1
19   },
20   "timestamps_start": [ 1661254995.22 ],
21   "timestamps_end": [ 1661255107.96 ],
22   "timestamps_elapsed": 112.74000000953674,
23   "countdowns_times_start": [],
24   "countdowns_times_left": [],
25   "countdowns_expired": [],
26   "accesses": 1,
27   "responses_retrieved": {
28     "data": [ ... ],
29     "amount": 15,
30     "groups": 1
31   },
32   "responses_selected": {
33     "data": [ ... ],
34     "amount": 1
35   }
36 }
```

Listing A.8: JSON data stored for the element of a HIT evaluated by a worker during a task deployed using Crowd_Frame.

details the cost of the data retrieval component in terms of HTTP requests. Equation A.3 shows the detailed computation of the data transfer cost component. Finally, Equation A.4 shows the overall cost estimation.

$$\begin{aligned}
 \text{Data Storage} &= \$0.023 * \text{Data Size (TB/Month)} \\
 &= \$0.023 * (0.00636/1024) \\
 &= \$0.023 * 0.00000621 = \$0.00000014 \quad (\$0 \iff \text{free tier}) \\
 &\text{storage class: standard} \\
 &\text{threshold: 50 TB/Month} \\
 &\text{free tier: 5 GB/Month}
 \end{aligned} \tag{A.1}$$

$$\begin{aligned}
 \text{Data Retrieval} &= \$0.0000004 * \text{Requests Amount} \\
 &= \$0.0000004 * (\text{Worker Accesses} * (\text{Files Retrieved})) \\
 &= \$0.0000004 * (214 * (7 + 3)) \\
 &= \$0.0000004 * 2140 = \$0.000856 \quad (\$0 \iff \text{free tier}) \\
 &\text{storage class: standard} \\
 &\text{threshold: 1000 requests} \\
 &\text{free tier: 2000 requests}
 \end{aligned} \tag{A.2}$$

$$\begin{aligned}
 \text{Data Transfer} &= \$0.09 * \text{Data Transfer Size} \\
 &= \$0.09 * [\text{Worker Accesses} * \text{Data Size (GB/Month)}] \\
 &= \$0.09 * [214 * 0.00636] \\
 &= \$0.09 * 1 = \$0.09 \quad (\$0 \iff \text{free tier}) \\
 &\text{storage class: standard} \\
 &\text{threshold: 10 TB/Month} \\
 &\text{free tier: 100 GB/Month}
 \end{aligned} \tag{A.3}$$

$$\begin{aligned}
 \text{S3} &= \text{Data Storage} + \text{Data Retrieval} + \text{Data Transfer} \\
 &= \$0.00000014 + \$0.000856 + \$0.09 = \$0.09085614
 \end{aligned} \tag{A.4}$$

The second service to consider is DynamoDB. The service offers two capacity modes¹² that come with specific billing options for processing read and write requests on the table. The on-demand capacity mode charges for the data reads and writes performed by an application on the tables. The provisioned capacity mode charges for the number of reads and writes that the application is expected to require. Crowd_Frame uses the on-demand mode since the amount of read and write requests depends on the number of workers

¹²<https://aws.amazon.com/dynamodb/pricing/>

recruited.

A write request unit (WRU) is the billing unit of a set of API calls used to write data to tables. A standard write request unit can store items up to 1 KB. Additional write request units are used if the item is larger than 1 KB. A transactional write requires two units. A read request unit (RRU) is the billing unit of API calls used to read data from tables. A strongly consistent read request of up to 4 KB requires one read request unit. Additional read request units are used if the item is larger than 4 KB. An eventually consistent read request requires one-half request unit, while a transactional read requires four read request units. The type of read request chosen has an impact of consistency.¹³ DynamoDB uses eventually consistent reads unless specified otherwise. Crowd_Frame sends read requests using the default parameters, thus relies on such type of reads. The response might not reflect the results of a recently completed write operation when the read request is eventually consistent. In such a case, a second read request is needed to retrieve the most up-to-date data. On the other hand, the usage of strongly consistent reads has some constraints and leads to using more throughput capacity, thus being more expensive.

To fee applied for each write or request unit depends on the storage class chosen for the tables when the on-demand capacity modes are used. DynamoDB offers a Standard table class and a Standard-infrequent access class. Crowd_Frame relies on the former class. The on-demand capacity mode charges \$1.25 for one million write request units and \$0.25 for one million read request units. Data storage is another table class dependent cost component that must be considered. The first 25 GB/Month are free using the Standard class, and a fee of \$0.25 is applied per GB/Month after that. There is not any free quota when using the Standard-infrequent table class. The fee is thus \$0.10 per GB/Month. The service applies additional fees for the usage of other features not required by Crowd_Frame.

The variant of the task deployed by Soprano et al. [395] and Draws et al. [111] considered uses two of the three tables shown in Figure A.11. The component uses only the access control list table and the data table. The log table is used by the Logger component. Its cost estimation process is described in Section A.3.5.5.

The data table has records of varying sizes. It depends on the type of data stored, as explained in Section A.3.3.4. In the variant of the task considered there are at minimum 13 records for each try of each worker, depending on how they behave during the task. These records are broken down into (at least) 8 document records, 3 questionnaire records, a single record containing the task setup and worker attributes, and a single record addressing quality checks. An additional record may appear if the worker provides a final comment to the requester. Table A.2 shows the parameters needed to estimate the cost for a single worker of the task considered. The data table for the whole 200 workers recruited for the task contains a total of 2183 document records, 787 questionnaire records, 262 general data records, 256 quality checks records and 41 comment records.

Equation A.5 shows the detailed computation for the write request units cost component of the data table. Equation A.6 shows the detailed computation for the read request units cost component of the data table. Equation A.7 shows the computation for the storage cost of the data written. Equation A.8 further summarizes the contribution of each cost

¹³<https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/HowItWorks.ReadConsistency.html>

Table A.2: Sample cost estimation parameters for the usage of the data table by the Skeleton component of Crowd_Frame.

Record Type	Average Amount	Average Size (KB)	Average WRUs	Average RRUs
document	11	182.85	183	23
questionnaire	4	14.48	15	2
data	1	38.86	39	5
checks	1	1.27	1	0.5
comment	1	0.04	1	0.5
Total	18	237.5	239	31

component for the usage of the data table by Crowd_Frame for the sample task considered.

$$\begin{aligned}
 \text{WRUs} &= \$1.25 * (\text{Workers Amount} * \text{Avg. Message Number}) * \\
 &\quad [\text{Avg. Payload Size (KB)} / \text{Unit Amount}] \\
 &= \$0.00000125 * (200 * 18) * \\
 &\quad [(182.85 + 14.48 + 38.86 + 1.27 + 0.04)/1] \tag{A.5} \\
 &= \$0.00000125 * 3600 * (238 \text{ Avg. WRUs} / \text{Record}) \\
 &= \$0.00000125 * 856800 \text{ Billable WRUs} = \$1.07 \\
 &\quad \text{threshold: 1 million WRUs}
 \end{aligned}$$

$$\begin{aligned}
 \text{RRUs} &= \$0.25 * (0.5 * (\text{Workers Amount} * \text{Avg. Message Number}) * \\
 &\quad [\text{Avg. Payload Size (KB)} / \text{Unit Amount}]) \\
 &= \$0.00000025 * (0.5 * (200 * 18)) * \\
 &\quad [182.85 + 14.48 + 38.86 + 1.27 + 0.04)/4] \tag{A.6} \\
 &= \$0.00000025 * (0.5 * (3600 * (60 \text{ Avg. RRUs} / \text{Record}))) \\
 &= \$0.00000025 * 108000 \text{ Billable RRUs} = \$0.03 \\
 &\quad \text{threshold: 1 million RRUs}
 \end{aligned}$$

$$\begin{aligned}
\text{Data Storage} &= \$0.25 * (\text{Message Number} * \text{Avg. Payload Size (GB)}) \\
&= \$0.25 * (200 * ((182.85 + 14.48 + 38.86 + 1.27 + 0.04)/1024/1024)) \\
&= \$0.25 * (200 * ((239/1024)/1024)) \\
&= \$0.25 * 0.00022793 \text{ GB/Month} = \$0.00005698 \quad (\$0 \iff \text{free tier}) \quad (\text{A.7}) \\
&\text{threshold: 1 GB/Month} \\
&\text{note: on-demand capacity mode, standard table class} \\
&\text{free tier: 25 GB/Month}
\end{aligned}$$

$$\begin{aligned}
\text{Data Table} &= \text{WCUs} + \text{RCUs} + \text{Data Storage} \\
&= \$1.07 + \$0.03 + \$0.00005698 = \$1.10005698 \quad (\text{A.8})
\end{aligned}$$

The access control list table has an average record size of roughly 800 B. Such an item consumes 0.5 RRUs when read with eventual consistency and 1 WRUs when being written. There is a total of 200 workers recruited and each worker has a single record. The total size of the table is roughly 160 KB, thus 160 WRUs are used. There may be a certain number of additional RRUs performed on the table. These RRUs take place when a worker does not complete the work assigned on time and the unit must be reallocated. In such a case, the table must be scanned to determine which unit to reallocate and update the affected workers' records. Such a scenario does not happen when considering the aforementioned variant. Otherwise, the RRUs computation depends on how many records were already present in the access control list when each additional worker was recruited. The impact of the access control list table is negligible due to its small number of records, request units and storage size.

$$\begin{aligned}
\text{WRUs} &= \$1.25 * (\text{Workers Amount} * \text{Avg. Message Number}) * \\
&\quad [\text{Avg. Payload Size (KB)} / \text{Unit Amount}] \\
&= \$0.00000125 * (200 * 1) * \\
&\quad [(800/1024)/1] \quad (\text{A.9}) \\
&= \$0.00000125 * 3600 * (1 \text{ Avg. WRUs} / \text{Record}) \\
&= \$0.00000125 * 3600 \text{ Billable WRUs} = \$0.045 \\
&\text{threshold: 1 million WRUs}
\end{aligned}$$

$$\begin{aligned}
\text{RRUs} &= \$0.25 * (0.5 * (\text{Workers Amount} * \text{Avg. Message Number}) * \\
&\quad [\text{Avg. Payload Size (KB)} / \text{Unit Amount}]) \\
&= \$0.00000025 * (0.5 * (200 * 1)) * \\
&\quad [(800/1024)/4] \tag{A.10} \\
&= \$0.00000025 * (0.5 * (3600 * (1 \text{ Avg. RRUs} / \text{Record}))) \\
&= \$0.00000025 * 1800 \text{ Billable RRUs} = \$0.00045 \\
&\quad \text{threshold: 1 million RRUs}
\end{aligned}$$

$$\begin{aligned}
\text{Data Storage} &= \$0.25 * (\text{Message Number} * \text{Avg. Payload Size (GB)}) \\
&= \$0.25 * (200 * (800/1024/1024/1024)) \\
&= \$0.25 * (200 * 0.00000075) \\
&= \$0.25 * 0.00014901 \text{ GB/Month} = \$0.00003725 \text{ } (\$0 \iff \text{free tier}) \tag{A.11} \\
&\quad \text{threshold: 1 GB/Month} \\
&\quad \text{note: on-demand capacity mode, standard table class} \\
&\quad \text{free tier: 25 GB/Month}
\end{aligned}$$

$$\begin{aligned}
\text{Table ACL} &= \text{WCUs} + \text{RCUs} + \text{Data Storage} \\
&= \$0.045 + \$0.00045 + \$0.00003725 = \$0.04548725 \tag{A.12}
\end{aligned}$$

Equation A.9 shows the detailed computation for the write request units cost component of the access control list table. Equation A.10 shows the detailed computation for the read request units cost component of the access control list table. Equation A.11 shows the computation for the storage cost of the records written. Equation A.12 further summarizes the contribution of each cost component for the usage of the access control list table by Crowd_Frame for the sample task considered.

Finally, Equation A.13 summarizes the cost of the usage of the Skeleton component of Crowd_Frame for the task considered. It must be noted that data are only written to the data table during the task and read-only afterwards while downloading results. The cost of the RCUs component for the data table can be charged at a later time. Furthermore, the cost must be intended on a per-month basis only for the storage components of both S3 and DynamoDB.

$$\begin{aligned}
\text{Skeleton Component} &= \text{S3} + \text{DynamoDB} \\
&= \text{S3} + \text{Table Data} + \text{Table ACL} \\
&= (\text{Storage} + \text{Data Retrieval} + \text{Data Transfer}) + \\
&\quad (\text{WRUs} + \text{RRUs} + \text{Data Storage}) + \\
&\quad (\text{WRUs} + \text{RRUs} + \text{Data Storage}) \\
&= (\$0.00000014 + \$0.000856 + \$0.09) + (\$1.07 + \$0.03 + \$0.00005698) + \\
&\quad (\$0.045 + \$0.00045 + \$0.00003725) \\
&= \$1.23640037
\end{aligned}
\tag{A.13}$$

A.3.4 Search Engine

Crowd_Frame implements a component that allows integrating a customizable search engine within the task body, as shown by Figure A.11.

A.3.4.1 Use Cases

The Search Engine component imitates the standard approach followed by the most popular search engines such as Google. Such an approach involves showing the search results found for a given query provided by a user.

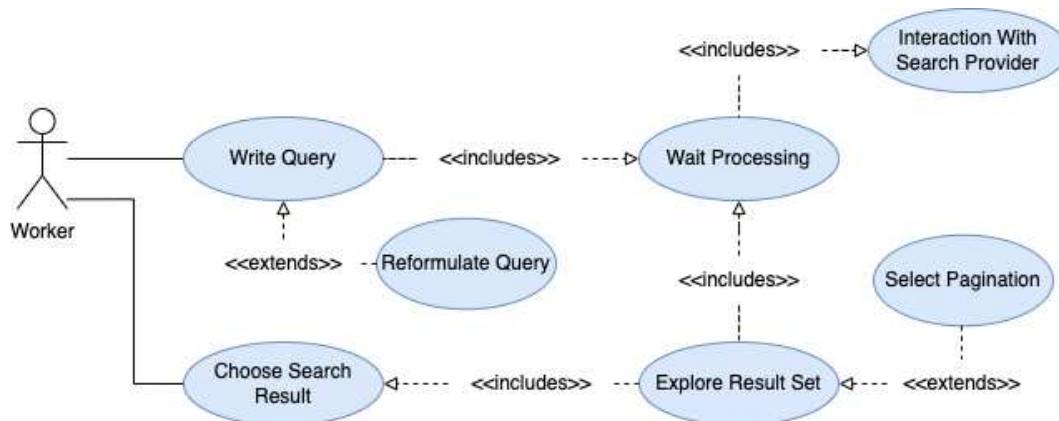


Figure A.17: Use case diagram for a worker that performs a task deployed using Crowd_Frame.

The diagram shown in Figure A.17 provides a high-level description of the interaction between a worker and the Search Engine component of Crowd_Frame. The worker writes a query in a simple text box shown by the user interface. The query is then sent to a search provider and the worker waits for its processing. The search provider processes the query

and returns a result set. The results are shown below the query box in a tabular format. Each search result is shown to the worker by providing its URL, page name and snippet. The result set is paginated by default. The worker can choose how many results to see for each page. The worker can reformulate the query and obtain a new result set at any time.

The approach followed by the component deviates from the standard one when exploring the result set. Integrating a custom search engine allows workers to provide a URL for certain evaluation dimensions. The underlying goal can be of various kinds. The requester, for instance, may want the worker to provide some kind of evidence, as done by Roitero et al. [361]. The user interface thus shows a button to the right of each search result shown. The worker explores each result and then finalizes the choice by clicking the corresponding button. Figure A.17 shows a sample of the user interface of the Search Engine component. In more detail, the figure shows a result set made of 48 elements split into pages of 5 elements each retrieved for the query having text Barack Obama.

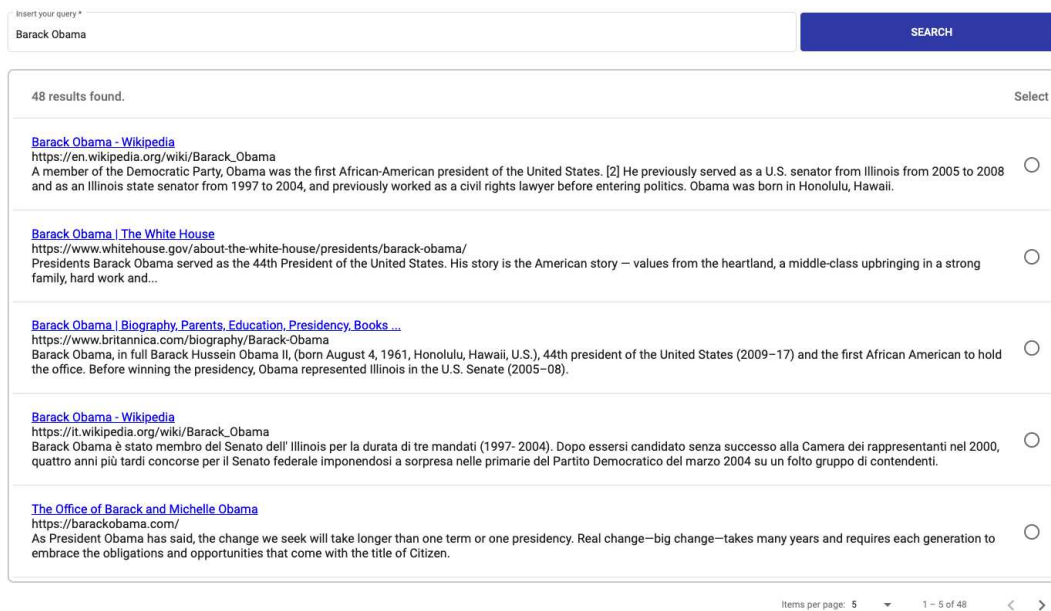


Figure A.18: Interface of the Search Engine component of Crowd_Frame.

A.3.4.2 Architecture

Figure A.15 details the interaction between the Search Engine component and the APIs of the search providers that a requester chooses while configuring a task. The query written by the worker is provided to a static method that encapsulates it in an HTTP message, along with the API Key of the search provider chosen. The component implements the API of three different search providers. The search provider processes the query and sends a second HTTP message to the component. The raw response is parsed using the corresponding model and then decoded using an interface that allows abstracting from the underlying data structure. The result set is composed of multiple base responses, depending on how

many search results the provider retrieves. Such a set is thus provided to the user interface that paginates and shows it to the worker. The worker is required to select one of the results provided by clicking the corresponding radio button.

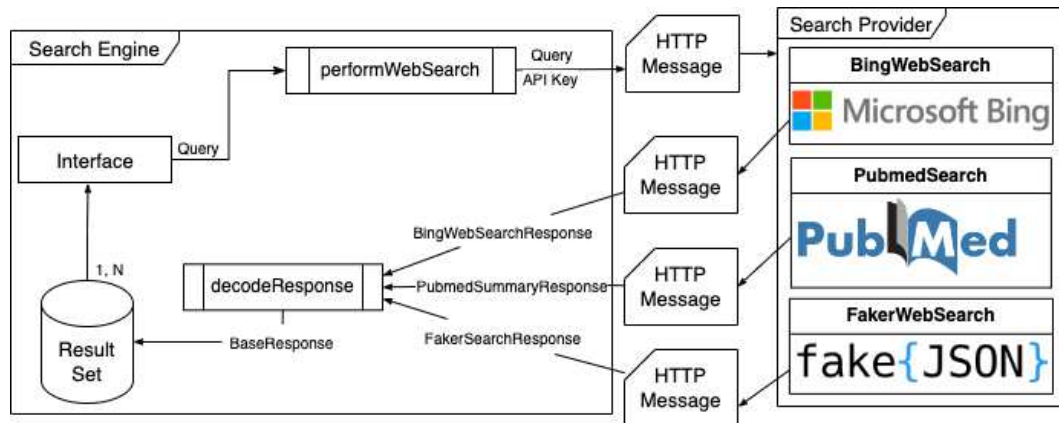


Figure A.19: Interface of the Search Engine component of Crowd_Frame.

The Search Engine component implements the API of three search providers, in sub-modules named BingWebSearch, PubmedSearch, and FakerWebSearch.

A.3.4.3 Microsoft Search API

The BingWebSearch provider implements the Bing Web Search API¹⁴, part of the Microsoft Search API. It is mainly designed to be used as a result of a direct user query or search, or as a result of an action within a system that can be logically interpreted as a user's search request. Acceptable search or search-like scenario thus include a user that provides a query directly into a search box, a user that requests "more information" about a text or image using some kind of user interface control, and so on. The API returns and rank implicitly whatever content is relevant to a query. It is possible to filter and control to some extent the results retrieved. For instance, it is possible to include or exclude specific types of results, return pages discovered within the last week, and so on.

Crowd_Frame relies on the v7.0 version of the Bing Web Search API. To use the API a developer must obtain a subscription key. Then, they can send HTTP GET messages to the API's endpoint¹⁵ and thus retrieve results. The subscription key must be inserted in a header of the HTTP request called `Ocp-Apim-Subscription-Key`. The GET parameter `q` is used to attach the user's query and must be URL-encoded. The component also provides four headers to improve the search experience for the worker. Table A.3 describes such additional headers and how the API uses them. Furthermore, the component captures three headers added to the response by the API. The headers include, for instance, `X-MSEdge-ClientID`, which must be attached to the subsequent request. Table A.4 described such headers and why they are captured.

¹⁴<https://www.microsoft.com/en-us/bing/apis/bing-web-search-api>

¹⁵<https://api.bing.microsoft.com/v7.0/search>

Table A.3: Headers provided while sending requests to the Bing Web Search API.

Header	Usage
User-Agent	Understand whether the worker is using Bing from a mobile or desktop device
X-MSEdge-ClientID	Provide continuity of search experience across multiple queries
X-MSEdge-ClientIP	Detect worker's location to improve location-aware queries
X-Search-Location	Detect worker's location to improve location-aware queries

Table A.4: Headers captured while receiving responses from the Bing Web Search API.

Header	Usage
X-MSEdge-ClientID	Provide continuity of search experience across multiple queries
BingAPIs-TraceId	Identifier of the request within the log data
BingAPIs-Market	Market used by Bing to process the request

There are various GET parameters¹⁶ that can be used to customize the results retrieved. The component attaches two additional parameters to the query's text when sending a request to the API. Table A.5 describes the GET parameters added while sending each query.

Table A.5: Query parameters provided when sending queries to the Bing Web Search API.

Parameters	Usage
count	Maximum number of results to return within a single request
offset	Handle results pagination within each query
mkt	Country from where the results come from. Typically, the country where the worker is making the query.

Listing A.9 shows a sample request sent to retrieve results for a query having text `microsoft devices` that contains the suggested headers and GET parameters. The request is reproduced using the cURL software. The response returned is a JSON object that contains various parameters and the array of web pages retrieved. Such an object is the one decoded by the component, as shown by Figure A.15.

The final result set for the query written by the worker and retrieved using the Bing Web Search API is then decoded using the static method shown in Figure A.19. The array of base responses is thus returned to the user interface of the Search Engine component of `Crowd_Frame`.

¹⁶<https://learn.microsoft.com/en-us/bing/search-apis/bing-web-search/reference/query-parameters>

```

1 | curl -H "Ocp-Apim-Subscription-Key: <yourkeygoeshere>" -H
   | ↪ "X-MSEdge-ClientID: 00B4230B74496E7A13CC2C1475056FF4" -H
   | ↪ "X-MSEdge-ClientIP: 11.22.33.44" -H "X-Search-Location:
   | ↪ lat:55;long:-111;re:22" -A "Mozilla/5.0 (X11; Linux x86_64)
   | ↪ AppleWebKit/537.36 (KHTML, like Gecko) Chrome/29.0.1547.65
   | ↪ Safari/537.36" https://api.bing.microsoft.com/v7.0/search?q=microsof
   | ↪ t+devices&mkt=en-us&count=10&offset=0

```

Listing A.9: Sample request that contains headers and query parameters sent to the Bing Web Search API.

A.3.4.4 Entrez Programming Utilities

The Entrez Programming Utilities¹⁷ (E-utilities) is a set of nine server-side programs that provide a stable interface into the Entrez query and database system at the National Center for Biotechnology Information (NCBI). Such programs require the usage of fixed endpoints whose URL-based syntax translates a set of input parameters into the values necessary for the underlying software components to search for and retrieve the data requested. In other words, the E-utilities are the structured interface to the Entrez system, which includes 38 databases covering a variety of biomedical data. Table A.6 provides a brief description of each E-utility. PubMed¹⁸ is a free search engine that access primarily the MEDLINE¹⁹ database which contains references to the biomedical literature. The United States National Library of Medicine (NLM) at the National Institutes of Health (NIH) maintain the database as part of the Entrez system.

The NCBI recommends using an API Key to access the E-utilities. Its usage helps to avoid overloading the underlying systems. Any IP address that sends more than 3 requests per second to the E-utilities without an API key receives an error message. On the other hand, IP addresses are allowed to send up to 10 requests per second when the key is provided. The API Key can be obtained from the NCBI account page.²⁰ Each request must be sent to the base endpoint <https://eutils.ncbi.nlm.nih.gov/entrez/eutils/>. The API key is added by appending its value to the `api_key` GET parameter. The E-utility wanted can be chosen by appending its lowercase name with the `.fcgi` suffix. The only exception is the `ECitMatch` utility, which requires the `.cgi` suffix. The usage of the `ESummary` utility, for instance, involves sending requests to the endpoint <https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esummary.fcgi>.

The PubMedSearch provider interacts with the `ESummary` and `ESearch` utilities of the Entrez system. The E-utilities must be combined to create an effective and useful data pipeline. The system provides the Entrez History server that simplifies transferring context and data between successive requests. The `ESearch` utility is the first block of the pipeline. The query provided by the worker is added as a parameter to the request. The request is sent to the PubMed search engine, which returns a list that contains the identifiers of items

¹⁷<https://www.ncbi.nlm.nih.gov/books/NBK25501/>

¹⁸<https://pubmed.ncbi.nlm.nih.gov/>

¹⁹https://www.nlm.nih.gov/medline/medline_overview.html

²⁰<https://www.ncbi.nlm.nih.gov/account/>

Table A.6: General description the of the nine E-utilities provided by the Entrez system.

E-utility	Goal	Description
EInfo	Database statistics	Provides the number of records indexed in each field of a given database, date of the last update and available links towards other Entrez databases.
ESearch	Text searches	Responds to a text query with the list of matching UIDs (identifiers) in a given database for later use in other E-utilities, along with the term translations of the query.
EPost	UID uploads	Accepts a list of UIDs from a given database and responds with the web environment of the uploaded dataset and its query key.
ESummary	Document summary download	Responds to a list of UIDs in a given database with the corresponding document summaries.
EFetch	Data record download	Responds to a list of UIDs in a given database with the corresponding data records in a specified format.
ELink	Entrez Links	Responds to a list of UIDs in a given database with either a list of related UIDs (and relevancy scores) in the same database or a list of linked UIDs in another Entrez database and more.
EGQuery	Global Query	Responds to a text query by providing the number of records matching the query in each Entrez database.
ESpell	Spelling Suggestions	Retrieves spelling suggestions for a text query in a given database
ECitMatch	PubMed Batch citation search	Retrieves PubMed IDs (PMIDs) corresponding to a set of input citation strings.

relevant to the query. There is a total of six GET parameters used to customize the results returned by the utility, shown in Table A.7.

Listing A.10 shows a sample initial request sent to the ESearch utility to retrieve a JSON set of identifiers for a query having text vaccines. The subsequent requests can use the WebEnv and query_key parameters if needed.

```
1 | curl https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?api_key=J
   | ↪ your_api_key&db=pubmed&term=vaccines&usehistory=y&retmode=json
```

Listing A.10: Sample initial request sent to the ESearch utility of the Entrez system.

The ESummary utility is the second and last block of the pipeline implemented by the search provider. The identifiers of each result item retrieved for the query written by the worker using the ESearch utility is used to fetch the details of the corresponding data

Table A.7: Query parameters provided when sending requests to the ESearch utility of the Entrez system.

Parameters	Usage
db	Database from which retrieve results, which is pubmed.
term	Text query for which retrieve results. All special characters must be URL encoded. Spaces may be replaced by + character.
usehistory	ESearch will post the UIDs resulting from the search operation in the History server to be used in the subsequent call, when set to y.
retmode	Format of the returned output. The JSON format is used by the search provided.
WebEnv	Web environment returned by a previous ESearch, EPost or ELink request. ESearch appends the result set retrieved to the one contained in the pre-existing environment. The usage of usehistory is required.
query_key	Integer query key returned by a previous ESearch, EPost or ELink request. ESearch will find the intersection of the result set identified by the key and the one retrieved for the current term. The usage of WebEnv is required.

records. The GET parameters used to customize the results returned are those shown in Table A.7. The only difference is that the term parameter is replaced with the id one. Such a parameter requires either a single UID or a comma-delimited list of UIDs. The search providers all those retrieved by sending the query to the ESearch utility. For instance, let us hypothesize that one of such UIDs retrieved for the request shown in Listing A.10 is 36511263. Listing A.11 shows the request sent to the ESummary utility to fetch the details of the data record.

```
1 | curl https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?api_key=
   | ↪ your_api_key&db=pubmed&id=36511263&retmode=json
```

Listing A.11: Sample initial request sent to the ESummary utility of the Entrez system.

The final result set for the query written by the worker and detailed using the pipeline implemented by the PubMedSearch provider is then decoded using the static method shown in Figure A.19 and the array of base responses are returned to the user interface of the Search Engine component of Crowd_Frame.

A.3.4.5 fakeJSON

The fakeJSON²¹ service provides a simple API for building mock data and helps the frontend development, end-to-end testing, data generation, etc. The underlying idea is that the developer sends a request directly to a single endpoint. In the request's body, they

²¹<https://fakejson.com/>

specify the format of the response desired. The service creates a JSON object populated with the data requested, initializing them with random values. The requests can be sent using any of the standard HTTP methods such as POST, PUT, DELETE, and PATCH. The service supports cross-origin resource sharing, thus allowing to receive asynchronous requests from anywhere. Every request must be sent to the endpoint `https://app.fakejson.com/q`.

The `FakerWebSearch` provider interacts with the service to retrieve fake search result data, to allow testing the search engine of `Crowd_Frame` while designing a task. Every request sent to the service to generate fake data must comply with the same payload format. The payload must be a JSON object that includes a unique token, an optional set of parameters and the data field. The unique token can be obtained on the account page.²² The parameters allow customizing the behavior of the response generated depending on the needs. For instance, the parameters allow delaying the response by a fixed amount of seconds, or to add a set of custom headers. Table A.8 shows the four parameters that can be included in the request's payload.

Table A.8: Payload parameters used to customize the behavior of the responses generated by the fakeJSON service.

Parameter	Usage
<code>code</code>	Specifying the response code number to fake.
<code>delay</code>	Delaying the response's sending by a fixed number of seconds.
<code>headers</code>	Adding additional custom headers in the response generated.
<code>consistent</code>	Disabling the service from returning the same fake dataset cached. The only possible value is <code>false</code> .

The data field of the request's payload is where the format of the responses to be returned is defined. It consists of fields, objects and arrays like any other valid JSON. It follows an attribute value pair format. The name of each key can be any string, however, the value must follow exactly the syntax of the corresponding data type defined by fakeJSON. The `_repeat` attribute can be added to the object to specify how many iterations of an object the service should return. Listing A.13 shows a sample request sent to the service using the cURL software.

```
1 | curl --request POST --url https://app.fakejson.com/q --header
   | ↪ 'content-type: application/json' --data 'your_data_object'
```

Listing A.12: Sample request sent to the fakeJSON service using CURL.

Listing A.13 shows the payload of the request sent by the search provided to generate fake data and allow the requester to test the search engine. The service is requested to generate a response with code 200. The response is made of 8 JSON objects. Each object has a `text` field that contains a long string. The `name` field contains a string of shorter length.

²²<https://app.fakejson.com/member/token>

The `url` field contains a random URL. The idea is imitating the results retrieved by a search engine, characterized by a page address, a page name and a text snippet that describes the page itself. In general, there are several formats²³ that can be used to mock data.

```
1  {
2    "token": "...",
3    "parameters": {
4      "code": 200
5    },
6    "data": {
7      "url": "internetUrl",
8      "name": "stringShort",
9      "text": "stringLong",
10     "_repeat": 8
11   }
12 }
```

Listing A.13: Payload of the request sent to the fakeJSON service to generate fake search results.

The fake search result data generated by the fakeJSON service encapsulated into the `FakerWebSearch` provider is then decoded using the static method shown in Figure A.19 and the array of base responses is returned to the user interface of the Search Engine component of `Crowd_Frame`.

A.3.4.6 Cost Estimation

A platform that provides some kind of search API usually relies on a payment model that considers usually the number of queries issued and their sending rate on a per-second basis as parameters to estimate the overall cost. In light of this, a more adequate estimation of the usage cost of the Search Engine component of `Crowd_Frame` can be performed by relying on the data gathered for tasks deployed in the past, as described in Section A.3.3.5 for the `Skeleton` component. The most recent variant of the task deployed by Soprano et al. [395] (described in Chapter 7) and Draws et al. [111] (described in Chapter 9) is thus used to estimate also such a usage cost.

The variant of the task considered publishes 200 HITs. The workers are required to provide a URL using the search engine for one of the evaluation dimensions. A total amount of 237 workers access the task deployed to try to complete the HIT assigned. The workers issue 3520 different queries. The minimum number of queries issued by a worker is 2. This worker abandons the task after a short while. The maximum number is 58. Workers with high amounts of queries issued perform multiple tries or write multiple queries for one or more elements of the HIT assigned. The workers issue on average roughly 15 queries each. The last query is issued after 2 days and 8 hours, 2 minutes, and 17 seconds from the first one. The queries can be split into 2498 unique transactions-per-second (TPS) blocks.

²³https://fakejson.com/documentation#request_data

A TPS block contains a set of queries sent within a single second. It is thus possible to detect 2498 seconds between the first and the last query. During such seconds, an arbitrary number of workers issued one or more queries. These 2498 seconds (roughly 42 minutes) happened between the timestamps of the first and last query. The smallest TPS size is 1. A block of such a size contains a single query issued by a single worker in a single second. The biggest TPS block size is 5. A block of such size contains five queries issued by one or more workers in a single second. The average TPS size is 1.40. The definition does not imply that the TPS blocks are consecutive. Splitting the queries into such blocks allows understanding the query throughput towards the platforms that provide the search APIs.

The variant of task considered uses the BingWebSearch provider, thus relying on the Microsoft Bing Search API. The pricing plans²⁴ of the API differs along three parameters. These parameters are the amount of TPS allowed, the type of search offered, and the cost of 1000 transactions. The last parameter differs when using or not the Japanese geographical market. Crowd_Frame relies on the Bing Web Search subset of the API for the United States market within the S3 pricing plan. The plan allows a maximum TPS size of 100. In other words, a maximum number of 100 queries can be received each second by the API. The usage of the service is throttled to be within the threshold if it is exceeded. This translates into a slower search experience for the user. The plan applies a fee of \$4 for a set of 1000 transactions. A transaction is a successful request received by the API. Usually, a single request is sent for a single query. The usage of advanced features such as auto-completion leads to more transactions to process a single query. However, the provided implemented by Crowd_Frame does not use such a feature. The fee is thus applied four times since there 3520 queries have been issued. The maximum TPS size obtained during the task is 5, far below the threshold of 100 transactions per second. This means that not any query saw its processing throttled. Equation A.14 shows the computation of the cost of the custom search engine used for the task considered.

$$\begin{aligned} \text{Bing Web Search} &= \$4 * [\text{Query Number} / \text{Billing Threshold}] \\ &= \$4 * [3520/1000] = \$4 * [3.52] = \$16 \end{aligned} \quad (\text{A.14})$$

threshold: 1000 transactions

The usage of the PubmedSearch provider does require any form of payment. The Entrez system and its E-utilities are publicly available for free. The only requirement is complying with the guidelines suggested by the NCBI. Similarly, the FakerWebSearch provided does not enforce any payment. The pricing model²⁵ of the fakeJSON service offers a free plan constrained to 1000 requests per day, which are usually enough to test the task's interface during its design. The service stops providing search results if the threshold is reached. Eventually, it is possible to upgrade the account to plans with a higher number of maximum requests per day. The cheapest plan is the Developer one, which costs \$12 per month and offers up to 50000 requests per day, along with other features.

²⁴<https://www.microsoft.com/en-us/bing/apis/pricing>

²⁵<https://fakejson.com/pricing>

A.3.5 Logger

Crowd_Frame implements a logging component that allows capturing worker behavior during the task, as shown by Figure A.11.

A.3.5.1 Architecture

The log messages produced when capturing user behavior events are stored using a cloud-based logging server which relies on the infrastructure provided by Amazon Web Services. Figure A.20 shows an overview of the whole logging pipeline.

API Gateway²⁶ is a service used to implement APIs for web applications or other AWS services. It addresses traffic management, cross-origin resource sharing (CORS) support, authorization and access control, request throttling, and monitoring of an API layer. In more detail, Crowd_Frame uses an HTTP API layer. Each user action is captured by the Logger component and sent to the layer using an HTTP message. The CORS is an HTTP-header based mechanism that allows a server to indicate any origins (in terms of port, domain or scheme) other than its own from which a browser should permit loading resources. The API layer's CORS is configured to allow receiving POST messages only. The layer receives the messages through a single endpoint provided by the application. The body of each message is redirected to a queue upon reception.

Simple Queue Service²⁷ (SQS) is a service used to create and manage queues of messages. It allows the creation of two types of queues, namely Standard or FIFO. The first type aims to ensure the best delivery sequence and each message may be delivered more than once if its processing fails or does not complete on time. The second type, on the other hand, ensures a first-in-first-out message delivery sequence where each message is processed exactly once. The main drawback is that the cost is higher. The usage of a message queue allows decoupling of the reception of the log requests from their processing. The payload of each log request has a sequence number. Each request is independent and the whole stream of messages can be reordered at a later time. Crowd_Frame thus uses a standard queue. The attached access policy allows only the gateway to send messages to the queue itself.

Lambda²⁸ is a service that provides serverless computing service that allows running source code without explicitly provisioning and maintaining servers of any kind. One of the core features of the service is the auto scalability feature of the instances depending on the size of the workload that the code must handle. In other words, Lambda allows running an algorithm only when needed, using a self-activating approach.

A serverless function deployed using Lambda polls the queue in search of new messages and collects batches of up to 100 messages. The function parses each log message and the user action is stored as a record in the DynamoDB log table shown in Figure A.11. The whole logging system back-end can be deployed also on a private server. The requester can set a custom endpoint to where each log message will be sent. There is a trigger between the function and the queue. The trigger is activated when the queue receives a new message.

²⁶<https://aws.amazon.com/api-gateway/>

²⁷<https://aws.amazon.com/sqs/>

²⁸<https://aws.amazon.com/lambda/>

In more detail, the lambda function provides a JavaScript algorithm. Initially, it performs a polling operation of a batch of log requests. The JSON payload of each message is parsed and the server time is added. Then, the payload is stored in the table. The function is configured to repeat the polling of new log requests every 20 seconds. It uses by default five parallel long polling connections. The service can reduce the interval between each polling operation and increase the number of instances assigned to the functions. Such a feature helps empty the queue when it is filling up, thus avoiding overload. In such a case, the messages would need to be enqueued again. There is a DynamoDB table for each batch deployed for each task. The Logger component targets each table accordingly.

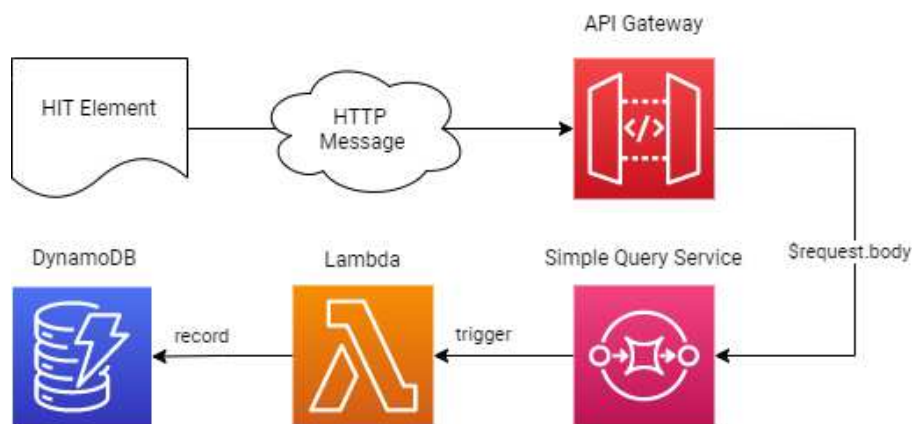


Figure A.20: Infrastructure of the Logger component of Crowd_Frame.

A.3.5.2 Event Handling

Angular uses custom markup elements. Such elements are defined using an extension of HTML that allows defining tags whose content is created and controlled by client-side scripting implemented using JavaScript or Typescript. In other words, these markup elements are useful in a dynamic environment, where flexibility is needed to create constantly changing web experiences. Angular developers were thus forced to provide a mechanism to bind event listeners to the custom elements used. Thus, in each custom element tag, additional syntax can be used to define the listener and specify the behavior when the corresponding event is triggered.

This technique is adequate when there are only a few elements. In a real-world scenario where tens if not hundreds of tags are generated, another approach is needed. Angular implements so-called Directives.²⁹ A Directive is a class that can dynamically add behavior to custom elements. A decorator is used to specify a CSS selector to target the custom elements. A constructor is used to pass a function that provides the implementation of the custom behavior. Furthermore, it attaches an event listener. The overall idea, thus, is that the set of markup elements to target depends on the type of event to log.

²⁹<https://angular.io/guide/attribute-directives>

There are several different events to monitor. Listing A.14 shows the generic JSON payload of each log request sent after having listened to a given event. The `detail` object contains the data logged depending on the type of the event. Each event detected is thus encapsulated by the `Logger` component into an instance of the base payload which is then enriched using the custom properties of the current event. Then, it is sent to the AWS infrastructure using an HTTP POST request to be processed. The server time is added when the infrastructure receives the message. Appendix A.3.6 describes the payload of each event monitored by the `Logger` component inserted into the `detail` object of the base payload. The list of events is obtained from the analysis of the HTML structure of `Crowd_Frame`.

```
1  {
2    "bucket": "string",
3    "worker": "string",
4    "task": "string",
5    "batch": "string",
6    "unitId": "string",
7    "try_current": "integer",
8    "type": "string",
9    "sequence": "integer",
10   "client_time": "string",
11   "server_time": "string",
12   "details": {
13     ...
14   }
15 }
```

Listing A.14: Payload of each log request sent by the `Logger` component of `Crowd_Frame`.

There is a directive available for each event. Every directive shares the same base template. Each event is monitored relative to a custom element selected using its CSS selector. A custom function to handle each event exists within the `Logger` component. Both the element and the logging function are thus provided to the directive constructor and the directive adds the custom behavior to the element targeted. Such an event listening behavior is attached to the corresponding markup element if and only if the `Logger` component is enabled by the requester and configured to log the corresponding event. The requester can enable or disable each event using the `Generator` component.

The event listeners debounce and optimize function calls to reduce the processing load triggered by spammable events such as mouse clicks, movements and scrolls. The `RxJS`³⁰ library is used to further enhance the logging capabilities. It allows the addition of the event listeners while piping a set of operations to extract data and perform various manipulations. The usage of such a library allows the composition of events. For instance, a text selection event starts only when the mouse button is pressed and held and terminates when the mouse button is released.

³⁰<https://rxjs.dev/>

A.3.5.3 Performance Evaluation

Locust³¹ is a Python-based performance testing tool. It allows the definition of custom user behavior and swarms a system with millions of simultaneous requests. The main script allows modelling a worker node. Each worker node, in this case, sends POST requests to the provided host address (i.e., the endpoint exposed by the API gateway). The cluster mode of Locust leads to having a master node that coordinates the worker nodes by declaring how many of them are required, the request spawn rate, and the duration of the test. The tool provides a file containing various statistics when the test ends.

The master node is spawned on a local machine. Each instance of the cluster runs two processes that spawn log requests, one for each instance's core, for a total starting amount of 200 processes. Each process sends a message every 10 ms. The test involves five rounds, where each round lasts for five minutes. The number of worker nodes doubles after each round, up to a total of 1600 nodes.

The pipeline has been initially tested on a dedicated server using an Amazon Elastic Compute Cloud³² instance based on a 2nd generation AMD EPYC processor with 4 cores/8 threads, running at frequencies up to 3.3 GHz, and having 16 GB of RAM. Figure A.21 shows that the server manages to process, without losing messages, up to 5000 requests/s. Over this limit, the server overloads thus failing progressively to accept new requests. The mean (median) time estimated between each log message to reach such a limit is 2.76 (1) seconds. Thus, at least 5000 workers working at the same time are required to overload the server. Such behavior is confirmed by looking at the requests managed per second, shown in Figure A.22. CPU and memory usages drop with more than 800 worker nodes, as shown in Figure A.23 and Figure A.24. The causes of such an overload cannot be determined precisely, since EC2 does not fully disclose the details of the underlying architecture. However, a hypothesis can be made in light of the data concerning the requests managed per second and CPU load. In more detail, it can be seen that the CPU is put under at nearly 100% of usage for a whole minute. Then, the performance drops abruptly. This could happen due to thermal throttling. The machine thus reaches the maximum number of manageable concurrent connections after one minute and then stops accepting new requests.

The performances of the AWS-based pipeline have been tested using a cluster of 100 EC2 instances. The chosen type of instance is a t3.micro, which features two cores from an Intel Xeon Platinum 8000 series processor with a clock speed of up to 3.1 GHz, 1 GB of RAM, and up to 10 Gb/s of network burst bandwidth. Each instance deployed the content of an image available on the Docker Hub.³³ The image is based on Ubuntu and contains a Locust script that allows interaction with the other nodes of the cluster. The management of the cluster of instances has been performed using Elastic Container Service.³⁴

The first round of the tests (i.e., the one with 200 worker nodes only) lead to 16538 Lambda invocations to manage the whole set of 1653801 requests sent by the cluster, in contrast with the 575000 requests processed by the dedicated server solution. The estimated mean number of requests processed each second by the pipeline is 5512. A negligible amount of messages

³¹<https://locust.io/>

³²<https://aws.amazon.com/ec2/>

³³<https://hub.docker.com/>

³⁴<https://aws.amazon.com/ecs/>

is lost, due to network communication. This confirms that the API layer of the infrastructure can manage each request rapidly and forward it to the message queue, preventing network congestion. The maximum default limit of concurrent Lambda instances is 1000. Such a limit has not been reached during the test. The maximum number of concurrent invocations is 206. In other words, the pipeline can scale in size.

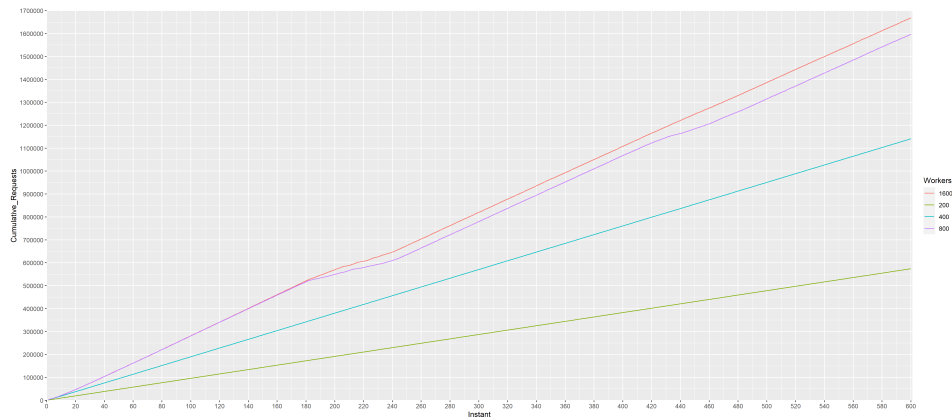


Figure A.21: Requests sent to evaluate the performances of the infrastructure during each round of the test.

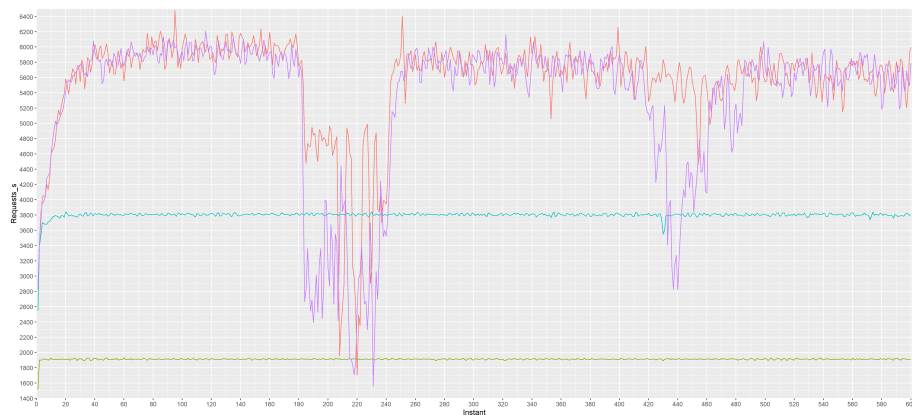


Figure A.22: Requests managed per second by the dedicated server solution during each round of the test.

A.3.5.4 Pilot Experiment

A variant of the task proposed by Roitero et al. [362] where workers are asked to evaluate the truthfulness of information items has been published on Amazon Mechanical Turk to gather real log data. The data is used to estimate the impact of the Logger component on the overall cost of a crowdsourcing task deployed using Crowd_Frame. The component has been configured to monitor every log event described in Appendix A.3.6.



Figure A.23: CPU usage of the dedicated server solution during each round of the test.

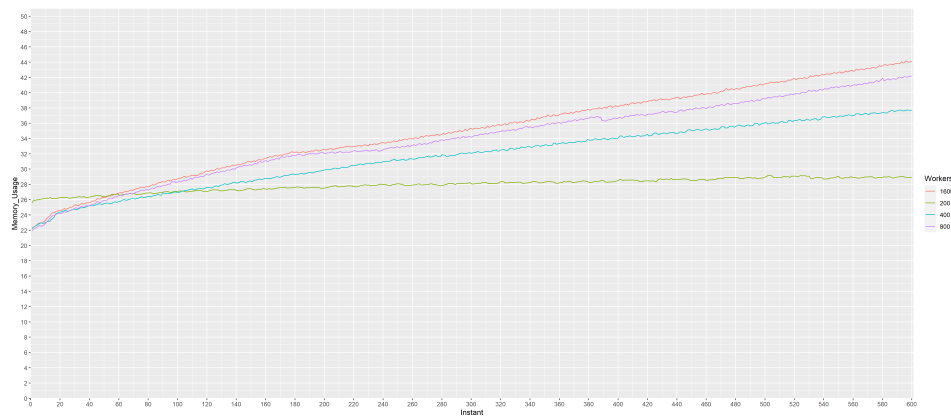


Figure A.24: Memory usage of the dedicated server solution during each round of the test.

The task involves 46 workers, who produce 12051 log requests. Each worker produces a mean value of 262 requests. The median value is 155. The minimum number of requests sent by a worker is 11, while the maximum is 1192. The mean task duration is 447 seconds and the median value is 230 seconds. The shortest try performed by a worker lasts for 2 seconds, while the longest lasts for 4588 seconds. The most frequent log event types collected are mouse movements, scrolls and clicks. Figure A.25 shows the distribution of event types detected by the component during the task. As one may expect, the most frequent event detected is mouse movement. Figure A.27 reconstructs the behavior of a worker that answers one of the questionnaires of the task deployed. The numbered points represent clicks, the blue line represents the movement trace, and the orange rectangles define a section of text selected. Figure A.26 reconstructs the behavior of a worker evaluating the first element of the HIT assigned.

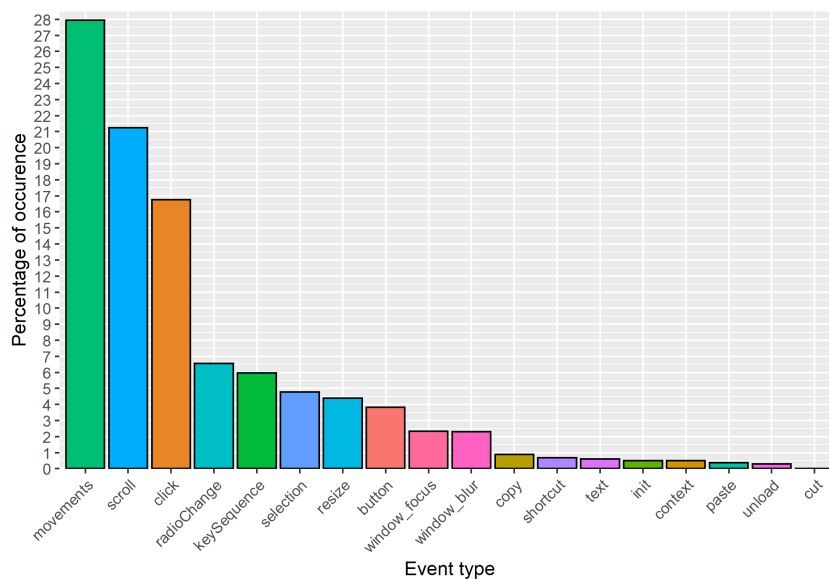


Figure A.25: Event type distribution of the pilot test deployed to gather real data using the Logger component.

A.3.5.5 Cost Estimation

Let us hypothesize a monthly stream of one million log messages where the average size of each message is the pilot test described in Section A.3.5.4. Each message thus contains, on average, a header of 550 bytes and a body of 529 bytes, for a total size of 1079 bytes. Let us recall the logging pipeline described in Figure A.20.

The first AWS service to consider is API Gateway. For an HTTP API, the service applies a fee³⁵ of \$1 for one million of request units up to 300 million a month, \$0.90 otherwise. The first million requests per month are free. Requests units are defined as having 512 KB of data. A request with 513 KB of data is counted as two separate units. The cost of

³⁵<https://aws.amazon.com/api-gateway/pricing/>

TASK INSTRUCTIONS
NO RETAKES

Q1 S1 S2 S3 S4 S5 S6 S7 S8 S9 S10 Outcome

Statement Nr. 1

Time left: 01:11

START

Statement: If I were not a Formula One pilot, I would be a football player.

Claimant: Abraham Lincoln

Instructions

Please, do the following three steps:

- Use the search engine below to retrieve some evidence about the truthfulness of the statement and select one of the retrieved URLs to justify the truthfulness value.
- Put as truthfulness value one of the six labels.
- Write a text justification. You can use your own words or, even better, copy/paste from the selected URL. You cannot use the selected search engine url as part of the justification.

Truthfulness

Insert your query

abraham lincoln 1

SEARCH

49 results found

[Abraham Lincoln - Wikipedia](#)
 https://en.wikipedia.org/wiki/Abraham_Lincoln
 Abraham Lincoln (/ ˈl ɪ ŋ ə ˈ l ɪ ŋ /; February 12, 1809 – April 15, 1865) was an American lawyer and statesman who served as the 16th president of the United States from 1861 to 1865. The assassination in 1865 led the nation through the American Civil War, and succeeded in preserving the Union, abolishing slavery, bolstering the federal government, and modernizing the U.S. economy. Select 1

[Abraham Lincoln | The White House](#)
 https://www.whitehouse.gov/about-the-white-house/presidents/abraham-lincoln/
 Abraham Lincoln became the United States 16th President in 1861, issuing the Emancipation Proclamation that declared forever free those slaves within the Confederacy in 1863. Select ○

[Abraham Lincoln - Facts, Birthday & Assassination - HISTORY](#)
 https://www.history.com/topics/us-presidents/abraham-lincoln
 Abraham Lincoln's Early Life. Lincoln was born on February 12, 1809 to Nancy and Thomas Lincoln in a one-room log cabin in Hardin County, Kentucky. His family moved to southern Indiana in 1816. Select 2

[President Abraham Lincoln Biography | American Battlefields](#)
 https://www.battlefields.org/learn/biographies/abraham-lincoln
 Abraham Lincoln, sixteenth President of the United States, was born near Hodgenville, Kentucky on February 12, 1809. His family moved to Indiana when he was seven and he grew up on the edge of the frontier. He had very little formal education, but read voraciously when not working on his father's farm. Select ○

[Abraham Lincoln Biography, History, and Facts](#)
 https://www.abrahamlincoln.net/
 Abraham Lincoln, born February 12, 1809, was the 16th President of the United States. Many historians and politicians believe he was the greatest president in terms of leadership, political acumen and character. Lincoln's biography is the stuff of legend. He rose from poverty to become a lawyer, leader and statesman, primarily ... Select ○

[Abraham Lincoln | Biography, Childhood, Quotes, Death ...](#)
 https://www.britannica.com/Biography/Abraham-Lincoln
 Abraham Lincoln, 16th U.S. president (1861–65), who preserved the Union during the American Civil War and brought about the emancipation of enslaved people in the United States. Among American heroes, Lincoln continues to have a unique appeal for his fellow countrymen and also for people of other lands. Select ○

[Abraham Lincoln - Quotes, Assassination & Height - Biography](#)
 https://www.biography.com/us-president/abraham-lincoln
 Abraham Lincoln was the 16th president of the United States. He preserved the Union during the U.S. Civil War and brought about the emancipation of slaves. Select ○

[Abraham Lincoln - History](#)
 https://kids.nationalgeographic.com/history/article/abraham-lincoln
 Abraham Lincoln was born in a log cabin in Kentucky on February 12, 1809, to parents who could neither read nor write. He went to school on and off for a total of about a year, but he educated himself by reading borrowed books. When Lincoln was nine years old, his mother died. Select ○

[Abraham Lincoln - Wikipedia](#)
 https://it.wikipedia.org/wiki/Abraham_Lincoln
 Abraham Lincoln, anche noto in italiano come Abramo Lincoln (Hodgenville, 12 febbraio 1809 – Washington, 15 aprile 1865), è stato un politico e avvocato statunitense. Fu il sedicesimo presidente degli Stati Uniti d'America, dal 4 marzo 1861 fino al suo assassinio avvenuto nel 15 aprile 1865. Select ○

[47 Interesting Facts About Abraham Lincoln - The Fact File](#)
 https://thefactfile.org/abraham-lincoln-facts/
 Abraham Lincoln, known for his determination and perseverance, is the most famous U.S. presidents in history. He became the 16th president of the U.S. on 4th March 1861. With these 47 interesting facts about Abraham Lincoln, let's learn about his life, career, politics, mission, philosophy, and death. Select ○

Items per page: 10 1 - 10 of 49 < >

Lie(The statement is not accurate and makes a ridiculous claim)

False(The statement is not accurate)

Mostly False(The statement contains an element of truth but ignores critical facts that would give a different impression)

Half True(The statement is accurate but needs clarification or additional information)

Mostly True(The statement contains an element of truth but ignores critical facts that would give a different impression)

True(The statement is accurate and there is nothing significant missing)

c. Your justification here (You can use your own words or, even better, copy/paste from the selected URL; you cannot use the selected search engine...)

This field is required

NEXT 3

Figure A.26: Worker behavior while evaluating a HIT's element reconstructed using the log data produced by the Logger component.

TASK INSTRUCTIONS NO RETAKES

1 2 3 4 5 6 7 8 9 10 11 12
Q1 S1 S2 S3 S4 S5 S6 S7 S8 S9 S10 Outcome

Please answer the following questions

What is your age range?

0-18

19-25

26-35

36-50

50-80

80

What is the highest level of school you have completed or the highest degree you have received?

High school incomplete or less

High school graduate or GED (includes technical/vocational training that doesn't towards college credit)

Some college (some community college, associate's degree)

Four year college degree/bachelor's degree

Some postgraduate or professional schooling, no postgraduate degree

Postgraduate or professional degree, including master's, doctorate, medical or law degree

Last year what was your total family income from all sources, before taxes?

Less than 10,000

10,000 to less than 20,000

20,000 to less than 30,000

30,000 to less than 40,000

40,000 to less than 50,000

50,000 to less than 75,000

75,000 to less than 100,000

100,000 to less than 150,000

150,000 or more

In general, would you describe your political views as

Very conservative

Conservative

Moderate

Liberal

Very liberal

In politics today, do you consider yourself a

Republican

Democrat

Independent

Something else

7 - Should the U.S. build a wall along the southern border. **8** - **9**

Agree

Disagree

No opinion either way

Should the government increase environmental regulations to prevent climate change?

Agree

Disagree

No opinion either way

Have you voted in the 2016 US presidential election?

Yes

No

Rather not answer

From what you know, do you agree or disagree with the Tea Party movement, or don't you have an opinion either way?

Agree

Disagree

No opinion either way

12 (*) you have to fill each field to proceed

Figure A.27: Worker behavior while answering a questionnaire reconstructed using the log data produced by the Logger component.

one million requests is thus expected to be \$1, given the average weight of the requests considered. Equation A.15 shows the detailed computation of each cost component.

$$\begin{aligned}
 \text{API Gateway} &= \$1 * (\text{Message Number} * \lceil \text{Average Message Size (KB)} / \text{Billing Factor} \rceil) \\
 &= \$1 * (1000000 * \lceil (1079/1024) / 512 \rceil) \\
 &= \$1 * 1 \text{ Billable Requests} = \$1 \quad (\$0 \iff \text{free tier}) \\
 &\quad \text{threshold: 300 million requests} \\
 &\quad \text{free tier: first million requests}
 \end{aligned}
 \tag{A.15}$$

The second service to consider is Simple Queue Service. The service applies a fee³⁶ of \$0.40 for one million of requests units a month, up to 100 billions. A message with a payload of size up to 64 KB is considered as a single request unit. The first million of requests per month is free. The service applies a fee also for data transfer, however inbound transfers are free of charge. The same holds also for outbound data towards a Lambda function in the same region. In light of this, the cost expected for data transfer is \$0. Equation A.16 shows the detailed computation of each cost component.

$$\begin{aligned}
 \text{Simple Queue Service} &= \$0.0000004000 * (\text{Multiplier} * \text{Requests Number (M)}) \\
 &= \$0.0000004000 * (\text{Multiplier} * (\text{Message Number} * \\
 &\quad (\text{Average Message Size (KB)} / \text{Max Payload Size}))) \\
 &= \$0.0000004000 * (1000000 * ((1079/1024)/64)) \\
 &= \$0.0000004000 * (1000000 * (0.016464)) \\
 &= \$0.0000004000 * 16464 \text{ Requests} \\
 &= \$0.0000004000 * 0 \text{ Billable Requests} = \$0 \\
 &\quad \text{threshold: 1 million requests} \\
 &\quad \text{free tier: first million requests}
 \end{aligned}
 \tag{A.16}$$

The third service to consider is Lambda. The pricing of a function depends on several factors such as the number of executions requested, the duration of each request and the amount of memory allocated for processing. The performance estimation described in Section A.3.5.3 allows understanding that in an ideal setting, each invocation of a Lambda function retrieves 100 requests and evaluates them in at most 5 seconds. On the other hand, in a not optimal setting, each function invocation retrieves up to 10 requests and executes them in around 500 milliseconds. The cost component³⁷ of the service to consider are thus three. The service applies a fee of \$0.20 for every million requests. Then, it applies a fee of \$0.0000166667 for every GB-second of execution of an x86-based architecture, for the first 6 billion GB-seconds. Lastly, it applies a fee of \$0.0000021 for every second of usage of 128 MB

³⁶<https://aws.amazon.com/sqs/pricing/>

³⁷<https://aws.amazon.com/lambda/pricing/>

of primary memory. In the ideal setting, 10000 function invocations lasting 5 milliseconds each are needed. In the not optimal setting, 10000000 invocations lasting 50 milliseconds each are needed. A final average estimation can thus be 5000000 function invocations lasting 250 milliseconds each for a total amount of 0.54 GB to process. Equation A.17 shows the detailed computation of each cost component.

$$\begin{aligned}
 \text{Lambda} &= \$0.0000166667 * \text{Max}(\text{Invocations Number} * \text{Avg. Proc. Time (s)} * \\
 &\quad (\text{Mem. Allocation (GB)}, 0) + \$0.0000002 * \text{Invocations Number} \\
 &= \$0.0000166667 * \text{Max}(500000 * (250/1000 * 0.125), 0) + \$0.0000002 * 500000 \\
 &= \$0.0000166667 * \text{Max}(125000 * 0.125), 0) + \$0.10 \\
 &= \$0.0000166667 * \text{Max}(15625 \text{ GB/s}, 0) + \$0.10 \\
 &= \$0.0000166667 * 15625 \text{ Billable GB/s} + \$0.10 = \$0.3603 \quad (\$0 \iff \text{free tier}) \\
 &\quad \text{thresholds: 1 million invocations, first 6 billions GB-Sec/Month} \\
 &\quad \text{free tier: subtract 40000 GB/s from Max() left hand side}
 \end{aligned}
 \tag{A.17}$$

The last service to consider is DynamoDB. Section A.3.3.5 describes the role of each cost component of DynamoDB. Equation A.18 shows the detailed computation of each cost component for the example considered.

$$\begin{aligned}
 \text{DynamoDB} &= \text{WRUs} + \text{RRUs} + \text{Data Storage} \\
 &= \$1.25 * (\text{Message Number} * [\text{Avg. Payload Size (KB)} / \text{Unit Amount}]) + \\
 &\quad \$0.25 * (0.5 * (\text{Message Number} * [\text{Avg. Payload Size (KB)} / \text{Unit Amount}])) + \\
 &\quad \$0.25 * (\text{Message Number} * \text{Avg. Payload Size (GB)}) \\
 &= \$0.00000125 * (1000000 * [(1079/1024)/1]) + \\
 &\quad \$0.00000025 * (0.5 * (1000000 * [(1079/1024)/4])) + \\
 &\quad \$0.25 * (1000000 * ((1079/1024)/1024)/1024) \\
 &= \$0.00000125 * (1000000 * 2 \text{ Billable WRUs}) + \\
 &\quad \$0.00000025 * (1000000 * 0.5 \text{ Billable RRUs}) + \\
 &\quad \$0.25 * (1 \text{ GB}) \\
 &= \$0.00000125 * 2000000 + \$0.00000025 * 500000 + \$0.25 * 1 \\
 &= \$2.50 + \$0.125 + \$0.25 = \$2.875 \quad (\$2.67 \iff \text{free tier}) \\
 &\quad \text{threshold: 1 million WRUs, 1 million RRUs} \\
 &\quad \text{note: on-demand capacity mode, standard table class} \\
 &\quad \text{free tier: 25 GB/Month data storage}
 \end{aligned}
 \tag{A.18}$$

Equation A.19 estimates the cost required to maintain the infrastructure of the Logger

component of Crowd_Frame hypothesizing that one million of log requests is processed each month. After a year of usage, the total cost of the Logger component would thus be roughly around \$51.

$$\begin{aligned}\text{Logger} &= \text{API Gateway} + \text{SQS} + \text{Lambda} + \text{DynamoDB} \\ &= \$1 + \$0 + \$2.875 + \$0.3603 \\ &= \$4.2353\end{aligned}\tag{A.19}$$

A.3.6 Log Events

List of events monitored by the Logger component of Crowd_Frame described in Section A.3.5. The payloads shown in the following are encapsulated in the detail object of the base payload shown in Listing A.14.

A.3.6.1 Context

The log message of `"type": "context"` contains information about the user agent and the IP address of the worker. It is the only not containing the section field.

```
1 | {
2 |   "ua": "string",
3 |   "ip": "string"
4 | }
```

A.3.6.2 Mouse Movements

The log message of `"type": "movements"` describes a mouse movement performed by the worker. The event is detected every 100 ms. The timestamp and (x, y) coordinates are mapped to a dictionary and buffered. The dictionaries contained in the buffer are pushed in an array of the details after a dwell time of 500 ms. Each dictionary is pushed in the points array.

```
1 | {
2 |   "section": "string",
3 |   "points": [
4 |     {
5 |       "timeStamp": "string",
6 |       "x": "integer",
7 |       "y": "integer"
8 |     },
9 |     {
10 |      "timeStamp": "string",
```

```
11     "x": "integer",
12     "y": "integer"
13   },
14   ...
15 ]
16 }
```

A.3.6.3 Mouse Clicks

The log message of `"type": "click"` describes a mouse click performed by the worker. The event is detected when the left or right mouse button is pressed. The event is detected after a fixed debounce time to prevent spamming too many log requests. The timestamps of the first and last clicks of each click sequence are stored. Furthermore, the (x, y) coordinates, the DOM element targeted and the number of clicks in the sequence are logged.

```
1  {
2  "section": "string",
3  "mouseButton": "right || left",
4  "startTime": "string",
5  "endTime": "string",
6  "x": "integer",
7  "y": "integer",
8  "target": "string",
9  "clicks": "integer"
10 }
```

A.3.6.4 Button Click

The log message of `"type": "button"` describes a mouse click performed by the worker on a button DOM element. This event has a debounce time like the base click event, but different information is extracted. The information include the DOM button targeted, the timestamp of the click, and the (x, y) coordinates.

```
1  {
2  "section": "string",
3  "timestamp": "string",
4  "button": "string",
5  "x": "integer",
6  "y": "integer"
7  }
```

A.3.6.5 Shortcuts

The log message of `"type": "shortcut"` a keystroke combination used by the worker. Only combinations that involve the keys CTRL or ALT are monitored.

Key combinations corresponding to shortcuts are monitored. From the event key pressed for the shortcut are extracted:

- `"ctrl": "boolean"` set to true if the CTRL or CMD key is pressed.
- `"alt": "boolean"` set to true if the ALT key is pressed.
- `"key": "string"` contains the value of the key pressed within the shortcut detected.

```
1 {
2   "section": "string",
3   "timestamp": "string",
4   "ctrl": "boolean",
5   "alt": "boolean",
6   "key": "string"
7 }
```

A.3.6.6 Keypress

The log message of `"type": "keySequence"` describes a sequence of keystrokes performed by the worker. Every keypress is stored as a dictionary in a buffer array. Each dictionary contains the timestamp and the key pressed. The full sentence is reconstructed. The event handling completes after a dwell time of 1 second.

```
1 {
2   "section": "string",
3   "keySequence": [
4     {
5       "timeStamp": "string",
6       "key": "string"
7     },
8     {
9       ...
10    }
11    ...
12  ],
13   "sentence": "string"
14 }
```

A.3.6.7 Selection

The log message of `"type": "selection"` describes a selection operation performed by the worker. The information logged includes the start and end timestamps and the content of the selection.

```
1 | {
2 |   "section": "string",
3 |   "startTime": "string",
4 |   "endTime": "string",
5 |   "selected": "string"
6 | }
```

A.3.6.8 Before Unload, Focus, and Blur

The log message of `"type": "unload || window_focus || window_blur"` describes the last log requests produced when the worker closes the page of the task or performs or the window loses the focus or blurs. The information includes only the corresponding timestamp.

```
1 | {
2 |   "section": "string",
3 |   "timestamp": "string"
4 | }
```

A.3.6.9 Scroll

The log message of `"type": "scroll"` describes a scroll operation performed by the worker. Similarly to mouse clicks and movements, a debounce time of 300ms is used to prevent spamming too many log requests. The information includes the start and end timestamps and the coordinates of the top left corner from which the scroll started. Scroll has a specific event listener and, like move movements, it needs a debouncing factor to prevent "spamming". For this event, the debounce time is set to 300 ms and the start timestamp, end timestamp, and (x, y) coordinates of the top left corner, are saved for logging.

```
1 | {
2 |   "section": "string",
3 |   "startTimestamp": "string",
4 |   "endTimestamp": "string",
5 |   "x": "integer",
6 |   "y": "integer"
7 | }
```

A.3.6.10 Resize

The log message of `"type": "resize"` is detected when the worker resizes the window of the task page. The information includes the start and end timestamp and the updated window size.

```
1 {
2   "section": "string",
3   "width": "integer",
4   "height": "integer",
5   "scrollWidth": "integer",
6   "scrollHeight": "integer",
7   "timestamp": "string"
8 }
```

A.3.6.11 Copy, Cut, and Paste

The log message of `"type": "copy || cut || paste"` is detected when the worker cuts/copies and pastes some content during the task. The information includes the timestamp of the event. The attribute `target` refers to the DOM element targeted by a copy or cut event. The attribute is substituted with the `text` one when a paste event is detected and contains the text pasted.

```
1 {
2   "section": "string",
3   "timestamp": "string",
4   "target": "string"
5 }
```

A.3.6.12 Text Input Backspace and Blur

The log message of `"type": "text"` is detected when the backspace key is pressed inside a text input. The event is detected also then the input is blurred. The information includes the timestamp and the text contained in the input.

```
1 {
2   "section": "string",
3   "timestamp": "string",
4   "text": "string"
5 }
```

A.3.6.13 Radio Group Input

The event of `"type": "radioChange"` is detected when the workers chooses a new value within a radio button control. The information includes the timestamp and the new value of the radio button.

```
1 | {
2 |   "section": "string",
3 |   "timestamp": "string",
4 |   "group": "string",
5 |   "value": "string"
6 | }
```

A.3.6.14 Search Engine Queries and Results

The event of `"type": "query || queryResults"` is detected when the worker queries the search engine and when results are retrieved. The information includes the text of the query in the former case, and the list of URL retrieved in the latter.

```
1 | {
2 |   "section": "string",
3 |   "query": "string"
4 | }
```

```
1 | {
2 |   "section": "string",
3 |   "urlArray": [
4 |     ...
5 |   ]
6 | }
```

A.3.6.15 System Usage

The GitHub repository contains detailed and updated instructions.³⁸ Four main prerequisites are required to start using Crowd_Frame. In more detail, these prerequisites are the AWS Command Line Interface³⁹ and distributions of Node.js⁴⁰ and Python.⁴¹ Docker⁴² may be optionally needed. Yarn⁴³ is used to manage the software dependencies.

³⁸https://github.com/Miccighe1/Crowd_Frame#readme

³⁹<https://docs.aws.amazon.com/cli/latest/userguide/getting-started-install.html>

⁴⁰<https://nodejs.org/it/download/>

⁴¹<https://www.python.org/downloads/>

⁴²<https://docs.docker.com/get-docker/>

⁴³<https://yarnpkg.com/>

A.4 Getting Started

There are 15 steps to follow to successfully initialize `Crowd_Frame` and the overall infrastructure. The first 13 steps must be performed only once. Then, the task requester can vary the configuration and repeat the steps #14 and #15.

1. Create a new Amazon AWS account.
2. Create a new IAM User⁴⁴ using a custom name such as `your_iam_user`.
3. Attach the `AdministratorAccess` policy (cfr. Listing A.15).
4. Generate a new access key pair.
5. Store the access key in the `credentials.json` file (cfr. Listing A.16).
6. Clone the repository in the local filesystem.
7. Enable the Yarn global library using the command: `corepack enable`.
8. Move to the repository's folder: `cd ~/path/to/project`.
9. Install the dependencies using the command: `yarn install`.
10. Move to the data folder using the command: `cd data`.
11. Create the environment file `.env` at path `your_repo_folder/data/.env`.
12. Add to the environment file the subset of required variables (cfr. Listing A.17).
13. Install the Python packages required (cfr. Listing A.18)
14. Run the Python script `init.py`. The script will:
 - read the environment variables;
 - setup the whole AWS infrastructure;
 - generate a sample task configuration;
 - deploy the sample source files on the public bucket.
15. Visit the task deployed using the link in the format:
 - `https://your_deploy_bucket.s3.your_aws_region.amazonaws.com/your_task_name/your_batch_name/index.html`

`Crowd_Frame` interacts with diverse Amazon Web Services to deploy crowdsourcing tasks, store the data produced and so on, as described in Section A.3. Each service used falls within the AWS Free Tier⁴⁵ program. The task requester can set the budget limit using the `budget_limit` environment variable. Thus, the usage of the services will be blocked if/when such a limit is surpassed.

⁴⁴<https://docs.aws.amazon.com/IAM/latest/UserGuide/id-users.html>

⁴⁵<https://aws.amazon.com/free/>

```
1 | {
2 |   "Version": "2012-10-17",
3 |   "Statement": [
4 |     {
5 |       "Effect": "Allow",
6 |       "Action": "*",
7 |       "Resource": "*"
8 |     }
9 |   ]
10 | }
```

Listing A.15: Input data specification for a project on Toloka.

```
1 | [your_iam_user]
2 | region = your_region
3 | aws_access_key_id=your_key
4 | aws_secret_access_key=your_secret
```

Listing A.16: Sample credentials.json file to store IAM user access key.

```
1 | mail_contact=your_email_address
2 | budget_limit=your_usd_budget_limit
3 | task_name=your_task_name
4 | batch_name=your_batch_name
5 | admin_user=your_admin_username
6 | admin_password=your_admin_password
7 | server_config=none
8 | aws_region=your_aws_region
9 | aws_private_bucket=your_private_bucket_name
10 | aws_deploy_bucket=your_deploy_bucket_name
```

Listing A.17: Subset of environment variables required by Crowd_Frame.

A.4.1 Environment Variables

Table A.9 describes each environment variable that can be set in the environment file to customize Crowd_Frame behavior.


```

1 boto3==1.26.37
2 ipapi==1.0.4
3 ipinfo==4.4.2
4 mako==1.2.4
5 chardet==5.1.0
6 docker==6.0.1
7 python-dotenv==0.21.0
8 rich==12.6.0
9 tqdm==4.64.1
10 scipy==1.9.3
11 pycountry==22.3.5
12 numpy==1.24.1
13 pandas==1.5.2
14 toloka-kit==1.0.2
15 python-on-whales==0.55.0
16 beautifulsoup4==4.11.1
17 aiohttp==3.8.3
18 datefinder==0.7.3

```

Listing A.18: Python packages required to initialize Crowd_Frame.

Table A.9: Environment variables used to customize Crowd_Frame.

Variable	Description	Required	Value
profile_name	Name of the IAM profile created during step 2 (cfr. Section A.4). If unspecified, the variable uses the value: default.	No	your_iam_user
mail_contact	Contact mail to receive AWS budgeting related communications	Yes	Valid email address
platform	Platform used to publish the crowdsourcing task. Set it to none to recruit the workers manually.	Yes	none, mturk, prolific or toloka
budget_limit	Maximum monthly amount of money allowed to operate in USD\$; e.g., 5.0.	Yes	Positive float number
task_name	Identifier of the crowdsourcing task.	Yes	Any string
batch_name	Identifier of the current batch.	Yes	Any string
batch_prefix	Prefix of the identifiers of one or more task's batches. The variable can be used to filter the final result set.	No	Any string

Continues in the next page

admin_user	Username of the admin user. Allows unlocking the Generator.	Yes	Any string
admin_password	Password of the admin user. Allows unlocking the Generator.	Yes	Any string
aws_region	Region of the AWS account.	Yes	Valid region identifier; e.g., us-east-1 ⁴⁶
aws_private_bucket	Name of the private S3 bucket in which store task configuration and data	Yes	String unique across AWS account
aws_deploy_bucket	Name of the public S3 bucket in which deploy the source code	Yes	String unique across AWS account
server_config	Used to specify where the worker behavior logging interface is. Set it to aws to deploy the AWS-based infrastructure. Set it to custom to provide a custom logging endpoint. Set it to none if logging worker behavior is not needed.	Yes	aws, custom or none
enable_crawling	Enables the crawling of the results retrieved by the search engine.	No	true or false
enable_solver	Allows to deploy the HITs solver locally. Allows to provide a set of elements to automatically allocated into a set of HITs. Requires the usage of Docker. Warning: the feature is <i>experimental</i> .	No	true or false
prolific_completion_code	Prolific study completion code. Provide here the code if you recruit crowd workers via Prolific . Required if the platform chosen is prolific.	No	Valid Prolific completion code
toloka_oauth_token	Token to access Toloka 's API. Required if the platform chosen is toloka.	No	Valid Toloka OAuth token
ip_info_token	API Key to use ipinfo.com tracking functionalities.	No	Valid IP Info API key

Continues in the next page

⁴⁶<https://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/Concepts.RegionsAndAvailabilityZones.html>

ip_geolocation_api_key	API Key to use ipinfo.com tracking functionalities.	No	Valid IP Info API key
ipapi_api_key	API Key to use ipgeolocation.info tracking functionalities.	No	Valid IP Geolocation API key
user_stack_token	API Key to use userstack.com tracking functionalities.	No	Valid Userstack API key
bing_api_key	API Key to use BingWebSearch search provider.	No	Valid Bing API Web Search API key
fake_json_token	API Key to use FakerWebSearch search provider. Returns dummy responses useful to test the search engine.	No	Valid fakeJSON.com API key

A.4.2 Build Output

Each execution of the initialization script of Crowd_Frame described in Section A.4 populates a build folder on the local filesystem, at path: `cd ~/path/to/project/build/`. The folder may contain up to 6 subfolders, depending on the crowdsourcing platforms used. Table A.10 provides a general description of the content of each subfolder.

Table A.10: Folder structure of the output of a Crowd_Frame build.

Sub Folder	Description
build/task/	Contains the configuration of the task to deploy.
build/config/	Contains the encrypted credentials used to unlock the Generator component.
build/environments/	Contains the development and production environments.
build/mturk/	Contains three files needed to publish the task using Amazon Mechanical Turk.
build/toloka/	Contains six files needed to publish the task using Toloka.
build/skeleton/	Contains a interface between the HITs and the application and a file used to implement quality checks.
build/deploy/	Contains the source files of the task to deploy.

A.4.2.1 build/task/

The folder contains the 7 configuration files of the task deployed. The Generator component described in Section A.3.2 populates one of these files along each step. In other words, each task deployed using Crowd_Frame is configured by using 8 special JSON files. Table A.11 provides a general description of the content of each configuration file.

Table A.11: Configuration files of a task deployed using Crowd_Frame.

File	Description
<code>hits.json</code>	Contains the whole set of HITs of the task.
<code>questionnaires.json</code>	Contains the definition of each questionnaire of the task.
<code>dimensions.json</code>	Contains the definitions of each evaluation dimension of the task.
<code>instructions_general.json</code>	Contains the general instructions of the task.
<code>instructions_evaluation.json</code>	Contains the evaluation instructions of the task.
<code>search_engine.json</code>	Contains the configuration of the custom search engine.
<code>task.json</code>	Contains several general settings of the task.
<code>workers.json</code>	Contains settings concerning worker access to the task.

A.4.2.2 `build/environments/`

The folder contains the development and production environments of Crowd_Frame. Each environment contains the values of the variables shown in Table A.9 along with additional data. The environments are overwritten during each execution of the initialization script to reflect changes in the environment variables.

A.4.2.3 `build/config/`

The folder contains the encrypted credentials used to unlock the access to the Generator component. There is a single file named `admin.json` that contains a hash generated using the HMAC [36] scheme. The hash is built using the values of the `admin_user` and `admin_password` variables shown in Table A.9.

A.4.2.4 `build/skeleton/`

The folder contains an interface called `Document` in a file named `document.ts`. Such an interface is a bridge between the Angular application and the HITs configured. The interface is generated during the execution of the initialization script if the HITs' attributes change. The folder contains also a file named `goldChecker.ts`. It provides a static method that a developer can implement to perform a custom quality check when enabled for one or more evaluation dimensions.

A.4.2.5 `build/deploy/`

The folder contains the three source files built by Angular that the initialization script deploys on the public S3 bucket. Table A.12 describes each of these source files. The role of the public bucket is simply making them available publicly on the Internet. The whole client-side code base is included. This means that a developer can deploy the files on a private server. The task will be initialized without issues.

Table A.12: Source files of a task deployed using Crowd_Frame.

File	Description
index.html	Markup of the task deployed.
styles.css	Styling of the task deployed.
scripts.js	Client side code of the task deployed.

A.4.2.6 build/mturk/

The folder contains the wrapper initialized to be deployed on Amazon Mechanical Turk in a file named `index.html`. The wrapper is initialized for the current task starting from a general model contained in the file named `model.html`. The initialization script uses a template engine called Mako⁴⁷ to initialize the wrapper. The folder contains also a file named `tokens.csv`. It contains the input and output tokens to be provided to the platform. Such tokens are those present in the file containing the definition of the whole set of HITs.

A.4.2.7 build/toloka/

The folder contains the wrapper initialized to be deployed on Toloka, the input and output data specification and the tokens to be provided. Table A.13 describes each file. It is generated using Mako from a model as for Amazon Mechanical Turk. However, the final result is split into three different files, due to Toloka's requester interface. The files `interface.html`, `interface.css`, and `interface.js` contain respectively the markup, the styling and the client-side code of the wrapper. The two JSON files provide the input and output data specification to be used for the task to be deployed. The TSV file contains the input and output tokens to be provided to the platform.

Table A.13: Content of the build folder to deploy a task on Toloka .

File	Description
index.html	Markup of the wrapper.
styles.css	Styling of the wrapper.
scripts.js	Client side code of the wrapper.
input_specification.json	Input data specification.
output_specification.json	Output data specification.
tokens.tsv	Tokens to be provided to the platform.

A.4.3 Task Configuration

The Generator component must be accessed to configure the crowdsourcing task deployed. This involves 4 steps:

⁴⁷<https://www.makotemplates.org/>

1. Open the administrator panel by appending the suffix `?admin=true` to the task's URL (cfr. step #15, Section A.4).
2. Click the **Generate** button to open the login form.
3. Input the admin credentials set in the corresponding environment variables (cfr. Table A.9).
4. Proceed through each configuration step and upload the final configuration.

The final configuration can be uploaded using the **Upload** button. Table A.1 provides a summary of each task configuration step. The initialization script synchronizes the local configuration and the remote one bidirectionally by downloading (uploading) the most recent.

A.4.4 HITs Allocation

The HITs for a crowdsourcing task designed and deployed using `Crowd_Frame` must be stored in a special JSON file. Such a file can be manually uploaded when configuring the crowdsourcing task itself, as described in Section A.3.2. The file must comply with a special format that satisfies 5 requirements:

1. There must be an array of HITs (also called *units*).
2. Each HIT must have a unique input/output token attribute pair.
3. The number of elements to assess must be specified for each HIT.
4. Each element must have an attribute named `id`.
5. Each element can have an arbitrary number of attributes.

Listing A.19 shows a valid set composed of a single HIT to be assessed within a crowdsourcing task configured using `Crowd_Frame`.

A.4.4.1 Manual Approach

The requester can build manually the set of HITs compliant with the format required by `Crowd_Frame`. Initially, the requester chooses an attribute whose values split the dataset into different classes. The core idea is to build pools of elements to allocate, one for each class. Four parameters are thus established. These parameters are the total number of elements to allocate in the whole set of HITs, the number of elements that each HIT must contain, the number of elements to allocate for each class, and the number of repetitions for each element. The pools of elements must thus be updated to include all the repetitions required. Each HIT is then built using a loop. It is useful to define a support function. The core idea is to sample the required number of elements for each class until a sample without duplicates is obtained. The elements are then removed from the pool of those still available if the condition is satisfied. The total number of HITs required depends on the parameters previously established. The lists of elements allocated in HITs can be serialized

```
1  [
2    {
3      "unit_id": "unit_0",
4      "token_input": "ABCDEFGHILM",
5      "token_output": "MNOPQRSTUVWXYZ",
6      "documents_number": 1,
7      "documents": [
8        {
9          "id": "identifier_1",
10         "text": "Lorem ipsum dolor sit amet"
11       }
12     ]
13   }
14 ]
```

Listing A.19: Valid set of one HIT for a task designed and deployed using Crowd_Frame.

for later reference. Using such an allocation matrix, the requester can finally build the set of HITs in the format required.

Algorithm A.1 provides a pseudocode that further details the allocation procedure. Let us hypothesize a requester that wants to determine the number m of HITs that they will publish on the crowdsourcing platform. Algorithm A.2 details the `SINGLEHIT(...)` sub-procedure used by the main algorithm to sample a set of elements without duplicates. The sample obtained is used to build a single HIT of the whole set. Let us hypothesize a requester that wants to allocate n elements in HITs made of k positions. Each element is repeated in p different HITs. The final number of HITs required is $m = (n * p)/k$.

A.4.4.2 Automatic Approach

Crowd_Frame provides a solution to allocate automatically the elements to evaluate in a set of HITs. It is experimental and works only when using Crowd_Frame within the local filesystem. Future versions of the software will consolidate and generalize such a feature. Ceschia et al. [64] optimize HITs construction by providing its formal definition and by applying a local search method to solve the formalized problem. Crowd_Frame allows deploying an implementation of their solver and provides a way to communicate with it. Docker needs to be installed in the local system since the usage of a container is required to allow software and solver to communicate. The container contains deployed using Docker contains two services. One of these services provides the implementation of the solver itself, while the other provides a reverse proxy based on the Nginx⁴⁸ web server. The reverse proxy forwards HTTP messages to the solver. The solver processes the messages and responds. Figure A.28 shows a deployment diagram which describes the interaction between Crowd_Frame, solver, and reverse proxy.

The requester can enable the feature using the `enable_solver` environment variable

⁴⁸<https://www.nginx.com/>

Algorithm A.1 Procedure to allocate a dataset into HITs using the format required

```

1: elementsFiltered ← FILTERELEMENTS(attribute, valuesChosen)
2: classes ← valuesChosen
3: pools ← LIST()
4: forEach class ∈ classes do
5:   elementsClass ← FINDERELEMENTS(elementsFiltered, class)
6:   pool ← UNIQUE(elementsClass)
7:   pools.APPEND(pool)
8: end for
9: totalElements ← LEN(elementsFiltered)
10: classElementsNumber ← LEN(classes)
11: hitElementsNumber ← k
12: repetitionsElement ← p
13: forEach pool ∈ pools do
14:   pool ← EXTENDPOOL(repetitionsElement)
15: end for
16: poolsDict ← MERGEPOOLS(pools, classes)
17: hits ← LIST()
18: forEach index ∈ RANGE((totalElements * repetitionsElement)/hitElementsNumber) do
19:   hitSample ← SINGLEHIT(poolsMerged)
20:   hitSample ← SHUFFLE(hitSample)
21:   hits.APPEND(hitSample)
22: end for
23: hits.SERIALIZE(pathAssignments)
24: hitsFinal ← LIST()
25: forEach hit ∈ hits do
26:   index ← INDEX(hit)
27:   unitId ← CONCAT("unit_", index)
28:   tokenInput ← RANDOMSTRING(11)
29:   tokenOutput ← RANDOMSTRING(11)
30:   hitObject ← BUILDJSON(unitId, tokenInput, tokenOutput, hitElementsNumber)
31:   forEach indexElem ∈ RANGE(hitElementsNumber) do
32:     hitObject["documents"] ← hits[indexElem]
33:   end for
34:   hitsFinal.APPEND(hitObject)
35: end for
36: hitsFinal.SERIALIZE(pathHits)

```

Algorithm A.2 Procedure to sample elements without duplicates for a single HIT

```

1: containsDuplicates ← True
2: while containsDuplicates do
3:   sample ← LIST()
4:   forEach class ∈ classes do
5:     forEach indexClass ∈ RANGE(classElementsNumber) do
6:       element ← RANDOM(poolsDict[class])
7:       sample.APPEND(element)
8:     end for
9:   end for
10:  if CHECKDUPLICATES(sample)==False then
11:    containsDuplicates ← False
12:  end if
13: end while
14: forEach s ∈ sample do
15:   forEach c ∈ classes do
16:    if s ∈ pool[c] then
17:      pool[c].REMOVE(s)
18:    end if
19:   end for
20: end for
21: return sample

```

shown in Table A.9. They can take advantage of the solver while configuring the task using the Generator component, during the sixth step of the configuration (cfr. Table A.1). The first step required to create the input data required by the solver involves uploading the elements to be allocated into a set of HITs. Each element must share the same set of attributes and the overall set must be provided in the form of a JSON array. In other words, the requester can upload the value of the `documents` object shown in Listing A.19, without writing any token or `unit_id`. Then, the requester can configure three parameters concerning the allocation. They thus configure the number of workers that evaluate each element and the overall number of workers among which the elements must be allocated. Lastly, the requester chooses the subset of attributes used to categorize the elements across different HITs. The requester must also indicate how many elements must be assigned to each worker for every possible value of the category chosen. For each category/element number pair the system verifies whether the two values are compatible. The minimum number of workers needed to evaluate the whole set of HITs is thus computed if the verification is successful. The requester can increase such a number as they prefer.

To provide an example of when such a verification can fail, let us hypothesize a requester who chooses as category an attribute named `A1` which has 2 different values among the elements to be evaluated. The requester requires that each worker evaluates 2 elements for each attribute's value. Then, a second attribute named `A2` is also chosen, which has 3 different values. The requester requires that each worker evaluates 1 element for each attribute's value. This means that according to the attribute `A1` each worker evaluates 4 elements, while according to `A2` each worker evaluates 3 elements. Such a selection of

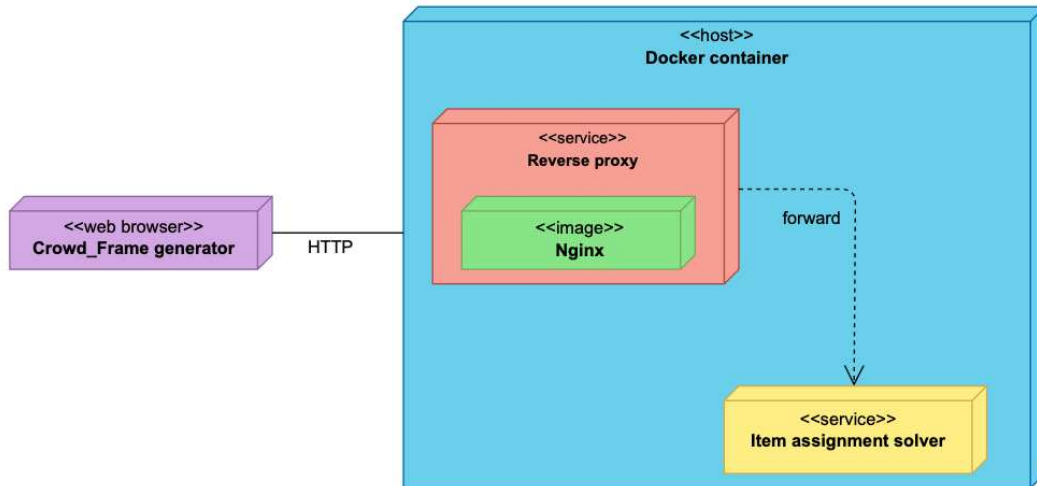


Figure A.28: Deploy diagram of the infrastructure used to allow Crowd_Frame and the solver communicating.

values is not allowed.

Figure A.29 shows a sample configuration for the subset of 120 statements sampled from PolitiFact [436] by Roitero et al. [361] and evaluated in their task. Listing A.20 shows a fragment of such elements. The requester, hence, uploads a JSON file containing 120 elements to allocate. They choose that each element must be assigned to 10 different workers. The attribute party is selected as a category. Each worker must evaluate 6 elements for each of the 2 values of the category. In other words, each worker must evaluate 12 different elements. The verification steps thus enforce a minimum number of 100 workers to recruit. The Generator allows selecting as categories only the attributes which are balanced with respect to the number of documents. In other words, those attributes are repeated across the same number of elements. Such a design choice is needed to provide input data to the solver compliant with the formalization of Ceschia et al. [64]. The request is now ready to send the request to the solver, which computes the allocation and returns a solution. Listing A.21 shows an example of the final allocation. The solution returned is then used by Crowd_Frame to build the corresponding set of HITs compliant with the format shown in Listing A.19.

A.4.5 Quality Checks

Crowd_Frame provides a way to manually define custom quality checks triggered for each evaluation dimension when the corresponding setting is enabled in the configuration. A custom quality check is obtained by providing an implementation for the static method `performGoldCheck`, as described in Section A.4.2. A custom quality check is triggered only for certain elements of HIT, with respect to a subset of the evaluation dimensions. An element can be marked for the quality check by prepending the string `GOLD` to its `id` attribute. Listing A.22 shows a single HIT where its second element is marked for the quality check.

Documents

SELECT DOCUMENTS FILE Filename: fake_news_docs.json, Documents detected: 120, Filesize: 47 Kb

Minimum documents repetitions in hits
10

Originated From	Number of values: 5	social media: 24 documents ad: 24 documents a speech: 24 documents a news relea.: 24 documents an interview: 24 documents	<input type="checkbox"/> Use as category	Worker assignments 0
Party	Number of values: 2	REP: 60 documents DEM: 60 documents	<input checked="" type="checkbox"/> Use as category	Worker assignments 6
Source	Number of values: 1	Politifact: 120 documents	<input type="checkbox"/> Use as category	Worker assignments 0

CHECK SELECTION **RESET SELECTION**

Workers number
100

SEND REQUEST TO THE SOLVER

Clone a deployed batch
Select configuration

Figure A.29: Sample configuration for the solver used to automatically allocate elements to be evaluated in a set of HITs.

```

1  [
2    {
3      "name": "REP_HALFTRUE_doc5",
4      "statement": "↵ The city of Houston now has more debt per capita than California." ↵
5      "claimant": "Rick Perry",
6      "date": "2010",
7      "originated_from": "ad",
8      "id_par": "1796.json",
9      "job": "Governor",
10     "party": "REP",
11     "source": "Politifact"
12   },
13   ...
14 ]

```

Listing A.20: Fragment of elements to be evaluated in the task published by Roitero et al. [361].

Listing A.23 shows the default implementation of the method generated by the initialization script. The document array provides the set of elements marked for the quality check. The answers array contains the answers provided by the worker for the evaluation

```

1  {
2  "finished": true,
3  "runner": "BSA",
4  "solution": {
5    "Instance_id": "1658421484781",
6    "Used_workers": 100,
7    "Workers": [
8      {
9        "Assignments": [
10       "REP_FALSE_doc8", "DEM_TRUE_doc5", "REP_BARELYTRUE_doc6",
11       "DEM_MOSTLYTRUE_doc5", "REP_TRUE_doc7", "DEM_LIE_doc9"
12     ], "Id": "W0"
13   },
14   {
15     "Assignments": [
16       "REP_TRUE_doc5", "DEM_TRUE_doc7", "REP_LIE_doc1",
17       "DEM_TRUE_doc1", "REP_HALFTRUE_doc2", "DEM_FALSE_doc6"
18     ], "Id": "W1"
19   },
20   ...
21   ]
22 },
23 "completed": "2022-07-21 16:37:56",
24 "started": "2022-07-21 16:37:56",
25 "submitted": "2022-07-21 16:37:56",
26 "task_id": "5447375273499815497"
27 }

```

Listing A.21: Fragment of the allocation built automatically using the solver integrated with Crowd_Frame.

dimensions that require the quality check. The check must be implemented among the two comments shown.

A.4.6 Local Development

Crowd_Frame provides a way to manually edit and test the configuration locally, without deploying the overall infrastructure. Enabling the local development capability involves 4 steps:

1. Move to the environments folder (Section A.4.2):
 - `cd your_repo_folder/data/build/environments/.`
2. Open the development environment file:
 - `environment.ts`

```
1  [
2    {
3      "unit_id": "unit_0",
4      "token_input": "ABCDEFGHILM",
5      "token_output": "MNOPQRSTUVWXYZ",
6      "documents_number": 1,
7      "documents": [
8        { "id": "identifier_1", "text": "Lorem ipsum dolor sit amet" },
9        { "id": "GOLD-identifier", "text": "Lorem ipsum dolor sit amet" }
10     ]
11   }
12 ]
```

Listing A.22: Valid set of one HIT with two elements, where one is used within a custom quality check.

3. Edit the variable `configuration_local` by setting the value `true`.
4. Run the command `ng serve`.

The requester may also skip the first three steps and simply run the `ng serve` command. In such a case, the source code can be modified locally while using the remote configuration to initialize the application. Listing A.24 shows a valid development environment file that allows testing the task configuration locally. It must be noted that every execution of the `init.py` script overwrites the environment files. Thus, the local testing capability must be enabled again if the requester deploys the task by running the script itself.

A.5 Task Performing

Publishing a crowdsourcing task configured using `Crowd_Frame` involves choosing the platform to recruit the human workforce, even though the requester can also manually recruit each worker needed. The process to publish and start the task deployed is slightly different depending on such a choice.

A.5.1 Manual Recruitment

A task requester that aims to manually recruit each worker to perform the task deployed must:

1. Set the environment variable `platform` (cfr. Table A.9) using the value `none`.
2. Generate ad assign each worker an alphanumeric identifier, such as `randomWorkerId`.
3. Append the identifier generated as a GET parameter to the task deploy link:
 - `?workerId=randomWorkerId`

```

1  export class GoldChecker {
2    static performGoldCheck(goldConfiguration : Array<Object>) {
3      let goldChecks = new Array<boolean>()
4      /* If there are no gold elements there is nothing to be checked */
5      if(goldConfiguration.length<=0) {
6        goldChecks.push(true)
7        return goldChecks
8      }
9      for (let goldElement of goldConfiguration) {
10       /* Element attributes */
11       let document = goldElement["document"]
12       /* Worker's answers for each gold dimensions */
13       let answers = goldElement["answers"]
14       /* Worker's notes*/
15       let notes = goldElement["notes"]
16       let goldCheck = true
17       /* CONTROL IMPLEMENTATION STARTS HERE */
18       /* The check for the current element holds if goldCheck remains
19        ↪ true */
19       ...
20       /* CONTROL IMPLEMENTATION ENDS HERE */
21       /* Push goldCheck inside goldChecks array for the current gold
22        ↪ element */
22       goldChecks.push(goldCheck)
23     }
24     return goldChecks
25   }
26 }
27

```

Listing A.23: Default implementation of the static method that performs custom quality checks in Crowd_Frame.

4. Provide each worker with the link to the task deployed.

- https://your_deploy_bucket.s3.your_aws_region.amazonaws.com/your_task_name/your_batch_name/index.html?workerID=randomWorkerId

5. Wait for task completion.

Steps #2 and #3 can be skipped because the task URL can be provided to a worker also without manually adding an identifier. In such a case, Crowd_Frame will automatically generate it.

A.5.2 Amazon Mechanical Turk

A task requester that aims to recruit each worker using Amazon Mechanical Turk must:

```
1  export const environment = {
2    production: false,
3    configuration_local: true,
4    platform: "mturk",
5    taskName: "your_task_name",
6    batchName: "your_batch_name",
7    region: "your_aws_region",
8    bucket: "your_private_bucket",
9    aws_id_key: "your_aws_key_id",
10   aws_secret_key: "your_aws_key_secret",
11   prolific_completion_code: false,
12   bing_api_key: "your_bing_api_key",
13   fake_json_key: "your_fake_json_key",
14   log_on_console: false,
15   log_server_config: "none",
16   table_acl_name: "Crowd_Frame-your_task_name_your_batch_name_ACL",
17   table_data_name: "Crowd_Frame-your_task_name_your_batch_name_Data",
18   table_log_name: "Crowd_Frame-your_task_name_your_batch_name_Logger",
19   hit_solver_endpoint: "None",
20 };
```

Listing A.24: Sample development environment of Crowd_Frame.

1. Create the task and set its general parameters and criterion (cfr. Figure A.1).
2. Move to the build output folder for the platform:
 - data/build/mturk/
3. Copy the code of the wrapper:
 - data/build/mturk/index.html
4. Paste everything into the Design Layout box.
5. Preview and save the task project.
6. Publish the task and recruit a batch of workers by uploading the file containing the input/output tokens:
 - data/build/mturk/tokens.csv
7. Review the status of each submission by using the Manage tab.

A.5.3 Toloka

A task requester that aims to recruit each worker using Toloka must:

1. Create the project and set its general parameters (cfr. Figure A.3).

2. Move to the build output folder for the platform:
 - `data/build/toloka/`
3. Copy the markup, JavaScript code, and CSS styles of the wrapper:
 - `data/build/toloka/interface.html`
 - `data/build/toloka/interface.js`
 - `data/build/toloka/interface.css`
4. Paste each source code into the `Task Interface` box, using the corresponding section of the `HTML/JS/ CSS` (cfr. Figure A.4).
5. Copy the input and output data specification (cfr. Listing A.3 and Listing A.4):
 - `data/build/toloka/input_specification.json`
 - `data/build/toloka/output_specification.json`
6. Paste each data specification into the `Data Specification` box, using the corresponding sections (cfr. Figure A.5).
7. Copy the text of the task general instructions:
 - `data/build/task/instructions_general.json`
8. Paste the texts into the `Instructions for Tolokers` box, using the source code edit modality.
9. Create a new pool of workers by defining the parameters of the audience and the reward (cfr. Figure A.6).
10. Publish the task and recruit the audience of workers for each pool by uploading the file containing the input/output tokens (cfr. Listing A.5).
 - `data/build/mturk/tokens.tsv`
11. Review the status of each submission by using the each pool's page.

A.5.4 Prolific

A task requester that aims to recruit each worker using Prolific must:

1. Create the study and set its general parameters (cfr. Figure A.7).
2. Set the data collection modality required (cfr. Figure A.8):
 - Choose `External study link` as the modality to collect data.
 - Provide the URL of the task deployed.
 - Choose using the URL parameters to record Prolific IDs.
 - Rename the `PROLIFIC_PID` parameter to `workerId`.
 - Choose to redirect the participants to confirm completion using a URL.

- Copy the completion code from the URL shown (i.e., the `cc` parameter).
 - Set the `prolific_completion_code` environment variable using the completion code found as value.
3. Configure the parameters and criterion of the audience of workers to recruit (cfr. Figure A.9).
 4. Set the overall study cost (cfr. Figure A.10).
 5. Review the status of each submission by using the study's page.

A.6 Task Results

The requester can download the final results of a crowdsourcing task deployed using `Crowd_Frame` by using the download script. This involves four steps:

1. Access the main folder: `cd ~/path/to/project/`.
2. Access the data folder: `cd data`.
3. Run the `download.py` script. The script will:
 - Download and store snapshots of the raw data produced by each worker.
 - Refine the raw data using a tabular format.
 - Download and store the configuration of the task deployed.
 - Build and store support files containing worker and user agent attributes.

The whole set of output data is stored in the results folder: `data/result/task_name/`, where `task_name` is the value of the corresponding environment variable shown in Table A.9. The folder is created by the download script if it does not exist. It contains 5 subfolders, one for each type of output data. Table A.14 describes each of these subfolders.

Table A.14: Structure of the results folder of `Crowd_Frame`.

Folder	Description
Data	Contains snapshots of the raw data produced by each worker.
Dataframe	Contains tabular based refined versions of the raw data.
Resources	Contains two support files for each worker with attribute about him/herself and the user agent.
Task	Contains a backup of the task's configuration.
Crawling	Contains a crawl of the pages retrieved while using the search engine.

A.6.1 result/Task/

The `Task` folder contains the backup of the task configuration. Table A.11 (Section A.4.2) describes its content.

A.6.2 result/Data/

The Data folder contains a snapshot of the data produced by each worker. A snapshot is a JSON dictionary. The top-level object is an array. The download script creates an object for each batch of workers recruited within a crowdsourcing task. Listing A.25 shows the snapshot for a worker with identifier ABEFLAGYVQ7IN4 who participates in the batch Your_Batch of the task Your_Task. This means that the snapshot contains an array with a single object. The source_* attributes represent the DynamoDB tables and the path on the local filesystem.

```

1  [
2    {
3      "source_data": "Crowd_Frame-Your-Task_Your-Batch_Data",
4      "source_acl": "Crowd_Frame-Your-Task_Your-Batch_ACL",
5      "source_log": "Crowd_Frame-Your-Task_Your-Batch_Logger",
6      "source_path": "result/Your_Task/Data/ABEFLAGYVQ7IN4.json",
7      "data_items": 1,
8      "task": {...},
9      "worker": {...},
10     "ip": {...},
11     "uag": {...},
12     "checks": [...],
13     "questionnaires_answers": [...],
14     "documents_answers": [...],
15     "comments": [...],
16     "logs": [],
17     "questionnaires": {...},
18     "documents": {...},
19     "dimensions": {...}
20   }
21 ]

```

Listing A.25: Snapshot of a worker who participates in a task with a single batch deployed using Crowd_Frame.

A.6.3 result/Resources/

The Resources folder contains two JSON files for each worker. Let us hypothesize a worker recruited using the identifier ABEFLAGYVQ7IN4. The two support files are named ABEFLAGYVQ7IN4_ip.json and ABEFLAGYVQ7IN4_uag.json. The former contains attributes obtained by performing the reverse lookup of the worker's IP addresses. The latter contains attributes obtained by analyzing the user agent strings. Since a worker could access a task from multiple locations and using multiple devices, the two files can contain multiple IP addresses and user agent strings. Listing A.26 and Listing A.27 show a subset of the information provided by the two support files.

```
1  {
2    "<ip_address_1>": {
3      "continent_code": "NA",
4      "continent_name": "North America",
5      "country_capital": "Washington D.C.",
6      "country_code_iso2": "US",
7      "country_code_iso3": "USA",
8      "country_currency_code_iso3": "USD",
9      "country_currency_name": "US Dollar",
10     "country_currency_numeric": "840",
11     "country_currency_symbol": "\$",
12     "country_flag_emoji": "...",
13     "country_flag_emoji_unicode": "...",
14     "country_flag_url": "...",
15     "country_is_eu": false,
16     "country_name": "United States",
17     "country_name_official": "United States of America",
18     "country_numeric": "840",
19     "hostname": "...",
20     "ip": "...",
21     "ip_address_type": "...",
22     "latitude": "...",
23     "location_calling_code": "...",
24     "location_coordinates": "...",
25     "location_geoname_id": "...",
26     "location_is_eu": false,
27     "location_languages": [...],
28     "location_name": "...",
29     "location_postal_code": "...",
30     "longitude": "...",
31     "provider_name": "...",
32     "region_code": "...",
33     "region_code_full": "...",
34     "region_name": "Louisiana",
35     "region_type": "State",
36     "timezone_name": "America/Chicago",
37     ...
38   },
39   ...
```

Listing A.26: Subset of the information obtained by perform the reverse lookup of the IP address of a worker.

A.6.4 result/Crawling/

The Crawling folder contains a crawl of the web pages retrieved by the search engine when queried by a worker. A task requester who deploys a crowdsourcing task

```
1  {
2  "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML,
   ↳ like Gecko) Chrome/105.0.5195.102 Safari/537.36": {
3    "browser_engine": "WebKit/Blink",
4    "browser_name": "Chrome",
5    "browser_version": "105.0.5195.102",
6    "browser_version_major": "105",
7    "device_is_crawler": false,
8    "device_is_mobile_device": false,
9    "device_type": "desktop",
10   "os_code": "windows_10",
11   "os_family": "Windows",
12   "os_family_code": "windows",
13   "os_family_vendor": "Microsoft Corporation.",
14   "os_icon": "https://assets.userstack.com/icon/os/windows10.png",
15   "os_icon_large": "...",
16   "os_name": "Windows 10",
17   "os_url": "https://en.wikipedia.org/wiki/Windows_10",
18   "ua": "...",
19   "ua_type": "browser",
20   "ua_url": "https://about.google/",
21   ...
22  },
23  ...
24 }
```

Listing A.27: Subset of the information obtained by analyzing the user agent string of a worker.

which uses the search engine within one or more evaluation dimensions can choose to enable the crawling by using the `enable_crawling` shown in Table A.9. The download script thus tries to crawl each web page if the variable is enabled. Initially, the download script creates two subfolders, namely `Metadata/` and `Source/`. Each web page is then assigned an UUID (Universally Unique Identifier). Let us hypothesize a page assigned with the UUID `59c0f70f-c5a6-45ec-ac90-b609e2cc66d7`, The script tries to download its source code. It is stored in the `Source` folder if the operation succeeds, in a file named `59c0f70f-c5a6-45ec-ac90-b609e2cc66d7_source`. The extension depends on the page's source code.

Then, the script stores some metadata about the crawling operation of the page in the `Metadata` folder, in a JSON file named `59c0f70f-c5a6-45ec-ac90-b609e2cc66d7_metadata.json`. It is possible to understand whether the operation succeeded or not and why (i.e., by acknowledging the HTTP response code) and to read the value of each HTTP header. Listing A.28 show an example of metadata produced by the download script while trying to crawl one of the pages retrieved.

```
1 {
2   "attributes": {
3     "response_uuid": "59c0f70f-c5a6-45ec-ac90-b609e2cc66d7",
4     "response_url": "...",
5     "response_timestamp": "...",
6     "response_error_code": null,
7     "response_source_path": "...",
8     "response_metadata_path": "...",
9     "response_status_code": 200,
10    "response_encoding": "utf-8",
11    "response_content_length": 125965,
12    "response_content_type": "text/html; charset=utf-8"
13  },
14  "data": {
15    "date": "Wed, 08 Jun 2022 22:33:12 GMT",
16    "content_type": "text/html; charset=utf-8",
17    "content_length": "125965",
18    "..."
19  }
20 }
```

Listing A.28: Metadata produced by the download script while trying to crawl a webpage retrieved by the search engine of Crowd_Frame.

A.6.5 result/Dataframe/

The Dataframe folder contains a refined version of the data stored within each worker snapshot. The download script inserts the raw data into structures called “DataFrame”. A DataFrame⁴⁹ is a two-dimensional data structure with labelled axes that contains heterogeneous data. Such structures may thus be implemented as two-dimensional arrays or tables with rows and columns. The download script refines the raw data into up to 10 tabular dataframe serialized into CSV files. The final amount of dataframes serialized in the Dataframe folder depends on the environment variables configured by the task requester.

Each DataFrame has a variable number of rows and columns. Part of them provides a general information, thus being composed of one row for each worker, such as `workers_info` and `workers_acl`. The remaining ones have higher granularity. For instance, a row of the `workers_url` dataframe contains a row for each result retrieved for each query submitted to the search engine while analyzing a single HIT’s element during a given try by a single worker. A row of the `workers_answers` dataframe contains the answers for the evaluation dimensions provided for a single HIT’s element during a given try by a single worker, and so on. The requester must thus be careful while exploring each DataFrame and properly understand what kind of data they is exploring and analyzing. Listing A.29 provides an example of the access control list for a task with a single worker recruited. Listing A.30 provide an example composed of the answers provided by a single worker for two elements

⁴⁹<https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.html>

Table A.15: DataFrame produced when downloading the final results of a task.

DataFrame	Description
workers_acl.csv	Access control list of the workers.
workers_ip_addresses.csv	Data concerning the IP addresses of the workers.
workers_user_agents.csv	Data concerning the User Agent of the workers.
workers_answers.csv	Answers provided for each evaluation dimension by workers.
workers_questionnaire.csv	Answers provided for each questionnaire by workers.
workers_dimensions_selection.csv	Temporal order along with each worker chooses a value for each evaluation dimension.
workers_notes.csv	Textual annotations provided by workers.
workers_urls.csv	Queries to the search engine provided by workers along with results retrieved.
workers_crawling.csv	Data concerning the crawling of web pages retrieved by the search engine.
workers_logs.csv	Log data produced by the Logger component while the workers perform the task.
workers_comments.csv	Final comments provided by workers to the requester at the end of the task.
workers_mturk_data.csv	Data concerning workers produced by Amazon Mechanical Turk.
workers_toloka_data.csv	Data concerning workers produced by Toloka.
workers_prolific_study_data.csv	Data concerning the study deployed on Prolific and its submissions.
workers_prolific_demographic_data.csv	Data concerning the demographics of the workers who participate in a study published on Prolific.

of the HIT assigned.

Each dataframe has its own characteristics and peculiarities. However, there are several rules of thumb that a requester should remember and eventually consider while they performs the analysis:

- The attribute `paid` is present in the whole set of dataframe. It can be used to split the data among the workers who completed or did not the task. The requester may want to explore the data of who failed the task.
- The attribute `batch_name` is present in a subset of dataframe. It can be used to split the data among the different batches of workers recruited. The requester may want to analyze separately each subset of data.
- The attributes `try_current` and `try_last` are present in a subset of dataframe. They can be used to split the data among each try performed by each worker. The latter attribute indicates the most recent try. The requester should not forget the possible

```

1 worker_id,generated,in_progress,paid,platform,task_name,batch_name,unit_
  ↪ id,token_input,token_output,try_current,try_last,try_left,tries_amou
  ↪ nt,status_code,access_counter,time_arrival,time_arrival_parsed,time_
  ↪ submit,time_submit_parsed,time_completion,time_completion_parsed,tim
  ↪ e_expiration_nearest,time_expiration_nearest_parsed,time_expiration_
  ↪ time_expiration_parsed,time_expired,time_removal,time_removal_parsed
  ↪ ,questionnaire_amount,questionnaire_amount_start,questionnaire_amoun
  ↪ t_end,dimensions_amount,documents_amount,ip_address,ip_source,user_a
  ↪ gent,user_agent_source,folder,source_path,source_acl,source_data,sou
  ↪ rce_log,study_id,session_id,n_submissions
2 ABEFLAGYVQ7IN4,False,False,False,mturk,Task-Sample,Batch-Sample,unit_0,K
  ↪ DSKUOHINM,VQULVJHRTOZ,1,1,10,10,,1,"Thu, 14 Apr 2022 12:37:47
  ↪ GMT",2022-04-14 12:37:47 00:00,"Thu, 14 Apr 2022 12:38:44
  ↪ GMT",2022-04-14 12:38:44 00:00,,,,,,True,"Thu, 14 Apr 2022 15:32:39
  ↪ GMT",2022-04-14 15:32:39
  ↪ 00:00,3,3,0,0,0,75.41.166.226,cf,"Mozilla/5.0 (Windows NT 10.0;
  ↪ Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko)
  ↪ Chrome/100.0.4896.75
  ↪ Safari/537.36",cf,Task-Sample/Batch-Sample/Data/A3CGQ0JC280VGN/,resu
  ↪ lt/Task-Sample/Data/A3CGQ0JC280VGN.json,Crowd_Frame-Task-Sample_Batc
  ↪ h-Sample_ACL,Crowd_Frame-Task-Sample_Batch-Sample_Data,,,,,,

```

Listing A.29: Example of the workers_acl dataframe produced by Crowd_Frame.

presence of multiple tries for each worker while analyzing the data.

- The attribute action is present in a subset of dataframe dataframe. It can be used to understand whether the worker proceeded to the previous/following HIT's element. The possible values are Back, Next and Finish. The Finish value indicates the last element evaluated before completing a given try. The requester should remember that only the rows with the latter two values describe the most recent answers for each element.
- The attribute index_selected is present in the workers_urls dataframe. It can be used to filter the results retrieved by the search engine. The results with a value different from -1 for the attribute have been selected by the worker on the user interface. If its value is equal to 4, three other results have been previously selected. If its value is equal to 7 six other results have been previously selected, and so on. The requester may want to simply analyze the results with whom the worker interacted.
- The attribute type is present in the workers_logs dataframe. It specifies the type of log record described by each row. The log records are generally sorted using the global timestamp. The requester can use the attribute to split the whole set of log records into subsets of the same type.
- The dataframe workers_acl contains several useful information about each worker. The requester may want to merge it with the rows of the other dataframe using the worker_id attribute as the key.
- The dataframe workers_urls contains the whole set of results retrieved by the search

```

1 | worker_id,paid,task_id,batch_name,unit_id,try_last,try_current,action,ti
   | me_submit,time_submit_parsed,doc_index,doc_id,doc_fact_check_ground_
   | truth_label,doc_fact_check_ground_truth_value,doc_fact_check_source,
   | doc_speaker_name,doc_speaker_party,doc_statement_date,doc_statement_
   | description,doc_statement_text,doc_truthfulness_value,doc_accesses,d
   | oc_time_elapsed,doc_time_start,doc_time_end,global_outcome,global_fo
   | rm_validity,gold_checks,time_spent_check,time_check_amount
2 | ABEFLAGYVQ7IN4,False,Task-Sample,Batch-Sample,unit_1,1,1,Next,"Wed, 09
   | Nov 2022 10:19:16 GMT",2022-11-09 10:19:16 00:00,0.0,conservative-ac
   | tivist-steve-lonegan-claims-social-,false,1,Politifact,Steve
   | Lonegan,REP,2022-07-12,"stated on October 1, 2011 in an interview on
   | News 12 New Jersey's Power & Politics show:", "Today, the Social
   | Security system is broke.",10,1,2.1,1667989144,1667989146.1,False,Fa
   | lse,False,False,False
3 | ABEFLAGYVQ7IN4,False,Task-Sample,Batch-Sample,unit_1,1,1,Next,"Wed, 09
   | Nov 2022 10:19:25 GMT",2022-11-09 10:19:25 00:00,1,yes-tax-break-ron
   | -johnson-pushed-2017-has-benefite,true,5,Politifact,Democratic Party
   | of Wisconsin,DEM,2022-04-29,"stated on April 29, 2022 in News
   | release:", "The tax carve out (Ron) Johnson spearheaded
   | overwhelmingly benefited the wealthiest, over small businesses.",100
   | ,1,10.27,1667989146.1,1667989156.37,False,False,False,False,False

```

Listing A.30: Example of the `workers_answers` dataframe produced by `Crowd_Frame`.

engine. The dataframe `workers_crawling` contains information about the crawling of each result. The requester may want to merge the rows of the two dataframe `response_uuid` attribute as the key.

- The dataframe `workers_dimensions_selection` shows the temporal ordering along with the workers' chosen answers for the evaluation dimensions. It is ordered using the global timestamp along with each worker making a choice. This means that the rows belonging to a worker may occur in different positions in the dataframe. This may happen if multiple workers perform the task at the same time. The requester should consider this aspect while exploring the dataframe.
- The dataframe `worker_comments` provides the final comments of the worker. The requester should remember that providing a final comment is not mandatory for the worker, thus the dataframe may be empty.

A.7 Conclusions

`Crowd_Frame` is a relatively young software which still has a long road ahead before becoming truly effective and usable by the whole population of task requesters. There is much that can be done to implement new features and strengthen the overall implementation. Despite everything, the software has been already used in the past to deploy several tasks, as sketched in Section A.2.

Roitero et al. [361] study whether crowdsourcing be reliably used to assess the truthfulness of information items and to create large-scale labelled collections for information credibility systems. They deploy a crowdsourcing task to collect thousands of truthfulness judgments over two sets of statements using judgment scales with various granularity levels. Roitero et al. [362] study whether crowdsourcing is an effective and reliable method to assess truthfulness during a pandemic, targeting statements related to COVID-19, thus addressing (mis)information that is both related to a sensitive and personal issue and very recent as compared to when the judgment is done. They deploy a crowdsourcing task where workers are asked to judge the truthfulness of a set of COVID-19 related statements and provide evidence for their judgments.

Soprano et al. [395] hypothesize that a unidimensional truthfulness scale is not enough to account for the subtle differences that exist among publicly available statements to fact check. They thus propose a multidimensional notion of truthfulness. They deploy a crowdsourcing task where workers are asked to judge a set of statements with respect to seven different dimensions of truthfulness. Soprano et al. [394] design a questionnaire that aims to understand how longitudinal studies are performed by crowd workers and which factors influence participation across different crowdsourcing platforms. They include the questionnaire within a crowdsourcing task deployed across Amazon Mechanical Turk, Toloka and Prolific to build a large-scale set of answers. Draws et al. [111] perform a systematic exploratory analysis of publicly available crowdsourced data to identify a set of potential systematic biases that may occur when crowd workers perform crowdsourcing tasks. Then, they deploy a crowdsourcing task to collect a novel set of truthfulness judgments to validate their hypotheses.

Brand et al. [48] generate veracity predictions for claims using machine learning models and jointly provide a human-readable explanation. Then, they deploy a crowdsourcing task to gather human evaluations on the impact of explanations generated. Ceolin et al. [61] introduce a rating system to identify review arguments and define weighted semantics using formal argumentation theory. Then, they identify and build the argumentation graph. They thus deploy crowdsourcing tasks to evaluate their contribution by comparing the results of the argumentation dataset with the upvotes received by the reviews.

The aforementioned crowdsourcing tasks deployed by several researchers and practitioners demonstrate the effectiveness of Crowd_Frame to gather crowd-powered data across multiple platforms and marketplaces.

A.8 Future Work

Crowd_Frame is still characterized by some limitations that need to be addressed in future versions. The requester needs to build manually the overall set to HITs to allocate the element that must be evaluated. The result must be compliant with the format described in Section A.4.4. The feature provided to allocate automatically the elements is still experimental. The first development direction is stabilizing such a feature. The requester must upload the set of elements manually anyway. The user interface of the Generator component may thus be enriched to facilitate such operation. The interface between Crowd_Frame and the solver used may be extended to specify attributes to characterize and differentiate each

worker, such as their level of expertise.

The Generator component is bundled and shipped together with the Skeleton component. The use case diagrams shown in Figure A.12 and Figure A.17 show that different actors interact with the two components. Such components, in other words, allow performing two different procedures. On top of that, the user interface of the generator needs to be improved to become truly effective. The two components should have a separate implementation in the future. The generator could be published as a standalone tool and should be able to interact with buckets of the requester. In the past, such a decision has been made to cope with some development difficulties. Initially, Crowd_Frame was meant to be completely client-side. Still today, there is not any kind of backend that gets initialized. The client-side codebase of the application communicates directly with Amazon Web Services. The various building blocks of Crowd_Frame can be put together to build complex tasks. However, it may be hard for inexperienced requesters to understand how to design effective crowdsourcing tasks. The system should offer a set of templates ready to be used to design and eventually customize the most common types of crowdsourcing tasks, as allowed by Amazon Mechanical Turk (Figure A.1).

Furthermore, Crowd_Frame does not have an interface to monitor the progress of a task deployed. A requester must use the user interface of the crowdsourcing platform chosen to understand and review the status of each HIT. Such an interface will be developed in the future. It will interact with the crowdsourcing platform chosen using the corresponding software development kit and will provide feedback within the system. The software development kits of each crowdsourcing platform allow, among everything, also to automatically publish the tasks and approve each worker submission. The usage of such features will prevent the requester from performing the phases described in Section A.5 and will speed up the whole workflow. In the future, such a possibility will be provided. However, Crowd_Frame needs a more stable implementation to provide such a feature to task requesters in a safe way.

Crowd_Frame relies on Amazon Web Services to deploy the whole backend infrastructure needed. However, task requesters may want to use different cloud providers for various reasons. In the future, the software will provide the possibility to rely on other providers such as Microsoft Azure⁵⁰ or the Google Cloud Platform.⁵¹ Furthermore, the deployment of the whole infrastructure on private servers will be facilitated.

⁵⁰<https://azure.microsoft.com/>

⁵¹<https://cloud.google.com/>

Chapter 4: Questionnaires And Statement List

The appendix reports the demographic questionnaire and the CRT tests used for the task described in Section 4.2.1. Such questionnaires are also employed for the tasks described in Chapter 5, Chapter 7, Chapter 9 and Section 11.2.4.

B.1 Demographic Questionnaire

Q1: What is your age range?

A1: 0–18

A2: 19–25

A3: 26–35

A4: 36–50

A5: 50–80

A6: 80+

Q2: What is the highest level of school you have completed or the highest degree you have received?

A1: High school incomplete or less,

A2: High school graduate or GED (includes technical/vocational training that doesn't towards college credit)

A3: Some college (some community college, associate's degree)

A4: Four year college degree/bachelor's degree

A5: Some postgraduate or professional schooling, no postgraduate degree

A6: Postgraduate or professional degree, including master's, doctorate, medical or law degree

Q3: Last year what was your total family income from all sources, before taxes?

- A1: Less than 10,000
- A2: 10,000 to less than 20,000
- A3: 20,000 to less than 30,000
- A4: 30,000 to less than 40,000
- A5: 40,000 to less than 50,000
- A6: 50,000 to less than 75,000
- A7: 75,000 to less than 100,000
- A8: 100,000 to less than 150,000
- A9: 150,000 or more

Q4: In general, would you describe your political views as

- A1: Very conservative
- A2: Conservative
- A3: Moderate
- A4: Liberal
- A5: Very liberal

Q5: In politics today, do you consider yourself a

- A1: Republican
- A2: Democrat
- A3: Independent
- A4: Something else

Q6: Should the U.S. build a wall along the southern border?

- A1: Agree
- A2: Disagree
- A3: No opinion either way

Q7: Should the government increase environmental regulations to prevent climate change?

- A1: Agree
- A2: Disagree
- A3: No opinion either way

B.2 Cognitive Reflection Test (CRT)

CRT1: If three farmers can plant three trees in three hours, how long would it take nine farmers to plant nine trees?

- Correct Answer: 3 hours
- Intuitive Answer: 9 hours

CRT2: Sean received both the 5th highest and the 5th lowest mark in the class. How many students are there in the class?

- Correct Answer: 9 students
- Intuitive Answer: 10 students

CRT3: In an athletics team, females are four times more likely to win a medal than males. This year the team has won 20 medals so far. How many of these have been won by males?

- Correct Answer: 4 medals
- Intuitive Answer: 5 medals

Chapter 5: Statements List

This appendix provides the list of statements used to perform the crowdsourcing experiment and the longitudinal study described in Chapter 5. The statements are sampled from the PolitiFact [436] dataset.

ID	Speaker	Date	Ground Truth	Statement
S1	Facebook User	2020-25-03	Pants-On-Fire	If your child gets this virus their going to hospital alone in a van with people they don't know... to be with people they don't know you will be at home without them in their time of need.
S2	Donald Trump	2020-30-03	Pants-On-Fire	We inherited a "broken test" for COVID-19.
S3	Facebook User	2020-19-03	Pants-On-Fire	Says "there is no" COVID-19 virus.
S4	Facebook User	2020-25-03	Pants-On-Fire	COVID literally stands for Chinese Originated Viral Infectious Disease.
S5	Bloggers	2020-26-02	Pants-On-Fire	A post say "hair weave and lace fronts manufactured in China may contain the coronavirus."
S6	Youtube Video	2020-29-02	Pants-On-Fire	A video says that the Vatican confirmed that Pope Francis and two aides tested positive for coronavirus.
S7	Ron Desantis	2020-09-04	Pants-On-Fire	This particular pandemic is one where I don't think nationwide, there's been a single fatality under 25.
S8	Facebook User	2020-16-03	Pants-On-Fire	The government is closing businesses to stop the spread of coronavirus even though "the numbers are nothing compared to H1N1 or Ebola. Everyone needs to realize our government is up to something . . ."
S9	Facebook User	2020-16-03	Pants-On-Fire	the U.S. is developing an "antivirus" that includes a chip to track your movement.
S10	Bloggers	2020-31-03	Pants-On-Fire	Italy arrested a doctor "for intentionally killing over 3,000 coronavirus patients."
S11	Facebook User	2020-28-03	False	Says COVID-19 remains in the air for eight hours and that everyone is now required to wear masks "everywhere."
S12	Facebook User	2020-27-03	False	Says to leave objects in the sun to avoid contracting the coronavirus.
S13	Facebook User	2020-23-03	False	"Slices of lemon in a cup of hot water can save your life. The hot lemon can kill the proliferation of" the novel coronavirus.
S14	Facebook User	2020-23-03	False	Says the CDC now says that the coronavirus can survive on surfaces for up to 17 days.
S15	Viral Image	2020-13-03	False	Drinking "water a lot and gargling with warm water & salt or vinegar eliminates" the coronavirus.

Continues on next page

ID	Speaker	Date	Ground Truth	Statement
S16	Snapchat	2020-23-03	False	Says "special military helicopters will spray pesticide against the Corona virus in the skies all over the country."
S17	Bloggers	2020-20-03	False	Says COVID-19 came to the United States in 2019.
S18	Bloggers	2020-09-04	False	Church services can't resume until we're all vaccinated, says Bill Gates.
S19	Facebook User	2020-10-04	False	Mass vaccination for COVID-19 in Senegal was started yesterday (4/8) and the first 7 CHILDREN who received it "DIED on the spot."
S20	Facebook User	2020-02-04	False	Says video shows "the Chinese are destroying the 5G poles as they are aware that it is the thing triggering the corona symptoms."
S21	Turning Point Usa	2020-25-03	Mostly-False	Says Nevada Governor Steve Sisolak "has banned the use of an anti-malaria drug that might help cure coronavirus."
S22	Marco Rubio	2020-19-03	Mostly-False	For coronavirus cases "in the U.S. 38% of those hospitalized are under 35."
S23	Image	2020-15-03	Mostly-False	COVID-19 started because we eat animals.
S24	Facebook User	2020-14-03	Mostly-False	Italy has decided not to treat their elderly for this virus.
S25	Viral Image	2020-13-03	Mostly-False	President Trump, COVID-19 coronavirus: U.S. cases 1,329; U.S. deaths, 38; panic level: mass hysteria. President Obama, H1N1 virus: U.S. cases, 60.8 million; U.S. deaths, 12,469; panic level: totally chill. Do you all see how the media can manipulate your life.
S26	Facebook User	2020-27-02	Mostly-False	Post says "the blood test for coronavirus costs \$3,200."
S27	Deanna Lorraine	2020-12-04	Mostly-False	Says of COVID-19 that Dr. Anthony Fauci "was telling people on February 29th that there was nothing to worry about and it posed no threat to the US public at large.
S28	Instagram Post	2020-18-03	Mostly-False	Bill Gates and other globalists, in collaboration with pharmaceutical companies, are reportedly working to push tracking bracelets and 'invisible tattoos' to monitor Americans during an impending lockdown.
S29	Facebook User	2020-28-03	Mostly-False	Says a "5G LAW PASSED while everyone was distracted" with the coronavirus pandemic and lists 20 symptoms associated with 5G exposure.
S30	Facebook User	2020-28-03	Mostly-False	Says for otherwise healthy people "experiencing mild to moderate respiratory symptoms with or without a COVID-19 diagnosis. . . only high temperatures kill a virus, so let your fever run high," but not over 103 or 104 degrees.
S31	Facebook User	2020-31-03	Half-True	Ron Johnson said Americans should go back to work, because "death is an unavoidable part of life."
S32	Jeff Jackson	2020-19-03	Half-True	North Carolina "hospital beds are typically 85% full across the state."
S33	Facebook User	2020-15-03	Half-True	So Oscar Health, the company tapped by Trump to profit from covid tests, is a Kushner company. Imagine that, profits over national safety.
S34	Brian Fitzpatrick	2020-23-03	Half-True	We've got to give the American public a rough estimate of how long we think this is going to take, based mostly on the South Korean model, which seems to be the trajectory that we are on, thankfully, and not the Italian model.
S35	Facebook User	2020-10-03	Half-True	Harvard scientists say the coronavirus is "spreading so fast that it will infect 70% of humanity this year."
S36	Drew Pinsky	2020-03-03	Half-True	You're more likely to die of influenza right now" than the 2019 coronavirus.
S37	Michael Bloomberg	2020-26-02	Half-True	Says of President Donald Trump's actions on the coronavirus: "No. 1, he fired the pandemic team two years ago. No. 2, he's been defunding the Centers for Disease Control."
S38	Joe Biden	2020-05-04	Half-True	45 nations had already moved "to enforce travel restrictions with China" before the president moved.
S39	Facebook User	2020-07-04	Half-True	Says Donald Trump "himself has a financial stake in the French company that makes the brand-name version of hydroxychloroquine."
S40	Facebook User	2020-01-04	Half-True	"Non-essential people get to file for unemployment and make two to three times more than normal," but essential workers still on the job get no pay raise.

Continues on next page

ID	Speaker	Date	Ground Truth	Statement
S41	Facebook User	2020-29-03	Mostly-True	Says a study projects Wisconsin's coronavirus cases will peak on April 26, 2020.
S42	Facebook User	2020-20-03	Mostly-True	Says truck drivers are being turned away from fast-food restaurants during the COVID-19 pandemic.
S43	Facebook User	2020-18-03	Mostly-True	2019 coronavirus can live for "up to 3 hours in the air, up to 4 hours on copper, up to 24 hours on cardboard up to 3 days on plastic and stainless steel."
S44	Facebook User	2020-15-03	Mostly-True	Bill Gates told us about the coronavirus in 2015.
S45	Chart	2020-09-03	Mostly-True	Says 80% of novel coronavirus cases are "mild."
S46	Lou Dobbs	2020-02-03	Mostly-True	The United States is "actually screening fewer people (for the coronavirus than other countries) because we don't have appropriate testing."
S47	Charlie Kirk	2020-24-02	Mostly-True	Three Chinese nationals were apprehended trying to cross our Southern border illegally. Each had flu-like symptoms. Border Patrol quickly quarantined them and assessed any threat of coronavirus.
S48	Bernie Sanders	2020-08-04	Mostly-True	It has been estimated that only 12% of workers in businesses that are likely to stay open during this crisis are receiving paid sick leave benefits as a result of the second coronavirus relief package.
S49	Viral Image	2020-08-04	Mostly-True	Says a California surfer was "alone, in the ocean," when he was arrested for violating the state's stay-at-home order.
S50	Dan Patrick	2020-31-03	Mostly-True	Says for the coronavirus, "the death rate in Texas, per capita of 29 million people, we're one of the lowest in the country."
S51	Facebook User	2020-02-04	True	On February 7, the WHO warned about the limited stock of PPE. That same day, the Trump administration announced it was sending 18 tons of masks, gowns and respirators to China.
S52	Pat Toomey	2020-28-03	True	My mask will keep someone else safe and their mask will keep me safe.
S53	Andrew Cuomo	2020-17-03	True	No city in the state can quarantine itself without state approval.
S54	Kelly Alexander	2020-14-03	True	Says "most" NC legislators are in the "high risk age group" for coronavirus
S55	Viral Image	2020-13-03	True	Says Spectrum will provide free internet to students during coronavirus school closures.
S56	Michael Dougherty	2020-12-03	True	Some states are only getting 50 tests per day, and the Utah Jazz got 58.
S57	Blog Post	2020-10-03	True	Whole of Italy goes into quarantine
S58	Dan Crenshaw	2020-13-03	True	Says longstanding Food and Drug Administration regulations "created barriers to the private industry creating a test quickly" for the coronavirus.
S59	Viral Image	2020-02-04	True	Photo shows a crowded New York City subway train during stay-at-home order.
S60	John Bel Edwards	2020-05-04	True	Says of the coronavirus threat, "there was not a single suggestion by anyone, a doctor, a scientist, a political figure, that we needed to cancel Mardi Gras."

Chapter 6: Survey Questions

This appendix provides the survey employed to investigate the barriers to running longitudinal tasks on crowdsourcing platforms. Section 6.2.1 provides the details concerning the overall design of the survey and the crowdsourcing task. Section D.1 reports the questions of the P1 part, while Section D.2 those of the P2 part. The text of each question is reported together with the expected answer. The questions are shown in order, as they were presented to the workers recruited. As described in the section, the number of questions shown for P1 depends on the answer provided for question 1.1. In more detail, if the worker reports a number of previous experiences $1 \leq n \leq 3$, the 1.1.X block of questions is repeated n times, one for each experience. As a last remark, only one question among 1.1.X.9.1 and 1.1.X.9.2 is shown, depending on the answer provided for question 1.1.X.9.

D.1 P1: Current Perception Of Longitudinal Studies

1: *Have you ever participated in a longitudinal study in the past, even if on other platforms?*

1.1: *How many?*

- Integer value in the $[0, 3]$ interval

1.1.X: *Describe your experience with the longitudinal study nr. 1*

1.1.X.1: *When was the study performed?*

- 1 month ago
- 2 months ago
- 3 to 5 months ago
- 6 to 12 months ago
- More than 1 year ago

1.1.X.2: *How many sessions did the longitudinal study have?*

- Any positive integer

1.1.X.3: *Which was the time interval between each session?*

- 1 day
- 2 to 4 days
- 5 to 9 days
- 10 to 14 days
- 15 to 20 days
- 20 to 24 days
- 25 to 1 month
- 2 months
- 3 months
- 4 months
- 5 to 6 months
- 7 to 12 months
- More than 1 year
- Other (please, specify)

1.1.X.4: *Which was the time interval between each session?*

- 15 minutes
- 30 minutes
- 45 minutes
- 60 minutes
- 1 hour
- 2 hours
- 3 hours
- More than 3 hours
- Other (please, specify)

1.1.X.5: *Which was the crowdsourcing platform?*

- Amazon Mechanical Turk
- Prolific
- Toloka
- Other (please, specify)

1.1.X.6: *Which was the payment model?*

- Payment after each session
- Single final reward
- Other (please, specify)

1.1.X.7: *How was your general satisfaction:*

1.1.1.X.1: *Would you participate in the same study again?*

- Yes
- No
- Other (please, specify)

1.1.1.X.2: *Please, tell us why*

- Text field

1.1.X.8: *What was the main incentives that convince you into participating in the longitudinal study?*

- Bonus
- Reward
- Interest on task
- Altruism (to help the research)
- Because the task was educative
- Other (please, specify)

1.1.X.9: *Did you complete the task?*

- Yes

1.1.X.9.1: *What were the main incentives that convinced you in completing the longitudinal study?*

- Bonus
- Reward
- Interest on task
- Altruism (to help the research)
- Because the task was educative

- No

1.1.X.9.2: *What are the reasons that made you dropout?*

- Text field

2: *Do you think this crowdsourcing platform is suitable to carry out longitudinal studies? Please, elaborate your answer*

- Text field

3: *Longitudinal studies are not very common in crowdsourcing yet. Which of these statements do you agree with?*

- Longitudinal studies are not optimally supported by current popular crowdsourcing platforms
- Workers do not like to commit on daily effort
- Reward and incentives are insufficient
- Requesters do not need longitudinal participation since most of the tasks work with static data to annotate
- Other (please, specify)

D.2 P2: Possible Participation And Commitment To Longitudinal Studies

1: *How many days would you be happy to commit to a longitudinal study (imagine a session of about 15 min per day)*

- Positive integer
- 2: *Which of the following would make you refuse participation in a longitudinal study?*
- Too frequent
 - Too long
 - Other (please, specify)
- 3: *What's your preferred frequency of participation in a longitudinal study?*
- Daily
 - Every other day
 - Weekly
 - Biweekly
 - Monthly
 - Every six months
 - Yearly
- 4: *What is your preferred session duration (in hours)?*
- Positive integer
- 5: *What do you consider an acceptable hourly payment for your work on this platform (in USD\$ dollars)?*
- Positive integer
- 6: *How much time would you be happy to allocate per day to work on longitudinal studies (in hours)?*
- Positive integer
- 7: *Which incentives would most motivate you to participate and engage in longitudinal studies?*
- Final bonus to be awarded after the last contribution
 - Payment after each session
 - Progressive increment of payment
 - Progressive decrement of payment
 - Being penalized when skipping working sessions
 - Work on different tasks type to increase engagement diversity
 - Experimental variants of the same tasks to reduce repeatability
 - Other (please, specify)
- 8: *What types of tasks would you like to perform in a longitudinal study?*
- Information Finding - Such tasks delegate the process of searching to satisfy one's information need to the workers in the crowd. For example, "Find information about a company in the UK".
 - Verification and Validation - These are tasks that require workers in the crowd to either verify certain aspects as per the given instructions, or confirm the validity of various kinds of content. For example, "Match the names of personal computers and verify corresponding information".

- Interpretation and Analysis - Such tasks rely on the wisdom of the crowd to use their interpretation skills during task completion. For example, "Choose the most suitable category for each URL".
- Content Creation - Such tasks usually require the workers to generate new content for a document or website. They include authoring product descriptions or producing question-answer pair. For example, "Suggest names for a new product".
- Surveys - Surveys about a multitude of aspects ranging from demographics to customer satisfaction are crowdsourced. For example, "Mother's Day and Father's Day Survey (18-29 year olds only)".
- Content Access - These tasks require the crowd workers to simply access some content. For example, "Click on the link and watch the video". Other (please, specify)

9: *What do you think are the benefits of being involved in longitudinal studies?*

- No need to spend time regularly searching for new tasks to perform
- No need to learn how to do the job (Learning curve)
- Better productivity (more operationale)
- Intermediate payments would increase trust on requester
- Other (please, specify)

10: *What do you think are the downsides that limit your interest in participating in longitudinal studies?*

- Lack of flexibility
- Long term commitment
- Reward assigned at the end
- Lack of diversity
- Other (please, specify)

11: *Do you have any additional suggestions for a requester who plans to design an attractive longitudinal study?*

- Text field

Chapter 7: Task Instructions

This appendix reports the instruction text provided to each worker before starting the task described in Chapter 7. The instructions contain descriptions of each truthfulness dimension as presented to the workers.

Task Instructions

In this task, you will be asked to assess the truthfulness of eight statements by means of seven specific quality dimensions.

First, you will be asked to fill in one questionnaire and to answer three questions. Then, we will show you 8 *statements* made by popular people (for example, political figures) together with the information of who made the statement and on which date. For each statement, we ask you to search for evidence using our custom search engine and to tell *how much do you agree with considering the statement true in general (as opposed to false)*; that is, its overall truthfulness. We also ask you to mark the evidence found in terms of an URL as well as your self-confidence about the topic, i.e., if you consider *yourself expert / knowledgeable about its topic (as opposed to novice/beginner)*.

Then, we ask you to assess seven specific *quality dimensions* by stating your level of agreement with them. All your answers are given on a 5 level scale, i.e., they must be selected among 5 different labels: (-2) Completely Disagree, (-1) Disagree, (0) Neither Agree Nor Disagree, (+1) Agree, (+2) Completely Agree. Each quality dimension is detailed in the following list. We provide a sample statement for each dimension so you can familiarize yourself with the seven dimensions. Please, note that there are some “positive” examples, (i.e., statements that completely agree with the current dimension), and “negative” examples, (i.e., statements that completely disagree with the current dimension). (Keep in mind that the examples are illustrative only, and it is likely that you may also need to use the rest of the labels in your answers). The seven dimensions we consider are the following:

- *Correctness*: the statement is expressed in an accurate way, as opposed to being incorrect and/or reporting mistaken information
 - Example (which label is: +2 Completely Agree): “It’s illegal to treat a minor without parental consent in the U.S. Even as hospitals are limiting visitors, minors will always be allowed to have one guardian present.”
- *Neutrality*: the statement is expressed in a neutral / objective way, as opposed to subjective / biased
 - Example(which label is: -2 Completely Disagree): “The Labor Party has repeatedly claimed the Coalition needs to make cuts of \$70 billion to vital services to balance the budget.”

- *Comprehensibility*: the statement is comprehensible / understandable / readable as opposed to difficult to understand
 - Example (which label is: +2 Completely Agree) “Florida ranks first among the nations for access to free prekindergarten.”
- *Precision*: the information provided in the statement is precise / specific, as opposed to vague
 - Example (which label is: -2 Completely Disagree): “There were more deaths after the gun bans from guns than there were in the three years before Port Arthur”
- *Completeness*: the information reported in the statement is complete as opposed to telling only a part of the story
 - Example (which label is: -2 Completely Disagree): “We inherited a broken test for COVID-19.”
- *Speaker’s trustworthiness*: the speaker is generally trustworthy / reliable as opposed to untrustworthy / unreliable / malicious
 - Example (which label is: -2 Completely Disagree): “Says video shows “the Chinese are destroying the 5G poles as they are aware that it is the thing triggering the corona symptoms.”
- *Informativeness*: the statement allows us to derive useful information as opposed to simply stating well known facts and/or tautologies.
 - Example (which label is: +2 Completely Agree): “2019 coronavirus can live for up to 3 hours in the air, up to 4 hours on copper, up to 24 hours on cardboard up to 3 days on plastic and stainless steel.”

If you wish to change a previously given judgment, you can use the Back and Next buttons to navigate the task and revisit your answers. Please note that the statements are not presented in any particular order. You might see many good statements, many bad ones, or any combination. Try not to anticipate, and simply rate each statement after reading it. Note that you’ll need to answer *all questions and fill in every field* in order to proceed in the task, otherwise you will not be able to proceed to the following steps. Note that there are some quality checks throughout the task, and if you do not perform these correctly you will not be able to terminate the task and get paid. The data from this task is being gathered for research purposes only. Participation is entirely voluntary, and you are free to leave the task at any point.

Appendix **F**

Chapter 8: PRISMA Checklists And List Of Cognitive Biases

This appendix reports the checklists used to characterize cognitive biases using the PRISMA-inspired methodology described in Section 8.2 and the full list of 220 cognitive biases found in the literature, from which the list of 39 that can manifest while performing the fact-checking process described in Section 8.3.1 is derived. The checklists start on the following page due to spacing issues.

F.1 PRISMA 2020 Checklist For Abstracts



PRISMA 2020 for Abstracts Checklist

Section and Topic	Item #	Checklist item	Reported (Yes/No)
TITLE			
Title	1	Identify the report as a systematic review.	Yes
BACKGROUND			
Objectives	2	Provide an explicit statement of the main objective(s) or question(s) the review addresses.	Yes
METHODS			
Eligibility criteria	3	Specify the inclusion and exclusion criteria for the review.	Yes
Information sources	4	Specify the information sources (e.g. databases, registers) used to identify studies and the date when each was last searched.	Not Relevant
Risk of bias	5	Specify the methods used to assess risk of bias in the included studies.	Not Relevant
Synthesis of results	6	Specify the methods used to present and synthesise results.	Yes
RESULTS			
Included studies	7	Give the total number of included studies and participants and summarise relevant characteristics of studies.	Yes
Synthesis of results	8	Present results for main outcomes, preferably indicating the number of included studies and participants for each. If meta-analysis was done, report the summary estimate and confidence/credible interval. If comparing groups, indicate the direction of the effect (i.e. which group is favoured).	Yes
DISCUSSION			
Limitations of evidence	9	Provide a brief summary of the limitations of the evidence included in the review (e.g. study risk of bias, inconsistency and imprecision).	Not Relevant
Interpretation	10	Provide a general interpretation of the results and important implications.	Yes
OTHER			
Funding	11	Specify the primary source of funding for the review.	No (specified at the end of the paper)
Registration	12	Provide the register name and registration number.	Not Relevant

From: Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021;372:n71. doi: 10.1136/bmj.n71

For more information, visit: <http://www.prisma-statement.org/>

F.2 PRISMA 2020 Checklist



PRISMA 2020 Checklist

Section and Topic	Item #	Checklist item	Location where item is reported
TITLE			
Title	1	Identify the report as a systematic review.	Title
ABSTRACT			
Abstract	2	See the PRISMA 2020 for Abstracts checklist.	Abstract, see PRISMA for Abstracts checklist
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of existing knowledge.	Section 1.3
Objectives	4	Provide an explicit statement of the objective(s) or question(s) the review addresses.	Section 8.1
METHODS			
Eligibility criteria	5	Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses.	Section 8.2.1
Information sources	6	Specify all databases, registers, websites, organisations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted.	Section 8.2.1
Search strategy	7	Present the full search strategies for all databases, registers and websites, including any filters and limits used.	Section 8.2.1
Selection process	8	Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in the process.	Section 8.2.2
Data collection process	9	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process.	Section 8.2.2
Data items	10a	List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (e.g. for all measures, time points, analyses), and if not, the methods used to decide which results to collect.	Section 8.3.1
	10b	List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information.	Section 8.3.1
Study risk of bias assessment	11	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and if applicable, details of automation tools used in the process.	Section 8.2.2
Effect measures	12	Specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the synthesis or presentation of results.	Not Relevant
Synthesis methods	13a	Describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention characteristics and comparing against the planned groups for each synthesis (item #5)).	Section 8.3.1
	13b	Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions.	Not Relevant
	13c	Describe any methods used to tabulate or visually display results of individual studies and syntheses.	Section 8.3.3
	13d	Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used.	Section 8.3.3
	13e	Describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression).	Not Relevant
	13f	Describe any sensitivity analyses conducted to assess robustness of the synthesized results.	Not Relevant
Reporting bias assessment	14	Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases).	Section 8.2.2
Certainty assessment	15	Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome.	Section 8.2.2
RESULTS			
Study selection	16a	Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review, ideally using a flow diagram.	Section 8.2.2 and Figure 8.1



PRISMA 2020 Checklist

Section and Topic	Item #	Checklist item	Location where item is reported
	16b	Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded.	Section 8.3.1
Study characteristics	17	Cite each included study and present its characteristics.	Section 8.3.1
Risk of bias in studies	18	Present assessments of risk of bias for each included study.	Not Relevant
Results of individual studies	19	For all outcomes, present, for each study: (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (e.g. confidence/credible interval), ideally using structured tables or plots.	Section 8.3.2
Results of syntheses	20a	For each synthesis, briefly summarise the characteristics and risk of bias among contributing studies.	Section 8.3.2
	20b	Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (e.g. confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect.	Not Relevant
	20c	Present results of all investigations of possible causes of heterogeneity among study results.	Table 8.1
	20d	Present results of all sensitivity analyses conducted to assess the robustness of the synthesized results.	Section 8.3.2
Reporting biases	21	Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed.	Not Relevant
Certainty of evidence	22	Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed.	Not Relevant
DISCUSSION			
Discussion	23a	Provide a general interpretation of the results in the context of other evidence.	Sections 8.3.3 and 8.3.4
	23b	Discuss any limitations of the evidence included in the review.	Section 8.3.4
	23c	Discuss any limitations of the review processes used.	Section 8.3.4
	23d	Discuss implications of the results for practice, policy, and future research.	Section 11.3.2
OTHER INFORMATION			
Registration and protocol	24a	Provide registration information for the review, including register name and registration number, or state that the review was not registered.	Not Relevant
	24b	Indicate where the review protocol can be accessed, or state that a protocol was not prepared.	Not Relevant
	24c	Describe and explain any amendments to information provided at registration or in the protocol.	Not Relevant
Support	25	Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review.	End of the paper, before references
Competing interests	26	Declare any competing interests of review authors.	End of the paper, before references
Availability of data, code and other materials	27	Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review.	The paper is self contained and all the material is included either in the paper or in the appendix.

F.3 The 220 Cognitive Biases

1. Action bias
2. Actor-observer bias
3. Additive bias
4. Agent detection bias
5. Ambiguity effect
6. Anchoring bias
7. Anthropocentric thinking
8. Anthropomorphism
9. Apophenia
10. Association fallacy
11. Assumed similarity bias
12. Attentional bias
13. Attribute substitution
14. Attribution bias
15. Authority bias
16. Automation bias
17. Availability bias
18. Availability cascade
19. Availability heuristic
20. Backfire effect
21. Bandwagon effect
22. Barnum effect or Forer effect
23. Base rate fallacy
24. Belief bias
25. Ben Franklin effect
26. Berkson's paradox
27. Bias blind spot
28. Bizarreness effect
29. Boundary extension
30. Cheerleader effect
31. Childhood amnesia
32. Choice-supportive bias
33. Cognitive dissonance
34. Commission bias
35. Compassion fade
36. Confirmation bias
37. Conformity
38. Congruence bias
39. Conjunction fallacy or the Linda problem
40. Conservatism bias or Regressive bias
41. Consistency bias
42. Context effect
43. Continued influence effect
44. Contrast effect
45. Courtesy bias
46. Cross-race effect
47. Cryptomnesia
48. Curse of knowledge
49. Declinism
50. Decoy effect
51. Default effect
52. Defensive attribution hypothesis
53. Denomination effect
54. Disposition effect
55. Distinction bias
56. Dread aversion
57. Dunning-Kruger effect
58. Duration neglect
59. Effort justification
60. Egocentric bias
61. End-of-history illusion
62. Endowment effect
63. Escalation of commitment, irrational escalation, or sunk cost fallacy
64. Euphoric recall
65. Exaggerated expectation
66. Experimenter's or expectation bias
67. Extension neglect
68. Extrinsic incentives bias
69. Fading affect bias
70. Fallacy of composition
71. Fallacy of division
72. False consensus effect
73. False memory
74. False uniqueness bias
75. Form function attribution bias

- | | | |
|--|--|------------------------------------|
| 76. Framing effect, frequency illusion, or Baader–Meinhof phenomenon | 100. Illusory correlation | 124. Moral credential effect |
| | 101. Illusory superiority | 125. Moral luck |
| | 102. Illusory truth effect | 126. Naïve cynicism |
| 77. Fundamental attribution error | 103. Impact bias | 127. Naïve realism |
| | 104. Implicit bias or association bias | 128. Negativity bias |
| 78. Gambler’s fallacy | | 129. Neglect of probability |
| 79. Gender bias | 105. Information bias | 130. Next-in-line effect |
| 80. Generation effect or Self-generation effect | 106. Ingroup bias | 131. Non-adaptive choice switching |
| 81. Google effect | 107. Insensitivity to sample size | 132. Normalcy bias |
| 82. Group attribution error | 108. Intentionality bias | 133. Not invented here syndrome |
| 83. Groupshift | 109. Interoceptive bias or Hungry judge effect | 134. Objectivity illusion |
| 84. Groupthink | | 135. Observer-expectancy effect |
| 85. Halo effect | 110. Just-world hypothesis | 136. Omission bias |
| 86. Hard–easy effect | 111. Lag effect | 137. Optimism bias |
| 87. Hindsight bias | 112. Less-is-better effect | 138. Ostrich effect |
| 88. Hostile attribution bias | 113. Leveling and sharpening | 139. Outcome bias |
| 89. Hot-cold empathy gap | 114. Levels-of-processing effect | 140. Outgroup homogeneity bias |
| 90. Hot-hand fallacy | 115. List-length effect | 141. Overconfidence effect |
| 91. Humor effect | 116. Logical fallacy | 142. Parkinson’s law of triviality |
| 92. Hyperbolic discounting | 117. Loss aversion | 143. Part-list cueing effect |
| 93. IKEA effect | 118. Memory inhibition | 144. Peak–end rule |
| 94. Illicit transference | 119. Mere exposure effect or familiarity principle | 145. Perky effect |
| 95. Illusion of asymmetric insight | 120. Misattribution | 146. Pessimism bias |
| 96. Illusion of control | 121. Modality effect | 147. Picture superiority effect |
| 97. Illusion of explanatory depth | 122. Money illusion | 148. Placement bias |
| 98. Illusion of transparency | 123. Mood-congruent memory bias | 149. Plan continuation bias |
| 99. Illusion of validity | | 150. Planning fallacy |

- | | | |
|---|--|--|
| 151. Plant blindness | 175. Salience bias | 199. System justification |
| 152. Positivity effect or Socioemotional selectivity theory | 176. Saying is believing effect | 200. Systematic bias |
| 153. Present bias | 177. Scope neglect | 201. Tachypsychia |
| 154. Prevention bias | 178. Selection bias | 202. Telescoping effect |
| 155. Primacy effect | 179. Self-relevance effect | 203. Testing effect |
| 156. Probability matching | 180. Self-serving bias | 204. Third-person effect |
| 157. Processing difficulty effect | 181. Semmelweis reflex | 205. Time-saving bias |
| 158. Pro-innovation bias | 182. Serial position effect | 206. Tip of the tongue phenomenon |
| 159. Projection bias | 183. Sexual overperception bias | 207. Trait ascription bias |
| 160. Proportionality bias | 184. Shared information bias | 208. Travis syndrome |
| 161. Prospect theory | 185. Social comparison bias | 209. Truth bias |
| 162. Pseudocertainty effect | 186. Social cryptomnesia | 210. Ultimate attribution error |
| 163. Puritanical bias | 187. Social desirability bias | 211. Unconscious bias or implicit bias |
| 164. Pygmalion effect | 188. Source confusion | 212. Unit bias |
| 165. Reactance Theory | 189. Spacing effect | 213. Verbatim effect |
| 166. Reactive devaluation | 190. Spotlight effect | 214. Von Restorff effect |
| 167. Recency effect | 191. Status quo bias | 215. Weber–Fechner law |
| 168. Recency illusion | 192. Stereotypical bias or stereotype bias | 216. Well travelled road effect |
| 169. Reminiscence bump | 193. Stereotyping subadditivity effect | 217. Women are wonderful effect |
| 170. Repetition blindness | 194. Spacing effect | 218. Worse-than-average effect |
| 171. Restraint bias | 195. Subjective validation | 219. Zero-risk bias |
| 172. Rhyme as reason effect | 196. Suffix effect | 220. Zero-sum bias |
| 173. Risk compensation or Peltzman effect | 197. Surrogation | |
| 174. Rosy retrospection | 198. Survivorship bias | |

Chapter 9: Questionnaires

This appendix reports the additional questionnaires used for the task described in Section 9.3.1, the “Belief in Science Scale” (BISS) questionnaire [92], and the generalized version of the “Citizen Trust in Government Organizations” (CTGO) questionnaire [165]. The BISS and CTGO questionnaires require an answer provided using a 5-level Likert scale, ranging from Completely Disagree (-2) to Completely Agree (+2). The task design includes also the questionnaires reported in Appendix B.1 and Appendix B.2.

G.1 Citizen Trust in Government Organizations (CTGO)

CTGO1: Politicians in general are capable.

CTGO2: Politicians in general are effective.

CTGO3: Politicians in general are skillful.

CTGO4: Politicians in general are expert.

CTGO5: Politicians in general carry out their duty very well.

CTGO6: If citizens need help, the politicians will do their best to help them.

CTGO7: Politicians in general act in the interest of citizens.

CTGO8: Politicians in general are genuinely interested in the well-being of citizens.

CTGO9: Politicians in general approach citizens in a sincere way.

CTGO10: Politicians in general approach are sincere.

CTGO11: Politicians in general keep their commitment.

CTGO12: Politicians in general are honest.

G.2 Belief in Science Scale (BISS)

BIS1: Science provides us with a better understanding of the universe than does religion.

BIS2: "In a demon-haunted world, science is a candle in the dark." (Carl Sagan)

BIS3: We can only rationally believe in what is scientifically provable.

BIS4: Science tells us everything there is to know about what reality consists of.

BIS5: All the tasks human beings face are soluble by science.

BIS6: The scientific method is the only reliable path to knowledge.

BIS7: The only real kind of knowledge we can have is scientific knowledge.

BIS8: Science is the most valuable part of human culture.

BIS9: Science is the most efficient means of attaining truth.

BIS10: Scientists and science should be given more respect in modern society.

Section 11.2.4: Multidimensional Scale For Reviews Quality Judgment

This appendix reports the adapted version of the multiple dimensions of truthfulness (Section 7.2.1) used to evaluate product reviews quality, as shown to the workers during the crowdsourcing task. The list should be compared to Appendix E.

- *Truthfulness*: measures the overall truthfulness and trustworthiness of the review.
- *Reliability*: the review is considered reliable, as opposed to reporting unreliable information. *Example (label: +2 Completely agree): "They fit great, look great, are quite comfortable and are just what I was looking for!"*.
- *Neutrality*: the review is expressed in objective terms, as opposed to resulting subjective or biased. *Example (label: -2 Completely disagree): "Love them!!"*
- *Comprehensibility*: the review is comprehensible/understandable/readable as opposed to difficult to understand. *Example (label: +2 Completely agree): "They run big. Order a full size smaller"*.
- *Precision*: the review is precise/specific, as opposed to vague. *Example (label: +2 Completely agree): They run big. Order a full size smaller.*
- *Completeness*: the review is complete as opposed to partial. *Example (label: +2 Completely agree): "I actually have 3 pairs of these trainers. They are very comfortable, there is a neoprene sleeve that goes around your ankle that makes them the most comfortable for me compared to normal athletic shoes. They run a little narrow - for me this is perfect, but you may want to round up on the size or try on in the store first if your feet are on the wider side."*
- *Informativeness*: The review allows deriving useful information as opposed to well-known facts and/or tautologies. *Example (label: +1 Agree): "Love these shoes! Needed new running shoes and these are perfect. Light weight and fit great!"*

Bibliography

- [1] Hervé Abdi and Lynne J Williams. Tukey’s honestly significant difference (HSD) test. In: *Encyclopedia of Research Design* 3.1 (2010), pp. 1–5. URL: <https://personal.utdallas.edu/~Herve/abdi-HSD2010-pretty.pdf>.
- [2] Alberto Acerbi. Cognitive Attraction And Online Misinformation. In: *Palgrave Communications* 5.1 (Feb. 2019), p. 15. ISSN: 2055-1045. DOI: 10.1057/s41599-019-0224-y.
- [3] Alan Agresti. *Analysis of Ordinal Categorical Data*. 2nd. Wiley Series in Probability and Statistics. Wiley, 2010. ISBN: 9780470082898. DOI: 10.1002/9780470594001.
- [4] Naser Ahmadi, Joohyung Lee, Paolo Papotti, and Mohammed Saeed. Explainable Fact Checking with Probabilistic Answer Set Programming. In: *Proceedings of the 2019 Truth and Trust Online Conference (TTO 2019)*. Ed. by Maria Liakata and Andreas Vlachos. London, UK, 2019. URL: https://truthandtrustonline.com/wp-content/uploads/2019/09/paper_15.pdf.
- [5] Lewis R. Aiken. *Rating scales and checklists: Evaluating behavior, personality, and attitudes*. Rating scales and checklists: Evaluating behavior, personality, and attitudes. Oxford, England: John Wiley & Sons, 1996. ISBN: 0-471-12787-6.
- [6] Firoj Alam, Shaden Shaar, Fahim Dalvi, Hassan Sajjad, Alex Nikolov, Hamdy Mubarak, Giovanni Da San Martino, Ahmed Abdelali, Nadir Durrani, Kareem Darwish, Abdulaziz Al-Homaid, Wajdi Zaghoulani, Tommaso Caselli, Gijs Danoe, Friso Stolk, Britt Bruntink, and Preslav Nakov. Fighting the COVID-19 Infodemic: Modeling the Perspective of Journalists, Fact-Checkers, Social Media Platforms, Policy Makers, and the Society. In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 611–649. DOI: 10.18653/v1/2021.findings-emnlp.56.
- [7] Nikolaos Aletras, Timothy Baldwin, Jey Han Lau, and Mark Stevenson. Evaluating Topic Representations for Exploring Document Collections. In: *Journal of the Association for Information Science and Technology* 68.1 (2017), pp. 154–167. DOI: 10.1002/asi.23574.

- [8] Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. Where is Your Evidence: Improving Fact-checking by Justification Modeling. In: *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 85–90. DOI: 10.18653/v1/W18-5513.
- [9] Jennifer Allen, Antonio A. Arechar, Gordon Pennycook, and David G. Rand. Scaling up fact-checking using the wisdom of crowds. In: *Science Advances* 7.36 (2021), eabf4393. DOI: 10.1126/sciadv.abf4393.
- [10] Jennifer Allen, Baird Howland, Markus Mobius, David Rothschild, and Duncan J. Watts. Evaluating the fake news problem at the scale of the information ecosystem. In: *Science Advances* 6.14 (2020). DOI: 10.1126/sciadv.aay3539.
- [11] Omar Alonso and Stefano Mizzaro. Using crowdsourcing for TREC relevance assessment. In: *Information Processing & Management* 48.6 (2012), pp. 1053–1066. DOI: 10.1016/j.ipm.2012.01.004.
- [12] Leila Amgoud and Claudette Cayrol. A Reasoning Model Based on the Production of Acceptable Arguments. In: *Annals of Mathematics and Artificial Intelligence* 34.1 (Mar. 2002), pp. 197–215. ISSN: 1573-7470. DOI: 10.1023/A:1014490210693.
- [13] Leila Amgoud and Srdjan Vesic. Two Roles of Preferences in Argumentation Frameworks. In: *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*. Ed. by Weiru Liu. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 86–97. DOI: 10.1007/978-3-642-22152-1_8.
- [14] Enrique Amigó, Jorge Carrillo de Albornoz, Irina Chugur, Adolfo Corujo, Julio Gonzalo, Tamara Martín, Edgar Meij, Maarten de Rijke, and Damiano Spina. Overview of RepLab 2013: Evaluating Online Reputation Monitoring Systems. In: *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*. Ed. by Pamela Forner, Henning Müller, Roberto Paredes, Paolo Rosso, and Benno Stein. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 333–352. ISBN: 978-3-642-40802-1. DOI: 10.1007/978-3-642-40802-1_31.
- [15] Enrique Amigó, Julio Gonzalo, Stefano Mizzaro, and Jorge Carrillo-de-Albornoz. An Effectiveness Metric for Ordinal Classification: Formal Properties and Experimental Results. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, July 2020, pp. 3938–3949. DOI: 10.18653/v1/2020.acl-main.363.
- [16] Enrique Amigó, Julio Gonzalo, and Felisa Verdejo. A General Evaluation Measure for Document Organization Tasks. In: *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '13*. Dublin, Ireland: Association for Computing Machinery, 2013, 643–652. ISBN: 9781450320344. DOI: 10.1145/2484028.2484081.
- [17] Corin R. Anderson, Pedro Domingos, and Daniel S. Weld. Relational Markov Models and Their Application to Adaptive Web Navigation. In: *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '02*. Edmonton, Alberta, Canada: Association for Computing Machinery, 2002, 143–152. ISBN: 158113567X. DOI: 10.1145/775047.775068.

- [18] Hal R Arkes and Catherine Blumer. The psychology of sunk cost. In: *Organizational Behavior and Human Decision Processes* 35.1 (1985), pp. 124–140. ISSN: 0749-5978. DOI: 10.1016/0749-5978(85)90049-4.
- [19] Pepa Atanasova, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, Georgi Karadzhov, Tsvetomila Mihaylova, Mitra Mohtarami, and James Glass. Automatic Fact-Checking Using Context and Discourse Information. In: *Journal of Data and Information Quality* 11.3 (May 2019). ISSN: 1936-1955. DOI: 10.1145/3297722.
- [20] Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. Generating Fact Checking Explanations. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 7352–7364. DOI: 10.18653/v1/2020.acl-main.656.
- [21] Katie Atkinson, Pietro Baroni, Massimiliano Giacomin, Anthony Hunter, Henry Prakken, Chris Reed, Guillermo Simari, Matthias Thimm, and Serena Villata. Towards Artificial Argumentation. In: *AI Magazine* 38.3 (Oct. 2017), pp. 25–36. DOI: 10.1609/aimag.v38i3.2704.
- [22] Elena M. Auer, Tara S. Behrend, Andrew B. Collmus, Richard N. Landers, and Ahleah F. Miles. Pay for performance, satisfaction and retention in longitudinal crowdsourced research. In: *PLOS ONE* 16.1 (Jan. 2021), pp. 1–17. DOI: 10.1371/journal.pone.0245460. URL: <https://doi.org/10.1371/journal.pone.0245460>.
- [23] Mamoun A. Awad and Issa Khalil. Prediction of User’s Web-Browsing Behavior: Application of Markov Model. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 42.4 (2012), pp. 1131–1142. DOI: 10.1109/TSMCB.2012.2187441.
- [24] Leif Azzopardi. Cognitive Biases in Search: A Review and Reflection of Cognitive Biases in Information Retrieval. In: *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*. CHIIR ’21. Canberra ACT, Australia: Association for Computing Machinery, 2021, 27–37. ISBN: 9781450380553. DOI: 10.1145/3406522.3446023.
- [25] Ricardo Baeza-Yates. Bias in Search and Recommender Systems. In: *Fourteenth ACM Conference on Recommender Systems*. New York, NY, USA: Association for Computing Machinery, 2020, p. 2. ISBN: 9781450375832. DOI: 10.1145/3383313.3418435.
- [26] Ricardo Baeza-Yates. Bias on the Web. In: *Communications of the ACM* 61.6 (May 2018), 54–61. ISSN: 0001-0782. DOI: 10.1145/3209581.
- [27] Yair Bar-Haim, Dominique Lamy, Lee Pergamin, Marian J Bakermans-Kranenburg, and Marinus H van IJzendoorn. Threat-related attentional bias in anxious and nonanxious individuals: a meta-analytic study. en. In: *Psychol. Bull.* 133.1 (Jan. 2007), pp. 1–24. DOI: 10.1037/0033-2909.133.1.1.
- [28] James H. Barnes. Cognitive Biases and Their Impact on Strategic Planning. In: *Strategic Management Journal* 5.2 (1984), pp. 129–137. ISSN: 01432095, 10970266. URL: <http://www.jstor.org/stable/2486172>.
- [29] Jonathan Baron and John C Hershey. Outcome bias in decision evaluation. In: *Journal of personality and social psychology* 54.4 (1988), p. 569. URL: <https://www.sas.upenn.edu/~baron/papers/outcomebias.pdf>.

- [30] P. Baroni, Martin Caminada, and M. Giacomin. “Abstract Argumentation Frameworks and their Semantics”. In: *Handbook of Formal Argumentation*. Ed. by P. Baroni, D. Gabbay, and M. Giacomin. College Publications, 2018. Chap. 4, pp. 159–236. ISBN: 978-1-84890-275-6. URL: <http://www.collegepublications.co.uk/downloads/handbooks00003.pdf>.
- [31] Pietro Baroni, Martin Caminada, and Massimiliano Giacomin. An introduction to argumentation semantics. In: *The Knowledge Engineering Review* 26.4 (2011), 365–410. DOI: 10.1017/S0269888911000166.
- [32] Alberto Barrón-Cedeño, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, Fatima Haouari, Nikolay Babulkov, Bayan Hamdan, Alex Nikolov, Shaden Shaar, and Zien Sheikh Ali. Overview of CheckThat! 2020: Automatic Identification and Verification of Claims in Social Media. In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Ed. by Avi Arampatzis, Evangelos Kanoulas, Theodora Tsirikika, Stefanos Vrochidis, Hideo Joho, Christina Lioma, Carsten Eickhoff, Aurélie Névél, Linda Cappellato, and Nicola Ferro. Cham: Springer International Publishing, 2020, pp. 215–236. ISBN: 978-3-030-58219-7. DOI: 10.1007/978-3-030-58219-7_17.
- [33] Carol L. Barry and Linda Schamber. Users’ criteria for relevance evaluation: A cross-situational comparison. In: *Information Processing & Management* 34.2 (1998), pp. 219–236. ISSN: 0306-4573. DOI: 10.1016/S0306-4573(97)00078-2.
- [34] Karen W. Bauer. Conducting longitudinal studies. In: *New Directions for Institutional Research* 2004.121 (2004), pp. 75–90. DOI: <https://doi.org/10.1002/ir.102>.
- [35] Tara S. Behrend, David J. Sharek, Adam W. Meade, and Eric N. Wiebe. The viability of crowdsourcing for survey research. In: *Behavior Research Methods* 43.3 (Mar. 2011), p. 800. ISSN: 1554-3528. DOI: 10.3758/s13428-011-0081-0.
- [36] Mihir Bellare, Ran Canetti, and Hugo Krawczyk. Keying Hash Functions for Message Authentication. In: *Advances in Cryptology — CRYPTO ’96*. Ed. by Neal Koblitz. Berlin, Heidelberg: Springer Berlin Heidelberg, 1996, pp. 1–15. ISBN: 978-3-540-68697-2. DOI: 10.1007/3-540-68697-5_1.
- [37] Trevor J. M. Bench-Capon. Persuasion in Practical Argument Using Value-based Argumentation Frameworks. In: *Journal of Logic and Computation* 13.3 (June 2003), pp. 429–448. ISSN: 0955-792X. DOI: 10.1093/logcom/13.3.429.
- [38] Trevor J. M. Bench-Capon. Value-based argumentation frameworks. In: *Proceedings of the 9th International Workshop on Non-Monotonic Reasoning*. Ed. by Salem Benferhat and Enrico Giunchiglia. Toulouse, France, 2002, pp. 443–454. URL: <https://nmr.cs.tu-dortmund.de/proceedings/NMR2002Proceedings.pdf>.
- [39] Fabrício Benevenuto, Tiago Rodrigues, Meeyoung Cha, and Virgílio Almeida. Characterizing User Behavior in Online Social Networks. In: *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement*. IMC ’09. Chicago, Illinois, USA: Association for Computing Machinery, 2009, 49–62. ISBN: 9781605587714. DOI: 10.1145/1644893.1644900.

- [40] Md Momen Bhuiyan, Amy X. Zhang, Connie Moon Sehat, and Tanushree Mitra. Investigating Differences in Crowdsourced News Credibility Assessment: Raters, Tasks, and Expert Criteria. In: *Proceedings of the ACM on Human-Computer Interaction* 4.CSCW2 (Oct. 2020). DOI: 10.1145/3415164.
- [41] Kourtney Bitterly. 1 in 4 Americans Own a Smart Speaker. What Does That Mean for News? In: *New York Times* (2019). URL: <https://open.nytimes.com/how-might-the-new-york-times-sound-on-smart-speakers-3b59a6a78ae3>.
- [42] J. Martin Bland and Douglas G. Altman. Statistics notes: Bootstrap resampling methods. In: *BMJ* 350 (2015). DOI: 10.1136/bmj.h2622.
- [43] Bryan Bollinger, Phillip Leslie, and Alan Sorensen. Calorie Posting in Chain Restaurants. In: *American Economic Journal: Economic Policy* 3.1 (2011), pp. 91–128. ISSN: 19457731, 1945774X. URL: <http://www.jstor.org/stable/41238086>.
- [44] Jose Borges and Mark Levene. Evaluating Variable-Length Markov Chain Models for Analysis of User Web Navigation Sessions. In: *IEEE Transactions on Knowledge and Data Engineering* 19.4 (2007), pp. 441–452. DOI: 10.1109/TKDE.2007.1012.
- [45] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, Sept. 2015, pp. 632–642. DOI: 10.18653/v1/D15-1075.
- [46] Tara Brabazon. The Google Effect: Googling, Blogging, Wikis and the Flattening of Expertise. In: 56.3 (2006), pp. 157–167. DOI: doi:10.1515/LIBR.2006.157.
- [47] Erik Brand, Kevin Roitero, Michael Soprano, and Gianluca Demartini. E-BART: Jointly Predicting and Explaining Truthfulness. In: *Proceedings of the 2021 Truth and Trust Online Conference*. Ed. by Isabelle Augenstein, Paolo Papotti, and Dustin Wright. Virtual Event, Oct. 2021, pp. 18–27. URL: https://truthandtrustonline.com/wp-content/uploads/2021/10/TT02021_paper_16-1.pdf.
- [48] Erik Brand, Kevin Roitero, Michael Soprano, Afshin Rahimi, and Gianluca Demartini. A Neural Model to Jointly Predict and Explain Truthfulness of Statements. In: *Journal of Data and Information Quality. Journal Rank: Scimago Q2 (2021)* (July 2022). ISSN: 1936-1955. DOI: 10.1145/3546917.
- [49] Jailson Brito, Vaninha Vieira, and Adolfo Duran. Towards a Framework for Gamification Design on Crowdsourcing Systems: The G.A.M.E. Approach. In: *2015 12th International Conference on Information Technology - New Generations*. 2015, pp. 445–450. DOI: 10.1109/ITNG.2015.78.
- [50] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei.

- Language Models are Few-Shot Learners. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R Hadsell, M.F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. URL: <https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf>.
- [51] Michael Buhrmester, Tracy Kwang, and Samuel D. Gosling. Amazon’s Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? In: *Perspectives on Psychological Science* 6.1 (2011), pp. 3–5. DOI: 10.1177/1745691610393980.
- [52] Joseph Bullock, Alexandra Luccioni, Katherine Hoffmann Pham, Cynthia Sin Nga Lam, and Miguel Luengo-Oroz. Mapping the Landscape of Artificial Intelligence Applications against COVID-19. In: *Journal of Artificial Intelligence Research* (Nov. 2020). DOI: 10.1613/jair.1.12162.
- [53] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. In: *Science* 356.6334 (2017), pp. 183–186. DOI: 10.1126/science.aal4230.
- [54] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-SNLI: Natural Language Inference with Natural Language Explanations. In: *Advances in Neural Information Processing Systems 31*. Ed. by Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett. Montréal, Canada: Curran Associates, Inc., 2018, pp. 9560–9572. URL: <https://proceedings.neurips.cc/paper/2018/file/4c7a167bb329bd92580a99ce422d6fa6-Paper.pdf>.
- [55] Andrea Carson, Andrew Gibbons, Aaron Martin, and Justin B. Phillips. Does Third-Party Fact-Checking Increase Trust in News Stories? An Australian Case Study Using the “Sports Rorts” Affair. In: *Digital Journalism* 10.5 (2022), pp. 801–822. DOI: 10.1080/21670811.2022.2031240.
- [56] J-P Caverni, J-M Fabre, and Michel Gonzalez. *Cognitive biases*. NY, USA.: Elsevier, 1990. ISBN: 9780080867229.
- [57] J.S. Caylor. *Methodologies for Determining Reading Requirements of Military Occupational Specialists*. Tech. rep. 1973. URL: <http://files.eric.ed.gov/fulltext/ED074343.pdf>.
- [58] Stephen J. Ceci and Wendy M. Williams. The Psychology of Fact-checking. In: *Scientific American* (2020), pp. 7–13. URL: <https://www.scientificamerican.com/article/the-psychology-of-fact-checking1/>.
- [59] Pew Research Center. *Most Border Wall Opponents, Supporters Say Shutdown Concessions Are Unacceptable*. Ed. by Pew Research Center, Washington, D.C. URL: <https://www.people-press.org/2019/01/16/most-border-wall-opponents-supporters-say-shutdown-concessions-are-unacceptable/>.
- [60] Davide Ceolin, Julia Noordegraaf, and Lora Aroyo. Capturing the Ineffable: Collecting, Analysing, and Automating Web Document Quality Assessments. In: *Knowledge Engineer*. Ed. by Eva Blomqvist, Paolo Ciancarini, Francesco Poggi, and Fabio Vitali. Cham: Springer International Publishing, 2016, pp. 83–97. ISBN: 978-3-319-49004-5. DOI: 10.1007/978-3-319-49004-5_6.

- [61] Davide Ceolin, Giuseppe Primiero, Michael Soprano, and Jan Wielemaker. Transparent assessment of information quality of online reviews using formal argumentation theory. In: *Information Systems* 110 (July 2022), p. 102107. ISSN: 0306-4379. DOI: 10.1016/j.is.2022.102107.
- [62] Davide Ceolin, Giuseppe Primiero, Jan Wielemaker, and Michael Soprano. Assessing the Quality of Online Reviews Using Formal Argumentation Theory. In: *Web Engineering*. Ed. by Marco Brambilla, Richard Chbeir, Flavius Frasinca, and Ioana Manolescu. Biarritz, France: Springer International Publishing, 2021, pp. 71–87. ISBN: 978-3-030-74296-6. DOI: 10.1007/978-3-030-74296-6_6.
- [63] Assunta Cerone, Elham Naghizade, Falk Scholer, Devi Mallal, Russell Skelton, and Damiano Spina. Watch 'n' Check: Towards a Social Media Monitoring Tool to Assist Fact-Checking Experts. In: *Proceedings of the 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*. 2020, pp. 607–613. DOI: 10.1109/DSAA49011.2020.00085.
- [64] Sara Ceschia, Kevin Roitero, Gianluca Demartini, Stefano Mizzaro, Luca Di Gaspero, and Andrea Schaerf. Task design in complex crowdsourcing experiments: Item assignment optimization. In: *Computers & Operations Research* 148 (2022), p. 105995. ISSN: 0305-0548. DOI: 0.1016/j.cor.2022.105995.
- [65] Praveen Chandar and Ben Carterette. Estimating Clickthrough Bias in the Cascade Model. In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. CIKM '18. Torino, Italy: Association for Computing Machinery, 2018, 1587–1590. ISBN: 9781450360142. DOI: 10.1145/3269206.3269315.
- [66] Jesse J. Chandler and Gabriele Paolacci. Lie for a Dime: When Most Prescreening Responses Are Honest but Most Study Participants Are Impostors. In: *Social Psychological and Personality Science* 8.5 (2017), pp. 500–508. DOI: 10.1177/1948550617698203.
- [67] Alessandro Checco and Gianluca Demartini. Pairwise, Magnitude, or Stars: What's the Best Way for Crowds to Rate? In: *Computing Research Repository* abs/1609.00683 (2016). arXiv: 1609.00683. URL: <http://arxiv.org/abs/1609.00683>.
- [68] Alessandro Checco, Kevin Roitero, Eddy Maddalena, Stefano Mizzaro, and Gianluca Demartini. Let's Agree to Disagree: Fixing Agreement Measures for Crowdsourcing. In: *Proceedings of the Fifth AAI Conference on Human Computation and Crowdsourcing*. Ed. by Steven Dow and Adam Tauman Kalai. AAI Press, 2017, pp. 11–20. URL: <https://aaai.org/ocs/index.php/HCOMP/HCOMP17/paper/view/15927>.
- [69] Charles Chen, Sungchul Kim, Hung Bui, Ryan Rossi, Eunye Koh, Branislav Kveton, and Razvan Bunescu. Predictive Analysis by Leveraging Temporal User Behavior and User Embeddings. In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. CIKM '18. Torino, Italy: Association for Computing Machinery, 2018, 2175–2182. ISBN: 9781450360142. DOI: 10.1145/3269206.3272032.
- [70] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. Bias and Debias in Recommender System: A Survey and Future Directions. In: *CoRR* abs/2010.03240 (2020). arXiv: 2010.03240. URL: <https://arxiv.org/abs/2010.03240>.

- [71] Xinran Chen, Sei-Ching Joanna Sin, Yin-Leng Theng, and Chei Sian Lee. Why Students Share Misinformation on Social Media: Motivation, Gender, and Study-level Differences. In: *The Journal of Academic Librarianship* 41.5 (2015), pp. 583–592. ISSN: 0099-1333. DOI: 10.1016/j.acalib.2015.07.003.
- [72] Fei-Fei Cheng and Chin-Shan Wu. Debiasing the framing effect: The effect of warning and involvement. In: *Decision Support Systems* 49.3 (2010), pp. 328–334. ISSN: 0167-9236. DOI: 10.1016/j.dss.2010.04.002.
- [73] Chun-Wei Chiang, Anna Kasunic, and Saiph Savage. Crowd Coach: Peer Coaching for Crowd Workers' Skill Growth. In: *Proceedings of the ACM on Human-Computer Interaction* 2.CSCW (Nov. 2018). DOI: 10.1145/3274306.
- [74] Wai Ki Ching, Eric S. Fung, and Michael K. Ng. Higher-order Markov chain models for categorical data sequences*. In: *Naval Research Logistics* 51.4 (2004), pp. 557–574. DOI: 0.1002/nav.20017.
- [75] Michael Chmielewski and Sarah C. Kucker. An MTurk Crisis? Shifts in Data Quality and the Impact on Study Results. In: *Social Psychological and Personality Science* 11.4 (2020), pp. 464–473. DOI: 10.1177/1948550619875149.
- [76] Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M. Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. Computational Fact Checking from Knowledge Networks. In: *PloS One* 10.6 (June 2015), pp. 1–13. DOI: 10.1371/journal.pone.0128193.
- [77] Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoti, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. The COVID-19 social media infodemic. In: *Scientific Reports* 10.1 (Oct. 2020), p. 16598. ISSN: 2045-2322. DOI: 10.1038/s41598-020-73510-5.
- [78] Anna E Clark and Yoshihisa Kashima. Stereotypes help people connect with others in the community: a situated functional analysis of the stereotype consistency bias in communication. In: *Journal of Personality and Social Psychology* 93.6 (Dec. 2007), pp. 1028–1039. DOI: 10.1037/0022-3514.93.6.1028.
- [79] Benjamin Y Clark, Nicholas Zingale, Joseph Logan, and Jeffrey Brudney. "A framework for using crowdsourcing in government". In: *Social Entrepreneurship: Concepts, Methodologies, Tools, and Applications*. 2019, pp. 405–425. DOI: 10.4018/978-1-5225-8182-6.ch020.
- [80] Charles L. A. Clarke, Saira Rizvi, Mark D. Smucker, Maria Maistro, and Guido Zuccon. Overview of the TREC 2020 Health Misinformation Track. In: *Proceedings of the Twenty-Ninth Text REtrieval Conference*. Ed. by Ellen M. Voorhees and Angela Ellis. Vol. 1266. NIST Special Publication. Gaithersburg, Maryland, USA: National Institute of Standards and Technology (NIST), 2020. URL: <https://trec.nist.gov/pubs/trec29/papers/OVERVIEW.HM.pdf>.
- [81] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Cambridge, MA, USA: Elsevier, 1977. ISBN: 978-0-12-179060-8. DOI: 10.1016/C2013-0-10517-X.

- [82] Meri Coleman and T. L. Liau. A computer readability formula designed for machine scoring. In: *Journal of Applied Psychology* 60 (1975), pp. 283–284. DOI: 10.1037/h0076540.
- [83] William Jay Conover, Armando Jesús Guerrero-Serrano, and Víctor Gustavo Tercero-Gómez. An update on ‘a comparative study of tests for homogeneity of variance’. In: *Journal of Statistical Computation and Simulation* 88.8 (2018), pp. 1454–1469. DOI: 10.1080/00949655.2018.1438437.
- [84] William Jay Conover, Mark E. Johnson, and Myrle M. Mohnson. A Comparative Study of Tests for Homogeneity of Variances, with Applications to the Outer Continental Shelf Bidding Data. In: *Technometrics* 23.4 (1981), pp. 351–361. DOI: 10.1080/00401706.1981.10487680.
- [85] Niall J Conroy, Victoria L Rubin, and Yimin Chen. Automatic Deception Detection: Methods for Finding Fake News. In: *Proceedings of the Association for Information Science and Technology* 52.1 (2015), pp. 1–4. DOI: 10.1002/pra2.2015.145052010082.
- [86] Jessica A Cooper, Marissa A Gorlick, Taylor Denny, Darrell A Worthy, Christopher G Beevers, and W Todd Maddox. Training attention improves decision making in individuals with elevated self-reported depressive symptoms. In: *Cognitive, Affective, & Behavioral Neuroscience* 14.2 (June 2014), pp. 729–741. DOI: 10.3758/s13415-013-0220-4.
- [87] Leda Cosmides and John Tooby. Better than Rational: Evolutionary Psychology and the Invisible Hand. In: *The American Economic Review* 84.2 (1994), pp. 327–332. ISSN: 00028282. URL: <http://www.jstor.org/stable/2117853>.
- [88] Sylvie Coste-Marquis, Sébastien Konieczny, Pierre Marquis, and Mohand Akli Ouali. Selecting Extensions in Weighted Argumentation Frameworks. In: *Proceedings of The 4th International Conference on Computational Models of Argument*. Ed. by Bart Verheij, Stefan Szeider, and Stefan Woltran. Vol. 245. Frontiers in Artificial Intelligence and Applications. Vienna, Austria: IOS Press, 2012, pp. 342–349. DOI: 10.3233/978-1-61499-111-3-342.
- [89] Sylvie Coste-Marquis, Sébastien Konieczny, Pierre Marquis, and Mohand Akli Ouali. Weighted Attacks in Argumentation Frameworks. In: *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*. KR’12. Rome, Italy: AAAI Press, 2012, 593–597. ISBN: 9781577355601. DOI: 10.5555/3031843.3031917.
- [90] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. An Experimental Comparison of Click Position-Bias Models. In: *Proceedings of the 2008 International Conference on Web Search and Data Mining*. WSDM ’08. Palo Alto, California, USA: Association for Computing Machinery, 2008, 87–94. ISBN: 9781595939272. DOI: 10.1145/1341531.1341545.
- [91] Mary Cummings. Automation Bias in Intelligent Time Critical Decision Support Systems. In: *AIAA 1st Intelligent Systems Technical Conference*. 2012. DOI: 10.2514/6.2004-6313.

- [92] Neil Dagnall, Andrew Denovan, Kenneth Graham Drinkwater, and Andrew Parker. An Evaluation of the Belief in Science Scale. In: *Frontiers in Psychology* 10 (2019). ISSN: 1664-1078. DOI: 10.3389/fpsyg.2019.00861.
- [93] Edgar Dale and Jeanne S. Chall. A Formula for Predicting Readability. In: *Educational Research Bulletin* 27.1 (1948), pp. 11–28. ISSN: 15554023. URL: <https://www.jstor.org/stable/1473169>.
- [94] Jeffrey Dalton, Sophie Fischer, Paul Owoicho, Filip Radlinski, Federico Rossetto, Johanne R. Trippas, and Hamed Zamani. Conversational Information Seeking: Theory and Application. In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '22. Madrid, Spain: Association for Computing Machinery, 2022, 3455–3458. ISBN: 9781450387323. DOI: 10.1145/3477495.3532678.
- [95] Timothy M. Daly and Rajan Natarajan. Swapping bricks for clicks: Crowdsourcing longitudinal data on Amazon Turk. In: *Journal of Business Research* 68.12 (2015), pp. 2603–2609. ISSN: 0148-2963. DOI: 10.1016/j.jbusres.2015.05.001.
- [96] T.K. Das and Bing-Sheng Teng. Cognitive Biases and Strategic Decision Processes: An Integrative Perspective. In: *Journal of Management Studies* 36.6 (1999), pp. 757–778. DOI: 10.1111/1467-6486.00157.
- [97] J. R. de Haan, R. Wehrens, S. Bauerschmidt, E. Piek, R. C. van Schaik, and L. M. C. Buydens. Interpretation of ANOVA models for microarray data using PCA. In: *Bioinformatics* 23.2 (Nov. 2006), pp. 184–190. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bt1572.
- [98] Gianluca Demartini, Djellel Eddine Difallah, Ujwal Gadiraju, and Michele Catasta. An Introduction to Hybrid Human-Machine Information Systems. In: *Found. Trends Web Sci.* 7.1 (Dec. 2017), 1–87. ISSN: 1555-077X. DOI: 10.1561/18000000025.
- [99] Gianluca Demartini, Stefano Mizzaro, and Damiano Spina. Human-in-the-loop Artificial Intelligence for Fighting Online Misinformation: Challenges and Opportunities. In: *IEEE Data Engineering Bulletin* 43.3 (2020), pp. 65–74. URL: <http://sites.computer.org/debull/A20sept/p65.pdf>.
- [100] Ronald Denaux and Jose Manuel Gomez-Perez. Linked Credibility Reviews for Explainable Misinformation Detection. In: *The Semantic Web – ISWC 2020*. Ed. by Jeff Z. Pan, Valentina Tamma, Claudia d’Amato, Krzysztof Janowicz, Bo Fu, Axel Polleres, Oshani Seneviratne, and Lalana Kagal. Cham: Springer International Publishing, 2020, pp. 147–163. ISBN: 978-3-030-62419-4. DOI: 10.1007/978-3-030-62419-4_9.
- [101] Aakash Desai, Jeremy Warner, Nicole Kuderer, Mike Thompson, Corrie Painter, Gary Lyman, and Gilberto Lopes. Crowdsourcing a Crisis Response for COVID-19 in Oncology. In: *Nature Cancer* 1.5 (Apr. 2020), pp. 473–476. DOI: 10.1038/s43018-020-0065-z.
- [102] Mukund Deshpande and George Karypis. Selective Markov Models for Predicting Web Page Accesses. In: *ACM Transactions on Internet Technology* 4.2 (May 2004), 163–184. ISSN: 1533-5399. DOI: 10.1145/990301.990304.

- [103] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: <https://aclanthology.org/N19-1423>.
- [104] A. Diaz. Through the Google Goggles. In: *Web Search: Multidisciplinary Perspectives*. Springer Berlin Heidelberg, 2008, pp. 11–34. DOI: 10.1007/978-3-540-75829-7_2.
- [105] Djellel Difallah, Elena Filatova, and Panos Ipeirotis. Demographics and Dynamics of Mechanical Turk Workers. In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. WSDM '18*. Marina Del Rey, CA, USA: Association for Computing Machinery, 2018, 135–143. ISBN: 9781450355810. DOI: 10.1145/3159652.3159661.
- [106] Djellel Difallah, Elena Filatova, and Panos Ipeirotis. Demographics and Dynamics of Mechanical Turk Workers. In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. WSDM '18*. Marina Del Rey, CA, USA: Association for Computing Machinery, 2018, 135–143. ISBN: 9781450355810. DOI: 10.1145/3159652.3159661.
- [107] Djellel Eddine Difallah, Michele Catasta, Gianluca Demartini, and Philippe Coudré-Maroux. Scaling-Up the Crowd: Micro-Task Pricing Schemes for Worker Retention and Latency Improvement. In: *Proceedings of the Second AAAI Conference on Human Computation and Crowdsourcing*. Ed. by Jeffrey P. Bigham and David C. Parkes. Pittsburgh, Pennsylvania, USA: AAAI, 2014. URL: <http://www.aaai.org/ocs/index.php/HCOMP/HCOMP14/paper/view/8958>.
- [108] Evanthia Dimara, Steven Franconeri, Catherine Plaisant, Anastasia Bezerianos, and Pierre Dragicevic. A Task-Based Taxonomy of Cognitive Biases for Information Visualization. In: *IEEE Transactions on Visualization and Computer Graphics* 26.2 (2020), pp. 1413–1432. DOI: 10.1109/TVCG.2018.2872577.
- [109] Xing Dongshan and Shen Junyi. A new Markov model for Web access prediction. In: *Computing in Science & Engineering* 4.6 (2002), pp. 34–39. DOI: 10.1109/MCISE.2002.1046594.
- [110] Tim Draws, Oana Inel, Nava Tintarev, Christian Baden, and Benjamin Timmermans. Comprehensive Viewpoint Representations for a Deeper Understanding of User Interactions With Debated Topics. In: *ACM SIGIR Conference on Human Information Interaction and Retrieval. CHIIR '22*. Regensburg, Germany: Association for Computing Machinery, 2022, 135–145. ISBN: 9781450391863. DOI: 10.1145/3498366.3505812.
- [111] Tim Draws, David La Barbera, Michael Soprano, Kevin Roitero, Davide Ceolin, Alessandro Checco, and Stefano Mizzaro. The Effects of Crowd Worker Biases in Fact-Checking Tasks. In: *2022 ACM Conference on Fairness, Accountability, and Transparency. FAccT '22*. Seoul, Republic of Korea: Association for Computing Machinery, 2022, 2114–2124. ISBN: 9781450393522. DOI: 10.1145/3531146.3534629.

- [112] Tim Draws, Alisa Rieger, Oana Inel, Ujwal Gadiraju, and Nava Tintarev. A Checklist to Combat Cognitive Biases in Crowdsourcing. In: *Proceedings of the Ninth AAAI Conference on Human Computation and Crowdsourcing 9.1* (Oct. 2021), pp. 48–59. DOI: 10.1609/hcomp.v9i1.18939.
- [113] Tim Draws, Nava Tintarev, Ujwal Gadiraju, Alessandro Bozzon, and Benjamin Timmermans. This Is Not What We Ordered: Exploring Why Biased Search Result Rankings Affect User Attitudes on Debated Topics. In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '21. Virtual Event, Canada: Association for Computing Machinery, 2021*, 295–305. ISBN: 9781450380379. DOI: 10.1145/3404835.3462851.
- [114] Angie Drobnic Holan. *The Principles of the Truth-O-Meter: PolitiFact's methodology for independent fact-checking*. (Accessed: 20.04.2021). 2021. URL: <https://www.politifact.com/article/2018/feb/12/principles-truth-o-meter-politifact-methodology-i/#Truth-O-Meter%20ratings>.
- [115] Beverly Dugan. Effects of assessor training on information use. In: *Journal of Applied Psychology* 73.4 (1988), pp. 743–748. DOI: 10.1037/0021-9010.73.4.743.
- [116] Anca Dumitrache, Oana Inel, Lora Aroyo, Benjamin Timmermans, and Chris Welty. CrowdTruth 2.0: Quality Metrics for Crowdsourcing with Disagreement. In: *Proceedings of the 1st Workshop on Subjectivity, Ambiguity and Disagreement in Crowdsourcing, and Short Paper Proceedings of the 1st Workshop on Disentangling the Relation Between Crowdsourcing and Bias Management co-located the 6th AAAI Conference on Human Computation and Crowdsourcing*. Ed. by Lora Aroyo, Anca Dumitrache, Praveen K. Paritosh, Alexander J. Quinn, Chris Welty, Alessandro Checco, Gianluca Demartini, Ujwal Gadiraju, and Cristina Sarasua. Vol. 2276. CEUR Workshop Proceedings. Zürich, Switzerland: CEUR-WS.org, 2018, pp. 11–18. URL: <http://ceur-ws.org/Vol-2276/paper2.pdf>.
- [117] Phan Minh Dung. On the acceptability of arguments and its fundamental role in non-monotonic reasoning, logic programming and n-person games. In: *Artificial Intelligence* 77.2 (1995), pp. 321–357. ISSN: 0004-3702. DOI: 10.1016/0004-3702(94)00041-X.
- [118] Olive Jean Dunn. Multiple Comparisons Using Rank Sums. In: *Technometrics* 6.3 (1964), pp. 241–252. DOI: 10.2307/1266041.
- [119] Paul E. Dunne, Anthony Hunter, Peter McBurney, Simon Parsons, and Michael Wooldridge. Weighted argument systems: Basic definitions, algorithms, and complexity results. In: *Artificial Intelligence* 175.2 (2011), pp. 457–486. ISSN: 0004-3702. DOI: 10.1016/j.artint.2010.09.005.
- [120] D. Dunning, D. W. Griffin, J. D. Milojkovic, and L. Ross. The overconfidence effect in social prediction. In: *Journal of Personality and Social Psychology* 58.4 (Apr. 1990), pp. 568–581. DOI: 10.1037/0022-3514.58.4.568.
- [121] David Dunning. The Dunning–Kruger Effect: On Being Ignorant of One's Own Ignorance. In: *Advances in Experimental Social Psychology* 44 (2011). Ed. by James M. Olson and Mark P. Zanna, pp. 247–296. ISSN: 0065-2601. DOI: 10.1016/B978-0-12-385522-0.00005-6.

- [122] Gregory Eady, Jonathan Nagler, Andy Guess, Jan Zilinsky, and Joshua A. Tucker. How Many People Live in Political Bubbles on Social Media? Evidence From Linked Survey and Twitter Data. In: *SAGE Open* 9.1 (2019), p. 2158244019832705. DOI: 10.1177/2158244019832705.
- [123] Ullrich K. H. Ecker, Stephan Lewandowsky, and David T. W. Tang. Explicit warnings reduce but do not eliminate the continued influence of misinformation. In: *Memory & Cognition* 38.8 (Dec. 2010), pp. 1087–1100. ISSN: 1532-5946. DOI: 10.3758/MC.38.8.1087.
- [124] Effectiviology. *The Bandwagon Effect: Why People Tend to Follow the Crowd*. <https://effectiviology.com/bandwagon/>. (Accessed: 15-12-2021). 2020.
- [125] J. Ehrlinger, W.O. Readinger, and B. Kim. “Decision-Making and Cognitive Biases”. In: *Encyclopedia of Mental Health (Second Edition)*. Ed. by Howard S. Friedman. Second Edition. Oxford: Academic Press, 2016, pp. 5–12. ISBN: 978-0-12-397753-3. DOI: 10.1016/B978-0-12-397045-9.00206-8.
- [126] Carsten Eickhoff. Cognitive Biases in Crowdsourcing. In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. WSDM '18. Marina Del Rey, CA, USA: Association for Computing Machinery, 2018, 162–170. ISBN: 9781450355810. DOI: 10.1145/3159652.3159654.
- [127] Carsten Eickhoff and Arjen P. de Vries. Increasing cheat robustness of crowdsourcing tasks. In: *Information Retrieval* 16.2 (Apr. 2013), pp. 121–137. ISSN: 1573-7659. DOI: 10.1007/s10791-011-9181-9.
- [128] Hillel J Einhorn and Robin M Hogarth. Confidence in judgment: Persistence of the illusion of validity. In: *Psychological Review* 85.5 (1978), pp. 395–416. DOI: 10.1037/0033-295X.85.5.395.
- [129] Tamer Elsayed, Preslav Nakov, Alberto Barrón-Cedeño, Maram Hasanain, Reem Suwaileh, Giovanni Da San Martino, and Pepa Atanasova. Overview of the CLEF-2019 CheckThat! Lab: Automatic Identification and Verification of Claims. In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Ed. by Fabio Crestani, Martin Braschler, Jacques Savoy, Andreas Rauber, Henning Müller, David E. Losada, Gundula Heinatz Bürki, Linda Cappellato, and Nicola Ferro. Cham: Springer International Publishing, 2019, pp. 301–321. ISBN: 978-3-030-28577-7. DOI: 10.1007/978-3-030-28577-7_25.
- [130] Susan E. Embretson. Item Response Theory Models and Spurious Interaction Effects in Factorial ANOVA Designs. In: *Applied Psychological Measurement* 20.3 (1996), pp. 201–212. DOI: 10.1177/014662169602000302.
- [131] Robert Epstein and Ronald E. Robertson. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. In: *Proceedings of the National Academy of Sciences* 112.33 (2015), E4512–E4521. DOI: 10.1073/pnas.1419828112.

- [132] Ziv Epstein, Gordon Pennycook, and David Rand. Will the Crowd Game the Algorithm? Using Layperson Judgments to Combat Misinformation on Social Media by Downranking Distrusted Sources. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2020, 1–11. ISBN: 9781450367080. DOI: 10.1145/3313831.3376232.
- [133] Edgar Erdfelder, Franz Faul, and Axel Buchner. GPOWER: A general power analysis program. In: *Behavior Research Methods, Instruments, & Computers* 28.1 (Mar. 1996), pp. 1–11. DOI: 10.3758/BF03203630.
- [134] Thomas Erickson. Some Thoughts on a Framework for Crowdsourcing. In: *CHI 2011 Workshop on Crowdsourcing and Human Computation*. 2011, pp. 1–4.
- [135] Pete Etchells. Declinism: is the world actually getting worse. In: *The Guardian* 15 (2015), pp. 1087–1089.
- [136] Nathan Evans, Darren Edge, Jonathan Larson, and Christopher White. News Provenance: Revealing News Text Reuse at Web-Scale in an Augmented News Search Experience. In: *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI EA '20. Honolulu, HI, USA: Association for Computing Machinery, 2020, 1–8. ISBN: 9781450368193. DOI: 10.1145/3334480.3375225.
- [137] FactCheck.org. *Our Process*. <https://www.factcheck.org/our-process/>. (Accessed: 15–12–2021). 2020.
- [138] Ju Fan, Guoliang Li, Beng Chin Ooi, Kian-lee Tan, and Jianhua Feng. ICrowd: An Adaptive Crowdsourcing Framework. In: *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. SIGMOD '15. Melbourne, Victoria, Australia: Association for Computing Machinery, 2015, 1015–1030. ISBN: 9781450327589. DOI: 10.1145/2723372.2750550.
- [139] Shaoyang Fan, Ujwal Gadiraju, Alessandro Checco, and Gianluca Demartini. CrowdCOOP: Sharing Risks and Rewards in Crowdsourcing. In: *Proceedings of ACM Human-Computer Interaction* 4.CSCW2 (Oct. 2020). DOI: 10.1145/3415203.
- [140] William Ferreira and Andreas Vlachos. Emergent: a novel data-set for stance classification. In: *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Kevin Knight, Ani Nenkova, and Owen Rambow. The Association for Computational Linguistics, 2016, pp. 1163–1168. DOI: 10.18653/v1/n16-1138.
- [141] Nicola Ferro, Yubin Kim, and Mark Sanderson. Using Collection Shards to Study Retrieval Performance Effect Sizes. In: *ACM Transactions on Information Systems* 37.3 (Mar. 2019). ISSN: 1046-8188. DOI: 10.1145/3310364.
- [142] Nicola Ferro and Gianmaria Silvello. A General Linear Mixed Models Approach to Study System Component Effects. In: *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '16. Pisa, Italy: ACM, 2016, 25–34. ISBN: 9781450340694. DOI: 10.1145/2911451.2911530.
- [143] Nicola Ferro and Gianmaria Silvello. Toward an Anatomy of IR System Component Performances. In: *Journal of the Association for Information Science and Technology* 69.2 (2018), pp. 187–200. DOI: 10.1002/asi.23910.

- [144] Catherine S. Fichten and Betty Sunerton. Popular Horoscopes and the “Barnum Effect”. In: *The Journal of Psychology* 114.1 (1983), pp. 123–134. DOI: 10.1080/00223980.1983.9915405.
- [145] Kenneth L. Fisher and Meir Statman. Cognitive Biases in Market Forecasts. In: *The Journal of Portfolio Management* 27.1 (2000), pp. 72–81. ISSN: 0095-4918. DOI: 10.3905/jpm.2000.319785.
- [146] Seth Flaxman, Sharad Goel, and Justin M. Rao. Filter Bubbles, Echo Chambers, and Online News Consumption. In: *Public Opinion Quarterly* 80.S1 (Mar. 2016), pp. 298–320. ISSN: 0033-362X. DOI: 10.1093/poq/nfw006.
- [147] Joseph L Fleiss and Jacob Cohen. The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability. In: *Educational and Psychological Measurement* 33.3 (1973), pp. 613–619. DOI: 10.1177/001316447303300309.
- [148] Shane Frederick. Cognitive Reflection and Decision Making. In: *Journal of Economic Perspectives* 19.4 (Dec. 2005), pp. 25–42. DOI: 10.1257/089533005775196732.
- [149] Norbert Fuhr. Some Common Mistakes In IR Evaluation, And How They Can Be Avoided. In: *SIGIR Forum* 51.3 (Feb. 2018), 32–41. ISSN: 0163-5840. DOI: 10.1145/3190580.3190586.
- [150] Adrian Furnham and Hua Chu Boo. A literature review of the anchoring effect. In: *The Journal of Socio-Economics* 40.1 (2011), pp. 35–42. ISSN: 1053-5357. DOI: 10.1016/j.socec.2010.10.008.
- [151] Ujwal Gadiraju, Sebastian Möller, Martin Nöllenburg, Dietmar Saupe, Sebastian Egger-Lampl, Daniel Archambault, and Brian Fisher. “Crowdsourcing Versus the Laboratory: Towards Human-Centered Experiments Using the Crowd”. In: *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments*. Ed. by Daniel Archambault, Helen Purchase, and Tobias Hoßfeld. Cham: Springer International Publishing, 2017, pp. 6–26. ISBN: 978-3-319-66435-4. DOI: 10.1007/978-3-319-66435-4_2.
- [152] Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. Clarity is a Worthwhile Quality: On the Role of Task Clarity in Microtask Crowdsourcing. In: *Proceedings of the 28th ACM Conference on Hypertext and Social Media*. HT '17. Prague, Czech Republic: Association for Computing Machinery, 2017, 5–14. ISBN: 9781450347082. DOI: 10.1145/3078714.3078715.
- [153] Riccardo Gallotti, Francesco Valle, Nicola Castaldo, Pierluigi Sacco, and Manlio De Domenico. *Assessing the risks of ‘infodemics’ in response to COVID-19 epidemics*. Dec. 2020. DOI: 10.1038/s41562-020-00994-6.
- [154] Pepa Gencheva, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, and Ivan Koychev. A Context-Aware Approach for Detecting Worth-Checking Claims in Political Debates. In: *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*. Varna, Bulgaria: INCOMA Ltd., Sept. 2017, pp. 267–276. DOI: 10.26615/978-954-452-049-6_037.

- [155] Amira Ghenai and Yelena Mejova. Catching Zika Fever: Application of Crowdsourcing and Machine Learning for Tracking Health Misinformation on Twitter. In: *2017 IEEE International Conference on Healthcare Informatics*. 2017, pp. 518–518. DOI: 10.1109/ICHI.2017.58.
- [156] Anindya Ghose and Panagiotis G. Ipeirotis. Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics. In: *IEEE Transactions on Knowledge and Data Engineering* 23.10 (2011), pp. 1498–1512. DOI: 10.1109/TKDE.2010.188.
- [157] Aniruddha Ghosh, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, John Barn- den, and Antonio Reyes. SemEval-2015 Task 11: Sentiment Analysis of Figurative Language in Twitter. In: *Proceedings of the 9th International Workshop on Semantic Eval- uation (SemEval 2015)*. Denver, Colorado: Association for Computational Linguistics, June 2015, pp. 470–478. DOI: 10.18653/v1/S15-2080.
- [158] Anastasia Giachanou and Paolo Rosso. The Battle Against Online Harmful Informa- tion: The Cases of Fake News and Hate Speech. In: *Proceedings of the 29th ACM In- ternational Conference on Information & Knowledge Management*. Virtual Event, Ireland: Association for Computing Machinery, 2020, pp. 3503–3504. ISBN: 9781450368599. DOI: 10.1145/3340531.3412169.
- [159] Gerd Gigerenzer and R. Selten. “Bounded and Rational”. In: *Philosophie: Grundla- gen und Anwendungen/Philosophy: Foundations and Applications*. Brill | mentis, 2008, pp. 233–257. ISBN: 9783969750056. DOI: 10.30965/9783969750056_016.
- [160] Thomas Gillier, Cédric Chaffois, Mustapha Belkhouja, Yannig Roth, and Barry L. Bayus. The effects of task instructions in crowdsourcing innovative ideas. In: *Tech- nological Forecasting and Social Change* 134 (2018), pp. 35–44. ISSN: 0040-1625. DOI: 10.1016/j.techfore.2018.05.005.
- [161] Yvette Graham, Timothy Baldwin, and Nitika Mathur. Accurate Evaluation of Segment- level Machine Translation Metrics. In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Tech- nologies*. Denver, Colorado: Association for Computational Linguistics, May 2015, pp. 1183–1191. DOI: 10.3115/v1/N15-1124.
- [162] Lucas Graves. *Understanding the Promise and Limits of Automated Fact-Checking*. 2018. URL: <https://reutersinstitute.politics.ox.ac.uk/our-research/understanding- promise-and-limits-automated-fact-checking>.
- [163] Mihajlo Grbovic, Nemanja Djuric, Vladan Radosavljevic, Fabrizio Silvestri, and Narayan Bhamidipati. Context- and Content-Aware Embeddings for Query Rewrit- ing in Sponsored Search. In: *Proceedings of the 38th International ACM SIGIR Confer- ence on Research and Development in Information Retrieval*. SIGIR ’15. Santiago, Chile: Association for Computing Machinery, 2015, 383–392. ISBN: 9781450336215. DOI: 10.1145/2766462.2767709.

- [164] Mihajlo Grbovic, Vladan Radosavljevic, Nemanja Djuric, Narayan Bhamidipati, Jaikit Savla, Varun Bhagwan, and Doug Sharp. E-Commerce in Your Inbox: Product Recommendations at Scale. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '15. Sydney, NSW, Australia: Association for Computing Machinery, 2015, 1809–1818. ISBN: 9781450336642. DOI: 10.1145/2783258.2788627.
- [165] Stephan Grimmelikhuijsen and Eva Knies. Validating a scale for citizen trust in government organizations. In: *International Review of Administrative Sciences* 83.3 (2017), pp. 583–601. DOI: 10.1177/0020852315585950.
- [166] David Groome and Michael Eysenck. *An Introduction to Applied Cognitive Psychology*. Psychology Press, 2016.
- [167] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. ICML'17. Sydney, NSW, Australia: JMLR.org, 2017, 1321–1330. DOI: 10.5555/3305381.3305518.
- [168] Matthew Haigh. Has the Standard Cognitive Reflection Test Become a Victim of Its Own Success? In: *Advances in Cognitive Psychology* 12.3 (2016), p. 145. DOI: 10.5709/acp-0193-5.
- [169] David L. Hamilton and Robert K. Gifford. Illusory correlation in interpersonal perception: A cognitive basis of stereotypic judgments. In: *Journal of Experimental Social Psychology* 12.4 (1976), pp. 392–407. ISSN: 0022-1031. DOI: 10.1016/S0022-1031(76)80006-6.
- [170] Aymen Hamrouni, Hakim Ghazzai, Mounir Frikha, and Yehia Massoud. A Spatial Mobile Crowdsourcing Framework for Event Reporting. In: *IEEE Transactions on Computational Social Systems* 7.2 (2020), pp. 477–491. DOI: 10.1109/TCSS.2020.2967585.
- [171] Lei Han, Alessandro Checco, Djellel Difallah, Gianluca Demartini, and Shazia Sadiq. Modelling User Behavior Dynamics with Embeddings. In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. CIKM '20. Virtual Event, Ireland: Association for Computing Machinery, 2020, 445–454. ISBN: 9781450368599. DOI: 10.1145/3340531.3411985.
- [172] Lei Han, Kevin Roitero, Ujwal Gadiraju, Cristina Sarasua, Alessandro Checco, Eddy Maddalena, and Gianluca Demartini. All Those Wasted Hours: On Task Abandonment in Crowdsourcing. In: *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. WSDM '19. Melbourne VIC, Australia: Association for Computing Machinery, 2019, 321–329. ISBN: 9781450359405. DOI: 10.1145/3289600.3291035.
- [173] Lei Han, Kevin Roitero, Eddy Maddalena, Stefano Mizzaro, and Gianluca Demartini. On Transforming Relevance Scales. In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. CIKM '19. Beijing, China: Association for Computing Machinery, 2019, 39–48. ISBN: 9781450369763. DOI: 10.1145/3357384.3357988.

- [174] Tom L. Han, Kevin Roitero, Ujwal Gadiraju, C. Sarasua, A. Checco, Eddy Maddalena, and Gianluca Demartini. The Impact of Task Abandonment in Crowdsourcing. In: *IEEE Transactions on Knowledge & Data Engineering* 1.1 (Oct. 2019), pp. 1–1. ISSN: 1558-2191. DOI: 10.1109/TKDE.2019.2948168.
- [175] Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. UKP-Athene: Multi-Sentence Textual Entailment for Claim Verification. In: *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 103–108. DOI: 10.18653/v1/W18-5516.
- [176] Casper Hansen, Christian Hansen, Stephen Alstrup, Jakob Grue Simonsen, and Christina Lioma. Neural Check-Worthiness Ranking with Weak Supervision: Finding Sentences for Fact-Checking. In: *Companion Proceedings of The 2019 World Wide Web Conference. WWW '19*. San Francisco, USA: Association for Computing Machinery, 2019, 994–1000. ISBN: 9781450366755. DOI: 10.1145/3308560.3316736.
- [177] Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P. Bigham. A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. CHI '18*. Montreal QC, Canada: Association for Computing Machinery, 2018, 1–14. ISBN: 9781450356206. DOI: 10.1145/3173574.3174023.
- [178] John H. Harvey, Jerri P. Town, and Kerry L. Yarkin. How fundamental is "the fundamental attribution error"? In: *Journal of Personality and Social Psychology* 40.2 (1981), pp. 346–349. DOI: 10.1037/0022-3514.40.2.346.
- [179] Martie G. Haselton and Daniel Nettle. The Paranoid Optimist: An Integrative Evolutionary Model of Cognitive Biases. In: *Personality and Social Psychology Review* 10.1 (2006), pp. 47–66. DOI: 10.1207/s15327957pspr1001_3.
- [180] Martie G. Haselton, Daniel Nettle, and Damian R. Murray. The Evolution of Cognitive Bias. In: *The Handbook of Evolutionary Psychology*. John Wiley & Sons, Ltd, 2015. Chap. 41, pp. 1–20. ISBN: 9781119125563. DOI: 10.1002/9781119125563.evpsych241.
- [181] Naeemul Hassan, Bill Adair, James T Hamilton, Chengkai Li, Mark Tremayne, Jun Yang, and Cong Yu. The Quest to Automate Fact-Checking. In: *Proceedings of the 2015 Computation + Journalism Symposium*. Oct. 2015.
- [182] Kenji Hata, Ranjay Krishna, Li Fei-Fei, and Michael S. Bernstein. A Glimpse Far into the Future: Understanding Long-Term Crowd Worker Quality. In: *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing. CSCW '17*. Portland, Oregon, USA: Association for Computing Machinery, 2017, 889–901. ISBN: 9781450343350. DOI: 10.1145/2998181.2998248.
- [183] Madeline E. Heilman. Gender stereotypes and workplace bias. In: *Research in Organizational Behavior* 32 (2012), pp. 113–135. ISSN: 0191-3085. DOI: 10.1016/j.riob.2012.11.003.
- [184] Danula Hettiachchi, Vassilis Kostakos, and Jorge Goncalves. A Survey on Task Assignment in Crowdsourcing. In: *ACM Computing Surveys* 55.3 (Feb. 2022). ISSN: 0360-0300. DOI: 10.1145/3494522.

- [185] Danula Hettiachchi, Mike Schaeckermann, Tristan J. McKinney, and Matthew Lease. The Challenge of Variable Effort Crowdsourcing and How Visible Gold Can Help. In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW2 (Oct. 2021). DOI: 10.1145/3476073.
- [186] M. Hilbert. Toward a synthesis of cognitive biases: how noisy information processing can bias human decision making. In: *Psychology Bulletin* 138.2 (Mar. 2012), pp. 211–237. DOI: 10.1037/a0025940.
- [187] Christopher J Holden, Trevor Dennie, and Adam D Hicks. Assessing the reliability of the M5-120 on Amazon’s mechanical Turk. In: *Computers in Human Behavior* 29.4 (2013), pp. 1749–1754. ISSN: 0747-5632. DOI: 10.1016/j.chb.2013.02.020.
- [188] Benjamin Horne and Sibel Adali. This Just In: Fake News Packs A Lot In Title, Uses Simpler, Repetitive Content in Text Body, More Similar To Satire Than Real News. In: *Proceedings of the International AAAI Conference on Web and Social Media* 11.1 (May 2017). URL: <https://ojs.aaai.org/index.php/ICWSM/article/view/14976>.
- [189] Jeff Howe. The rise of crowdsourcing. In: *Wired Magazine* 14.6 (2006), pp. 1–4. URL: <https://www.wired.com/2006/06/crowds/>.
- [190] Hsiu-Fang Hsieh and Sarah E. Shannon. Three Approaches to Qualitative Content Analysis. In: *Qualitative Health Research* 15.9 (2005), pp. 1277–1288. DOI: 10.1177/1049732305276687.
- [191] Yingxiang Huang, Wentao Li, Fima Macheret, Rodney A Gabriel, and Lucila Ohno-Machado. A tutorial on calibration measurements and calibration models for clinical prediction models. In: *Journal of the American Medical Informatics Association* 27.4 (Feb. 2020), pp. 621–633. ISSN: 1527-974X. DOI: 10.1093/jamia/ocz228.
- [192] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. Understanding and Mitigating Worker Biases in the Crowdsourced Collection of Subjective Judgments. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI ’19. Glasgow, Scotland UK: Association for Computing Machinery, 2019, 1–12. ISBN: 9781450359702. DOI: 10.1145/3290605.3300637.
- [193] International Organization for Standardization. *ISO/IEC 25012:2008 Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model*. Tech. rep. ISO, 2008. URL: <https://www.iso.org/standard/35736.html>.
- [194] Lilly C. Irani and M. Six Silberman. Turkopticon: Interrupting Worker Invisibility in Amazon Mechanical Turk. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’13. Paris, France: Association for Computing Machinery, 2013, 611–620. ISBN: 9781450318990. DOI: 10.1145/2470654.2470742.
- [195] Jennifer Jerit. Issue Framing and Engagement: Rhetorical Strategy in Public Policy Debates. In: *Political Behavior* 30.1 (2008), pp. 1–24. ISSN: 01909320, 15736687. URL: <http://www.jstor.org/stable/40213302> (visited on 07/26/2022).

- [196] Jiepu Jiang, Daqing He, and James Allan. Comparing In Situ and Multidimensional Relevance Judgments. In: *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '17. Shinjuku, Tokyo, Japan: Association for Computing Machinery, 2017, 405–414. ISBN: 9781450350228. DOI: 10.1145/3077136.3080840.
- [197] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. Accurately Interpreting Clickthrough Data as Implicit Feedback. In: *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '05. Salvador, Brazil: Association for Computing Machinery, 2005, 154–161. ISBN: 1595930345. DOI: 10.1145/1076034.1076063.
- [198] Dominic D.P. Johnson, Daniel T. Blumstein, James H. Fowler, and Martie G. Haselton. The evolution of error: error management, cognitive constraints, and adaptive decision-making biases. In: *Trends in Ecology & Evolution* 28.8 (2013), pp. 474–481. ISSN: 0169-5347. DOI: 10.1016/j.tree.2013.05.014.
- [199] Emily L. Jones. The courtesy bias in South-East Asian surveys. In: *Social Research in Developing Countries: Surveys and Censuses in the Third World*. 1993, pp. 253–9. URL: <https://unesdoc.unesco.org/ark:/48223/pf0000016829>.
- [200] Garth S Jowett and Victoria O'Donnell. *Propaganda & persuasion*. Sage Publications, 2018. ISBN: 9781506371344. URL: <https://us.sagepub.com/en-us/nam/propaganda-persuasion/book250869>.
- [201] Wen-Hua Ju and Yehuda Vardi. A Hybrid High-Order Markov Chain Model for Computer Intrusion Detection. In: *Journal of Computational and Graphical Statistics* 10.2 (2001), pp. 277–295. ISSN: 10618600. URL: <http://www.jstor.org/stable/1391012>.
- [202] Mahdi Kafaee, Hanie Marhamati, and Shahriar Gharibzadeh. “Choice-supportive bias” in science: Explanation and mitigation. In: *Accountability in Research* 28.8 (2021), pp. 528–543. DOI: 10.1080/08989621.2021.1872377.
- [203] Beverly K. Kahn, Diane M. Strong, and Richard Y. Wang. Information Quality Benchmarks: Product and Service Performance. In: *Communications of the ACM* 45.4 (Apr. 2002), 184–192. ISSN: 0001-0782. DOI: 10.1145/505248.506007.
- [204] Daniel Kahneman. *Thinking, fast and slow*. New York: Macmillan, 2011. ISBN: 9780374275631 0374275637.
- [205] Daniel Kahneman and Shane Frederick. Representativeness revisited: Attribute substitution in intuitive judgment. In: *Heuristics and biases: The psychology of intuitive judgment* (2002), pp. 49–81. DOI: 10.1017/CB09780511808098.004.
- [206] Daniel Kahneman and Amos Tversky. On the psychology of prediction. In: *Psychological review* 80.4 (1973), pp. 237–251. DOI: 10.1037/h0034747.
- [207] Noriko Kando, ed. *Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access, NTCIR-7*. National Institute of Informatics (NII), 2008. ISBN: 978-4-86049-044-7. URL: <http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings7>.

- [208] Alireza Karduni, Ryan Wesslen, Sashank Santhanam, Isaac Cho, Svitlana Volkova, Dustin Arendt, Samira Shaikh, and Wenwen Dou. Can You Verifi This? Studying Uncertainty and Decision-Making about Misinformation in Visual Analytics. In: *Proceedings of the Twelfth International AAAI Conference on Web and Social Media*. 2018. URL: <https://ojs.aaai.org/index.php/ICWSM/article/view/15014/14864>.
- [209] Niklas Karlsson, George Loewenstein, and Duane Seppi. The ostrich effect: Selective attention to information. In: *Journal of Risk and Uncertainty* 38.2 (Apr. 2009), pp. 95–115. ISSN: 1573-0476. DOI: 10.1007/s11166-009-9060-6.
- [210] Alan E Kazdin. Artifact, bias, and complexity of assessment: The ABCs of reliability. In: *Journal of Applied Behavior Analysis* 10.1 (1977), pp. 141–150. DOI: 10.1901/jaba.1977.10-141.
- [211] Melissa G. Keith, Louis Tay, and Peter D. Harms. Systems Perspective of Amazon Mechanical Turk for Organizational Research: Review and Recommendations. In: *Frontiers in Psychology* 8 (2017). Ed. by Darren C. Treadway, p. 1359. DOI: 10.3389/fpsyg.2017.01359.
- [212] Diane Kelly and Leif Azzopardi. How Many Results per Page? A Study of SERP Size, Search Behavior and User Experience. In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '15. Santiago, Chile: Association for Computing Machinery, 2015, 183–192. ISBN: 9781450336215. DOI: 10.1145/2766462.2767732.
- [213] Ryan Kennedy, Scott Clifford, Tyler Burleigh, Philip D. Waggoner, Ryan Jewell, and Nicholas J. G. Winter. The shape of and solutions to the MTurk quality crisis. In: *Political Science Research and Methods* 8.4 (2020), 614–629. DOI: 10.1017/psrm.2020.6.
- [214] Johannes Kiesel, Damiano Spina, Henning Wachsmuth, and Benno Stein. The Meant, the Said, and the Understood: Conversational Argument Search and Cognitive Biases. In: *Proceedings of the 3rd Conference on Conversational User Interfaces*. CUI '21. Bilbao (online), Spain: Association for Computing Machinery, 2021. ISBN: 9781450389983. DOI: 10.1145/3469595.3469615.
- [215] Jooyeon Kim, Dongkwan Kim, and Alice Oh. Homogeneity-Based Transmissive Process to Model True and False News in Social Networks. In: *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. WSDM '19. Melbourne VIC, Australia: Association for Computing Machinery, 2019, 348–356. ISBN: 9781450359405. DOI: 10.1145/3289600.3291009.
- [216] Seonhoon Kim, Inho Kang, and Nojun Kwak. Semantic Sentence Matching with Densely-Connected Recurrent and Co-Attentive Information. In: *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*. AAAI'19/IAAI'19/EAAI'19. Honolulu, Hawaii, USA: AAAI Press, 2019. ISBN: 978-1-57735-809-1. DOI: 10.1609/aaai.v33i01.33016586.
- [217] Sunghan Kim, David Goldstein, Lynn Hasher, and Rose T Zacks. Framing effects in younger and older adults. In: *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 60.4 (2005), P215–P218. DOI: 10.1093/geronb/60.4.p215.

- [218] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. *Derivation Of New Readability Formulas (Automated Readability Index, Fog Count And Flesch Reading Ease Formula) For Navy Enlisted Personnel*. Research Branch Report 8-75. Chief of Naval Technical Training; Naval Air Station Memphis, 1975. URL: <https://stars.library.ucf.edu/istlibrary/56>.
- [219] Áron Kiss and Gábor Simonovits. Identifying the bandwagon effect in two-round elections. In: *Public Choice* 160.3 (Sept. 2014), pp. 327–344. ISSN: 1573-7101. DOI: 10.1007/s11127-013-0146-y.
- [220] Aniket Kittur, Ed H. Chi, and Bongwon Suh. Crowdsourcing user studies with Mechanical Turk. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '08. Florence, Italy: ACM, 2008, pp. 453–456. ISBN: 9781605580111. DOI: 10.1145/1357054.1357127.
- [221] Aniket Kittur, Boris Smus, Susheel Khamkar, and Robert E. Kraut. CrowdForge: Crowdsourcing Complex Work. In: *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*. UIST '11. Santa Barbara, California, USA: Association for Computing Machinery, 2011, 43–52. ISBN: 9781450307161. DOI: 10.1145/2047196.2047202.
- [222] Nikolaos Korfiatis, Elena García-Bariocanal, and Salvador Sánchez-Alonso. Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content. In: *Electronic Commerce Research and Applications* 11.3 (2012), pp. 205–217. ISSN: 1567-4223. DOI: 10.1016/j.e1erap.2011.10.003.
- [223] Travis Kriplean, Caitlin Bonnar, Alan Borning, Bo Kinney, and Brian Gill. Integrating On-demand Fact-checking with Public Dialogue. In: *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*. CSCW '14. Baltimore, Maryland, USA: Association for Computing Machinery, 2014, 1188–1199. ISBN: 9781450325400. DOI: 10.1145/2531602.2531677.
- [224] Klaus Krippendorff. Computing Krippendorff's Alpha-Reliability. In: *UPENN Libraries* 1 (2008), p. 43. URL: https://repository.upenn.edu/asc_papers/43.
- [225] William H Kruskal and W Allen Wallis. Use of Ranks in One-criterion Variance Analysis. In: *Journal of the American statistical Association* 47.260 (1952), pp. 583–621. DOI: 10.2307/2280779.
- [226] Paul Kubicek. The Commonwealth of Independent States: an example of failed regionalism? In: *Review of International Studies* 35.S1 (2009), 237–256. DOI: 10.1017/S026021050900850X.
- [227] Timur Kuran and Cass R Sunstein. Availability cascades and risk regulation. In: *Stanford Law Review* 51 (4 1998). URL: <https://ssrn.com/abstract=138144>.
- [228] Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. From Word Embeddings to Document Distances. In: *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*. ICML'15. Lille, France: JMLR.org, 2015, 957–966. DOI: 10.5555/3045118.3045221.

- [229] Mucahid Kutlu, Tyler McDonnell, Yasmine Barkallah, Tamer Elsayed, and Matthew Lease. Crowd vs. Expert: What Can Relevance Judgment Rationales Teach Us About Assessor Disagreement? In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. SIGIR '18. Ann Arbor, MI, USA: Association for Computing Machinery, 2018, 805–814. ISBN: 9781450356572. DOI: 10.1145/3209978.3210033.
- [230] Mucahid Kutlu, Tyler McDonnell, Tamer Elsayed, and Matthew Lease. Annotator Rationales for Labeling Tasks in Crowdsourcing. In: *Journal of Artificial Intelligence Research* 69 (2020), pp. 143–189. DOI: 10.1613/jair.1.12012.
- [231] Tarald O Kvålseth. Note on Cohen's kappa. In: *Psychological reports* 65.1 (1989), pp. 223–226. DOI: 10.2466/pr0.1989.65.1.223.
- [232] David La Barbera, Kevin Roitero, Gianluca Demartini, Stefano Mizzaro, and Damiano Spina. Crowdsourcing Truthfulness: The Impact of Judgment Scale and Assessor Bias. In: *Advances in Information Retrieval*. Ed. by Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins. Cham: Springer International Publishing, 2020, pp. 207–214. ISBN: 978-3-030-45442-5. DOI: 10.1007/978-3-030-45442-5_26.
- [233] Klodiana Lanaj, Russell E Johnson, and Christopher M Barnes. Beginning the work-day yet already depleted? Consequences of late-night smartphone use and sleep. In: *Organizational Behavior and Human Decision Processes* 124.1 (2014), pp. 11–23. DOI: 10.1016/j.obhdp.2014.01.001.
- [234] Justin F Landy, Miaolei Liam Jia, Isabel L Ding, Domenico Viganola, Warren Tierney, Anna Dreber, Magnus Johannesson, Thomas Pfeiffer, Charles R Ebersole, Quentin F Gronau, Alexander Ly, Don van den Bergh, Maarten Marsman, Koen Derks, Eric-Jan Wagenmakers, Andrew Proctor, Daniel M Bartels, Christopher W Bauman, William J Brady, Felix Cheung, Andrei Cimpian, Simone Dohle, M Brent Donnellan, Adam Hahn, Michael P Hall, William Jiménez-Leal, David J Johnson, Richard E Lucas, Benoît Monin, Andres Montealegre, Elizabeth Mullen, Jun Pang, Jennifer Ray, Diego A Reinero, Jesse Reynolds, Walter Sowden, Daniel Storage, Runkun Su, Christina M Tworek, Jay J Van Bavel, Daniel Walco, Julian Wills, Xiaobing Xu, Kai Chi Yam, Xiaoyu Yang, William A Cunningham, Martin Schweinsberg, Molly Urwitz, The Crowdsourcing Hypothesis Tests Collaboration, and Eric L Uhlmann. Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. In: *Psychological Bulletin* 146.5 (May 2020), pp. 451–479. DOI: 10.1037/bu10000220.
- [235] *Läsbarhet:*
- [236] John Lawrence and Chris Reed. Argument Mining: A Survey. In: *Computational Linguistics* 45.4 (Dec. 2019), pp. 765–818. DOI: 10.1162/coli_a_00364.
- [237] Quoc Le and Tomas Mikolov. Distributed Representations of Sentences and Documents. In: *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*. ICML'14. Beijing, China: JMLR.org, 2014, II-1188–II-1196. DOI: 10.5555/3044805.3045025.

- [238] Eun-Ju Lee. That's Not the Way It Is: How User-Generated Comments on the News Affect Perceived Media Bias. In: *Journal of Computer-Mediated Communication* 18.1 (2012), pp. 32–45. DOI: 10.1111/j.1083-6101.2012.01597.x.
- [239] Jacqueline P. Leighton and Robert J. Sternberg. *The Nature of Reasoning*. UK: Cambridge University Press, 2004.
- [240] Patrick J. Leman and Marco Cinnirella. A major event has a major cause: Evidence for the role of heuristics in reasoning about conspiracy theories. In: *Social Psychological Review* 9.2 (2007), pp. 18–28.
- [241] Melvin J. Lerner and Dale T. Miller. Just world research and the attribution process: Looking back and ahead. In: *Psychological Bulletin* 85.5 (1978), pp. 1030–1051. DOI: 10.1037/0033-2909.85.5.1030.
- [242] Gabriel Shing-Koon Leung, Vincent Cho, and CH Wu. Crowd Workers' Continued Participation Intention in Crowdsourcing Platforms: An Empirical Study in Compensation-Based Micro-Task Crowdsourcing. In: *Journal of Global Information Management* 29.6 (2021), pp. 1–28.
- [243] Stephan Lewandowsky, Ullrich K. H. Ecker, Colleen M. Seifert, Norbert Schwarz, and John Cook. Misinformation and Its Correction: Continued Influence and Successful Debiasing. In: *Psychological Science in the Public Interest* 13.3 (2012), pp. 106–131. DOI: 10.1177/1529100612451018.
- [244] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 7871–7880. DOI: 10.18653/v1/2020.acl-main.703.
- [245] Guohui Li, Ming Dong, Fuming Yang, Jun Zeng, Jiansen Yuan, Congyuan Jin, Nguyen Quoc Viet Hung, Phan Thanh Cong, and Bolong Zheng. Misinformation-oriented expert finding in social networks. In: *World Wide Web* 23.2 (Mar. 2020), pp. 693–714. ISSN: 1573-1413. DOI: 10.1007/s11280-019-00717-6.
- [246] Hang Li. Learning to Rank for Information Retrieval and Natural Language Processing. In: *Synthesis Lectures on Human Language Technologies*. Synthesis Lectures on Human Language Technologies. Springer Cham, 2014. ISBN: 978-3-031-01027-9. DOI: 10.1007/978-3-031-02155-8.
- [247] Ming Li, Jian Weng, Anjia Yang, Wei Lu, Yue Zhang, Lin Hou, Jia-Nan Liu, Yang Xiang, and Robert H. Deng. CrowdBC: A Blockchain-Based Decentralized Framework for Crowdsourcing. In: *IEEE Transactions on Parallel and Distributed Systems* 30.6 (2019), pp. 1251–1266. DOI: 10.1109/TPDS.2018.2881735.
- [248] Ting Li, Ning Liu, Jun Yan, Gang Wang, Fengshan Bai, and Zheng Chen. A Markov Chain Model for Integrating Behavioral Targeting into Contextual Advertising. In: *Proceedings of the Third International Workshop on Data Mining and Audience Intelligence for Advertising*. ADKDD '09. Paris, France: Association for Computing Machinery, 2009, 1–9. ISBN: 9781605586717. DOI: 10.1145/1592748.1592750.

- [249] Filip Lievens. Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. In: *Journal of Applied Psychology* 86.2 (2001), p. 255. DOI: 10.1037/0021-9010.86.2.255.
- [250] Rensis Likert. A technique for the measurement of attitudes. In: *Archives of Psychology* 22.140 (1932), pp. 55–55.
- [251] Chloe Lim. Checking how fact-checkers check. In: *Research & Politics* 5.3 (2018), p. 2053168018786848. DOI: 10.1177/2053168018786848.
- [252] Sora Lim, Adam Jatowt, Michael Färber, and Masatoshi Yoshikawa. Annotating and Analyzing Biased Sentences in News Articles using Crowdsourcing. English. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 1478–1484. ISBN: 979-10-95546-34-4. URL: <https://aclanthology.org/2020.lrec-1.184>.
- [253] Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 74–81. URL: <https://aclanthology.org/W04-1013>.
- [254] Kristian Linnet. Nonparametric estimation of reference intervals by simple and bootstrap-based procedures. In: *Clinical chemistry* 46.6 (2000), pp. 867–869. ISSN: 0009-9147. DOI: 10.1093/clinchem/46.6.867.
- [255] Leib Litman, Aaron Moss, Cheskie Rosenzweig, and Jonathan Robinson. Reply to MTurk, Prolific or panels? Choosing the right audience for online research. In: *Social Science Research Network* (Feb. 2021). DOI: 10.2139/ssrn.3775075.
- [256] Leib Litman and Jonathan Robinson. *Conducting Online Research on Amazon Mechanical Turk and Beyond*. 1st ed. Thousand Oaks, California, Dec. 2021. DOI: 10.4135/9781506391151.
- [257] Leib Litman, Jonathan Robinson, and Tzvi Abberbock. TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. In: *Behavior Research Methods* 49.2 (Apr. 2017), pp. 433–442. ISSN: 1554-3528. DOI: 10.3758/s13428-016-0727-z.
- [258] Dong C. Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. In: *Mathematical Programming* 45.1 (Aug. 1989), pp. 503–528. ISSN: 1436-4646. DOI: 10.1007/BF01589116. URL: <https://doi.org/10.1007/BF01589116>.
- [259] Sophia B. Liu. Crisis Crowdsourcing Framework: Designing Strategic Configurations of Crowdsourcing for the Emergency Management Domain. In: *Computer Supported Cooperative Work* 23.4 (Dec. 2014), pp. 389–443. ISSN: 1573-7551. DOI: 10.1007/s10606-014-9204-3.
- [260] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. Multi-Task Deep Neural Networks for Natural Language Understanding. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 4487–4496. DOI: 10.18653/v1/P19-1441.

- [261] Yang Liu and Yi-Fang Brook Wu. FNED: A Deep Network for Fake News Early Detection on Social Media. In: *ACM Transactions on Information Systems* 38.3 (May 2020). ISSN: 1046-8188. DOI: 10.1145/3386253.
- [262] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. In: *CoRR* abs/1907.11692 (2019). arXiv: 1907.11692. URL: <http://arxiv.org/abs/1907.11692>.
- [263] Kevin G. Love. Comparison of peer assessment methods: Reliability, validity, friendship bias, and user reaction. In: *Journal of Applied Psychology* 66.4 (1981), p. 451. DOI: 10.1037/0021-9010.66.4.451.
- [264] Guo Ying Luo. Conservatism Bias and Asset Price Overreaction or Underreaction to New Information in a Competitive Securities Market. In: *Asset Price Response to New Information: The Effects of Conservatism Bias and Representativeness Heuristic*. New York, NY: Springer New York, 2014, pp. 5–14. ISBN: 978-1-4614-9369-3. DOI: 10.1007/978-1-4614-9369-3_2.
- [265] Eddy Maddalena, Marco Basaldella, Dario De Nart, Dante Degl’Innocenti, Stefano Mizzaro, and Gianluca Demartini. Crowdsourcing Relevance Assessments: The Unexpected Benefits of Limiting the Time to Judge. In: *Proceedings of the Fourth AAAI Conference on Human Computation and Crowdsourcing*. Texas, USA: AAAI Press, 2016, pp. 129–138. URL: <http://aaai.org/ocs/index.php/HCOMP/HCOMP16/paper/view/14040>.
- [266] Eddy Maddalena, Davide Ceolin, and Stefano Mizzaro. Multidimensional News Quality: A Comparison of Crowdsourcing and Nichesourcing. In: *Proceedings of the CIKM 2018 Workshops co-located with 27th ACM International Conference on Information and Knowledge Management*. Torino, Italy, Oct. 2018. URL: <http://ceur-ws.org/Vol-2482/paper17.pdf>.
- [267] Eddy Maddalena, Stefano Mizzaro, Falk Scholer, and Andrew Turpin. On Crowdsourcing Relevance Magnitudes for Information Retrieval Evaluation. In: *ACM Transactions on Information Systems* 35.3 (Jan. 2017), 19:1–19:32. ISSN: 1046-8188. DOI: 10.1145/3002172.
- [268] Eddy Maddalena, Stefano Mizzaro, Falk Scholer, and Andrew Turpin. On Crowdsourcing Relevance Magnitudes for Information Retrieval Evaluation. In: *ACM Transactions On Information Systems* 35.3 (Jan. 2017). ISSN: 1046-8188. DOI: 10.1145/3002172. URL: <https://doi.org/10.1145/3002172>.
- [269] Eddy Maddalena, Kevin Roitero, Gianluca Demartini, and Stefano Mizzaro. Considering Assessor Agreement in IR Evaluation. In: *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*. ICTIR ’17. Amsterdam, The Netherlands: Association for Computing Machinery, 2017, 75–82. ISBN: 9781450344906. DOI: 10.1145/3121050.3121060.
- [270] David J. Malenka, John A. Baron, Sarah Johansen, Jon W. Wahrenberger, and Jonathan M. Ross. The framing effect of relative and absolute risk. In: *Journal of General Internal Medicine* 8.10 (Oct. 1993), pp. 543–548. ISSN: 1525-1497. DOI: 10.1007/BF02599636.

- [271] E. Manavoglu, D. Pavlov, and C.L. Giles. Probabilistic user behavior models. In: *Third IEEE International Conference on Data Mining*. 2003, pp. 203–210. DOI: 10.1109/ICDM.2003.1250921.
- [272] Diego C. Martínez, Alejandro Javier García, and Guillermo Ricardo Simari. An Abstract Argumentation Framework with Varied-Strength Attacks. In: *Principles of Knowledge Representation and Reasoning: Proceedings of the Eleventh International Conference*. Ed. by Gerhard Brewka and Jérôme Lang. Sydney, Australia: AAAI Press, 2008, pp. 135–144. URL: <https://www.aaai.org/Papers/KR/2008/KR08-014.pdf>.
- [273] Nitika Mathur, Timothy Baldwin, and Trevor Cohn. Sequence Effects in Crowdsourced Annotations. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 2860–2865. DOI: 10.18653/v1/D17-1306.
- [274] Helena Matute, Ion Yarritu, and Miguel A. Vadillo. Illusions of causality at the heart of pseudoscience. In: *British Journal of Psychology* 102.3 (2011), pp. 392–405. DOI: 10.1348/000712610X532210.
- [275] Panagiotis Mavridis, Owen Huang, Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. Chatterbox: Conversational Interfaces for Microtask Crowdsourcing. In: *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*. UMAP '19. Larnaca, Cyprus: Association for Computing Machinery, 2019, 243–251. ISBN: 9781450360210. DOI: 10.1145/3320435.3320439.
- [276] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. Image-Based Recommendations on Styles and Substitutes. In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '15. Santiago, Chile: Association for Computing Machinery, 2015, 43–52. ISBN: 9781450336215. DOI: 10.1145/2766462.2767755.
- [277] Peter McCullagh. *Generalized Linear Models*. Routledge, 2018. ISBN: 9780412317606.
- [278] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. Reliability and Inter-Rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. In: *Proceedings of ACM Human-Computer Interaction* 3.CSCW (Nov. 2019). DOI: 10.1145/3359174.
- [279] Tyler McDonnell, Matthew Lease, Mucahid Kutlu, and Tamer Elsayed. Why Is That Relevant? Collecting Annotator Rationales for Relevance Judgments. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 4.1 (Sept. 2016), pp. 139–148. URL: <https://ojs.aaai.org/index.php/HCOMP/article/view/13287>.
- [280] G. Harry McLaughlin. SMOG Grading — a New Readability Formula. In: *Journal of Reading* 8 (1969), 639–646. URL: https://ogg.osu.edu/media/documents/health_lit/WRRSMOG_Readability_Formula_G._Harry_McLaughlin__1969_.pdf.
- [281] Yelena Mejova and Kyriaki Kalimeri. *Advertisers Jump on Coronavirus Bandwagon: Politics, News, and Business*. 2020. arXiv: 2003.00923. URL: <https://arxiv.org/abs/2003.00923>.

- [282] Alexandra M. Mellis and Warren K. Bickel. Mechanical Turk data collection in addiction research: utility, concerns and best practices. In: *Addiction* 115.10 (2020), pp. 1960–1968. DOI: 10.1111/add.15032.
- [283] Paul Mena. Principles and Boundaries of Fact-checking: Journalists' Perceptions. In: *Journalism Practice* 13.6 (2019), pp. 657–672. DOI: 10.1080/17512786.2018.1547655.
- [284] Ethan Mendes, Yang Chen, Alan Ritter, and Wei Xu. Human-in-the-loop Evaluation for Early Misinformation Detection: A Case Study of COVID-19 Treatments. In: (2022). DOI: 10.48550/ARXIV.2212.09683.
- [285] Tsvetomila Mihaylova, Georgi Karadjov, Pepa Atanasova, Ramy Baly, Mitra Mohitarami, and Preslav Nakov. SemEval-2019 Task 8: Fact Checking in Community Question Answering Forums. In: *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, June 2019, pp. 860–869. DOI: 10.18653/v1/S19-2149.
- [286] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and Their Compositionality. In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'13. Lake Tahoe, Nevada: Curran Associates Inc., 2013, 3111–3119. DOI: 10.5555/2999792.2999959.
- [287] Sanjay Modgil. Reasoning about preferences in argumentation frameworks. In: *Artificial Intelligence* 173.9 (2009), pp. 901–934. ISSN: 0004-3702. DOI: 10.1016/j.artint.2009.02.001.
- [288] Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. SemEval-2016 Task 6: Detecting Stance in Tweets. In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. San Diego, California: Association for Computational Linguistics, June 2016, pp. 31–41. DOI: 10.18653/v1/S16-1003.
- [289] David Moher, Alessandro Liberati, Jennifer Tetzlaff, and Douglas G Altman. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. In: *BMJ* 339 (2009). DOI: 10.1136/bmj.b2535.
- [290] Susan Morgan. Fake news, disinformation, manipulation and online tactics to undermine democracy. In: *Journal of Cyber Policy* 3.1 (2018), pp. 39–43. DOI: 10.1080/23738871.2018.1462395.
- [291] Donald F. Morrison. Multivariate Analysis of Variance. In: (2005). DOI: 10.1002/0470011815.b2a13045.
- [292] Howard R Moskowitz. Magnitude Estimation: Notes on What, How, When, and Why to Use It. In: *Journal of Food Quality* 1.3 (1977), pp. 195–227. DOI: 10.1111/j.1745-4557.1977.tb00942.x.
- [293] Howard R. Moskowitz. Magnitude Estimation: Notes On What, How, When, And Why To Use It. In: *Journal of Food Quality* 1.3 (1977), pp. 195–227. DOI: 10.1111/j.1745-4557.1977.tb00942.x.

- [294] Abbe Mowshowitz and Akira Kawaguchi. Measuring search engine bias. In: *Information Processing & Management* 41.5 (2005), pp. 1193–1205. ISSN: 0306-4573. DOI: 10.1016/j.ipm.2004.05.005.
- [295] Lev Muchnik, Sinan Arel, and Sean J. Taylor. Social Influence Bias: A Randomized Experiment. In: *Science* 341.6146 (2013), pp. 647–651. DOI: 10.1126/science.1240466.
- [296] Brian Mullen, Rupert Brown, and Colleen Smith. Ingroup bias as a function of salience, relevance, and status: An integration. In: *European Journal of Social Psychology* 22.2 (), pp. 103–122. DOI: 10.1002/ejsp.2420220202.
- [297] Thomas Mussweiler, Fritz Strack, and Tim Pfeiffer. Overcoming the Inevitable Anchoring Effect: Considering the Opposite Compensates for Selective Accessibility. In: *Personality and Social Psychology Bulletin* 26.9 (2000), pp. 1142–1150. DOI: 10.1177/01461672002611010.
- [298] Mahdi Pakdaman Naeni, Gregory F. Cooper, and Milos Hauskrecht. Obtaining Well Calibrated Probabilities Using Bayesian Binning. In: AAI'15. Austin, Texas: AAAI Press, 2015, 2901–2907. ISBN: 0262511290. DOI: 10.5555/2888116.2888120.
- [299] Preslav Nakov, Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Lluís Màrquez, Wajdi Zaghouni, Pepa Atanasova, Spas Kyuchukov, and Giovanni Da San Martino. Overview of the CLEF-2018 CheckThat! Lab on Automatic Identification and Verification of Political Claims. In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Cham: Springer International Publishing, 2018, pp. 372–387. ISBN: 978-3-319-98932-7. DOI: 10.1007/978-3-319-98932-7_32.
- [300] Preslav Nakov, Alberto Barrón-Cedeño, Giovanni da San Martino, Firoj Alam, Julia Maria Struß, Thomas Mandl, Rubén Míguez, Tommaso Caselli, Mucahid Kutlu, Wajdi Zaghouni, Chengkai Li, Shaden Shaar, Gautam Kishore Shahi, Hamdy Mubarak, Alex Nikolov, Nikolay Babulkov, Yavuz Selim Kartal, Michael Wiegand, Melanie Siegel, and Juliane Köhler. Overview of the CLEF-2022 CheckThat! Lab on Fighting the COVID-19 Infodemic and Fake News Detection. In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Ed. by Alberto Barrón-Cedeño, Giovanni Da San Martino, Mirko Degli Esposti, Fabrizio Sebastiani, Craig Macdonald, Gabriella Pasi, Allan Hanbury, Martin Potthast, Guglielmo Faggioli, and Nicola Ferro. Cham: Springer International Publishing, 2022, pp. 495–520. ISBN: 978-3-031-13643-6. DOI: 10.1007/978-3-031-13643-6_29.
- [301] Preslav Nakov, Giovanni Da San Martino, Tamer Elsayed, Alberto Barrón-Cedeño, Rubén Míguez, Shaden Shaar, Firoj Alam, Fatima Haouari, Maram Hasanain, Nikolay Babulkov, Alex Nikolov, Gautam Kishore Shahi, Julia Maria Struß, and Thomas Mandl. The CLEF-2021 CheckThat! Lab on Detecting Check-Worthy Claims, Previously Fact-Checked Claims, and Fake News. In: *Advances in Information Retrieval*. Ed. by Djoerd Hiemstra, Marie-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani. Cham: Springer International Publishing, 2021, pp. 639–649. ISBN: 978-3-030-72240-1. DOI: 10.1007/978-3-030-72240-1_75.

- [302] Randolph M. Nesse. Natural selection and the regulation of defenses: A signal detection analysis of the smoke detector principle. In: *Evolution and Human Behavior* 26.1 (2005), pp. 88–105. ISSN: 1090-5138. DOI: 10.1016/j.evolhumbehav.2004.08.002.
- [303] Randolph M Nesse. Natural selection and the regulation of defensive responses. In: *Annals of the New York Academy of Sciences* 935 (2001), pp. 75–85. URL: <https://pubmed.ncbi.nlm.nih.gov/11411177/>.
- [304] Eryn. J. Newman, Madeline C. Jalbert, Norbert Schwarz, and Deva P. Ly. Truthiness, the illusory truth effect, and the role of need for cognition. In: *Consciousness and Cognition* 78 (2020), p. 102866. ISSN: 1053-8100. DOI: 10.1016/j.concog.2019.102866.
- [305] C. Thi Nguyen. Echo Chambers and Epistemic Bubbles. In: *Episteme* 17.2 (2020), pp. 141–161. DOI: 10.1017/epi.2018.32.
- [306] Thanh Tam Nguyen, Matthias Weidlich, Hongzhi Yin, Bolong Zheng, Quang Huy Nguyen, and Quoc Viet Hung Nguyen. FactCatch: Incremental Pay-as-You-Go Fact Checking with Minimal User Effort. In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '20. Virtual Event, China: Association for Computing Machinery, 2020, 2165–2168. ISBN: 9781450380164. DOI: 10.1145/3397271.3401408.
- [307] Feng Ni, David Arnott, and Shijia Gao. The anchoring effect in business intelligence supported decision-making. In: *Journal of Decision Systems* 28.2 (2019), pp. 67–81. DOI: 10.1080/12460125.2019.1620573.
- [308] Raymond S. Nickerson. Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. In: *Review of General Psychology* 2.2 (1998), pp. 175–220. DOI: 10.1037/1089-2680.2.2.175.
- [309] Yixin Nie, Haonan Chen, and Mohit Bansal. Combining Fact Extraction and Verification with Neural Semantic Matching Networks. In: *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*. Honolulu, Hawaii, USA: AAAI Press, 2019. ISBN: 978-1-57735-809-1. DOI: 10.1609/aaai.v33i01.33016859.
- [310] Yixin Nie, Haonan Chen, and Mohit Bansal. Combining Fact Extraction and Verification with Neural Semantic Matching Networks. In: *AAAI'19/IAAI'19/EAAI'19*. Honolulu, Hawaii, USA: AAAI Press, 2019. ISBN: 978-1-57735-809-1. DOI: 10.1609/aaai.v33i01.33016859.
- [311] Gerardo Ocampo Diaz and Vincent Ng. Modeling and Prediction of Online Product Review Helpfulness: A Survey. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 698–708. DOI: 10.18653/v1/P18-1065.

- [312] Anne Oeldorf-Hirsch and Christina L. DeVoss. Who Posted That Story? Processing Layered Sources in Facebook News Posts. In: *Journalism & Mass Communication Quarterly* 97.1 (2020), pp. 141–160. DOI: 10.1177/1077699019857673.
- [313] IFCN Code of Principles. *PolitiFact IFCN principles*. (Accessed: 20.04.2021). 2021. URL: <https://ifcncodeofprinciples.poynter.org/profile/politifact>.
- [314] Stephen F. Olejnik and James Algina. Generalized Eta and Omega Squared Statistics: Measures of Effect Size for Some Common Research Designs. In: *Psychological Methods* 8 (Jan. 2004), pp. 434–47. DOI: 10.1037/1082-989X.8.4.434.
- [315] Jahna Otterbacher, Jo Bates, and Paul Clough. Competent Men and Warm Women: Gender Stereotypes and Backlash in Image Search Results. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI '17. Denver, Colorado, USA: Association for Computing Machinery, 2017, 6620–6631. ISBN: 9781450346559. DOI: 10.1145/3025453.3025727.
- [316] Jahna Otterbacher, Alessandro Checco, Gianluca Demartini, and Paul Clough. Investigating User Perception of Gender Bias in Image Search: The Role of Sexism. In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. SIGIR '18. Ann Arbor, MI, USA: Association for Computing Machinery, 2018, 933–936. ISBN: 9781450356572. DOI: 10.1145/3209978.3210094.
- [317] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, Luke A McGuinness, Lesley A Stewart, James Thomas, Andrea C Tricco, Vivian A Welch, Penny Whiting, and David Moher. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. In: *BMJ* 372 (2021). DOI: 10.1136/bmj.n71.
- [318] Stefan Palan and Christian Schitter. Prolific.ac—A subject pool for online experiments. In: *Journal of Behavioral and Experimental Finance* 17 (2018), pp. 22–27. ISSN: 2214-6350. DOI: <https://doi.org/10.1016/j.jbef.2017.12.004>.
- [319] Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. Running experiments on Amazon Mechanical Turk. In: *Judgment and Decision Making* 5.5 (2010), pp. 411–419. URL: <https://journal.sjdm.org/10/10630a/jdm10630a.pdf>.
- [320] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. ACL '02. Philadelphia, Pennsylvania: Association for Computational Linguistics, 2002, 311–318. DOI: 10.3115/1073083.1073135.
- [321] Eli Pariser. *The filter bubble: What the Internet is hiding from you*. Penguin UK, 2011. ISBN: 9780141969923.

- [322] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. In: *Journal of Experimental Social Psychology* 70 (2017), pp. 153–163. ISSN: 0022-1031. DOI: 10.1016/j.jesp.2017.01.006.
- [323] Eyal Peer, David Rothschild, Andrew Gordon, Zak Evernden, and Ekaterina Damer. Data quality of platforms and panels for online behavioral research. In: *Behavior Research Methods* 54.4 (Aug. 2022), pp. 1643–1662. ISSN: 1554-3528. DOI: 10.3758/s13428-021-01694-3.
- [324] Godfrey Pell, Matthew S. Homer, and Trudie E. Roberts. Assessor training: its effects on criterion-based assessment in a medical context. In: *International Journal of Research & Method in Education* 31.2 (2008), pp. 143–154. DOI: 10.1080/17437270802124525.
- [325] Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A. Arechar, Dean Eckles, and David G. Rand. Shifting attention to accuracy can reduce misinformation online. In: *Nature* 592.7855 (Apr. 2021), pp. 590–595. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03344-2.
- [326] Gordon Pennycook and David G Rand. Fighting Misinformation on Social Media Using Crowdsourced Judgments of News Source Quality. In: *Proceedings of the National Academy of Sciences* 116.7 (2019), pp. 2521–2526. DOI: 10.1073/pnas.1806781116.
- [327] Gordon Pennycook and David G. Rand. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. In: *Cognition* 188 (2019). *The Cognitive Science of Political Thought*, pp. 39–50. ISSN: 0010-0277. DOI: 10.1016/j.cognition.2018.06.011.
- [328] Gordon Pennycook and David G Rand. Who Falls for Fake News? The Roles of Bullshit Receptivity, Overclaiming, Familiarity, and Analytic Thinking. In: *Journal of Personality* 88.2 (2020), pp. 185–200. DOI: 10.1111/jopy.12476.
- [329] L. S. Penrose. The Elementary Statistics of Majority Voting. In: *Journal of the Royal Statistical Society* 109.1 (1946), pp. 53–57. ISSN: 09528385. DOI: 10.2307/2981392.
- [330] Sarah Perez. COVID-19 quarantine boosts smart speaker usage among U.S. adults, particularly younger users. In: *TechCrunch* (2020). URL: <https://techcrunch.com/2020/04/30/covid-19-quarantine-boosts-smart-speaker-usage-among-u-s-adults-particularly-younger-users/>.
- [331] Jonathan Pilault, Amine Elhattami, and Christopher J. Pal. Conditionally Adaptive Multi-Task Learning: Improving Transfer Learning in NLP Using Fewer Parameters & Less Data. In: *Proceedings of the 9th International Conference on Learning Representations*. Virtual Event, Austria, 2021. URL: <https://openreview.net/pdf?id=de11dbHzAMF>.
- [332] Marcos Rodrigues Pinto, Yuri Oliveira de Lima, Carlos Eduardo Barbosa, and Jano Moreira de Souza. Towards Fact-Checking through Crowdsourcing. In: *2019 IEEE 23rd International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. 2019, pp. 494–499. DOI: 10.1109/CSCWD.2019.8791903.

- [333] John Pitts, Colin Coles, Peter Thomas, and Frank Smith. Enhancing reliability in portfolio assessment: discussions between assessors. In: *Medical Teacher* 24.2 (2002), pp. 197–201. DOI: 10.1080/01421590220125321.
- [334] John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: *Advances in Large Margin Classifiers*. Vol. 10. 3. MIT Press, 2000, pp. 61–74. ISBN: 9780262283977.
- [335] Robert E. Ployhart and Anna-Katherine Ward. The “Quick Start Guide” for Conducting and Publishing Longitudinal Research. In: *Journal of Business and Psychology* 26.4 (Dec. 2011), pp. 413–422. ISSN: 1573-353X. DOI: 10.1007/s10869-011-9209-6.
- [336] Frances A. Pogacar, Amira Ghenai, Mark D. Smucker, and Charles L.A. Clarke. The Positive and Negative Influence of Search Results on People’s Decisions about the Efficacy of Medical Treatments. In: *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*. ICTIR ’17. Amsterdam, The Netherlands: Association for Computing Machinery, 2017, 209–216. ISBN: 9781450344906. DOI: 10.1145/3121050.3121074.
- [337] Politifact. *The Principles of the Truth-O-Meter*. <https://www.politifact.com/article/2018/feb/12/principles-truth-o-meter-politifacts-methodology-i/#Truth-0-Meter%20ratings>. (Accessed: 15–12–2021). 2020.
- [338] Chrisa D Pornari and Jane Wood. Peer and cyber aggression in secondary school students: The role of moral disengagement, hostile attribution bias, and outcome expectancies. In: *Aggressive Behavior: Official Journal of the International Society for Research on Aggression* 36.2 (2010), pp. 81–94. DOI: 10.1002/ab.20336.
- [339] Beatrice Portelli, Jason Zhao, Tal Schuster, Giuseppe Serra, and Enrico Santus. Distilling the Evidence to Augment Fact Verification Models. In: *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*. Online: Association for Computational Linguistics, July 2020, pp. 47–51. DOI: 10.18653/v1/2020.fever-1.7.
- [340] Nicolas Pröllochs. Community-Based Fact-Checking on Twitter’s Birdwatch Platform. In: (2022). Ed. by Ceren Budak, Meeyoung Cha, and Daniele Quercia, pp. 794–805. URL: <https://ojs.aaai.org/index.php/ICWSM/article/view/19335>.
- [341] Rehab Qarout, Alessandro Checco, Gianluca Demartini, and Kalina Bontcheva. Platform-Related Factors in Repeatability and Reproducibility of Crowdsourcing Tasks. In: *Proceedings of the Seventh AAI Conference on Human Computation and Crowdsourcing* 7.1 (Oct. 2019), pp. 135–143. DOI: 10.1609/hcomp.v7i1.5264.
- [342] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. TickTalkTurk: Conversational Crowdsourcing Made Easy. In: *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*. CSCW ’20 Companion. Virtual Event, USA: Association for Computing Machinery, 2020, 53–57. ISBN: 9781450380591. DOI: 10.1145/3406865.3418572.
- [343] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. *Improving Language Understanding by Generative Pre-Training*. 2018. URL: <https://openai.com/blog/language-unsupervised/>.

- [344] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 2931–2937. DOI: 10.18653/v1/D17-1317.
- [345] Torsten Reimer, Andrea Reimer, and Uwe Czienskowski. Decision-Making Groups Attenuate the Discussion Bias in Favor of Shared Information: A Meta-Analysis. In: *Communication Monographs* 77.1 (2010), pp. 121–142. DOI: 10.1080/03637750903514318.
- [346] Julio C. S. Reis, André Correia, Fabrício Murai, Adriano Veloso, and Fabrício Benvenuto. Explainable Machine Learning for Fake News Detection. In: *Proceedings of the 10th ACM Conference on Web Science*. ACM, 2019, 17–26. ISBN: 9781450362023. DOI: 10.1145/3292522.3326027.
- [347] Huorong Ren, Zhixing Ye, and Zhiwu Li. Anomaly detection based on a dynamic Markov model. In: *Information Sciences* 411 (2017), pp. 52–65. ISSN: 0020-0255. DOI: 10.1016/j.ins.2017.05.021.
- [348] Flavio P. Ribeiro, Dinei A. F. Florêncio and Cha Zhang, and Michael L. Seltzer. CROWDMOS: An approach for crowdsourcing mean opinion score studies. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. Prague Congress Center, Prague, Czech Republic: IEEE, 2011, pp. 2416–2419. DOI: 10.1109/ICASSP.2011.5946971.
- [349] Alisa Rieger, Tim Draws, Mariët Theune, and Nava Tintarev. This Item Might Reinforce Your Opinion: Obfuscation and Labeling of Search Results to Mitigate Confirmation Bias. In: *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*. HT '21. Virtual Event, USA: Association for Computing Machinery, 2021, 189–199. ISBN: 9781450385510. DOI: 10.1145/3465336.3475101.
- [350] Al Ries. Understanding marketing psychology and the halo effect. In: *Advertising Age* 17 (2006). URL: <https://adage.com/article/al-ries/understanding-marketing-psychology-halo-effect/108676>.
- [351] RMIT ABC. *Fact Check*. <https://www.abc.net.au/news/factcheck/about/?nw=0>. (Accessed: 15–12–2021). 2021.
- [352] Kirk Roberts, Tasmee Alam, Steven Bedrick, Dina Demner-Fushman, Kyle Lo, Ian Soboroff, Ellen Voorhees, Lucy Lu Wang, and William R Hersh. TREC-COVID: rationale and structure of an information retrieval shared task for COVID-19. In: *Journal of the American Medical Informatics Association* (July 2020). ocaa091. ISSN: 1527-974X. DOI: 10.1093/jamia/ocaa091.
- [353] Jonathan Robinson, Cheskie Rosenzweig, Aaron J. Moss, and Leib Litman. Tapped out or barely tapped? Recommendations for how to harness the vast and largely unused potential of the Mechanical Turk participant pool. In: *PLOS ONE* 14.12 (Dec. 2019), pp. 1–29. DOI: 10.1371/journal.pone.0226394.
- [354] Neal J. Roese and Kathleen D. Vohs. Hindsight Bias. In: *Perspectives on Psychological Science* 7.5 (2012), pp. 411–426. DOI: 10.1177/1745691612454303.

- [355] Kevin Roitero, Ben Carterette, Rishabh Mehrotra, and Mounia Lalmas. Leveraging Behavioral Heterogeneity Across Markets for Cross-Market Training of Recommender Systems. In: *Companion Proceedings of the Web Conference 2020*. WWW '20. Taipei, Taiwan: ACM, 2020, 694–702. ISBN: 9781450370240. DOI: 10.1145/3366424.3384362.
- [356] Kevin Roitero, Gianluca Demartini, Stefano Mizzaro, and Damiano Spina. How Many Truth Levels? Six? One Hundred? Even More? Validating Truthfulness of Statements via Crowdsourcing. In: *Proceedings of the CIKM 2018 Workshops co-located with 27th ACM International Conference on Information and Knowledge Management*. Torino, Italy, Oct. 2018. URL: <http://ceur-ws.org/Vol-2482/paper38.pdf>.
- [357] Kevin Roitero, David La Barbera, Michael Soprano, Gianluca Demartini, Stefano Mizzaro, and Tetsuya Sakai. How Many Assessors Do I Need? On Statistical Power When Crowdsourcing Relevance Judgments. In: *ACM Transactions on Information Systems* (2023). Under Review.
- [358] Kevin Roitero, Eddy Maddalena, Gianluca Demartini, and Stefano Mizzaro. On Fine-Grained Relevance Scales. In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. SIGIR '18. Ann Arbor, MI, USA: ACM, 2018, pp. 675–684. ISBN: 978-1-4503-5657-2. DOI: 10.1145/3209978.3210052.
- [359] Kevin Roitero, Eddy Maddalena, Stefano Mizzaro, and Falk Scholer. On the effect of relevance scales in crowdsourcing relevance assessments for Information Retrieval evaluation. In: *Information Processing & Management* 58.6 (2021), p. 102688. ISSN: 0306-4573. DOI: 10.1016/j.ipm.2021.102688.
- [360] Kevin Roitero, Michael Soprano, Davide Ceolin, David La Barbera, Damiano Spina, Gianluca Demartini, and Stefano Mizzaro. Fact-Checking: A Systematic Review of Cognitive Biases and Effective Countermeasures. In: *Information Processing & Management* (2023). Under Review.
- [361] Kevin Roitero, Michael Soprano, Shaoyang Fan, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. Can The Crowd Identify Misinformation Objectively? The Effects of Judgment Scale and Assessor's Background. In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '20. Xi'an, China (Virtual): Association for Computing Machinery, July 2020, 439–448. ISBN: 9781450380164. DOI: 10.1145/3397271.3401112.
- [362] Kevin Roitero, Michael Soprano, Beatrice Portelli, Massimiliano De Luise, Damiano Spina, Vincenzo Della Mea, Giuseppe Serra, Stefano Mizzaro, and Gianluca Demartini. Can The Crowd Judge Truthfulness? A Longitudinal Study On Recent Misinformation About COVID-19. In: *Personal and Ubiquitous Computing* (Sept. 2021). ISSN: 1617-4917. DOI: 10.1007/s00779-021-01604-6.
- [363] Kevin Roitero, Michael Soprano, Beatrice Portelli, Damiano Spina, Vincenzo Della Mea, Giuseppe Serra, Stefano Mizzaro, and Gianluca Demartini. The COVID-19 Infodemic: Can the Crowd Judge Recent Misinformation Objectively? In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. CIKM '20. Virtual Event, Ireland: Association for Computing Machinery, 2020, 1305–1314. ISBN: 9781450368599. DOI: 10.1145/3340531.3412048.

- [364] Sigrid Rouam. False Discovery Rate (FDR). In: (2013). Ed. by Werner Dubitzky, Olaf Wolkenhauer, Kwang-Hyun Cho, and Hiroki Yokota, pp. 731–732. DOI: 10.1007/978-1-4419-9863-7_223.
- [365] Zick Rubin and Letitia Anne Peplau. Who Believes in a Just World? In: *Journal of Social Issues* 31.3 (1975), pp. 65–89. DOI: 10.1111/j.1540-4560.1975.tb00997.x.
- [366] Natali Ruchansky, Sungyong Seo, and Yan Liu. CSI: A Hybrid Deep Model for Fake News Detection. In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. CIKM '17*. Singapore, Singapore: Association for Computing Machinery, 2017, 797–806. ISBN: 9781450349185. DOI: 10.1145/3132847.3132877.
- [367] Graeme D Ruxton and Guy Beauchamp. Time for Some A Priori Thinking About Post Hoc Testing. In: *Behavioral Ecology* 19.3 (Feb. 2008), pp. 690–693. DOI: 10.1093/beheco/arn020.
- [368] Niloufar Salehi, Lilly C. Irani, Michael S. Bernstein, Ali Alkhatib, Eva Ogbe, Kristy Milland, and Clickhappier. We Are Dynamo: Overcoming Stalling and Friction in Collective Action for Crowd Workers. In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. CHI '15*. Seoul, Republic of Korea: Association for Computing Machinery, 2015, 1621–1630. ISBN: 9781450331456. DOI: 10.1145/2702123.2702508.
- [369] Parnia Samimi and Sri Devi Ravana. Agreement between Crowdsourced Workers and Expert Assessors in Making Relevance Judgment for System Based IR Evaluation. In: *Recent Advances on Soft Computing and Data Mining*. Ed. by Tutut Herawan, Rozaida Ghazali, and Mustafa Mat Deris. Springer International Publishing, 2014, pp. 399–407. ISBN: 978-3-319-07692-8. DOI: 10.1007/978-3-319-07692-8_38.
- [370] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In: *CoRR abs/1910.01108* (2019). URL: <http://arxiv.org/abs/1910.01108>.
- [371] Franklin E. Satterthwaite. An Approximate Distribution of Estimates of Variance Components. In: *Biometrics Bulletin* 2.6 (1946), pp. 110–114. DOI: 10.2307/3002019.
- [372] Yaacov Schul. When Warning Succeeds: The Effect of Warning on Success in Ignoring Invalid Information. In: *Journal of Experimental Social Psychology* 29.1 (1993), pp. 42–62. ISSN: 0022-1031. DOI: 10.1006/jesp.1993.1003.
- [373] Brian B Schultz. Levene's Test for Relative Variation. In: *Systematic Zoology* 34.4 (1985), pp. 449–456.
- [374] R. Sethi and R. Rangaraju. Extinguishing the Backfire Effect: Using Emotions in Online Social Collaborative Argumentation for Fact Checking. In: *2018 IEEE International Conference on Web Services (ICWS)*. Vol. 1. 2018, pp. 363–366. DOI: 10.1109/ICWS.2018.00062.
- [375] R. J. Sethi. Spotting Fake News: A Social Argumentation Framework for Scrutinizing Alternative Facts. In: *2017 IEEE International Conference on Web Services (ICWS)*. Vol. 1. 2017, pp. 866–869. DOI: 10.1109/ICWS.2017.108.

- [376] Ricky J. Sethi. Crowdsourcing the Verification of Fake News and Alternative Facts. In: *Proceedings of the 28th ACM Conference on Hypertext and Social Media*. HT '17. Prague, Czech Republic: Association for Computing Machinery, 2017, 315–316. ISBN: 9781450347082. DOI: 10.1145/3078714.3078746.
- [377] Ricky J. Sethi, Raghuram Rangaraju, and Bryce Shurts. Fact Checking Misinformation Using Recommendations from Emotional Pedagogical Agents. In: *Intelligent Tutoring Systems*. Ed. by Andre Coy, Yugo Hayashi, and Maiga Chang. Cham: Springer International Publishing, 2019, pp. 99–104. ISBN: 978-3-030-22244-4. DOI: 10.1007/978-3-030-22244-4_13.
- [378] Shaban Shabani and Maria Sokhn. Hybrid Machine-Crowd Approach for Fake News Detection. In: *2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC)*. 2018, pp. 299–306. DOI: 10.1109/CIC.2018.00048.
- [379] Danielle N Shapiro, Jesse Chandler, and Pam A Mueller. Using Mechanical Turk to Study Clinical Populations. In: *Clinical Psychological Science* 1.2 (2013), pp. 213–220. DOI: 10.1177/2167702612469015.
- [380] Samuel S Shapiro and RS Francia. An Approximate Analysis of Variance Test for Normality. In: *Journal of the American Statistical Association* 67.337 (1972), pp. 215–216. DOI: 10.1080/01621459.1972.10481232.
- [381] Tali Sharot. The optimism bias. In: *Current Biology* 21.23 (2011), R941–R945. ISSN: 0960-9822. DOI: 10.1016/j.cub.2011.10.030.
- [382] Hongzhou Shen, Junpeng Shi, and Yihan Zhang. CrowdEIM: Crowdsourcing emergency information management tasks to mobile social media users. In: *International Journal of Disaster Risk Reduction* 54 (2021), p. 102024. ISSN: 2212-4209. DOI: 10.1016/j.ijdr.2020.102024.
- [383] Sam Shleifer and Alexander M Rush. Pre-trained Summarization Distillation. In: *CoRR* abs/2010.13002 (2020). arXiv: 2010.13002. URL: <https://arxiv.org/abs/2010.13002>.
- [384] Anu Shrestha and Francesca Spezzano. Textual Characteristics of News Title and Body to Detect Fake News: A Reproducibility Study. In: *Advances in Information Retrieval*. Ed. by Djoerd Hiemstra, Marie-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani. Cham: Springer International Publishing, 2021, pp. 120–133. ISBN: 978-3-030-72240-1. DOI: 10.1007/978-3-030-72240-1_9.
- [385] Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. DEFEND: Explainable Fake News Detection. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '19. Anchorage, AK, USA: Association for Computing Machinery, 2019, 395–405. ISBN: 9781450362016. DOI: 10.1145/3292500.3330935. URL: <https://doi.org/10.1145/3292500.3330935>.
- [386] Diwakar Shukla and Rahul Singhai. Analysis of Users Web Browsing Behavior Using Markov chain Model. In: *International Journal of Advanced Networking and Applications* 2 (Mar. 2011), pp. 824–830. DOI: 10.35444/IJANA.2021.

- [387] Sneha Singhania, Nigel Fernandez, and Shrisha Rao. 3HAN: A Deep Neural Network for Fake News Detection. In: *Neural Information Processing*. Cham: Springer International Publishing, 2017, pp. 572–581. ISBN: 978-3-319-70096-0. DOI: 10.1007/978-3-319-70096-0_59.
- [388] Paul Slovic, Melissa L. Finucane, Ellen Peters, and Donald G. MacGregor. The affect heuristic. In: *European Journal of Operational Research* 177.3 (2007), pp. 1333–1352. ISSN: 0377-2217. DOI: 10.1016/j.ejor.2005.04.006.
- [389] E. A. Smith and R. J. Senter. Automated readability index. In: *AMRL TR* (May 1967), pp. 1–14. URL: <https://apps.dtic.mil/sti/pdfs/AD0667273.pdf>.
- [390] Mark Snaith, John Lawrence, Alison Pease, and Chris Reed. A Modular Platform for Argument and Dialogue. In: *Proceedings of The 8th International Conference on Computational Models of Argument*. Ed. by Henry Prakken, Stefano Bistarelli, Francesco Santini, and Carlo Taticchi. Vol. 326. Frontiers in Artificial Intelligence and Applications. Perugia, Italy: IOS Press, Sept. 2020, pp. 473–474. DOI: 10.3233/FAIA200540.
- [391] Amir Soleimani, Christof Monz, and Marcel Worring. BERT for Evidence Retrieval and Claim Verification. In: *Advances in Information Retrieval*. Ed. by Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins. Cham: Springer International Publishing, 2020, pp. 359–366. ISBN: 978-3-030-45442-5. DOI: 10.1007/978-3-030-45442-5_45.
- [392] Michael Soprano. In *Crowd Veritas: Leveraging Human Intelligence To Fight Misinformation*. Apr. 2023. DOI: 10.17605/OSF.IO/JR6VC.
- [393] Michael Soprano, Kevin Roitero, Francesco Bombassei De Bona, and Stefano Mizzaro. Crowd_Frame: A Simple and Complete Framework to Deploy Complex Crowdsourcing Tasks Off-the-Shelf. In: *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. WSDM '22. Virtual Event, AZ, USA: Association for Computing Machinery, 2022, 1605–1608. ISBN: 9781450391320. DOI: 10.1145/3488560.3502182.
- [394] Michael Soprano, Kevin Roitero, Ujwal Gadiraju, Eddy Maddalena, and Gianluca Demartini. Longitudinal Loyalty: Understanding the Barriers to Running Longitudinal Studies on Crowdsourcing Platforms. In: Under Review. May 2023.
- [395] Michael Soprano, Kevin Roitero, David La Barbera, Davide Ceolin, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. The many dimensions of truthfulness: Crowdsourcing misinformation assessments on a multidimensional scale. In: *Information Processing & Management* 58.6 (2021), p. 102710. ISSN: 0306-4573. DOI: 10.1016/j.ipm.2021.102710.
- [396] Dominik Stambach and Elliott Ash. e-FEVER: Explanations and Summaries for Automated Fact Checking. In: *Proceedings of the 2020 Truth and Trust Online Conference*. 2020, p. 32. URL: <https://truthandtrustonline.com/wp-content/uploads/2020/10/TT004.pdf>.

- [397] Dominik Stambach and Guenter Neumann. Team DOMLIN: Exploiting Evidence Enhancement for the FEVER Shared Task. In: *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 105–109. DOI: 10.18653/v1/D19-6616.
- [398] Neil Stewart, Christoph Ungemach, Adam J. L. Harris, Daniel M. Bartels, Ben R. Newell, Gabriele Paolacci, and Jesse Chandler. The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. In: *Judgment and Decision Making* 10.5 (2015), pp. 479–491. URL: <https://journal.sjdm.org/14/14725/jdm14725.pdf>.
- [399] Robert C. Streijl, Stefan Winkler, and David S. Hands. Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives. In: *Multimedia Systems* 22.2 (Mar. 2016), pp. 213–227. ISSN: 1432-1882. DOI: 10.1007/s00530-014-0446-1.
- [400] Justin C. Strickland and William W. Stoops. Feasibility, acceptability, and validity of crowdsourcing for collecting longitudinal alcohol use data. In: *Journal of the Experimental Analysis of Behavior* 110.1 (2018), pp. 136–153. DOI: 10.1002/jeab.445.
- [401] Justin C Strickland and William W Stoops. The use of crowdsourcing in addiction science research: Amazon Mechanical Turk. In: *Experimental and Clinical Psychopharmacology* 27.1 (2019), p. 1. DOI: 10.1037/pha0000235.
- [402] John A. Swets, Robyn M. Dawes, and John Monahan. Psychological Science Can Improve Diagnostic Decisions. In: *Psychological Science in the Public Interest* 1.1 (2000), pp. 1–26. DOI: 10.1111/1529-1006.001.
- [403] Wen-Ying Sylvia Chou, Anna Gaysynsky, and Joseph N. Cappella. Where We Go From Here: Health Misinformation on Social Media. In: *American Journal of Public Health* 110.S3 (2020), S273–S275. DOI: 10.2105/AJPH.2020.305905.
- [404] Michelle Y. Szpara and E. Caroline Wylie. National Board for Professional Teaching Standards Assessor Training: Impact of Bias Reduction Exercises. In: *Teachers College Record* 107.4 (2005), pp. 803–841. DOI: 10.1177/016146810510700410.
- [405] Jenny Tang, Eleanor Birrell, and Ada Lerner. How Well Do My Results Generalize Now? The External Validity of Online Privacy and Security Surveys. In: (2022). DOI: 10.48550/ARXIV.2202.14036.
- [406] Amanda Taub. *The Real Story About Fake News Is Partisanship*. (Accessed: 20.04.2021). 2017. URL: <https://www.nytimes.com/2017/01/11/upshot/the-real-story-about-fake-news-is-partisanship.html>.
- [407] Andon Tchechmedjiev, Pavlos Fafalios, Katarina Boland, Malo Gasquet, Matthäus Zloch, Benjamin Zapilko, Stefan Dietze, and Konstantin Todorov. ClaimsKG: A Knowledge Graph of Fact-Checked Claims. In: *The Semantic Web – ISWC 2019*. Ed. by Chiara Ghidini, Olaf Hartig, Maria Maleshkova, Vojtěch Svátek, Isabel Cruz, Aidan Hogan, Jie Song, Maxime Lefrançois, and Fabien Gandon. Cham: Springer International Publishing, 2019, pp. 309–324. ISBN: 978-3-030-30796-7. DOI: 10.1007/978-3-030-30796-7_20.

- [408] The Verge. *Snopes forced to scale back fact-checking*. <https://www.theverge.com/2020/3/24/21192206/snopes-coronavirus-covid-19-misinformation-fact-checking-staff>. (Accessed: 15-12-2021). 2020.
- [409] Oliver Thomas. Two decades of cognitive bias research in entrepreneurship: What do we know and where do we go from here? In: *Management Review Quarterly* 68.2 (2018), pp. 107–143. DOI: 10.1007/s11301-018-0135-9.
- [410] Charles P. Thompson, John J. Skowronski, and D. John Lee. Telescoping in dating naturally occurring events. In: *Memory & Cognition* 16.5 (Sept. 1988), pp. 461–468. ISSN: 1532-5946. DOI: 10.3758/BF03214227.
- [411] James Thorne and Andreas Vlachos. Automated Fact Checking: Task Formulations, Methods and Future Directions. In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 3346–3359. URL: <https://aclanthology.org/C18-1283>.
- [412] James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. The Fact Extraction and VERification (FEVER) Shared Task. In: *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 1–9. DOI: 10.18653/v1/W18-5501.
- [413] Peter M. Todd and Gerd Gigerenzer. Précis of Simple heuristics that make us smart. In: *Behavioral and Brain Sciences* 23.5 (2000), 727–741. DOI: 10.1017/S0140525X00003447.
- [414] Ehsan Toreini, Mhairi Aitken, Kovila Coopamootoo, Karen Elliott, Carlos Gonzalez Zelaya, and Aad van Moorsel. The Relationship between Trust in AI and Trustworthy Machine Learning Technologies. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. FAT* '20*. Barcelona, Spain: Association for Computing Machinery, 2020, 272–283. ISBN: 9781450369367. DOI: 10.1145/3351095.3372834.
- [415] Stephen E Toulmin. *The Uses of Argument*. Cambridge University Press, 2012. DOI: 10.1017/CB09780511840005.
- [416] Carlos Toxtli, Siddharth Suri, and Saiph Savage. Quantifying the Invisible Labor in Crowd Work. In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW2 (Oct. 2021). DOI: 10.1145/3476060.
- [417] Cecilie Steenbuch Traberg and Sander van der Linden. Birds of a feather are persuaded together: Perceived source credibility mediates the effect of political bias on misinformation susceptibility. In: *Personality and Individual Differences* 185 (2022), p. 111269. ISSN: 0191-8869. DOI: 10.1016/j.paid.2021.111269.
- [418] Thanh Tran, Kyumin Lee, Yiming Liao, and Dongwon Lee. Regularizing Matrix Factorization with User and Item Embeddings for Recommendation. In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management. CIKM '18*. Torino, Italy: Association for Computing Machinery, 2018, 687–696. ISBN: 9781450360142. DOI: 10.1145/3269206.3271730.

- [419] Sebastian Tschitschek, Adish Singla, Manuel Gomez Rodriguez, Arpit Merchant, and Andreas Krause. Detecting fake news in social networks via crowdsourcing. In: *CoRR* abs/1711.09025 (2017). arXiv: 1711.09025. URL: <http://arxiv.org/abs/1711.09025>.
- [420] A Tversky and D Kahneman. Judgment under Uncertainty: Heuristics and Biases. In: *Science* 185.4157 (Sept. 1974), pp. 1124–1131. DOI: 10.1126/science.185.4157.1124.
- [421] Amos Tversky and Daniel Kahneman. Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. In: *Psychological review* 90.4 (1983), p. 293. DOI: 10.1037/0033-295X.90.4.293.
- [422] Petter Törnberg. Echo chambers and viral misinformation: Modeling fake news as complex contagion. In: *PLOS ONE* 13.9 (Sept. 2018), pp. 1–21. DOI: 10.1371/journal.pone.0203958.
- [423] Sagar Uprety, Prayag Tiwari, Shahram Dehdashti, Lauren Fell, Dawei Song, Peter Bruza, and Massimo Melucci. Quantum-Like Structure in Multidimensional Relevance Judgements. In: *Advances in Information Retrieval*. Ed. by Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins. Cham: Springer International Publishing, 2020, pp. 728–742. ISBN: 978-3-030-45439-5. DOI: 10.1007/978-3-030-45439-5_48.
- [424] Laurens Van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. In: *Journal of Machine Learning Research* 9.86 (2008), pp. 2579–2605. URL: <http://jmlr.org/papers/v9/vandermaaten08a.html>.
- [425] Jiří Vaníček. Software and data quality. In: *Agricultural Economics* 52 (2006), pp. 138–146. DOI: 10.17221/5007-AGRICON.
- [426] Slavena Vasileva, Pepa Atanasova, Lluís Màrquez, Alberto Barrón-Cedeño, and Preslav Nakov. It Takes Nine to Smell a Rat: Neural Multi-Task Learning for Check-Worthiness Prediction. In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*. Varna, Bulgaria: INCOMA Ltd., Sept. 2019, pp. 1229–1239. DOI: 10.26615/978-954-452-056-4_141.
- [427] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In: *Advances in Neural Information Processing Systems* 30. Ed. by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Long Beach, CA, USA: Curran Associates, Inc., 2017, pp. 5998–6008. URL: <https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>.
- [428] Karin Verspoor, Kevin Bretonnel Cohen, Mark Dredze, Emilio Ferrara, Jonathan May, Robert Munro, Cecile Paris, and Byron Wallace, eds. *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. Online: Association for Computational Linguistics, July 2020. URL: <https://aclanthology.org/2020.nlpCOVID19-acl.0>.
- [429] Jacky Visser, John Lawrence, and Chris Reed. Reason-Checking Fake News. In: *Communications of the ACM* 63.11 (Oct. 2020), 38–40. ISSN: 0001-0782. DOI: 10.1145/3397189.

- [430] Andreas Vlachos and Sebastian Riedel. Fact Checking: Task Definition And Dataset Construction. In: *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*. Baltimore, MD, USA: Association for Computational Linguistics, June 2014, pp. 18–22. doi: 10.3115/v1/W14-2508.
- [431] Nguyen Vo and Kyumin Lee. The Rise of Guardians: Fact-checking URL Recommendation to Combat Fake News. In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. SIGIR '18. Ann Arbor, MI, USA: Association for Computing Machinery, 2018, 275–284. isbn: 9781450356572. doi: 10.1145/3209978.3210037.
- [432] Soroush Vosoughi, Deb Roy, and Sinan Aral. The Spread of True and False News Online. In: *Science* 359.6380 (2018), pp. 1146–1151. doi: 10.1126/science.aap9559.
- [433] Maja Vukovic. Crowdsourcing for Enterprises. In: *2009 Congress on Services - I*. 2009, pp. 686–692. doi: 10.1109/SERVICES-I.2009.56.
- [434] Silvio Waisbord. Truth is What Happens to News. In: *Journalism Studies* 19.13 (2018), pp. 1866–1878. doi: 10.1080/1461670X.2018.1492881.
- [435] Richard Y. Wang and Diane M. Strong. Beyond Accuracy: What Data Quality Means to Data Consumers. In: *Journal of Management Information Systems* 12.4 (Mar. 1996), 5–33. issn: 0742-1222. doi: 10.1080/07421222.1996.11518099.
- [436] William Yang Wang. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Ed. by Regina Barzilay and Min-Yen Kan. Vol. 4. Association for Computational Linguistics, July 2017, pp. 422–426. doi: 10.18653/v1/P17-2067.
- [437] Xiaohui Wang, Dion Hoe-Lian Goh, and Ee-Peng Lim. Understanding Continuance Intention toward Crowdsourcing Games: A Longitudinal Investigation. In: *International Journal of Human-Computer Interaction* 36.12 (2020), pp. 1168–1177. doi: 10.1080/10447318.2020.1724010.
- [438] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. EANN: Event Adversarial Neural Networks for Multi-modal Fake News Detection. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '18. London, United Kingdom: Association for Computing Machinery, 2018, 849–857. isbn: 9781450355520. doi: 10.1145/3219819.3219903.
- [439] Yongqiao Wang, Lishuai Li, and Chuangyin Dang. Calibrating Classification Probabilities with Shape-Restricted Polynomial Regression. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.8 (2019), pp. 1813–1827. doi: 10.1109/TPAMI.2019.2895794.
- [440] Apurva Wani, Isha Joshi, Snehal Khandve, Vedangi Wagh, and Raviraj Joshi. Evaluating Deep Learning Approaches for Covid19 Fake News Detection. In: *Combating Online Hostile Posts in Regional Languages during Emergency Situation*. Ed. by Tanmoy Chakraborty, Kai Shu, H. Russell Bernard, Huan Liu, and Md Shad Akhtar. Cham: Springer International Publishing, 2021, pp. 153–163. isbn: 978-3-030-73696-5. doi: 10.1007/978-3-030-73696-5_15.

- [441] C. Nadine Wathen and Jacquelyn Burkell. Believe it or not: Factors influencing credibility on the Web. In: *Journal of the American Society for Information Science and Technology* 53.2 (2002), pp. 134–144. DOI: 10.1002/asi.10016.
- [442] Margaret A. Webb and June P. Tangney. Too Good to Be True: Bots and Bad Data From Mechanical Turk. In: *Perspectives on Psychological Science* (Nov. 2022), p. 17456916221120027. DOI: 10.1177/17456916221120027.
- [443] William Webber, Alistair Moffat, and Justin Zobel. A Similarity Measure for Indefinite Rankings. In: *ACM Transactions on Information Systems* 28.4 (Nov. 2010). ISSN: 1046-8188. DOI: 10.1145/1852102.1852106.
- [444] David Weiss and Benjamin Taskar. Structured Prediction Cascades. In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Yee Whye Teh and Mike Titterton. Vol. 9. Proceedings of Machine Learning Research. Chia Laguna Resort, Sardinia, Italy: PMLR, Apr. 2010, pp. 916–923. URL: <https://proceedings.mlr.press/v9/weiss10a.html>.
- [445] Matthew B. Welsh and Daniel J. Navarro. Seeing is believing: Priors, trust, and base rate neglect. In: *Organizational Behavior and Human Decision Processes* 119.1 (2012), pp. 1–14. ISSN: 0749-5978. DOI: 10.1016/j.obhdp.2012.04.001.
- [446] R. Wesslen, S. Santhanam, A. Karduni, I. Cho, S. Shaikh, and W. Dou. Investigating Effects of Visual Anchors on Decision-Making about Misinformation. In: *Computer Graphics Forum* 38.3 (2019), pp. 161–171. DOI: 10.1111/cgf.13679.
- [447] Ryen W. White and Eric Horvitz. Belief Dynamics and Biases in Web Search. In: *ACM Transactions on Information Systems* 33.4 (May 2015). ISSN: 1046-8188. DOI: 10.1145/2746229.
- [448] Colin Wilkie and Leif Azzopardi. Best and Fairest: An Empirical Analysis of Retrieval System Bias. In: *Advances in Information Retrieval*. Ed. by Maarten de Rijke, Tom Kenter, Arjen P. de Vries, ChengXiang Zhai, Franciska de Jong, Kira Radinsky, and Katja Hofmann. Cham: Springer International Publishing, 2014, pp. 13–25. DOI: 10.1007/978-3-319-06028-6_2.
- [449] Alex C. Williams, Gloria Mark, Kristy Milland, Edward Lank, and Edith Law. The Perpetual Work Life of Crowdworkers: How Tooling Practices Increase Fragmentation in Crowdwork. In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW (Nov. 2019). DOI: 10.1145/3359126.
- [450] Wired. *When It Comes to Gorillas, Google Photos Remains Blind*. <https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/>. (Accessed: 15-12-2021). 2018.
- [451] Thomas Wood and Ethan Porter. The Elusive Backfire Effect: Mass Attitudes’ Steadfast Factual Adherence. In: *Political Behavior* 41.1 (Mar. 2019), pp. 135–163. ISSN: 1573-6687. DOI: 10.1007/s11109-018-9443-y.

- [452] Lianwei Wu, Yuan Rao, Xiong Yang, Wanzhen Wang, and Ambreen Nazir. Evidence-Aware Hierarchical Interactive Attention Networks for Explainable Claim Verification. In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*. Ed. by Christian Bessiere. Main track. International Joint Conferences on Artificial Intelligence Organization, July 2020, pp. 1388–1394. DOI: 10.24963/ijcai.2020/193.
- [453] Lianwei Wu, Yuan Rao, Yongqiang Zhao, Hao Liang, and Ambreen Nazir. DTCA: Decision Tree-based Co-Attention Networks for Explainable Claim Verification. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 1024–1035. DOI: 10.18653/v1/2020.acl-main.97.
- [454] Meng-Han Wu and Alexander Quinn. Confusing the crowd: Task instruction quality on amazon mechanical turk. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*. Vol. 5. USA: AAAI, 2017, pp. 206–215. DOI: 10.4018/JGIM.20211101.oa13.
- [455] Philip Fei Wu, Hans van der Heijden, and Nikolaos Korfiatis. The Influences of Negativity and Review Quality on the Helpfulness of Online Reviews. In: *Proceedings of the International Conference on Information Systems*. Ed. by Dennis F. Galletta and Ting-Peng Liang. Shanghai, China: Association for Information Systems, 2011. URL: <https://ssrn.com/abstract=1937664>.
- [456] Adam Wyner, Jodi Schneider, Katie Atkinson, and Trevor Bench-Capon. Semi-Automated Argumentative Analysis of Online Product Reviews. In: *Proceedings of The 4th International Conference on Computational Models of Argument*. Frontiers in Artificial Intelligence and Applications. Vienna, Austria: IOS Press, 2012, pp. 43–50. DOI: 10.3233/978-1-61499-111-3-43.
- [457] Yunjie (Calvin) Xu and Zhiwei Chen. Relevance Judgment: What Do Information Users Consider beyond Topicality? In: *J. Am. Soc. Inf. Sci. Technol.* 57.7 (May 2006), 961–973. ISSN: 1532-2882. DOI: 10.5555/1133031.1133039.
- [458] Kai-Cheng Yang, Christopher Torres-Lugo, and Filippo Menczer. *Prevalence of Low-Credibility Information on Twitter During the COVID-19 Outbreak*. 2020. URL: <https://arxiv.org/abs/2004.14484>.
- [459] Ziyang Yang, Alistair Moffat, and Andrew Turpin. Pairwise crowd judgments: Preference, absolute, and ratio. In: *Proceedings of the 23rd Australasian Document Computing Symposium*. New York, NY, USA: ACM, 2018, pp. 1–8. ISBN: 9781450365499. DOI: 10.1145/3291992.3291995. URL: 10.1145/3291992.3291995.
- [460] Cheng Ye, Joseph Coco, Anna Epishova, Chen Hajaj, Henry Bogardus, Laurie Novak, Joshua Denny, Yevgeniy Vorobeychik, Thomas Lasko, Bradley Malin, et al. A Crowdsourcing Framework for Medical Data Sets. In: *AMIA Summits on Translational Science Proceedings 2018* (2018), p. 273. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5961774/pdf/2840819.pdf>.
- [461] Pinar Yildirim, Esther Gal-Or, and Tansev Geylani. User-Generated Content and Bias in News Media. In: *Management Science* 59.12 (2013), pp. 2655–2666. ISSN: 00251909, 15265501. URL: <http://www.jstor.org/stable/42919501> (visited on 10/21/2022).

- [462] Takuma Yoneda, Jeff Mitchell, Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. UCL Machine Reading Group: Four Factor Framework For Fact Finding (HexaF). In: *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 97–102. DOI: 10.18653/v1/W18-5515.
- [463] Di You, Nguyen Vo, Kyumin Lee, and Qiang LIU. Attributed Multi-Relational Attention Network for Fact-Checking URL Recommendation. In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. CIKM '19. Beijing, China: Association for Computing Machinery, 2019, 1471–1480. ISBN: 9781450369763. DOI: 10.1145/3357384.3358006.
- [464] Yisong Yue, Rajan Patel, and Hein Roehrig. Beyond Position Bias: Examining Result Attractiveness as a Source of Presentation Bias in Clickthrough Data. In: *Proceedings of the 19th International Conference on World Wide Web*. WWW '10. Raleigh, North Carolina, USA: Association for Computing Machinery, 2010, 1011–1018. ISBN: 9781605587998. DOI: 10.1145/1772690.1772793.
- [465] Bianca Zadrozny and Charles Elkan. Obtaining Calibrated Probability Estimates from Decision Trees and Naive Bayesian Classifiers. In: *Proceedings of the Eighteenth International Conference on Machine Learning*. ICML '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, 609–616. ISBN: 1558607781. DOI: 10.5555/645530.655658.
- [466] Fabio Zampieri, Kevin Roitero, J. Shane Culpepper, Oren Kurland, and Stefano Mizzaro. On Topic Difficulty in IR Evaluation: The Effect of Systems, Corpora, and System Components. In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR'19. Paris, France: ACM, 2019, 909–912. ISBN: 9781450361729. DOI: 10.1145/3331184.3331279.
- [467] Antonia Zapf, Stefanie Castell, Lars Morawietz, and André Karch. Measuring interrater reliability for nominal data – which coefficients and confidence intervals are appropriate? In: *BMC Medical Research Methodology* 16.1 (Aug. 2016), p. 93. ISSN: 1471-2288. DOI: 10.1186/s12874-016-0200-9.
- [468] Yinglong Zhang, Jin Zhang, Matthew Lease, and Jacek Gwizdka. Multidimensional Relevance Modeling via Psychometrics and Crowdsourcing. In: *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*. SIGIR '14. Gold Coast, Queensland, Australia: Association for Computing Machinery, 2014, 435–444. ISBN: 9781450322577. DOI: 10.1145/2600428.2609577.
- [469] Zhuosheng Zhang, Yuwei Wu, Zuchao Li, and Hai Zhao. Explicit Contextual Semantics for Text Comprehension. In: *CoRR* abs/1809.02794 (2018). arXiv: 1809.02794. URL: <http://arxiv.org/abs/1809.02794>.
- [470] Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. Semantics-Aware BERT for Language Understanding. In: vol. 34. 05. Apr. 2020, pp. 9628–9635. DOI: 10.1609/aaai.v34i05.6510.

- [471] Lanqin Zheng, Panpan Cui, Xin Li, and Ronghuai Huang. Synchronous discussion between assessors and assessees in web-based peer assessment: impact on writing performance, feedback quality, meta-cognitive awareness and self-efficacy. In: *Assessment & Evaluation in Higher Education* 43.3 (2018), pp. 500–514. DOI: 10.1080/02602938.2017.1370533.
- [472] Yanmengqian Zhou and Lijiang Shen. Confirmation bias and the persistence of misinformation on climate change. In: *Communication Research* 49.4 (2022), pp. 500–523. DOI: 10.1177/00936502211028049.
- [473] Fabiana Zollo. Dealing with digital misinformation: a polarised context of narratives and tribes. In: *EFSA Journal* 17.S1 (2019), e170720. DOI: 10.2903/j.efsa.2019.e170720.
- [474] Arkaitz Zubiaga and Heng Ji. Tweet, but Verify: Epistemic Study of Information Verification on Twitter. In: *Social Network Analysis and Mining* 4.1 (2014), pp. 1–12. DOI: 10.1007/s13278-014-0163-y.
- [475] Arkaitz Zubiaga, Maria Liakata, Rob Procter, Kalina Bontcheva, and Peter Tolmie. Crowdsourcing the Annotation of Rumourous Conversations in Social Media. In: *Proceedings of the 24th International Conference on World Wide Web. WWW '15 Companion*. Florence, Italy: Association for Computing Machinery, 2015, 347–353. ISBN: 9781450334730. DOI: 10.1145/2740908.2743052.
- [476] Guido Zuccon, Teerapong Leelanupab, Stewart Whiting, Emine Yilmaz, Joemon M. Jose, and Leif Azzopardi. Crowdsourcing Interactions: Using Crowdsourcing for Evaluating Interactive Information Retrieval Systems. In: *Information Retrieval* 16.2 (Apr. 2013), 267–305. ISSN: 1386-4564. DOI: 10.1007/s10791-012-9206-z.

Analytical Index

A

ABC Fact Check, 3, 24, 26, 31, 32, 34–36, 38–43, 45, 49, 50, 127, 128, 130, 133–137, 140, 144, 146–150, 170, 180, 181

Affect Heuristic, 158, 163, 172, 174, 176, 184, 210, 211

Agree, 130, 135, 209, 215

Amazon

- API Gateway, 264, 270
- DynamoDB, 237, 249, 274
- Elastic Compute Cloud, 267
- Lambda, 264, 273
- Mechanical Turk, 6, 7, 16, 21, 23, 29, 32, 35, 55, 56, 62, 90–93, 95–107, 109, 111, 116, 123, 125, 127, 130, 131, 177, 197, 214, 225–228, 231, 233, 234, 237, 244, 245, 268, 285, 287, 296, 304, 307, 308, 318
- Review Dataset, 214
- S3, 237, 247
- Simple Queue Service, 264, 273

Anchoring Bias, 159, 163–167, 211, 220

Angular, 265

ARI, 214

Attentional Bias, 159, 163, 166

Authority Bias, 159, 163

Automation Bias, 159, 163, 164, 167

Availability

- Cascade, 159, 163
- Heuristic, 159, 163, 165, 167

B

Backfire Effect, 159, 163, 165, 167

Bandwagon Effect, 159, 163, 165–167

Barnum Effect, 159, 163

BART, 188–191, 193, 195, 196, 203

- Large, 188, 191
- Separate, 195–197

Base Rate Fallacy, 159, 163, 165–167

Batch

- all, 56, 57, 71, 73, 74
- 1, 56, 57, 70–74, 77–79, 81, 83, 85
- 2, 56, 57, 70–74, 77–79, 81, 83
 - from1, 57, 81, 85
- 3, 56, 57, 70–74, 77–79, 81, 83
 - from1or2, 57, 81, 86
 - from1, 57
 - from2, 57
- 4, 56, 57, 70–74, 77–79, 81, 83, 85
 - from1or2or3, 57, 81, 86
 - from1, 57
 - from2, 57
 - from3, 57

Belief Bias, 159, 163

BERT, 189

- BASED, 193
- Sem, 194

BLEU, 194

C**C**

2, 62, 63
 3, 62, 63
 6, 55, 61–63, 66
 CA-MTL, 194
 CEM, 65, 66, 68, 81, 82
 Choice-Supportive Bias, 160, 163
 Compassion Fade, 160, 163
 Completely
 Agree, 130, 135, 209, 214, 215, 333
 Disagree, 130, 214, 215, 333
 Completeness, 127, 129, 139, 140, 146, 153, 214
 Comprehensibility, 127, 129, 134, 138–141, 144, 146, 152, 184, 214, 215
 Confidence, 130, 139–141, 171, 174, 211
 Confirmation Bias, 160, 163, 172, 176, 184
 Conjunction Fallacy, 160, 163
 Conservatism Bias, 160, 163, 165
 Conservative, 174
 Consistency Bias, 160, 162
 Contradiction, 29, 30
 Correctness, 127, 129, 134, 136, 138–141, 144, 146, 152, 174, 180, 181, 218
 Courtesy Bias, 160, 162
 COVID-19, iii, 1, 2, 9, 14, 15, 25, 51, 53–55, 58, 60, 62, 64, 87, 207, 209, 307
 Crowd_Frame, ix, 10, 91, 93, 225, 237–251, 253–257, 260–266, 268, 275, 280–283, 285–289, 292, 294–297, 299, 300, 303, 305–308

D

D-CRCo-AN, 194
 Declinism, 160, 163
 Democrat, 26, 173–175, 177
 Disagree, 130, 172, 174, 175, 215
 Docker, 280
 DOMLIN, 192, 193
 Dunning-Kruger Effect, 160, 163, 167

E**E**

2, 61, 62, 64
 3, 61, 62, 64
 6, 25, 61–64
 Search, 259
 Summary, 259
 E-BART, 188–198, 201–205, 208, 221
 Full, 191–193, 202–205
 Small, 191–194
 E-FEVER
 Full, 191, 193
 Small, 191, 193, 196, 197
 e-FEVER, 25, 28, 29, 191–194, 196, 202, 203
 e-INFERSENT, 194
 e-SNLI, 25, 29, 30, 191, 192, 194, 196, 203
 eAE, 171, 172
 eE, 171, 177, 180, 182, 183
 eMAE, 171–174, 177, 180, 183, 184
 eME, 171–174, 177, 180, 182
 Entailment, 29, 30, 195
 Entrez Programming Utilities, 258

F

False, 4, 14, 25, 26, 34, 38, 43, 49, 58, 60–62, 65, 73, 75, 76, 83–85, 87, 170, 180, 209, 313, 314
 FEVER, 25, 27, 28, 191–194, 197, 198
 Framing Effect, 160, 163–167
 Fundamental Attribution Error, 160, 163

G

G*Power, 176
 Google Effect, 161, 163

H

Half-True, 14, 25, 26, 34, 75, 76, 83, 84, 137, 170, 314
 High, 36

Hindsight Bias, 161, 163
 Hostile Attribution Bias, 161, 163

I

iE, 171, 177
 Illusion of Validity, 161, 163, 165, 167
 Illusory
 Correlation, 161, 163–165, 167
 Truth Effect, 161, 163
 iMe, 171, 174, 177, 184
 In-Between, 4, 26, 27, 32, 34, 38, 49, 170
 Independent, 173, 174, 177
 Infodemic, 1
 Informativeness, 127, 129, 133, 139, 140, 144,
 146, 153
 Ingroup Bias, 161–163

J

Joint Prediction Head, 188–191, 203
 JSONL, 27
 Just-World Hypothesis, 161, 163

K

Krippendorff's α , 41, 137, 180, 201

L

Labor, 27, 177
 Liberal, 27, 177
 Lie, 34, 49
 LMTransformer, 194
 Locust, 267
 Low, 36

M

MEDLINE, 258
 Mixed, 14

Mostly-False, 14, 25, 34, 38, 60, 75, 76, 83, 84,
 137, 170, 314
 Mostly-True, 14, 25, 32, 34, 60, 75, 76, 83, 84,
 170, 315
 MT-DNN, 194

N

Negative, 170
 Neither Agree Nor Disagree, 130, 215
 Neutral, 29, 30, 170, 195
 Neutrality, 127, 129, 134, 138–141, 152, 184
 NodeJS, 280
 NOT ENOUGH INFO, 27, 28, 191, 192
 null, 191

O

Optimism Bias, 161, 163
 Ostrich Effect, 161, 163, 164, 167
 Other, 14
 Outcome Bias, 161, 163
 Overall Truthfulness, 128–130, 134, 136–142,
 144–146, 149, 151, 153, 171, 174–176,
 180, 181, 184, 214, 218
 Overconfidence Effect, 162, 163, 167, 174,
 176, 184, 210

P

P
 1, 90–92, 95, 97, 109, 111, 114, 116, 317
 2, 90–92, 95, 103, 109, 112, 114, 118, 317,
 319
 Pants-On-Fire, 25, 34, 38, 43, 58, 60–62, 65,
 69, 73, 75, 76, 83–85, 87, 170, 180,
 209, 313
 PolitiFact, 3, 4, 14, 25, 26, 31–36, 38–43, 45, 49,
 50, 55, 58–61, 67, 71, 73, 74, 77, 83,
 84, 127, 128, 130, 133–138, 140, 142,
 144–149, 154, 170, 171, 180, 181, 215,
 292
 Positive, 170

Precision, 127, 129, 134, 136–141, 146, 152, 180, 181, 215
 Prolific, 6, 22, 90–93, 95–107, 109–113, 116, 119, 120, 123–126, 215, 225, 232–237, 244, 247, 284, 298, 304, 307, 318
 Proportionality Bias, 162
 Python, 280

R

REFUTES, 27–29, 192, 194, 197, 205
 Republican, 26, 172–175, 177
 Rouge, 193, 196

S

S
 100, 34–44, 46–48, 50
 3, 34–48, 50
 6, 34–48, 50
 Salience Bias, 162, 163, 166
 SJRC, 194
 SNLI, 29
 Speaker's Trustworthiness, 127, 129, 139–141, 146, 153, 174
 Stereotypical Bias, 162, 163
 SUPPORTS, 27–29, 194, 197, 204, 205

T

Task
 1, 197–199, 201
 2, 197–199, 201
 3, 198, 199, 201
 4, 198, 200, 201
 Telescoping Effect, 162, 163
 Toloka, 6, 90–93, 95–107, 109–113, 116, 118, 119, 122–124, 126, 225, 228–233, 237, 244, 282, 284, 287, 297, 304, 307, 318
 Transformer, 188
 True, 4, 14, 25–27, 34, 49, 58, 60, 65, 69, 73, 75, 76, 83–85, 128, 170, 180, 197, 315

U

UCL MR, 193
 UKP-Athene, 192
 UNC, 192

Y

Yarn, 280