DiLBERT: Cheap Embeddings for Disease Related Medical NLP

Published
DOI:10.1109/ACCESS.2021.3131386

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 2, 2021, accepted November 17, 2021. Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2021.3131386

DiLBERT: Cheap Embeddings for Disease Related Medical NLP

KEVIN ROITERO, BEATRICE PORTELLI, MIHAI HORIA POPESCU, AND VINCENZO DELLA MEA

Department of Mathematics, Computer Science and Physics, University of Udine, 33100 Udine, Italy

Corresponding author: Vincenzo Della Mea (vincenzo.dellamea@uniud.it)

ABSTRACT Electronic Health Records include health-related information, among which there is text mentioning health conditions and diagnoses. Usually, text is also coded using appropriate terminologies and classifications. The act of coding is time consuming and prone to mistakes. Consequently, there is increasing demand for clinical text mining tools to help coding. In last few years Natural Language Processing (NLP) models has been shown to be effective in sentence-level tasks. Taking advantage from the transfer learning capabilities of those models, a number of biomedicine and health specific models have been also developed. However, also biomedical models can be seen as too general for some specific area like diagnostic expressions. In this paper, we describe a BERT model specialized on tasks related to diagnoses and health conditions. To obtain a disease-related language model, we created a pre-training corpora starting from ICD-11 entities, and enriched them with documents selected by querying PubMed and Wikipedia with entity names. Fine-tuning has been carried out towards three downstream tasks on two different datasets. Results show that our model, besides being trained on a much smaller corpora than state-of-the-art algorithms, leads to comparable or higher accuracy scores on all the considered tasks, in particular 97.53% accuracy on death certificate coding, and 81.32% on clinical document coding, which are both slightly higher than other models. To summarize the practical implications of our work, we pre-trained and fine-tuned a domain specific BERT model on a small corpora, with comparable or better performance than state-of-the-art models. This approach may also simplify the development of models for languages different from English, due to the minor quantity of data needed for training.

INDEX TERMS Natural language processing, language models, embeddings, disease, transformer, ICD-11.

I. INTRODUCTION

The electronic health records (EHR) are the digital version of patients paper chart which are more and more used in the management of health care issues at different levels. EHRs include information like demographic and social aspects, signs and symptoms, exam reports, health conditions and diagnoses, medical procedures, functioning aspects. This information can be structured or expressed in free text form, the latter sometimes as narrative, sometimes as short textual descriptions. Usually, text is also coded and classified using appropriate terminologies and classifications, like ICD, ICF, SNOMED-CT, CPT, etc. The process of coding is used to associate some standard meaning to the textual expressions, so then data can be aggregated and interpreted. (e.g. EHR contains the diagnosis that are filled in the death certificates coded with the ICD codes, which then can be used for mortality and morbidity statistics.). This allows for the rapid retrieval of data, as well as the implementation of decision support, surveillance systems, epidemiological applications. A notable application of coding is related to the financial aspects of health care: billing and reimbursements are calculated from coded information extracted from specific health documents, like the hospital discharge letter. The act of coding can be done by the health professional entering the text, or afterwards by professional coders. This task is time consuming and prone to mistakes, thus ending in possibly low quality information. Furthermore, coding is not always made for all the information. Consequently, there is increasingly more demand for accurate clinical text mining tools for extracting information from the EHR and its clinical documents.

Taking into account recent achievements in the field of NLP, the application of NLP techniques was shown to be

The associate editor coordinating the review of this manuscript and approving it for publication was Wenbing Zhao.

beneficial for the understanding of semantics in unstructured medical data [1]–[3]. In the past, many medical researchers have used NLP to solve biomedical text mining problems, such as summarizing a long paragraph like clinical notes or academic journal articles, by identifying keywords or concepts in free-form texts [2].

Recently, NLP models have become more advanced in the extraction of meanings from unstructured health data, and computers will then increasingly take on more repetitive work, which before could only be done by humans [3], [4]. In last few years NLP models has been shown to be effective for improving sentence-level tasks, which was made possible by the use of pre-trained models [5]. Those techniques use the transfer of learning and contextual word embedding models, such as ELMo, ULMFiT and BERT (Bidirectional Encoder Representations from Transformers) [5]–[7]. However, those models are primarily trained and tested on datasets containing general domain texts. Taking advantage from the transfer learning capabilities of those models, a number of biomedicine and health specific models have been also developed starting from BERT, like BioClinicalBERT [8] and BioBert [9]. For biomedical and clinical purposes, BioClinicalBERT [8] was shown to reach state of the art in some specific tasks such as medical natural language inference and two well established clinical Named Entity Recognition (NER) tasks [8].

However, even behind an apparently specific domain like biomedicine and health there is further room for specialization. For example, correct interpretation of expressions mentioning diseases and diagnoses might one area where specialized models might be useful to better accomplish some tasks. The automatic assignment of ICD codes [10] to diagnostic expressions is one of such tasks. This task is needed in two traditional applications: coding of the hospital discharge letter details, and coding of the death certificate. Furthermore, it could be also at the basis of surveillance and decision support systems. Transfer learning and contextual word embedding with bioclinical domain-specific models was shown to improve performance for such task [4]. However, general biomedical documents and diagnostic expressions have a different word distribution, and this suggest to investigate the development of models tailored on the description of diagnoses and health conditions.

In this study, we hypothesize that current models for the generic biomedical domain could be outperformed, in disease-related tasks, by models specifically trained on a disease-related corpus. For obtaining such corpus, we investigate the construction of a training set originating from the ICD-11 classification, whose terms act as seeds to obtain disease-related texts. We demonstrate that the disease-related model improves the performance of generic and biomedical BERT-based models, with a significantly smaller training set.

We like to remark that the usage of less resources, and in particular a significantly smaller training set is a direction that many researchers are trying to address in the case of deep learning models (see for example [11]).

Summarizing, the main contributions of our work are the following:

- We provide a methodology to build a pre-trained and fine-tuned domain specific BERT model.
- We experimentally validate the proposed approach against existing models and we show that our model has both a comparable or better performance than state-of-the-art models and is trained on a significantly much smaller corpora.
- We publicly release the code and the pre-trained and fine-tuned models to the research community.

II. RELATED WORKS

In the following we report related work on diseases. The International Statistical Classification of Diseases and Related Health Problems is a medical classification of diseases maintained by the World Health Organization (WHO) [10]. ICD can be defined as a system of categories to which morbid entities are assigned according to established criteria. The classification contains categories for the universe of diseases, health related conditions, and external causes of illness or death. The aim of ICD is to support the systematic recording, analysis, interpretation and comparison of mortality and morbidity data collected in different countries or areas and at different times. While most of the countries are using the 10 Revision, the 9 Revision it is still used. With its adoption by the 72th World Health Assembly in May 2019, ICD-11 (International Classification of Diseases, Revision 11) has become the forthcoming standard for coding diseases and health problems [12]. While many countries has stated its adoption, the new classification will need time for its implementation and transition from the currently used revision.

There are 26 chapters in ICD-11 plus two special sections for functioning assessment and extension codes. Table 1 shows the chapter structure in ICD-11. The codes of the ICD-11 are alphanumeric, hierarchical, and cover the range from 1A00.00 to ZZ9Z.ZZ.

Extensive work has been done on automatic coding and automatic assignment of ICD codes [13] from traditional approaches range from rule based and dictionary look ups [14] to machine learning models [15] and standard deep learning architectures such as Convolutional Neural Network (CNN) [16], Recurrent Neural Network (RNN) [17] and Long Short-Term Memory Network (LSTM) [18]. Many techniques have been proposed using deep learning and hybrid systems. Zhong and Yi use NLP and deep learning algorithms to categorize patients diseases where through comparative studies shows that CNN model achieves better performance than the RNN-based LSTM, gated recurrent unit (GRU) models and traditional machine learning model such as support vector machine (SVM) [19]. Wang *et al.* [1] employ an RNN approach for classifying diseases of ICD-10-CM based on a large free-text medical notes of hospital data, with prediction results that reach F1-score of 0.62.

TABLE 1. Chapter structure of ICD-11.

Chapter	Title
01	Certain infectious or parasitic diseases
02	Neoplasms
03	Diseases of the blood or blood-forming organs
04	Diseases of the immune system
05	Endocrine, nutritional or metabolic diseases
06	Mental, behavioural or neurodevelopmental disorders
07	Sleep–wake disorders
08	Diseases of the nervous system
09	Diseases of the visual system
10	Diseases of the ear or mastoid process
11	Diseases of the circulatory system
12	Diseases of the respiratory system
13	Diseases of the digestive system
14	Diseases of the skin
15	Diseases of the musculoskeletal system or connective tissue
16	Diseases of the genitourinary system
17	Conditions related to sexual health
18	Pregnancy, childbirth or the puerperium
19	Certain conditions originating in the perinatal period
20	Developmental anomalies
21	Symptoms, signs or clinical findings, not elsewhere classified
22	Injury, poisoning or certain other consequences of external causes
23	External causes of morbidity or mortality
24	Factors influencing health status or contact with health services
25	Codes for special purposes
26	Supplementary Chapter Traditional Medicine Conditions
V	Supplementary section for functioning assessment
X	Extension Codes

Névéol *et al.* [20] show the report on Task 1 of the 2018 CLEF eHealth evaluation lab, which extended the previous information extraction tasks of ShARe/CLEF eHealth evaluation labs for automatic coding of causes of death in death certificates where different approaches to the task were used to archive the goal. In particular, Flicoteaux uses shallow CNN and a dictionary-based lexical matching to improve its predictions for rare labels [21]. Ive *et al.* [22] chose to work with two Romance languages and treated the task as a Sequence-to-sequence (seq2seq) prediction problem [23], using an encoder-decoder architecture, with CNN based on character embeddings as encoders and RNN decoders, which allow the model to generalise across two genealogically related languages. Seva *et al.* [24] approach builds on two recurrent neural networks models, to extract and classify causes of death from the death certificates for the CLEF eHealth 2018 Task 1. The first model is LSTM-based sequence-to-sequence model and it is used to obtain a death cause from each death certificate line, then a bidirectional LSTM model with attention mechanism is used to predict the respective ICD-10 codes to the received death cause description. Atutxa *et al.* [25] also uses sequence-to-sequence framework with promising results, well above the average results for the task.

Xie and Xing [26] propose a tree-of-sequences LSTM architecture to simultaneously capture the hierarchical relationship among codes and the semantics of each code using the MIMIC-III dataset for the evaluations.

To better utilize information buried in longer input sentence, Baumel *et al.* [27] have proposed a technique based on a Hierarchical Attention bidirectional Gated Recurrent Unit (HA-GRU) architecture.

Word-level vector representations has been well established within the NLP community. Unsupervised methods such as word2vec [28] and GloVe [29] can express all possible meanings of a word as a single vector representation but fail to incorporate wider context into account, in learning representations of words. Over the last years, there have been several approaches developed to learn unsupervised encoders that produce contextualized word embedding which have been shown to substantially improve performance on many NLP tasks. ELMo [6] and BERT [5] derived from pre-trained bidirectional language models (biLMs) and presents strong solutions. ELMo was pre-trained by a vast text corpus as a language model and is able to create a context-sensitive embedding for each word in a given sentence, which will be fed into downstream tasks. BERT uses a similar approach as ELMo, but is deeper and contains more parameters, which make the model possess greater representative power. Compared to ELMo, BERT can be incorporated into a downstream task and get fine-tuned as an integrated task-specific architecture instead of simply providing word embeddings as features. Such models, that result from the fine-tuning of the pre-trained ones on a down-stream supervised task, have been shown to achieve better results with minimal effort when compared to CNN and RNN based methods [30]. BERT has also been proven to be better to ELMo and far superior to non-contextual embeddings in a variety of tasks [31]. After the release of BERT, multiple models which use a transformer based architecture have been released. Lample and Conneau [32] developed XLM, an extension of the pre-training of language models capable of working on multiple languages, Yang *et al.* [33] developed XLNET, making use of an autoregressive pre-training method to overcome the limitations of previous models, Liu *et al.* developed RoBERTa [34], a model that optimizes the training of BERT by relying on large mini batches, Sanh *et al.* [35] developed DistilBERT, a distilled version of BERT capable or retaining 97% of the original effectiveness while producing a model which is 40% smaller than the original one.

Inspired by the success in other domains, BERT has been widely used in biomedical studies and tasks [30], [36], [37]. Della Mea *et al.* [4] employ BERT and other biomedical BERT models (BioBert [9], BioClinicalBERT [8]) to identify the underlying cause of death from death certificates, by reversing coded certificates back to text from the ICD-10 title of the code. Shang *et al.* [38] proposed G-BERT, a BERT style model for medicine recommendation by learning embeddings for ICD codes. Sänger *et al.* [39] use BERT and BioBERT as base models for ICD code prediction. Transformer based architectures have led to a large increase in performance on biomedical/clinical tasks. BioBERT has been pre-trained by [9] with the same structure as BERT including a biomedical corpora. In [8] a BERT model is fine-tuned on MIMIC III notes and discharge summaries and apply to downstream tasks. Zhang *et al.* [40] trained a BERT model (BERT-XML) from scratch on EHR notes, creating a more suitable vocabulary for EHR tasks adapting

the BERT architecture for ICD coding with multi-label attention.

The main differences between our model and existing ones if that our model is trained on a different, domain specific, and much smaller set of data with respect to state-of-the-art models. Such difference lead to significant advantages, as detailed in the following.

III. METHODS

A. PRE-TRAINING CORPORA

As a general purpose language representation model, BERT is pre-trained on English text corpora, mainly on Wikipedia and BooksCorpus [5]. Specialized models are pre-trained using a similar English text corpora with the addition of a considerable number of domain specific texts; for example, many models have been specialized for medical purposes, such as BioBERT or BioClinicalBERT which include clinical and medical corpora gathered from Pubmed [41] and PMC Full-text articles. Those domain specific models have substantially increased the effectiveness of NLP models in biomedical text mining tasks [9].

In this work, we follow the same approach, with some specific choice. To train our model on disease-related texts, we decided to start the construction of a suitable corpora from the ICD-11 classification, which for the most part deals with diseases and health conditions. Using ICD-11 as a stem, we then collected text documents from other well known sources like PubMed and Wikipedia. Details on the process are described in the rest of this section, while Figure 1 summarizes the process.

The ICD corpus used was collected from the latest WHO revision available in September 2020. The classification contains 34,679 entities, where each entity is described with a title, none or multiple index terms (31,164 entities contain at least one) and can contain a description, when the title is not self-describing (7,383 entities contain a short or detailed description). ICD-11 index terms are synonyms and narrower terms of the entity.

Since only 21.29% of the entities has an associated description, we decided to enrich the dataset with other sources, which we choose to be Wikipedia and PubMed.

The corpora include ICD-11 entities (title, description and index terms), a sub-set of PubMed records(title and abstract), and Wikipedia articles (title and summary). The PubMed documents were collected using the official API, by searching for ICD-11 titles and index terms.

We have selected 20 articles of PubMed for each ICD-11 entity title and index term related to the article title, abstract, or Medical Subject Headings (MeSH) Terms. The PubMed documents were collected using the official API, by searching for ICD-11 titles and index terms. The Wikipedia documents were collected from the wiki API, again by searching for ICD-11 titles and index terms.

After the collection we have cleaned the dataset by removing the duplicate articles (the latter issue due to the fact that some searches may have returned some common document). Then we have cleaned the text of the articles by removing some elements such as markdown syntax (URLs and text formatting), tables and lists, converting the information in plain text.

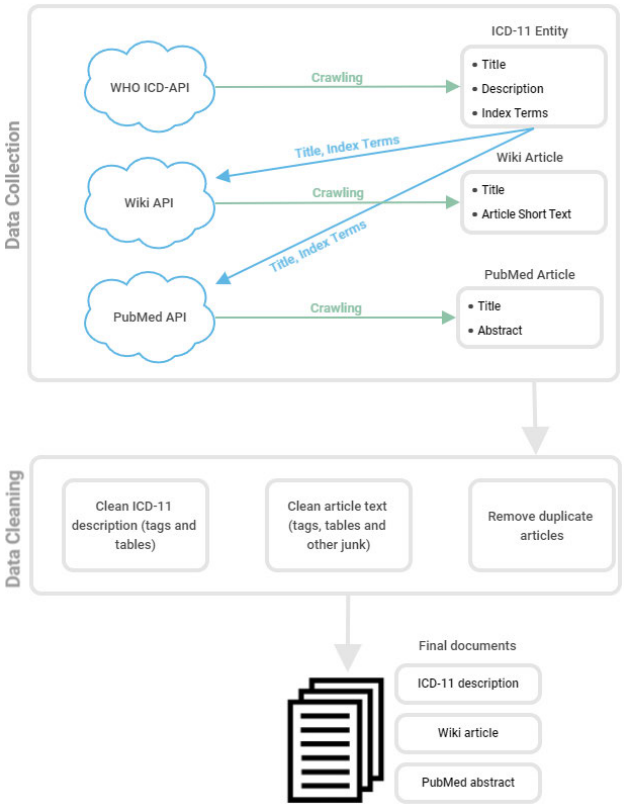

FIGURE 1. Schema for the dataset construction process.

In Table 2 the text corpora with the total number of words and number of documents are listed.

TABLE 2. Corpora composition.

Source	Documents	Words
ICD-11 descriptions	34,676	1.0 million
PubMed Title and Abstracts	852,550	184.6 million
Wikipedia pages	37,074	6.1 million

B. PRE-TRAINING SETUP

The text of the various corpora are pre-processed to obtain a new set of text entries suitable for the pre-training of the model. The final textual documents contain:

- ICD-11 descriptions: the title of each entry, followed by the list of its index terms (if present) and the description of the entry (if present);
- PubMed entities: the name of the ICD-11 entity related to the PubMed article (either an ICD-11 title or index-term), the title of the PubMed article, and the abstract of the article;

- Wikipedia articles: the name of the ICD-11 entity related to the Wikipedia article (which usually coincides with the title of the article) and the text of the Wikipedia article.

Figure 2 shows an example of the documents used, with an indication on the origin of each part of the text. Note that if a PubMed (or Wikipedia) article is associated with more than one ICD-11 entity, it will appear just once in the final pre-training corpus. This was chosen because in this case the text samples shown to the model would only differ slightly and this would not impact the model positively, as the objective of the pre-training is to learn the embeddings of the words by seeing them in different contexts.

ICD-11	
Title	Respiratory failure.
Index Terms	Respiratory failure; lung failure NOS; pulmonary failure.
Description	Respiratory failure is a life-threatening impairment of oxygenation or CO2 elimination. Respiratory failure may occur because of impaired [. . .]
PubMed	
ICD-11 Entity	Respiratory failure.
Title	Respiratory characteristics and related intraoperative ventilatory management for patients with COVID-19 pneumonia.
Abstract	A substantial proportion of patients with coronavirus disease 19 (COVID-19) develop severe respiratory failure. [. . .]
Wikipedia	
ICD-11 Entity	Respiratory failure.
Text	Respiratory failure results from inadequate gas exchange by the respiratory system, meaning that the arterial oxygen, carbon dioxide or both cannot be kept at normal levels. [. . .]

FIGURE 2. Examples of text entries of the final corpus coming from the three original sources: ICD-11 descriptions, PubMed abstracts, and Wikipedia pages.

BERT-based models usually employ WordPiece tokenization [42], which allows the tokenizer vocabulary to be composed of both words and sub-word units. The creation of the vocabulary is dependent on the training corpora, and happens through the following steps: (1) define an upper bound D to the dimension of the vocabulary; (2) initialize the vocabulary with one entry for each character in the corpus; (3) for all the possible pairs of tokens already in the vocabulary, calculate which pair appears more frequently in the corpus; (4) create a new entry in the vocabulary by merging the two tokens found at step 3; (5) repeat steps 3 and 4 until the vocabulary reaches size D (or the vocabulary contains all the words in the corpus).

Since the pre-training of the model was performed entirely from scratch, the first step is to create the vocabulary itself. Our aim is to create a model which is as comparable as possible with the original BERT (base) model and its derivatives, so we choose the same maximum dimension (30,522 tokens). The vocabulary is initialized with all the characters in the pre-training corpus and the original BERT special tokens (`[CLS]`, `[PAD]`, `[SEP]`, `[UNK]`, `[MASK]`) which are needed for the pre-training and fine-tuning procedures. It is then fit on the whole pre-training corpus, resulting in the creation of 30,522 unique tokens.

The architecture of the BERT model has the same characteristics of the original BERT base model: it is composed of 12 stacked BERT encoders, with 12 attention heads; the hidden size of the model (and the embedding dimension) is 768; the maximum input sequence length is 512 tokens.

The pre-training is performed using a Masked Language Modeling task, which is commonly used for the pre-training of BERT-based models. The objective is to create context sensitive embeddings for each word in the vocabulary. In order to do so, the model is given one sentence at the time and each word in the sentence has a 15% probability to become masked (replaced with the `[MASK]` token). The model calculates the output (contextual) embeddings using the 12 stacked encoder with multiple-head attention. In the case of the masked words, the output embedding is completely dependent on the other words in the sentence since the `[MASK]` token provides no information. After the embedding process has taken place, the model has to guess the original value of the masked word, based on its embedding alone. The MLM loss measures the degree of confidence and accuracy with which the masked words can be reconstructed.

The model was pre-trained for 50 epochs on 90% of the final corpus (172.5 million words), with a batch size of 32. The remaining 10% of the documents was used as evaluation set to test the performance of the model on unseen documents. For a quick comparison we report here the training setup of the original BERT model: circa 50 epochs, on a 3.3 billion words corpus, with a batch size of 256 [5]. The pre-training task is summarized in Figure 3. The pre-training was carried out on a the Google-cloud infrastructure with a machine having 16 vcpu, 128 GB vram and a Nvidia V100 GPU.

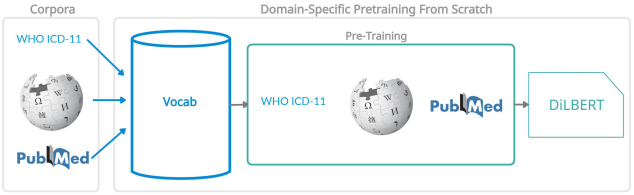

FIGURE 3. Schema for the pre-training of DiLBERT.

C. FINE-TUNING TASK AND DATASET

The death certificates dataset used for the fine-tuning tasks are provided by the U.S. National Center for Health Statistics [43], making them available for statistical and analytical research. The dataset is composed by a total of 12,919,268 records for the years 2014-2017. We used a subset of the available data, composed of 500,000 records: 400,000 for

TABLE 3. State-of-the-art accuracy scores of the selected models (from [4]).

Model	Accuracy
XLNET	97.46%
XLM	97.32%
BioClinicalBERT	97.14%
RoBERTa	97.10%
BioBERT	97.10%
BERT	97.05%
DistilBERT	96.96%

training and validation; 100,000 for testing. The task is framed as multi-class classification: the objective is to predict the main reason of death (among a finite set of possible answers) given some information about the patient. The source of information that were used are: sex, age, and conditions appearing on both Part 1 and Part 2 of the certificate. Since the certificates were not containing the textual information, we reversed the codes to the corresponding text. The reverse coding consists in substituting the ICD codes with their narrative title and by an additional encoding which dealt with writing in an explicit form for the administrative data (e.g., Female, 39y old). On the contrary, the CLEF dataset (CepiDC Causes of Death Corpus) comes from "Task 1: Multilingual Information Extraction - ICD-10 coding" [44] which already includes the textual descriptions of the causes of death together with the coded conditions. For this dataset we have used the plain text conditions directly, plus we have encoded the administrative data as for the U.S. National Center for Health Statistics certificates.

The train and test samples were extracted from all the records after we applied random shuffling and stratified sampling (either by year or by code). The selected dataset was used for a direct comparison with the work of [4], where different Transformer models are fine-tuned and tested on this dataset, leading to the current state-of-the-art results (see Table 3). In particular, we compare the performance of DiLBERT with the following language models: XLNET [33], XLM [32], RoBERTa [34], BioClinicalBERT [8], BioBERT [9], BERT (base uncased) [5], DistilBERT [35].

In order to train the models for the classification fine-tuning task, we use the corresponding variant of the BERT architecture. Recall that BERT-based models take a sequence of tokens as input, and output a sequence of embeddings. A linear layer is added on top of the first embedding, which is typically the embedding of the special `[CLS]` (classification) token. The linear layer projects the embedding to a vector with one position for each of output class. This can be easily converted into a probability distribution over the possible causes of death.

As regards our language model, we fine-tune several checkpoints which were saved at different epochs during pre-training, to investigate the how the performance on the downstream task varies depending on the number of pre-training epochs. This is also useful to evaluate which of the checkpoints is the best-performing one. We can expect the best checkpoint to be one of the last 10 (epoch 40-50) for the model pre-trained on the whole dataset, as these are the epochs where the two losses converge in Figure 4.

D. FINE-TUNING SETUP

The text of the documents is processed as follows: lower-casing; tokenization using the appropriate WordPiece tokenizer; addition of the special tokens `[CLS]` (at the beginning of the sequence) and `[SEP]` (at the end); padding of each sequence to 256 tokens (using the `[PAD]` token). The cause of death (output label) is numericalized, and the linear layer of the new architecture is initialized according to the number of unique output labels found in the 400,000 training samples.

The model is fine-tuned for 4 epochs (which is the standard amount of epochs needed for most fine-tuning tasks [5]) on the 400,000 training samples, with a batch size of 16, cross-entropy loss, Adam optimizer, learning rate $3e{-}5$.

At the end of the training, the model is used to predict the main cause of death of each of the 100,000 test samples. Performances are evaluated using accuracy, both to compare with state-of-the-art results and because we are interested in identifying the correct code.

The fine-tuning was carried out on a server infrastructure equipped with an Intel Xeon E5-1620 CPU, 64 GB ram and a Nvidia Titan Xp GPU.

IV. RESULTS

A. PRE-TRAINING RESULTS

In this section we discuss the results obtained by the Masked Language Modeling pre-training, which has the objective of learning in-domain embeddings effectively. The most straightforward way to evaluate the success of this procedure is to monitor the trend of the MLM loss across the epochs, which measures the error of the model in reconstructing the original words of the texts. Monitoring the MLM loss on a validation set allows us to check that the knowledge of the learned embeddings can be transferred to another set of unseen documents.

Figure 4 shows the development of the loss during the pre-training. The x-axis shows the epoch number (up to 50 epochs) and the y-axis reports the value of the Masked Language Modeling loss. The two series show the values of the loss on the training set (90% of the corpus) and the validation set (10% of the corpus). We can see that the pre-training loss decreases smoothly stabilizing around epoch 40-50. The two losses converge around epoch 45, showing no signs of over-fit and forecasting the creation of embeddings with a good degree of generalization. We save one version of the pre-trained model (checkpoint) every epoch to further evaluate the quality of the embeddings on downstream tasks.

We perform a second pre-training procedure to gain further insight on the interaction between the composition of the

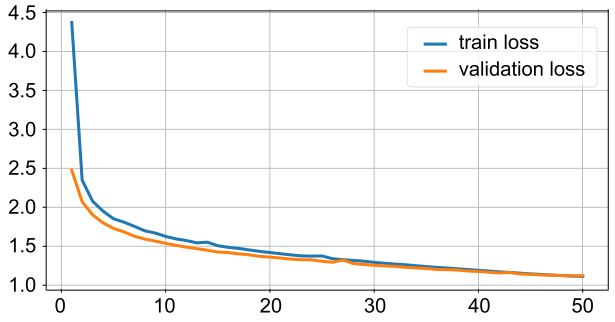

FIGURE 4. Plot of the training an evaluation loss during pre-training. Pre-training is performed using documents from all three data sources: ICD-11, PubMed and Wikipedia.

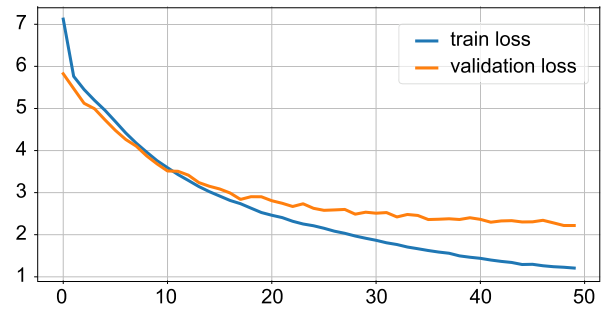

FIGURE 5. Training an evaluation loss during pre-training. Pre-training is performed using documents from one source: ICD-11 only.

corpus and the quality of the embeddings. We pre-train the same architecture using the ICD-11 descriptions only. The plot in Figure 5 shows that the loss decreases more slowly, and that the model shows signs of over-fit starting from epoch 10-15: the evaluation and train losses diverge noticeably. Furthermore, the validation loss settles at a value over 2.0, which is noticeably higher than the value reached by the first pre-training procedure in Figure 4. This should signify that these embeddings are of a lesser quality, as they show lower generalization capabilities. As in the previous case, we save the language models generated by the ICD-11-only pre-training at every epoch to evaluate them on downstream tasks and compare them with the other embeddings.

B. FINE-TUNING RESULTS

The plots in Figures 6, 7, 8 compare the fine-tuning results of our DiLBERT at varying pre-training steps against various Transformer baselines with and without in-domain medical pre-training. The Transformer baselines are the same state-of-the-art models tested in [4], for which we replicated the results. Summarizing, we are testing our approach with the current state-of-the-art models (both in terms of training data and effectiveness) on the considered task. The x-axis represents the number of pre-training epochs of the DiLBERT checkpoint, while the y-axis reports the classification accuracy on the 100,000 test samples. The performance of the baselines is indicated by the horizontal dashed lines, while

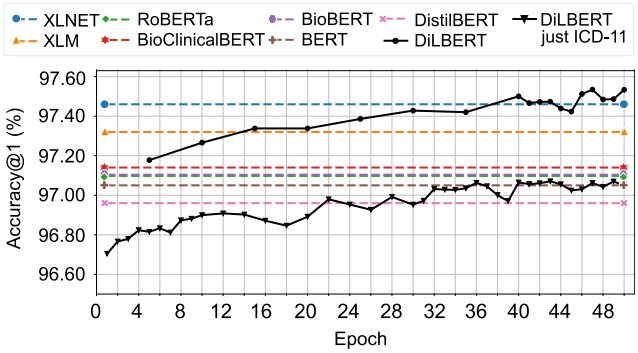

FIGURE 6. Results for the finetuned models, tested on 100K instances, stratified by year.

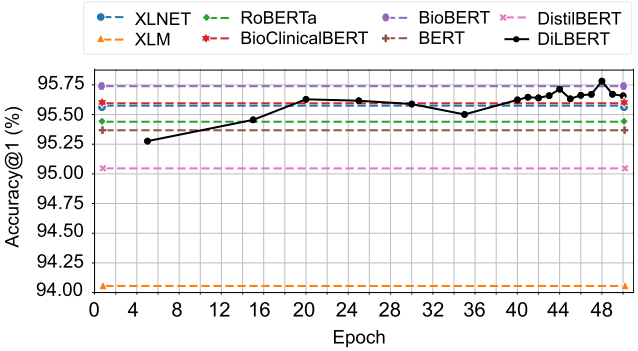

FIGURE 7. Results for the finetuned models, tested on 100K instances, stratified by code.

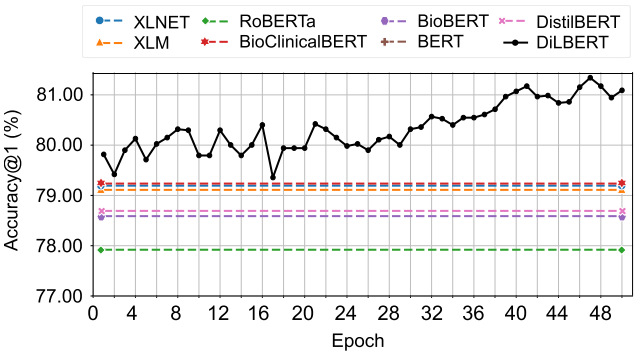

FIGURE 8. Results for the finetuned models, tested on the CLEF dataset.

the accuracy of DiLBERT is represented by the black solid line.

In general, our custom language models matches or surpasses the other Transformer variants on all the fine-tuning tasks, despite it being trained on way smaller corpus (20 times smaller than the one used to train the original BERT).

Figure 6 shows the results on the data stratified by year (which is the original stratification used in [4]) and contains two series of results for DiLBERT: one for the model trained on the whole dataset (black circles) and one for the model trained solely on the ICD-11 documents (black triangles). As forecasted by the pre-training results, the accuracy of the

ICD-11-only model is clearly lower than the one of the full DiLBERT model at all epochs, proving that its embeddings are of a lesser quality. We can also observe that its accuracy stabilizes after epoch 30, which is roughly the point at which the validation loss stabilized in Figure 5.

DiLBERT trained on ICD-11-only reaches a maximum accuracy of 97.06% at the end of its pre-training (epochs 43-49), obtaining the same performance as the basic BERT model. As regards the full DiLBERT, we see that its best checkpoints (47 and 50) achieve an accuracy of 97.53%, surpassing XLNET (97.46%) and all the other baseline models. Its accuracy seem to peak and stabilize between epoch 40 and 50, which is the point where the pre-training losses converged and stabilized (see Figure 4). This confirms that the last 10 checkpoints are the most interesting ones to investigate.

We can also note that the language model pretrained solely on ICD-11 texts achieve performances which are comparable with the ones of a BERT base uncased, despite it being trained on a 1 million words corpus instead of a 3.3 billion words one (over 3,000 times smaller). This goes to show the great data efficiency that in-domain pre-training has in topics with highly specialized language, and how it is still a relevant and powerful alternative to general-domain pre-training.

We then tested all the models on the data stratified by code (Figure 7). In general, the performance of all the models drops by at least 2 percentage points. While this small drop in effectiveness can be caused by statistical fluctuation and thus be not significant, it can be also caused by the dataset composition: the least used codes in the training set are those more difficult to correctly identify. The order of the best-performing baselines gets disrupted quite significantly, too (e.g. XLM goes from second place to last place, while BioBERT and BioClinicalBERT go from mid to top-tier). Despite that, DiLBERT maintains a stable performance, and its checkpoints 44-50 match the results of the two best baselines (BioBERT 95.74% and BioClinicalBERT 95.59%).

Finally, we tested all models on the CLEF dataset. All baseline models experience a further drop in accuracy (about 15 percentage points). This is due because instances of the CLEF dataset are real ones, and thus include typographic errors, synonyms and abbreviations, etc. The order of the best-performing baseline Transformers changes once again, and at this point the only ones which showed a consistently good performance on all datasets are XLNET and BioClinicalBERT. In this case our language models greatly outperforms all the other baselines, obtaining results which are always above the accuracy of BioClinicalBERT (79.24%), and reaching its absolute best performance at epoch 47, with 81.32% accuracy.

Summarizing, our DiLBERT model consistently matched or outperformed the other baseline language models, which were pre-trained either on general-purpose or in-domain corpora containing approximately 20 times more data. Any of the last 10-5 checkpoints seems to be a good candidate to produce high-quality embeddings for NLP tasks which revolve around ICD-11 terminology. Furthermore, it is interesting to notice that even when pre-trained on ICD-11 documents only (3,000 times less data), DiLBERT still matches the performance of a the general-purpose BERT model. This shows the critical role of in-domain pre-training in this particular domain.

V. DISCUSSION

As we can derive from the experimental results, we found that our model outperforms the state-of-the-art models on the fine-tuning task where we can have direct comparability (i.e., cause of death identification with different stratification, and on the CLEF dataset). Regarding the other tasks, we obtain comparable or better performance, but we cannot state the superiority of the model because of the different datasets.

Despite the difference in the accuracy scores is not always striking, we remark that our model is trained on a corpora which is at least 20 times smaller than the one used to train the rest of the algorithms. Furthermore, the algorithm variant trained solely on the ICD-11 text, despite being trained on a corpora 3,000 times smaller, achieves similar performance as the ones obtained with the base version of BERT.

To summarize the practical implications of our work, our contribution is the following. We pre-trained and fine-tuned a domain specific BERT model which both has comparable or better performance than state-of-the-art models and is trained on a much smaller corpora, thus with more limited training time and data. This may also simplify the development of models for languages different from English, due to the minor quantity of data needed for training. However, still the issue of finding something equivalent to Pubmed but in other languages remains. On the other side, the model trained only on ICD-11 contents achieved acceptable performance; ICD-11 is expected to be translated to a lot of languages, due to its mandatory usage for mortality coding. This, together with the exploitation of national Wikipedia articles could help in enabling disease-specific NLP also in less widespread languages. Among the foreseen future work there is the generation of a model for Spanish, since this is one of the languages in which ICD-11 has already been translated.

One limitation of our work is that it has not been tested on tasks related to classifications or terminologies different from ICD, e.g., SNOMED-CT that has too an important role in EHR coding. However, in SNOMED-CT diseases are only a small portion of the overall represented entities, and as far as we know, there aren't available datasets with SNOMED-CT encoded diagnostic expressions. However, as soon as datasets become available, we plan to experiment our approach also on them.

Furthermore, to correctly compare with previously validated models, in the current paper we stick with the one stage finetuning process. However, alternatives such as a two steps finetuning process are possible [45]–[49]; such approaches are out of scope for the current work, as they might introduce another set of effects and complications to the training of our model. For this reason, we plan to address the two steps finetuning process of our model in future work.

For the same set of reasons, we leave for future work also the application and comparison of deep learning models compression strategies to our model, as well as the in depth-comparison between pruning and compressing strategies of larger models and a data-centric approach [50], [51].

The code used to perform the experiments and our models can be found at: https://github.com/KevinRoitero/dilbert.

ACKNOWLEDGMENT

The authors thank Robert N. Anderson at the Mortality Statistics Branch, U.S. National Center for Health Statistics, for having provided the death certificates data set. They also thank the anonymous reviewers for providing insightful comments which helped to improve the overall quality of the article.

REFERENCES

[1] S.-M. Wang, Y.-H. Chang, L.-C. Kuo, F. Lai, Y.-N. Chen, F.-Y. Yu, C.-W. Chen, Z.-W. Li, and Y. Chung, "Using deep learning for automatic ICD-10 classification from free-text data," *Eur. J. Biomed. Informat.*, vol. 16, no. 1, pp. 1–10, 2020.

[2] R. Pivovarov and N. Elhadad, "Automated methods for the summarization of electronic health records," *J. Amer. Med. Inform. Assoc.*, vol. 22, no. 5, pp. 938–947, Sep. 2015.

[3] D. J. Feller, J. Zucker, M. T. Yin, P. Gordon, and N. Elhadad, "Using clinical notes and natural language processing for automated HIV risk assessment," *J. Acquired Immune Deficiency Syndromes*, vol. 77, no. 2, pp. 160–166, 2018.

[4] V. D. Mea, M. H. Popescu, and K. Roitero, "Underlying cause of death identification from death certificates using reverse coding to text and a NLP based deep learning approach," *Informat. Med. Unlocked*, vol. 21, Jan. 2020, Art. no. 100456. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352914820306067

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2019, *arXiv:1810.04805*.

[6] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," 2018, *arXiv:1802.05365*.

[7] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," 2018, *arXiv:1801.06146*.

[8] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and B. A. M McDermott, "Publicly available clinical BERT embeddings," 2019, *arXiv:1904.03323*.

[9] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Sep. 2019, doi: 10.1093/bioinformatics/btz682.

[10] World Health Organization. (2016). *International Statistical Classification of Diseases and Related Health Problems*. Accessed: Feb. 9, 2021. [Online]. Available: https://icd.who.int/browse10/Content/statichtml/ICD10Volume2_en_2016.pdf

[11] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big," in *Proc. ACM Conf. Fairness, Accountability, Transparency*, Mar. 2021, pp. 610–623.

[12] World Health Organization. (2019). *International Statistical Classification of Diseases and Related Health Problems*. Accessed: Feb. 9, 2021. [Online]. Available: https://icd.who.int/icd11refguide/en/index.html

[13] K. Crammer, M. Dredze, K. Ganchev, P. P. Talukdar, and S. Carroll, "Automatic code assignment to medical text," in *Proc. Workshop BioNLP Biol., Transl., Clin. Lang. Process.*, 2007, pp. 129–136.

[14] R. Bounaama and A. Mohammed El Amine, "Tlemcen university at CELF eHealth 2018 team techno: Multilingual information extraction-ICD10 coding," in *Proc. CLEF*, Sep. 2018, pp. 1–8.

[15] J. Gobeill and P. Ruch, "Instance-based learning for ICD10 categorization," in *Proc. Conf. Labs Eval. Forum*, vol. 2125, L. Cappellato, N. Ferro, J. Nie, and L. Soulier, Eds., Avignon, France, Sep. 2018, pp. 1–7. [Online]. Available: http://ceur-ws.org/Vol-2125/paper_149.pdf

[16] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 649–657.

[17] K. Cho, B. Van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," 2014, *arXiv:1406.1078*.

[18] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[19] J. Zhong and X. Yi, "Categorizing patient disease into ICD-10 with deep learning for semantic text classification," in *Recent Trends in Computational Intelligence*, A. Sadollah and T. S. Sinha, Eds. Rijeka, Croatia: IntechOpen, 2020, ch. 6, doi: 10.5772/intechopen.91292.

[20] A. Névéol, A. Robert, F. Grippo, C. Morgand, C. Orsi, L. Pelikan, L. Ramadier, G. Rey, and P. Zweigenbaum, "CLEF eHealth 2018 multilingual information extraction task overview: ICD10 coding of death certificates in French, Hungarian and Italian," in *Proc. CLEF*, 2018, pp. 1–18.

[21] R. Flicoteaux, "ECSTRA-APHP CLEF eHealth2018-task 1: ICD10 code extraction from death certificates," in *Proc. CLEF*, 2018, pp. 1–7.

[22] J. Ive, N. Viani, D. Chandran, A. Bittar, and S. Velupillai, "KCL-health-NLP CLEF eHealth 2018 task 1: ICD-10 coding of French and Italian death certificates with character-level convolutional neural networks," in *Proc. CLEF*, 2018, pp. 1–13.

[23] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," 2014, *arXiv:1409.3215*.

[24] J. Seva, M. Sänger, and U. Leser, "WBI at CLEF eHealth 2018 task 1: Language-independent ICD-10 coding using multi-lingual embeddings and recurrent neural networks," in *Proc. CLEF*, 2018, pp. 1–4.

[25] A. Atutxa, A. Casillas, N. Ezeiza, V. Fresno-Fernández, I. Goenaga, K. Gojenola, R. Martínez, M. O. Anchordoqui, and O. P. De Vinaspre, "IxaMed at CLEF eHealth 2018 task 1: ICD10 coding with a sequence-to-sequence approach," in *Proc. CLEF*, 2018, pp. 1–9.

[26] P. Xie and E. Xing, "A neural architecture for automated ICD coding," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, Melbourne, VIC, Australia, Jul. 2018, pp. 1066–1076. [Online]. Available: https://www.aclweb.org/anthology/P18-1098

[27] T. Baumel, J. Nassour-Kassis, R. Cohen, M. Elhadad, and N. Elhadad, "Multi-label classification of patient notes: Case study on ICD code assignment," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2017, pp. 409–416.

[28] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 135–146, Dec. 2017. [Online]. Available: https://www.aclweb.org/anthology/Q17-1010

[29] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Doha, Qatar, Oct. 2014, pp. 1532–1543. [Online]. Available: https://www.aclweb.org/anthology/D14-1162

[30] S. Amin, G. Neumann, K. Dunfield, A. Vechkaeva, K. Chapman, and M. Wixted, "MLT-DFKI at CLEF eHealth 2019: Multi-label classification of ICD-10 codes with BERT," in *Proc. CLEF*, Sep. 2019, pp. 1–15.

[31] Y. Si, J. Wang, H. Xu, and K. Roberts, "Enhancing clinical concept extraction with contextual embeddings," *J. Amer. Med. Inform. Assoc.*, vol. 26, no. 11, pp. 1297–1304, Jul. 2019, doi: 10.1093/jamia/ocz096.

[32] G. Lample and A. Conneau, "Cross-lingual language model pretraining," 2019, *arXiv:1901.07291*.

[33] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," 2019, *arXiv:1906.08237*.

[34] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[35] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," 2019, *arXiv:1910.01108*.

[36] F. Li, Y. Jin, W. Liu, B. P. S. Rawat, P. Cai, and H. Yu, "Fine-tuning bidirectional encoder representations from transformers (BERT)–based models on large-scale electronic health record notes: An empirical study," *JMIR Med. Inf.*, vol. 7, no. 3, Sep. 2019, Art. no. e14830.

[37] H. Liu, Y. Perl, and J. Geller, "Concept placement using BERT trained by transforming and summarizing biomedical ontology structure," *J. Biomed. Informat.*, vol. 112, Dec. 2020, Art. no. 103607.

[38] J. Shang, T. Ma, C. Xiao, and J. Sun, "Pre-training of graph augmented transformers for medication recommendation," 2019, *arXiv:1906.00346*.

[39] M. Sänger, L. Weber, M. Kittner, and U. Leser, "Classifying German animal experiment summaries with multi-lingual BERT at CLEF eHealth 2019 task 1," in *Proc. CLEF*, Aug. 2019, pp. 1–12.
[40] Z. Zhang, J. Liu, and N. Razavian, "BERT-XML: Large scale automated ICD coding using BERT pretraining," in *Proc. 3rd Clin. Natural Lang. Process. Workshop*, Nov. 2020, pp. 24–34. [Online]. Available: https://www.aclweb.org/anthology/2020.clinicalnlp-1.3
[41] K. Canese and S. Weis, "PUBMED: The bibliographic database," in *The NCBI Handbook*. Bethesda, MD, USA: National Center for Biotechnology Information, 2013.
[42] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, and W. Macherey, "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016, *arXiv:1609.08144*.
[43] National Center for Health Statistics. (2011). *Public Birth, Period Linked Birth—Infant Death, Birth Cohort Linked Birth—Infant Death, Mortality Multiple Cause, and Fetal Death Data Files are Available for Independent Research and Analyses*. Accessed: Apr. 23, 2021. [Online]. Available: https://www.cdc.gov/nchs/data_access/vitalstatsonline.htm
[44] A. Névéol, R. N. Anderson, K. B. Cohen, C. Grouin, T. Lavergne, G. Rey, A. Robert, and P. Zweigenbaum, "CLEF eHealth 2017 multilingual information extraction task overview: ICD10 coding of death certificates in English and French," in *Proc. CLEF eHealth Eval.*, Dublin, Ireland, Sep. 2017, pp. 1–17.
[45] Y. Hao, L. Dong, F. Wei, and K. Xu, "Visualizing and understanding the effectiveness of BERT," 2019, *arXiv:1908.05620*.
[46] M. Mosbach, M. Andriushchenko, and D. Klakow, "On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines," 2020, *arXiv:2006.04884*.
[47] A. Rogers, O. Kovaleva, and A. Rumshisky, "A primer in bertology: What we know about how BERT works," *Trans. Assoc. Comput. Linguistics*, vol. 8, pp. 842–866, Dec. 2020.
[48] Y. Pruksachatkun, J. Phang, H. Liu, P. M. Htut, X. Zhang, R. Y. Pang, C. Vania, K. Kann, and S. R. Bowman, "Intermediate-task transfer learning with pretrained models for natural language understanding: When and why does it work?" 2020, *arXiv:2005.00628*.
[49] L. Zhang and Y. Hu, "A fine-tuning approach research of pre-trained model with two stage," in *Proc. IEEE Int. Conf. Power Electron., Comput. Appl. (ICPECA)*, Jan. 2021, pp. 905–908.
[50] T. Chen, J. Frankle, S. Chang, S. Liu, Y. Zhang, Z. Wang, and M. Carbin, "The lottery ticket hypothesis for pre-trained BERT networks," 2020, *arXiv:2007.12223*.
[51] M. A. Gordon, K. Duh, and N. Andrews, "Compressing BERT: Studying the effects of weight pruning on transfer learning," 2020, *arXiv:2002.08307*.

KEVIN ROITERO is currently a Postdoctoral Research Fellow with the University of Udine, Italy. He visited and collaborated with multiple universities, where he worked on the development of crowdsourcing tasks with the aim of understanding and predicting user features, such as user engagement, user agreement, and bias. His research interests include information retrieval evaluation, crowdsourcing, data mining and analysis, machine learning, and statistical modeling.

BEATRICE PORTELLI is currently pursuing the Ph.D. degree with the University of Udine, Italy. She works employing deep learning and natural language processing techniques. Her current research interests include the study and development of fact verification models and the automatic extraction of adverse drug events from social media texts.

MIHAI HORIA POPESCU is currently pursuing the Ph.D. degree with the University of Udine, Italy. He collaborated with multiple organizations in the domain of mortality and morbidity, where he worked on the development of tools supporting automated coding. His research interests include explainable artificial intelligence, natural language processing, computer vision, data mining and analysis, machine learning, and statistical modeling over medical domain.

VINCENZO DELLA MEA is currently an Associate Professor of medical informatics and advanced web technologies with the University of Udine, Italy. He is also the Head of the Medical Informatics, Telemedicine and eHealth Lab (MITEL). He was a National Delegate for the COST Action "EUROTELEPATH" (from 2008 to 2011) and a local responsible for the EU Marie Curie Project "AIDPATH" (from 2013 to 2017). He also participated in other national projects. He served as the WHO Informatics and Terminologies Committee Co-Chair for four years in the WHO Network of the Family of International Classifications. In the same network, he was also a member of the Joint Task Force on ICD-11 until the end of mission. He is also a member of the ICHI Task Force (International Classification of Health Interventions).