Combining Large Language Models and Crowdsourcing for Hybrid Human-AI Misinformation Detection

Xia Zeng*
x.zeng@qmul.ac.uk
Queen Mary University of London
London, U.K.

David La Barbera*
david.labarbera@uniud.it
University of Udine
Udine, Italy

Kevin Roitero
kevin.roitero@uniud.it
University of Udine
Udine, Italy

Arkaitz Zubiaga
a.zubiaga@qmul.ac.uk
Queen Mary University of London
London, U.K.

Stefano Mizzaro
mizzaro@uniud.it
University of Udine
Udine, Italy

ABSTRACT

Research on misinformation detection has primarily focused either on furthering Artificial Intelligence (AI) for automated detection or on studying humans' ability to deliver an effective crowdsourced solution. Each of these directions however shows different benefits. This motivates our work to study hybrid human-AI approaches jointly leveraging the potential of large language models and crowdsourcing, which is understudied to date. We propose novel combination strategies Model First, Worker First, and Meta Vote, which we evaluate along with baseline methods such as mean, median, hard- and soft-voting. Using 120 statements from the PolitiFact dataset, and a combination of state-of-the-art AI models and crowdsourced assessments, we evaluate the effectiveness of these combination strategies. Results suggest that the effectiveness varies with scales granularity, and that combining AI and human judgments enhances truthfulness assessments' effectiveness and robustness.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; • **Human-centered computing**;

KEYWORDS

Misinformation Detection, AI-Crowdsourcing Integration, LLM.

ACM Reference Format:
Xia Zeng, David La Barbera, Kevin Roitero, Arkaitz Zubiaga, and Stefano Mizzaro. 2024. Combining Large Language Models and Crowdsourcing for Hybrid Human-AI Misinformation Detection. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3626772.3657965

1 INTRODUCTION

Misinformation poses a significant societal issue with wide-ranging implications in various domains, including politics and public health. This complex landscape underscores the urgent need for reliable methods able to detect and accurately classify the veracity of information. Both Artificial Intelligence (AI) methods, particularly Large Language Models (LLMs) [2, 8, 17, 27, 35], and crowdsourced approaches [1, 3, 9, 33] have shown distinct strengths and limitations in the task of misinformation and veracity assessment. LLMs have significantly enhanced our ability to process and analyze large volumes of natural language data; however they often struggle with the complex, nuanced nature of misinformation, and are susceptible to biases [5, 26]. Similarly, crowdsourced approaches can take advantage of the valuable human insight and judgment, yet they too can be affected by biases [7, 13]. This mixed landscape demands innovative solutions that effectively combine the precision of AI with the discernment of human analysis, aiming to improve the effectiveness and reliability of misinformation detection.

Different from approaches detecting misinformation by relying on classification outputs generated by either LLMs or crowdsourcing, this paper explores the integration of LLMs and crowdsourcing (henceforth LLM+Crowd) for the task of misinformation detection, and providing insights into how to best achieve an effective outcome. By introducing Model First, Worker First, and Meta Vote, three novel combination methods that leverage the strengths of both AI and human classifiers and through experimentation, we propose the first practical framework to achieve hybrid human–AI verification, and we develop a hybrid methodology that, under carefully defined settings, improves the effectiveness of the individual approaches. We focus on three Research Questions (RQ):

RQ1: What are the individual performance of LLMs and Crowd in misinformation detection? What are similarities and differences in their performance?
RQ2: When and how does the LLM+Crowd combination outperform individual approaches in misinformation detection?
RQ3: How do advanced combination techniques and judgment scales affect the effectiveness, reliability, and robustness of integrating LLM+Crowd for misinformation detection?

2 BACKGROUND AND RELATED WORK

AI, especially LLMs, emerged as a powerful tool against misinformation. Many studies investigated the task of misinformation detection, with some focusing on automated claim verification providing evidence alongside claims, while others perform misinformation detection where only the content to be verified is available. Few-shot research predominantly relies on prompting generative LLMs like LLaMA 2 and GPT-4 [19], yielding comparable results to SEED [29] and MAPLE [31], despite not leveraging ample training data. Most AI methods excel with fully supervised training, such as fine-tuning BERT for text classification on misinformation datasets [15, 23, 28, 32]. While Zeng and Zubiaga [30] broaden the scope of misinformation detection by prioritizing data annotation, there is a lack of research on integrating AI and crowd.

*The first two authors contributed equally to this research.

SIGIR '24, July 14–18, 2024, Washington, DC, USA

ACM ISBN 979-8-4007-0431-4/24/07.
https://doi.org/10.1145/3626772.3657965

Focusing on crowdsourced approaches to misinformation, La Barbera et al. [13] found evidence of worker biases when performing verification tasks. Roitero et al. [20] and later Soprano et al. [22] expanded this line of research, using different truthfulness scales and introducing a multidimensional notion of truthfulness. Draws et al. [7] validated results from Soprano et al. [22] identifying potential and systematic biases. Allen et al. [1] found correlation between politically balanced crowds and experts on news headlines, while Saeed et al. [21] found similar patterns on Twitter data.

Considering the intersection between AI and crowd for misinformation detection, the development of Human-In-The-Loop (HITL) systems gained research attention [2, 3]. Qu et al. [18] investigated the use of self-reported confidence as an indicator for accurate predictions from both crowd and LLMs. La Barbera et al. [14] proposed a theoretical HITL framework to efficiently integrate AI and crowdsourced workers; this approach, differently from the one proposed in this paper, focused on the conceptual aspects of combining crowd and AI resources. Some concrete HITL approaches exist: Dong et al. [6] proposed a decentralized HITL method for misinformation detection integrating user feedback, and Yang et al. [25] developed a hybrid system to accelerate misinformation detection by grouping similar social media through semantic analysis and clustering.

3 METHODOLOGY

3.1 Data Source

Building on data from published works [7, 13, 20, 22], this study uses a subset of 120 statements from the extensive PolitiFact dataset [24], which comprises more than 10 thousand statements from US politicians, public figures, and social media posts. We manually selected 20 statements for each of the six ground truth levels, ranging from most false (Pants-On-Fire) to most true (True), from the PolitiFact website for statements made in 2022. This selection aims at providing crowd workers with a current and relevant representation of misinformation statements. In the following we will refer to PolitiFact scale as S6. This approach aligns with the task design used by Soprano et al. [22] and Draws et al. [7].

Using the Prolific[1] platform, we collect 1,200 assessments (i.e., 20 statements for each PolitiFact categories × 6 PolitiFact categories × 10 assessments per statement) from US-based workers. Each participant sees 8 statements, 1 for each ground truth level plus two gold questions for quality assurance. Consistently with previous studies [7, 13, 20, 22], we focus on the individual worker's reported truthfulness for each statement (on the same six-level scale used by PolitiFact), their confidence in each assessment, and the internal agreement among participants quantified as Krippendorff's α [12].

Complementary to the S6 scale, we binarize the crowd judgments and the corresponding ground truth values for each statement, which we refer to as S2. Thus, statements categorized as Pants-On-Fire, False, and Mostly-False are grouped under a singular False label, and Half-True, Mostly-True, and True under the True label. This simplification allows us to compare our results with the state of the art reporting binary performance. Moreover, S2 allows to avoid issues with ordinal classification metrics: the distance between categories is equal, thus is not subject to misinterpretation, eliminating the ambiguity associated with a multi-level scale.

[1] https://www.prolific.com/

3.2 Large Language Models

We employ three state-of-the-art LLMs: BERT-large [4], DeBERTa-large [10], and RoBERTa-large [16]. To tailor them for misinformation detection, we fine-tune on the statements from the PolitiFact dataset, excluding the 120 statements used in the crowdsourcing task used as test set. All of the models are trained for 3 epochs as the vast amount of training data allows early convergence. Hyperparameter tuning on batch size and learning rate are performed with optuna using 10% of the train set. Learning rate search range is between $1e^{-6}$ and $1e^{-4}$. Batch size is among 16, 64, 128. We use the soft-max scores computed in the last layer of the neural network as model confidence.

3.3 Combination Strategies

Following independent classifications from both the AI models and the crowd workers, we propose various combination techniques to combine the results to boost performance and gain insight into the potential of model and crowd contributions.

We leverage mean and median as foundational techniques, providing a baseline comparison with previous works [13, 20, 22]. We consider also two other standard techniques: the hard-voting (mode) [11, 34] and the soft-voting (argmax on the weighted average of all prediction probabilities) [11, 34]. As weights for the soft-voting we use internal agreement for workers and prediction probability for models. Moreover, we introduce three advanced methods: Model First, Worker First, and Meta Vote. **Model First** involves a two-step process where we first apply hard-voting to aggregate individual predictions from the workers into a interim vote for all workers and obtain the internal agreement as the confidence for this vote. We then combine the interim vote with all predictions from the models to reach a final prediction using soft-voting. This method emphasizes the input from the models as the final prediction is based on multiple votes from models and only a single combined vote from the workers. **Worker First** starts with applying hard-voting to individual models' predictions to obtain a interim vote. It then reachs a final prediction by employing hard-voting on the interim vote and all predictions from the workers. Worker First gives precedence to worker insights, emphasizing human intuition and expertise, as the decisive step employs hard-voting on the aggregated worker-centric data for the final judgment, thus the **Meta Vote** combination relies on subgroup and vote: first the voting members are grouped into all possible subgroups from size 2 to all; then the voting decisions with both soft-voting and hard-voting are computed within each subgroup. Lastly these voting decisions are used to reach the final decision through hard-voting.

3.4 Evaluation Metrics

We evaluate the performance of different combination strategies with three metrics: (i) accuracy, suitable for measuring effectiveness on a balanced dataset; (ii) mean squared error (MSE), since in our context it might be important to emphasizes the severity of errors (e.g., a False as a True); and (iii) mean absolute error (MAE), as a more balanced measure of the average error magnitude, presenting an easy-to-interpret picture of typical prediction errors.[2]

[2] F1 score is not reported given its similarity with accuracy on balanced data.

Table 1: Accuracy, MSE, and MAE for the combined models (M), crowd (C), and both together (M+C) for S2 (top part; note that MSE = MAE due to the binary nature of the S2 scale) and S6 (bottom) scales. H-V = hard-voting, S-V = soft-voting, μ = Mean, Me = Median, M-V = Meta Vote, M-F = Model First, W-F = Worker First.

Metric	Who	H-V	S-V	μ	Me	M-V	M-F	W-F
ACC	M	.700	.700	.667	.700	.692	–	–
	C	.816	.775	.816	.816	.816	–	–
	M+C	.791	.766	.791	.791	**.875**	.775	.800
MSE/	M	.300	.300	.332	.300	.308	–	–
MAE	C	.183	.225	.183	.183	.183	–	–
	M+C	.208	.233	.208	.208	**.125**	.225	.200
ACC	M	.325	.325	.283	.325	.316	–	–
	C	.425	.425	.275	.408	.408	–	–
	M+C	.408	.366	.283	**.441**	.400	.425	.383
MSE	M	2.242	2.242	2.250	2.216	2.325	–	–
	C	2.233	1.983	1.583	1.516	3.092	–	–
	M+C	2.075	2.375	1.550	**1.383**	2.425	1.833	2.233
MAE	M	1.092	1.092	1.133	1.083	1.125	–	–
	C	.983	.933	.983	.850	1.142	–	–
	M+C	.958	1.075	.967	**.783**	1.025	.900	1.017

4 RESULTS

We discuss each of the RQs in the following three subsections.

4.1 RQ1: Individual Approaches, LLM vs. Crowd

Figure 1 shows the results for accuracy (left) and MSE (right) for the two scales (top for S2, bottom for S6). Table 1 report accuracy, MSE, and MAE for the aggregated models, the crowd, and the LLM+Crowd combination, (top for S2, bottom for S6). Considering S2 (Figure 1 top row, Table 1 top) we first assess individual model contributions (M in the table, blue circles in the figure) and crowd workers (C in the table, orange squares in the figure). We can see consistent accuracy values around 0.7, with a slight decrease when applying mean (0.667). Similarly, MSE and MAE scores show little variation, indicating stable but not optimal performance. Similarly, crowd judgments show little variation with higher accuracy of 0.816 when compared to the models for all the aggregation techniques except for soft-voting and notably lower error rates (MSE and MAE) compared to the models. For the S6 scale (Figure 1 bottom row, Table 1 bottom) the models achieve the best accuracy (0.325) when applying hard-voting, soft-voting, and Median. Conversely, the crowd demonstrates higher accuracy than the models, achieving the highest values with hard-voting and soft-voting (0.425), and lower error rates. Nevertheless, crowd accuracy significantly drops when aggregating using the mean, achieving scores that are similar to the ones achieved by the models despite showing a lower error rate, as by MSE and MAE. This suggests that aggregation methods like the mean function may not effectively leverage the complementary strengths of human and models, highlighting the necessity for more sophisticated combination strategies.

4.2 RQ2: Combination, LLM + Crowd

We now consider RQ2, identifying crowd and models combinations (M+C in the tables, green Xs in Figure 1) that outperform individual approaches. Considering the S2 scale (top row in Figure 1, Table 1 top) we see that the Meta Vote (M-V column in the table) method delivers the highest accuracy (0.875) and the lowest error rates (MSE and MAE at 0.125), both combined and individual.

Considering results for S6 (bottom row in Figure 1, Table 1 bottom), among the combination strategies, the best accuracy is achieved by the median Me (0.441), along with a reduction in MSE and MAE compared to model- and crowd-only scores. Differently from S2, Model First (M-V) achieve an accuracy (0.425) comparable to the best crowd performance, albeit showing lower error rates. Hence, while the use of a complex combination method such as Meta Vote leads to high accuracy and low error rates for coarser grained scales such as S2, for finer grained scales such as S6 the best results are achieved using the Median. These results corroborate a critical aspect of misinformation detection: the effectiveness of aggregation strategies is highly dependent on the scale and nature of the task. While complex methods like Meta Vote perform better when measuring performance on coarser scales, simpler approaches like Median are more effective for fine-grained ones. Overall, these results suggest that there is no one-fits-all solution in the realm of aggregation strategies for misinformation detection, and hint for the necessity of a further studies to develop a selection approach for the selection of the optimal aggregation strategy, which should be aligned with the specific task and scale being used.

4.3 RQ3: Effectiveness, Reliability, and Robustness of Combination Strategies

In answering RQ3, we look at classification differences over the possible labels computing confusion matrices. In Figure 2 we report the confusion matrices considering the Meta Vote, for the S6 scale. The others, omitted, show a similar pattern. Considering the first two plots, we see that models provide more consistent classifications on middle scale values such as Mostly-False (MF), Half-True (HT), and Mostly-True (MT), while struggle with extreme categories, Pants-On-Fire (PF) and True (T). Conversely, the crowd can better identify those two extreme categories, reflecting a better understanding than models to the difference of truthfulness in statements.

Finally considering models and crowd combinations (Figure 2 third plot), it is clear that the combination of both provide more balanced judgments across S6 truthfulness spectrum. Thus, while as shown for RQ2 the combined approaches do not clearly outperform the others in terms of accuracy and error rates, these combinations allow to preserve crowd's contextual sensitivity to the scale differences and the models' classification consistency. Thus, for applications requiring nuanced understanding of truthfulness, a hybrid approach that leverages both human judgment and automated models can provide a more robust and accurate classification than relying on either source alone. Having identified the best combination techniques and delved into their performances across single labels, we now complete our analysis examining the robustness of these metrics to ensure a comprehensive evaluation. Thus, we iteratively change judgments from crowd or models, recompute the aggregation technique and the consequent accuracy.

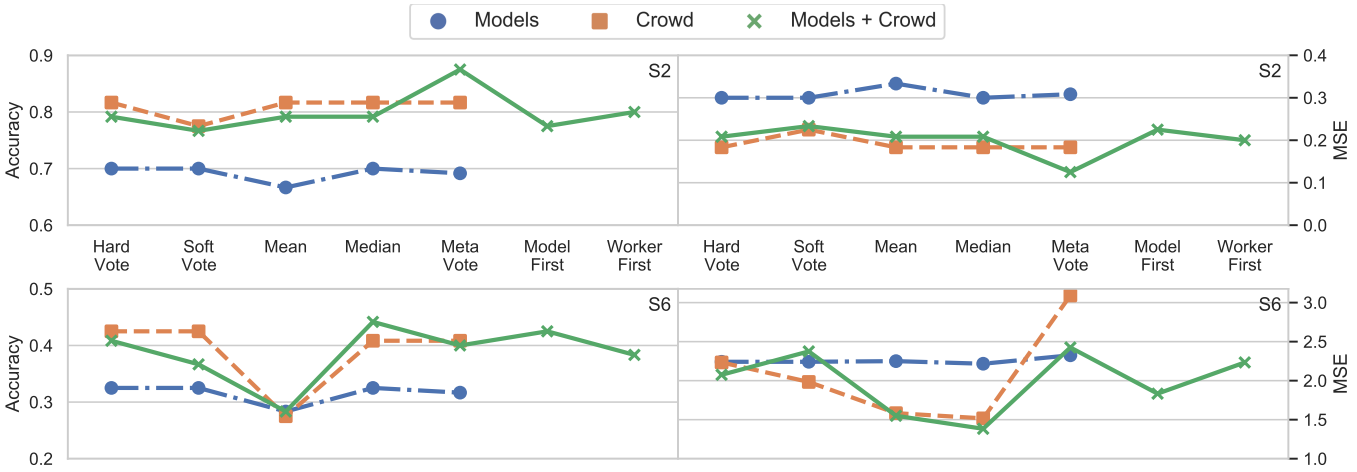

Figure 1: Accuracy (left) and MSE (right) for aggregated models, crowd and their combination. Top for S2, bottom for S6.

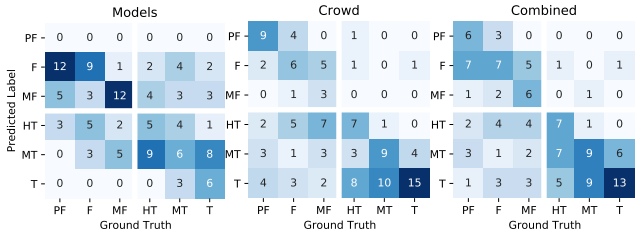

Figure 2: Confusion matrices for meta voting combination, S6 scale. PF: Pants-On-Fire, F: Mostly-False, MF: Mostly-False, HT: Half-True, MT: Mostly-True, and T: Half-True.

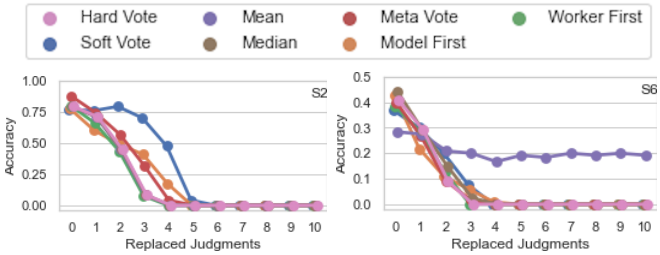

Figure 3: Variations in accuracy for the combined metrics when replacing single judgments. S2 left, S6 right.

Figure 3 shows the results on the aggregation metrics for the two scales: accuracy degradation with increasing replacements is more pronounced for S2 (left), despite being evident for both scales, as most values get close to 0 with 3 to 5 judgments replaced. Notably, the soft-voting method for S2 shows more resilience to few data manipulations. This suggests robustness to minor errors but also vulnerability to extensive manipulations. Conversely, the Mean on S6 shows consistent performance, despite originally being the worst performances metric. This indicates a resilience that, while not improving metric's performance, prevents a sharp decline highlighting its potential reliability in more granular assessments.

5 CONCLUSIONS AND FUTURE WORK

In this work, we study the integration of crowd and LLM to get the best of both worlds for misinformation detection. We introduce novel combination strategies including Worker First, Model First and Meta Vote, which we evaluate on a sample of the PolitiFact dataset. Our findings show that (RQ1) while crowd performs better than LLMs, they show a similar and consistent pattern in terms of accuracy across the considered aggregation functions, despite the crowd having lower error rates. Moreover, (RQ2) we found evidence of the influence of the used scale on the effectiveness of aggregation methods. Thus, the used scale should be taken into account when selecting an aggregation function: while Meta Vote method excels for less fine grained scales (S2), the Median shows better scores for more complex scales, (S6). Finally, (RQ3) we found that combining crowd and LLM results in more balanced assessments across the truthfulness spectrum, suggesting that a hybrid approach for nuanced truthfulness understanding offers greater accuracy and robustness. In future work, we plan to use emerging methods such as Retrieval Augmented Generation or chat-based crowd-LLM interactions, and to expand to other datasets.

Acknowledgments. This research is supported by the NextGenerationEU PRIN 2022 project "20227F2ZN3_001 MoT–The Measure of Truth: An Evaluation-Centered Machine-Human Hybrid Framework for Assessing Information Truthfulness CUP G53D23002800006" and by the Strategic Plan of the University of Udine–Interdepartment Project on Artificial Intelligence (2020-25). Xia Zeng is funded by China Scholarship Council (CSC). Arkaitz Zubiaga acknowledges support from the EU and UKRI under Grant No. 101073351 as part of MSCA Hybrid Intelligence to monitor, promote, and analyze transformations in good democracy practices. This research utilised Queen Mary's Apocrita HPC facility, supported by QMUL Research-IT. http://doi.org/10.5281/zenodo.438045

REFERENCES

[1] Jennifer Allen, Antonio A. Arechar, Gordon Pennycook, and David G. Rand. 2021. Scaling up fact-checking using the wisdom of crowds. *Science Advances* 7, 36 (2021), eabf4393. https://doi.org/10.1126/sciadv.abf4393

[2] Anubrata Das, Houjiang Liu, Venelin Kovatchev, and Matthew Lease. 2023. The State of Human-centered NLP Technology for Fact-checking. *Information Processing & Management* 60, 2 (2023), 103219. https://doi.org/10.1016/j.ipm.2022.103219

[3] Gianluca Demartini, Stefano Mizzaro, and Damiano Spina. 2020. Human-in-the-loop Artificial Intelligence for Fighting Online Misinformation: Challenges and Opportunities. *IEEE Data Engineering Bulletin* 43, 3 (2020), 65–74. http://sites.computer.org/debull/A20sept/p65.pdf

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). arXiv:1810.04805

[5] Harnoor Dhingra, Preetiha Jayashanker, Sayali Moghe, and Emma Strubell. 2023. Queer People are People First: Deconstructing Sexual Identity Stereotypes in Large Language Models. arXiv:2307.00101 [cs.CL]

[6] Xishuang Dong, Shouvon Sarker, and Lijun Qian. 2022. Integrating Human-in-the-loop into Swarm Learning for Decentralized Fake News Detection. In *International Conference on Intelligent Data Science Technologies and Applications (IDSTA)*. IEEE, San Antonio, TX, USA, 46–53. https://doi.org/10.1109/IDSTA55301.2022.9923043

[7] Tim Draws, David La Barbera, Michael Soprano, Kevin Roitero, Davide Ceolin, Alessandro Checco, and Stefano Mizzaro. 2022. The Effects of Crowd Worker Biases in Fact-Checking Tasks. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, Seoul, Republic of Korea, 2114–2124. https://doi.org/10.1145/3531146.3534629

[8] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics* 10 (2022), 178–206. https://doi.org/10.1162/tacl_a_00454

[9] Bing He, Yibo Hu, Yeon-Chang Lee, Soyoung Oh, Gaurav Verma, and Srijan Kumar. 2023. A Survey on the Role of Crowds in Combating Online Misinformation: Annotators, Evaluators, and Creators. arXiv:2310.02095 [cs.SI]

[10] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. arXiv:2006.03654

[11] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas. 1998. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 3 (1998), 226–239. https://doi.org/10.1109/34.667881

[12] Klaus Krippendorff. 2008. Computing Krippendorff's Alpha-Reliability. *UPENN Libraries* 1 (2008), 43. https://repository.upenn.edu/asc_papers/43

[13] David La Barbera, Kevin Roitero, Gianluca Demartini, Stefano Mizzaro, and Damiano Spina. 2020. Crowdsourcing Truthfulness: The Impact of Judgment Scale and Assessor Bias. In *Advances in Information Retrieval*, Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins (Eds.). Springer International Publishing, Cham, 207–214. https://doi.org/10.1007/978-3-030-45442-5_26

[14] David La Barbera, Kevin Roitero, and Stefano Mizzaro. 2022. A Hybrid Human-In-The-Loop Framework for Fact Checking. In *Proceedings of the 6th Workshop on Natural Language for Artificial Intelligence (NL4AI '22)*. CEUR-WS.org, Udine, Italy, 1–10. https://ceur-ws.org/Vol-3287/paper4.pdf

[15] Xiangci Li, Gully Burns, and Nanyun Peng. 2021. A Paragraph-level Multi-task Learning Model for Scientific Fact-Verification. *arXiv* abs/2012.14500 (Jan. 2021). http://arxiv.org/abs/2012.14500

[16] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* abs/1907.11692 (2019). arXiv:1907.11692

[17] Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated Fact-Checking for Assisting Human Fact-Checkers. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, Zhi-Hua Zhou (Ed.). International Joint Conferences on Artificial Intelligence Organization, 4551–4558. https://doi.org/10.24963/ijcai.2021/619 Survey Track.

[18] Yunke Qu, Kevin Roitero, David La Barbera, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. 2022. Combining Human and Machine Confidence in Truthfulness Assessment. *Journal of Data and Information Quality*. 15, 1, Article 5 (dec 2022), 17 pages. https://doi.org/10.1145/3546916

[19] Dorian Quelle and Alexandre Bovet. 2023. The Perils & Promises of Fact-checking with Large Language Models. arXiv:2310.13549 [cs.CL]

[20] Kevin Roitero, Michael Soprano, Shaoyang Fan, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. 2020. Can The Crowd Identify Misinformation Objectively? The Effects of Judgment Scale and Assessor's Background. In *Proceedings of the 43st International ACM SIGIR Conference on Research and Development in Information Retrieval*. (Xi'an, China (Virtual)) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 439–448. https://doi.org/10.1145/3397271.3401112

[21] Mohammed Saeed, Nicolas Traub, Maelle Nicolas, Gianluca Demartini, and Paolo Papotti. 2022. Crowdsourced Fact-Checking at Twitter: How Does the Crowd Compare With Experts?. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management* (Atlanta, GA, USA) *(CIKM '22)*. Association for Computing Machinery, New York, NY, USA, 1736–1746. https://doi.org/10.1145/3511808.3557279

[22] Michael Soprano, Kevin Roitero, David La Barbera, Davide Ceolin, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. 2021. The Many Dimensions of Truthfulness: Crowdsourcing Misinformation Assessments on a Multidimensional Scale. *Information Processing & Management* 58, 6 (2021), 102710. https://doi.org/10.1016/j.ipm.2021.102710

[23] David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2022. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In *Findings of the Association for Computational Linguistics: NAACL 2022*. Association for Computational Linguistics, Seattle, United States, 61–76. https://doi.org/10.18653/v1/2022.findings-naacl.6

[24] William Yang Wang. 2017. "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*, Regina Barzilay and Min-Yen Kan (Eds.), Vol. 4. Association for Computational Linguistics, Vancouver, Canada, 422–426. https://doi.org/10.18653/v1/P17-2067

[25] Jing Yang, Didier Vega-Oliveros, Tais Seibt, and Anderson Rocha. 2021. Scalable Fact-checking with Human-in-the-Loop. In *2021 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, Montpellier, France, 1–6. https://doi.org/10.1109/WIFS53200.2021.9648388

[26] Kai-Ching Yeh, Jou-An Chi, Da-Chen Lian, and Shu-Kai Hsieh. 2023. Evaluating Interfaced LLM Bias. In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, Jheng-Long Wu and Ming-Hsiang Su (Eds.). The Association for Computational Linguistics and Chinese Language Processing (ACLCLP), Taipei City, Taiwan, 292–299. https://aclanthology.org/2023.rocling-1.37

[27] Xia Zeng, Amani S. Abumansour, and Arkaitz Zubiaga. 2021. Automated fact-checking: A survey. *Language and Linguistics Compass* 15, 10 (2021), e12438. https://doi.org/10.1111/lnc3.12438

[28] Xia Zeng and Arkaitz Zubiaga. 2021. QMUL-SDS at SCIVER: Step-by-Step Binary Classification for Scientific Claim Verification. In *Proceedings of the Second Workshop on Scholarly Document Processing*, Iz Beltagy, Arman Cohan, Guy Feigenblat, Dayne Freitag, Tirthankar Ghosal, Keith Hall, Drahomira Herrmannova, Petr Knoth, Kyle Lo, Philipp Mayr, Robert M. Patton, Michal Shmueli-Scheuer, Anita de Waard, Kuansan Wang, and Lucy Lu Wang (Eds.). Association for Computational Linguistics, Online, 116–123. https://doi.org/10.18653/v1/2021.sdp-1.15

[29] Xia Zeng and Arkaitz Zubiaga. 2022. Aggregating pairwise semantic differences for few-shot claim verification. *PeerJ Computer Science* 8 (Oct. 2022), e1137. https://doi.org/10.7717/peerj-cs.1137 Publisher: PeerJ Inc..

[30] Xia Zeng and Arkaitz Zubiaga. 2023. Active PETs: Active Data Annotation Prioritisation for Few-Shot Claim Verification with Pattern Exploiting Training. In *Findings of the Association for Computational Linguistics: EACL 2023*. Association for Computational Linguistics, Dubrovnik, Croatia, 190–204. https://aclanthology.org/2023.findings-eacl.14

[31] Xia Zeng and Arkaitz Zubiaga. 2024. MAPLE: Micro Analysis of Pairwise Language Evolution for Few-Shot Claim Verification. In *Findings of the Association for Computational Linguistics: EACL 2024*, Yvette Graham and Matthew Purver (Eds.). Association for Computational Linguistics, St. Julian's, Malta, 1177–1196. https://aclanthology.org/2024.findings-eacl.79

[32] Zhiwei Zhang, Jiyi Li, Fumiyo Fukumoto, and Yanming Ye. 2021. Abstract, Rationale, Stance: A Joint Model for Scientific Claim Verification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 3580–3586. https://doi.org/10.18653/v1/2021.emnlp-main.290

[33] Andy Zhao and Mor Naaman. 2023. Insights from a Comparative Study on the Variety, Velocity, Veracity, and Viability of Crowdsourced and Professional Fact-Checking Services. *Journal of Online Trust & Safety* 2, 1 (2023), eabf4393. https://doi.org/10.54501/jots.v2i1.118

[34] Zhi-Hua Zhou. 2012. *Ensemble Methods: Foundations and Algorithms* (1st ed.). Chapman & Hall/CRC.

[35] Arkaitz Zubiaga. 2024. Natural language processing in the era of large language models. *Frontiers in Artificial Intelligence* 6 (2024), 1350306.