Journal of Healthcare Informatics Research (2023) 7:433–446
https://doi.org/10.1007/s41666-023-00140-7

RESEARCH ARTICLE

BioBERTurk: Exploring Turkish Biomedical Language Model Development Strategies in Low-Resource Setting

Hazal Türkmen[1] · Oğuz Dikenelli[1] · Cenk Eraslan[2] · Mehmet Cem Çallı[2] · Süha Süreyya Özbek[2]

Received: 14 October 2022 / Revised: 6 March 2023 / Accepted: 28 July 2023 /
Published online: 19 September 2023

Abstract

Pretrained language models augmented with in-domain corpora show impressive results in biomedicine and clinical Natural Language Processing (NLP) tasks in English. However, there has been minimal work in low-resource languages. Although some pioneering works have shown promising results, many scenarios still need to be explored to engineer effective pretrained language models in biomedicine for low-resource settings. This study introduces the BioBERTurk family and four pretrained models in Turkish for biomedicine. To evaluate the models, we also introduced a labeled dataset to classify radiology reports of head CT examinations. Two parts of the reports, impressions and findings, are evaluated separately to observe the performance of models on longer and less informative text. We compared the models with the Turkish BERT (BERTurk) pretrained with general domain text, multilingual BERT (mBERT), and LSTM+attention-based baseline models. The first model initialized from BERTurk and then further pretrained with biomedical corpus performs statistically better than BERTurk, multilingual BERT, and baseline for both datasets. The second model continues to pretrain the BERTurk model by using only radiology Ph.D. theses to test the effect of task-related text. This model slightly outperformed all models on the impression dataset and showed that using only radiology-related data for continual pre-training could be effective. The third model continues to pretrain by adding radiology theses to the biomedical corpus but does not show a statistically meaningful difference for both datasets. The final model combines radiology and biomedicine corpora with the corpus of BERTurk and pretrains a BERT model from scratch. This model is the worst-performing model of the BioBERT family, even worse than BERTurk and multilingual BERT.

✉ Hazal Türkmen
hazal.turkmen@ege.edu.tr

Extended author information available on the last page of the article

Keywords Biomedicine · Pretrained language model · Transformer · Radiology reports

1 Introduction

After the impressive performance of Bidirectional Encoder Representations from Transformers (BERT) [1] in several downstream Natural Language Processing (NLP) tasks, the use of pretrained language models has become the standard engineering approach for NLP systems. These models were trained on public domain corpora, such as Wikipedia and the Book Corpus, to ensure sufficient generality. A natural research question is whether the use of domain-specific text corpora improves the performance of these models in domain-specific tasks. Biomedicine is the domain most likely to benefit from resources, such as PubMed, and MIMIC III provides readily available, high-volume, and high-quality data for generating such models. A recent survey found 13 models based on PubMed, 12 based on MIMIC, and 16 different models using private data in other languages [2], highlighting the critical engineering decisions that need to be made when analyzing the proposed pretrained language models for the biomedicine domain: the pre-training approach and the corpus selection for pre-training.

Continual pre-training is the first approach attempted in the literature to create domain-specific models. In this process, a new model is initialized from an existing model, such as BERT, and is then further pretrained using the domain-specific corpus. BioBERT [3] is the first model to demonstrate the effectiveness of continuous pre-training. It was initialized from the general BERT version and further trained on PubMed abstracts and full-text articles. The inclusion of PubMed data through continual pre-training improved the performance over BERT for all tasks (Named Entity Recognition (NER), Relation Extraction, and Question Answering) in 15 open biomedical datasets. Clinical BERT [4] is another work that evaluates continual pre-training in different settings. The authors used all MIMIC notes, discharge summaries, and continued pre-training, initializing from both general BERT and BioBERT. The results showed that versions initialized from BioBERT performed better in three out of five clinical tasks than BERT and BioBERT, with very similar performances in the remaining tasks. In other words, the use of MIMIC data via continual pre-training improved performance in clinical tasks.

An alternative approach to contiunal pre-training is pre-training from scratch. PubMedBERT [5] evaluated this approach by pre-training a BERT model and creating vocabulary from scratch using PubMed abstracts. They created a new benchmark that included a set of biomedical NLP tasks from publicly available datasets. Although the results were close and slightly better than BioBERT and significantly better than ClinicalBERT, it should not be concluded that creating a model from scratch always yields better performance. For instance, the study presented FS-BERT, a BERT model built from scratch using 3.8 million unstructured radiology reports in German [6]. However, it performed worse than RAD-BERT, which was initialized from the general German BERT and continued pre-training using the corpus of FS-BERT. These

results suggest that the comprehensiveness of the domain data plays a critical role in pre-training from scratch.

While selecting a pre-training approach is important, the success of a new domain-specific model also largely depends on the careful selection of the corpus used for pre-training. The primary strategy for corpus selection is to integrate general domain knowledge with in-domain knowledge, which is achieved through continual pre-training by transferring model weights. BioBERT, ClinicalBERT, and BlueBERT [7] are examples of combining in-domain corpora with a general model via continual pre-training. All models performed better than the baseline models. However, the relevance of the added in-domain corpus to the task domain seems to affect the performance. For example, adding MIMIC data for pre-training resulted in better performance in clinical tasks, as demonstrated in ClinicalBERT and BlueBERT. The use of in-domain data is another alternative for corpus selection. As evidenced by PubMedBERT, if there is a large, comprehensive, and quality dataset like PubMed in a domain, using it exclusively to generate vocabulary and model can be effective.

Although pre-training BERT models is a common approach that can enhance performance in various biomedical NLP tasks, it requires substantial domain-specific data for pre-training. Biomedical unlabeled text data is not as readily available as general-domain data, and in some languages, it cannot be sourced from a single repository, such as the PubMed database. Consequently, there may be situations where a low-resource setting is encountered with only a small in-domain corpus available and insufficient pre-training data to train a language model. In such cases, one solution is to mix it with a general domain corpus via continual pre-training. BioBERTpt [8] evaluated this situation in Portuguese using a small corpus that included clinical notes and abstracts of scientific papers. It performed slightly better than both multilingual BERT and Portuguese BERT in two NER tasks. The researchers also examined the impact of using only clinical data versus abstracts and found that both approaches led to a slight improvement in performance. ABioNER [9] demonstrated similar results for Arabic, which was initialized from a general Arabic BERT and further pretrained with a small biomedical corpus.

To the best of our knowledge, only a single study exists on Turkish biomedical text classification [10]. In this study, the authors utilized the existing Turkish BERT (BERTurk) [11] and multilingual BERT (mBERT)[1] to classify Turkish medical abstracts into disease categories. The primary objective of our study is to develop pre-trained language models for the biomedical domain, which is distinct from previous study. These models can potentially enhance the performance of various biomedical applications. We present four pretrained language models for the Turkish biomedicine domain and investigate the impact of different corpus selection and pre-training techniques. Owing to limited resources for collecting the Turkish biomedical corpus, our corpora provide a constrained resource for language model training. We also created a labeled dataset to classify Head Computed Tomography (CT) radiology reports to evaluate these models. The main contributions of this study are as follows:

- We compiled two in-domain corpora by collecting open full-text Turkish scientific papers in biomedicine and theses on radiology. We then constructed four

[1] https://github.com/google-research/bert/blob/master/multilingual.md.

domain-specific pretrained language models using these corpora, and publicly released both corpora and models for the first time in Turkish.
- We introduced a text classification task for head CT radiology reports in Turkish for the first time and evaluated two different parts of the reports, *impressions* and *findings*. To the best of our knowledge, this work is the first to evaluate the performance of pretrained language models in Turkish clinical texts.
- The existing literature has demonstrated the beneficial impact of task-specific corpora on model performance. Similar results were observed in the Turkish biomedical corpus. We also evaluated the effect of pre-training with the radiology theses corpus and the pre-training-from-scratch approach on the radiology report classification task for the first time.

2 Materials and Method

This section provides detailed information about the four pretrained language models developed in this study, as well as the characteristics of the domain-specific corpora used to generate these models. The first model, BioBERTurk$_{con}$(+trM), uses only Turkish biomedical text and applies continual pre-training approach, initializing weights from the publicly available general Turkish BERTurk [11]. This model was designed to test the hypothesis that the use of a biomedical corpus via continual pre-training enhances the performance of biomedical and clinical tasks, and whether this holds true for the Turkish language. The second model, known as BioBERTurk$_{con}$(trR), used only the radiology theses corpus for continual pre-training. This was done to understand the impact of using a task-related corpus more comprehensively. Furthermore, the third model, BioBERTurk$_{con}$(+trM+trR), incorporated a corpus that includes radiology theses along with Turkish biomedical text. This model evaluates the impact of a task-related corpus on model performance via continual pre-training. In the model names, "trM" and "trR" refer to biomedical and radiology theses corpora, respectively. Finally, we trained a BERT model from scratch, called BioBERTurk$_{sc}$(+trW+trM+trR), to evaluate the pre-training from scratch approach in a low-resource setting. This model utilized a mixed corpus comprising the collected Turkish biomedical and radiology theses corpora and a general domain corpus. For a fair comparison, we used the same general domain corpus on which BERTurk was trained for the pre-training from scratch approach. We have made the models and Turkish biomedical corpora available in a public Github repository.

2.1 Building Domain-Specific Corpora

The initial step in developing BioBERTurk$_{con}$(+trM) is to gather text data in the biomedicine domain. Owing to the limited availability of Turkish abstracts in PubMed, we had to find alternative resources to build a corpus of meaningful size. We turned to Dergipark,[2] a platform developed and managed by Ulakbim[3] (Turkish Academic

[2] www.dergipark.com.tr.

[3] https://ulakbim.tubitak.gov.tr/

Table 1 Corpora statistics

Corpus	N. tokens	Size (GB)	Source	Domain
(trW)Turkish Web Corpus	4,404,976,662	35	sources like Wikipedia etc	General
(trM)Turkish Medical articles	60,318,554	0,48	www.dergipark.com.tr	Biomedical
(trR)Turkish Radiology thesis	15,268,779	0,11	www.tez.yok.gov.tr	Radiology

Network and Information Center), which serves as a gateway to periodic refereed journals. To assemble the corpus, we built a crawler application to visit all biomedical journals in Dergipark and collect all full-text PDF articles published in those journals. Following the collection of these PDF documents, we scraped the data based on heuristic rules similar to those used by ABioNER [9]. For instance, one rule identifies essential sections of articles, such as the starting section should be "özet" (abstract), and the ending section should be "referanslar" (references) in Turkish articles. However, defining all the rules for extracting text from unstructured PDFs is challenging and time consuming. After retrieving the necessary text, we applied a cleaning pipeline with custom steps to the raw text data. First, we combined all data into one large text file with one sentence per line. Next, we aggressively processed the files using language detection and hand-written heuristic rules. These rules identified suspicious patterns, such as too high a ratio of digits or punctuations, non-Turkish alphabet characters, or low average token numbers. Finally, to avoid repetitive content, the remaining corpora were deduplicated.

The second corpus was created to assess the effects of the task-related texts. Given that we used a classification task for head CT radiology reports, we searched for open-domain text in radiology. The Turkish Council of Higher Education provides a website[4] for searching and accessing all open Ph.D. theses. We filtered all theses conducted in the radiology departments of the medical schools. We then combined all the collected theses and applied the cleaning pipeline, thereby building a corpus on radiology. To the best of our knowledge, this is the first attempt to use Ph.D. theses as a task-related domain corpus in pre-training. The statistics of the final pre-training data produced during the cleaning steps are summarized in Table 1.

2.2 Analyzing Domain Similarity

Before embarking on the pre-training of our BERT models, our goal was to gauge the similarity between our BERT domain and the target task domains. The domain similarities were assessed by computing the ratio of intersections among their respective vocabularies. The underlying assumption of this approach is that the quantity of vocabulary words shared between domains should provide a measure of their similarity [12]. We considered the most frequently used 10k unigram as the domain vocabulary after excluding stopwords, punctuations, and numbers. We also utilized 100k sentences from random document samples in each BERT domain corpus to generate vocabularies. For the task vocabulary, we relied on 50k radiology report impressions given their

[4] www.tez.yok.gov.tr/

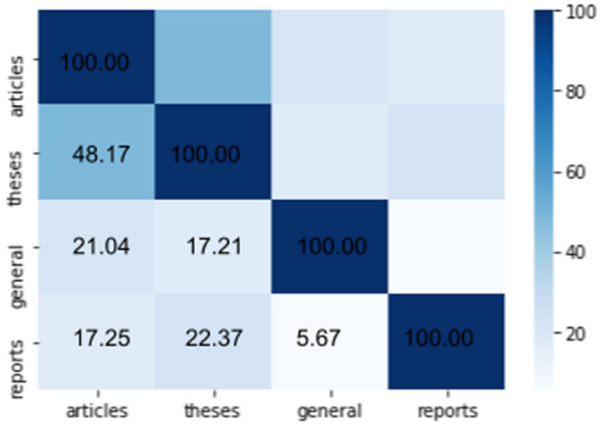

Fig. 1 Vocabulary overlap ratio (%) between domains

brevity. Figure 1 illustrates the ratio of the shared vocabulary between domains. The calculated measures revealed a substantial overlap of vocabulary between the domains of Turkish medical articles and Turkish radiology theses. Although the article domain is broader than the thesis domain, they are similar because they have a similar tenor. We also noted that the target domain shares the greatest similarity with these domains (%22.37) because of their mutual focus on radiology, whereas it shares the least similarity with the general domain (%5.67). Thus, the assembled thesis corpus appears to be more appropriate (%22.37) for studying the effects of task-related corpus usage in pretrained language model development, even if it does not entirely share the same tenor as the reports.

2.3 Data Preprocessing

Turkish is a morphologically rich language with unique characteristics, owing to its agglutinative structure. Turkish's rich morphology can generate words with many different meanings from a single root. From an NLP perspective, this linguistic feature leads to a high rate of out-of-vocabulary (OOV) problems and reduces training accuracy. Wordpiece tokenization is a powerful approach for mitigating the challenging OOV problem, and has been proven to provide the highest performance in several Turkish NLP tasks [13].

Given this context, we adopted the Wordpiece vocabulary from BERTurk to preprocess the inputs for both the pre-training and fine-tuning of BioBERTurk$_{con}$. Conversely, we constructed a new Wordpiece vocabulary for preprocessing BioBERTurk$_{con}$. We also used the tokenizer library from HuggingFace[5] to build an uncased vocabulary and set the vocabulary size at 32k to align with the size defined in the BERTurk configuration file. We then used the official *create_pre-training_data.py* script provided by the Google AI Research team to convert all raw BERT inputs into structured TensorFlow examples.

[5] https://huggingface.co/docs/tokenizers/python/latest/

2.4 Pretraining Process

We conducted a series of experiments using two pre-training approaches for our models. BioBERTurk$_{sc}$ was pretrained from scratch using mixed corpora, whereas BioBERTurk$_{con}$ variants were initialized with a TensorFlow version of the BERTurk checkpoints to continue pre-training. To train our BERT variants, we followed the same procedure as the BERTurk training. Each model was trained for 1 M steps, with a maximum sequence length of 512 and a batch size of 128. We used Adam with a learning rate of 1e-4 and warming up for 10K steps. We trained all models using open-source training scripts available in the official BERT Github repository, utilizing V3 TPUs with eight cores from Google Cloud Compute Services.

2.5 Model Baseline

Following [1], we implemented a fully connected layer on top of BERT for classification tasks. We also established a baseline model, as presented in [14] to provide a comparative analysis of classification performance. This model was used to classify radiology reports of head CT examinations in a non-English language (Hebrew), mirroring the task used in our experiments. We selected the best-performing model, which achieved a classification accuracy of 90.8, and performed significantly better than both Logistic Regression and Gradient Boosting. The model incorporated an LSTM layer stacked with an attention layer and a fully connected layer on top. It accepted as input the Word2Vec embedding derived from our Turkish biomedical corpora. We refer to this baseline model as the LSTM-attn-wvc.

3 Experiments

3.1 Classification of Radiology Reports

Turkish is a low-resource language that lacks a labeled clinical dataset for constructing NLP tasks. To evaluate our models within the context of a text classification task at the document level, we created two datasets based on different sections of radiology reports: one containing *findings* and the other containing *impressions*.

To curate the radiology datasets, we used an in-house corpus of 45,304 de-identified Turkish CT head radiology examinations for patients aged 8 years and above from the neurology and emergency departments at Ege University Hospital, Turkey. The reports cover the period January 2016 to June 2018. The same individual's report was used to separate the *findings* and *impressions* datasets. Prior to the data analysis, we filtered out texts with fewer than 300 characters and removed newlines and domain-specific encodings. The final dataset included 5514 reports. Following the preparation of the dataset, we performed an annotation process to create the text classification task. The annotation schema incorporated three classes: presence of intracranial pathology (abnormal), no intracranial pathology (normal), and out of series. These were used to indicate the presence or absence of intracranial pathology. The annotation process

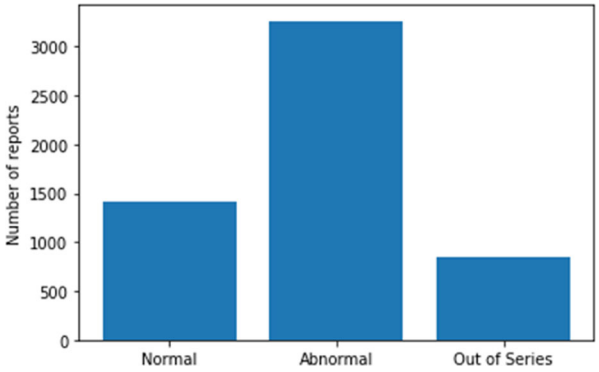

Fig. 2 Dataset class distribution

was conducted in three phases by three radiologists (C. E., M. C. C., and S. S. O.), each with years of experience in radiology reporting. In each phase, two annotators (C. E. and M. C. C.) independently labeled a subset of the reports. Subsequently, a third annotator (S. S. O.) reviewed these annotations to identify conflicting ones. At the end of each phase, all three annotators reached consensus by creating fully agreed annotations. The annotation task was performed using a spreadsheet file to facilitate the annotators' work. We then divided the final annotated dataset into two datasets, *impressions* and *findings*, for separate evaluation. Our *impressions* dataset comprises 28,704 sentences and 13,892 tokens; meanwhile, our *findings* dataset contains 78,939 sentences and 17,348 tokens. The *findings* part of a report typically includes longer texts, which allows us to assess the performance of models with more extended texts. The annotated datasets were then randomly divided into test (%10), validation (%10), and training (%80) sets for fine tuning. The class distributions of the two datasets are the same, as illustrated in Fig. 2. As shown in Fig. 2, the datasets exhibited an unbalanced distribution, which is a common characteristic of text processing in the radiology domain [15].

3.2 Experimental Setup

The pretrained model fine-tuning was done using the same architecture and optimization method as in [1]. For each model, we performed hyperparameters searches for learning rate values ϵ {2e-4, 3e-5, 5e-5}, max sequence length ϵ {128, 256, 512}, batch size ϵ {16, 32} and the number of the training epoch ϵ {3, 4, 5}. Batch size 64 was not utilized due to memory limitations. Adam optimizer also was employed in all experiments. Fine-tuning was executed with NVIDIA Quadro RTX 8000 graphic cards, and each experiment took approximately 10 min.

For the baseline, we used 200 dimensional word2vec vectors as stated in the study [14]. The vectors were also trained by the Gensim framework using the CBOW architecture [16]. We evaluated a range of parameter combinations for our baseline model, selecting the maximum length for 128, batch size for 16, and training for 25 epochs.

3.3 Evaluation Criteria

The performance of the models was evaluated using precision, recall, and F1-score. For a more detailed analysis, the performance of each class was evaluated separately. In addition to the precision, recall, and F1-score, a *t*-test [17] was conducted to determine whether there were any statistically significant differences between the models. A threshold of 0.05 was used to determine whether the results were statistically significant.

4 Experimental Results

We conducted experiments on the *impressions* and *findings* datasets separately. All scores were presented under the optimal hyperparameter settings for each model. Table 2 presents the average F1-scores over ten runs for the *impressions* dataset. According to the results, all the BERT variants significantly outperformed the baseline model (LSTM-attn-wvc). Moreover, our in-domain model BioBERTurk$_{con}$+(trR) achieved a statistically higher F1-score than BioBERTurk$_{con}$+(trM+trR), BERTurk, BioBERTurk$_{sc}$, and multilingual BERT (mBERT). Although BioBERTurk$_{con}$+(trR) performed better than BioBERTurk$_{con}$+(trM), there was no statistical difference between the models (*P* value 0.59). We also compared all Turkish BERT models with mBERT to measure the effect of language on document level text classification in radiology reports. Although some studies have shown that mBERT performs more robustly than monolingual BERT models for certain tasks [8, 18], our study shows that BioBERTurk$_{con}$ variants and BERTurk classfied Turkish radiology reports more accurately. For a detailed analysis, per class F1-scores are also reported in Table 3. While the highest F1-score for the "normal" and "out of series" classes was obtained by BioBERTurk$_{con}$+(trR), the "abnormal" class had the highest score with BioBERTurk$_{con}$+(trM). We also observed that the top-performing model,

Table 2 Precision, recall, and F1-score of radiology report classification experiments based on *impressions* test set

Model	Precision	Recall	F1-score	*P*-value compared to BioBERTurk$_{con}$ *+trR(c)*
BERTurk *+trW(c)*[1]	91.88%	91.87%	91.86%	2.46E-05*
BioBERTurk$_{con}$ *+trM(c)*	93.00%	93.02%	92.99%	0.59
BioBERTurk$_{con}$ *+(trM+trR)(c)*	92.74%	92.77%	92.75%	0.001*
BioBERTurk$_{con}$ *+trR(c)*	**93.13%**	**93.14%**	**93.13%**	
BioBERTurk$_{sc}$ *+(trW+trM+trR)(u)*[2]	89.52%	89.51%	89.48%	1.77E-11*
mBERT(c)	91.45%	91.43%	91.42%	9.12E-07*
LSTM-attn-wvc	80.80%	82.00%	80.72%	1.04E-17*

The best scores are in bold
[1] refers cased model
[2]refers uncased model
*indicates a statistically significant difference ($p < 0.05$)

Table 3 Per-label F1-score on *impressions* test set

Model	Normal	Abnormal	Out of series
BERTurk *+trW(c)*	94.24%	90.61%	85.39%
BioBERTurk$_{con}$ *+trM(c)*	94.97%	**93.02%**	85.91%
BioBERTurk$_{con}$ *+(trM+trR)(c)*	94.80%	92.84%	85.29%
BioBERTurk$_{con}$ *+trR(c)*	**95.11%**	92.17%	**87.59%**
BioBERTurk$_{sc}$ *+(trW+trM+trR)(u)*[2]	92.13%	89.33%	80.33%
mBERT(c)	93.63%	91.06%	84.15%
LSTM-attn-wvc	88.40%	84.71%	47.75%

The best scores are in bold

BioBERTurk$_{con}$+(trR), obtained higher precision and recall than the other models (Table 2).

Table 4 displays the average F1-scores over ten runs for the *findings* dataset. The initial clear observation from these experiments is that all models perform less effectively on the *findings* data. Similar results were observed in English [19], which is expected because *findings* are longer and less informative regarding classification than *impressions*. In the *findings* dataset, BioBERTurk$_{con}$+(trM) delivered the highest F1-score of 89.97%, followed by BioBERTurk$_{con}$+(trM+trR) (P value 0.02), with no statistically significant difference, and all BERT variants significantly outperformed our baseline model. Upon examining the other metrics from Table 5, it's evident that the BioBERTurk$_{con}$+(trM) model performed highly effectively for the "normal" class but surprisingly not as well for the "abnormal" and "out of series" classes.

5 Discussions

By broadly evaluating the experiments, we can deduce several conclusions from our study. First, our results demonstrate that all BioBERTurk$_{con}$ variants yield better results

Table 4 Precision, recall, and F1-score of radiology report classification experiments based on *findings* test set

Model	Precision	Recall	F1-score	P-value compared to BioBERTurk$_{con}$ *+trM(c)*
BERTurk *+trW(c)*[1]	89.00%	88.55%	88.60%	1.97E-11*
BioBERTurk$_{con}$ *+(trM)*	**90.34%**	**89.98%**	**89.97%**	
BioBERTurk$_{con}$ *+(trM+trR)(c)*	88.93%	89.35%	89.38%	0.02
BioBERTurk$_{con}$ *+trR(c)*	88.61%	88.76%	88.75%	1.28E-9*
LSTM-attn-wvc	82.49%	83.01%	82.61%	1.87E-15*

The best scores are in bold
[1]refers cased model
*indicates a statistically significant difference ($p < 0.05$)

Table 5 Per-label F1-score on *findings* test set

Model	Normal	Abnormal	Out of series
BERTurk *+trW(c)*	89.55%	91.57%	76.22%
BioBERTurk$_{con}$ *+(trM)*	**92.75%**	91.17%	77.94%
BioBERTurk$_{con}$ *+(trM+trR)(c)*	89.85%	**92.25%**	**78.17%**
BioBERTurk$_{con}$ *+trR(c)*	89.17%	91.88%	76.69%
LSTM-attn-wvc	84.71%	88.49%	57.83%

The best scores are in bold

in both datasets than the existing generic BERT model and the traditional baseline model. This aligns with observations made in English, where in-domain models outperform generic ones [3, 4]. However, in our case, continuing pre-training with a rather small in-domain corpus compared to the generic corpus still proved to be highly effective in clinical text classification. We observed similar results with the medical articles and radiology theses corpora, despite these corpora containing noisier data than the PubMed abstracts.

Another pivotal observation is the impact of these corpora on continual pre-training. The theses corpus is significantly smaller compared to the medical article's corpus (0.11 GB vs 0.48 GB). The BioBERTurk$_{con}$+(trR) model, which was continuously trained solely with the theses corpus, outperformed the other models. These results demonstrate that a corpus slightly similar to the task domain can be exceptionally effective, even with small-sized and noisy text data. This model outperformed the other models in classifying the out of series and "normal" labels in the impression dataset. When the theses corpus was combined with medical articles, the resulting model (BioBERTurk$_{con}$+(trM+trR)) performed efficiently, especially in classifying the Abnormal and Out of Series labels of the *findings* dataset. Thus, we can conclude that continual pre-training with small task-related data led to improved accuracy for low-frequency label (Out of Series) classification in Turkish radiology reports.

Finally, we compare the results of BioBERTurk$_{sc}$ with those of other models to investigate the pre-training technique. Our model, BioBERTurk$_{sc}$, provides poor classification accuracy for both datasets, except for our baseline model. Therefore, combining a very small domain data with large generic data is not an effective approach, at least not in the creation of Turkish domain-oriented pretrained models from scratch. In light of these results, we demonstrate that if the target domain is dramatically different from the general domain (the similarity of the Turkish general domain and Turkish clinical domain is 9%), using task-related data for continual pre-training can enhance classification performance.

There has also been a previous study that applied mBERT, which has significant zero-shot cross-lingual transfer abilities for low-resource languages [20]. When we compare the F1 score of mBERT with other models, our monolingual models developed using continual pre-training outperform in the Turkish clinical task. In summary, the success of our in-domain models shows that continual pre-training of biomedical

articles can improve model performance on a clinical task in Turkish, even when available language resources are restricted.

Finally, and most importantly, we introduce the first Turkish biomedical resources and make them available to the NLP community.

Our study has several limitations. Since there are no NLP-shared tasks in Turkish for the medical domain, we evaluated our in-domain models for a single clinical task in Turkish. Secondly, we reported character sizes longer than 512, which exceeds the input size limit required by the BERT model.

6 Conclusion

In this study, we introduced the BioBERTurk family-four pretrained biomedical language models-and evaluated them for classifying Turkish radiology reports. Our work demonstrates that further pre-training models with a small-scale radiology corpus, specifically our domain-specific BioBERTurk$_{con}$ variant, outperforms out-of-the-box BERT embeddings in classifying Turkish radiology reports. In future work, we aim to investigate different pre-training and fine-tuning approaches in low-resource settings for clinical Turkish domains, and we plan to evaluate our model for[6] different tasks in clinical NLP.

Acknowledgements We would like to acknowledge the TPU Research Cloud program (TRC) and the Google's CURe program in providing access to TPUv3 units and GCP credits, respectively.

Author Contribution S.S.O., C.E, O.D, and H.T were responsible for the study design and writing of the manuscript. H.T implemented algorithms and conducted data analysis. C.E., M.C.C, and S.S.O. labeled dataset.

Funding The study was supported by the TPU Research Cloud program (TRC) and the Google's CURe program.

Data Availability The models and public datasets are available in our GitHub repository: https://github.com/hazalturkmen/BioBERTurk. The models also were developed using Tensorflow 2 library with Keras and Pytorch library. The source code can be found on GitHub https://github.com/hazalturkmen/RADBERTurk

Declarations

Ethics Approval The study was approved by the Ege University Ethical Committee under study number UH150040389 and conducted in accordance with the Declaration of Helsinki.

Consent to Participate Consent to participate was waived as data were anonymized and collected retrospectively.

Consent for Publication Consent to publish was waived as data were anonymized and collected retrospectively.

Conflicts of Interest The authors declare no competing interests.

[6] https://sites.research.google/trc/about/

References

1. Devlin J, Chang M-W, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805
2. Kalyan KS, Rajasekharan A, Sangeetha S (2021) Ammu: a survey of transformer-based biomedical pretrained language models. J Biomed Inform 103982
3. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J (2020) Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36(4):1234–1240
4. Alsentzer E, Murphy JR, Boag W, Weng W-H, Jin D, Naumann T, McDermott M (2019) Publicly available clinical bert embeddings. arXiv preprint arXiv:1904.03323
5. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, Naumann T, Gao J, Poon H (2021) Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH) 3(1):1–23
6. Bressem KK, Adams LC, Gaudin RA, Tröltzsch D, Hamm B, Makowski MR, Schüle C-Y, Vahldiek JL, Niehues SM (2020) Highly accurate classification of chest radiographic reports using a deep learning natural language model pre-trained on 3.8 million text reports. Bioinformatics 36(21):5255–5261
7. Peng Y, Yan S, Lu Z (2019) Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets. arXiv preprint arXiv:1906.05474
8. Schneider ETR, Souza JVA, Knafou J, Oliveira LES, Copara J, Gumiel YB, Oliveira LFA, Paraiso EC, Teodoro D, Barra CMCM (2020) Biobertpt-a portuguese neural language model for clinical named entity recognition. In: Proceedings of the 3rd Clinical Natural Language Processing Workshop, pp 65–72
9. Boudjellal N, Zhang H, Khan A, Ahmad A, Naseem R, Shang J, Dai L (2021) Abioner: a bert-based model for arabic biomedical named-entity recognition. Complexity 2021
10. Çelıkten A, Bulut H (2021) Turkish medical text classification using bert. In: 2021 29th Signal Processing and Communications Applications Conference (SIU), pp 1–4. IEEE
11. Schweter S (2020) BERTurk - BERT models for Turkish. Zenodo. https://doi.org/10.5281/zenodo.3770924
12. Dai X, Karimi S, Hachey B, Paris C (2019) Using similarity measures to select pretraining data for ner. arXiv preprint arXiv:1904.00585
13. Toraman C, Yilmaz EH, Şahinuç F, Ozcelik O (2022) Impact of tokenization on language models: an analysis for turkish. arXiv preprint arXiv:2204.08832
14. Barash Y, Guralnik G, Tau N, Soffer S, Levy T, Shimon O, Zimlichman E, Konen E, Klang E (2020) Comparison of deep learning models for natural language processing-based classification of non-english head ct reports. Neuroradiology 62(10):1247–1256
15. Qu W, Balki I, Mendez M, Valen J, Levman J, Tyrrell PN (2020) Assessing and mitigating the effects of class imbalance in machine learning with application to x-ray imaging. Int J CARS 15(12):2041–2048
16. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781
17. Dietterich TG (1998) Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput 10(7):1895–1923
18. Muller B, Elazar Y, Sagot B, Seddah D (2021) First align, then predict: understanding the cross-lingual ability of multilingual bert. arXiv preprint arXiv:2101.11109
19. Gundogdu B, Pamuksuz U, Chung JH, Telleria JM, Liu P, Khan F, Chang PJ (2021) Customized impression prediction from radiology reports using bert and lstms. IEEE Transactions on Artificial Intelligence
20. Wu S, Dredze M (2019) Beto, bentz, becas: the surprising cross-lingual effectiveness of bert. arXiv preprint arXiv:1904.09077

Authors and Affiliations

Hazal Türkmen[1] · Oğuz Dikenelli[1] · Cenk Eraslan[2] · Mehmet Cem Çallı[2] · Süha Süreyya Özbek[2]

Oğuz Dikenelli
oguz.dikenelli@ege.edu.tr

Cenk Eraslan
cenk.eraslan@ege.edu.tr

Mehmet Cem Çallı
cem.calli@ege.edu.tr

Süha Süreyya Özbek
sureyya.ozbek@ege.edu.tr

1 Department of Computer Engineering, Ege University, 35100 İzmir, Turkey

2 Department of Radiology, Ege University, 35100 İzmir, Turkey