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

Hazal Turkmen (hazal.turkmen@ege.edu.tr)
Ege University
Oguz Dikenelli
Ege University
Cenk Eraslan
Ege University
Mehmet Cem Callı
Ege University

Research Article

Keywords: biomedicine, pretrained language model, transformer, radiology reports

Posted Date: October 21st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2165226/v1

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

Hazal Turkmen[1*], Oguz Dikenelli[1], Cenk Eraslan[2], Mehmet Cem Callı[2] and Suha Sureyya Ozbek[2]

[1*]Computer Engineering, Faculty of Engineering, Ege University, İzmir, Turkey.
[2]Radiology, Faculty of Medicine, Ege University, İzmir, Turkey.

*Corresponding author(s). E-mail(s): hazal.turkmen@ege.edu.tr;
Contributing authors: oguz.dikenelli@ege.edu.tr;
cenk.eraslan@ege.edu.tr; cem.calli@ege.edu.tr;
sureyya.ozbek@ege.edu.tr;

Abstract

Pretrained language models elevated with in-domain corpora show impressive results in biomedicine and clinical NLP tasks in English. However, there is minimal work in low resource languages. Although some pioneering works show promising results, many scenarios still need to be explored to engineer effective pretrained language models in biomedicine for low resource settings. This work introduces the BioBERTurk family, four pretrained models in Turkish for biomedicine. To evaluate models, we also introduce a labeled dataset to classify radiology reports of head CT exams. Two different parts of the reports, impressions, and findings, are evaluated separately to observe the performance of models on longer and less informative text. We compare models with the Turkish BERT-BERTurk pretained with general domain text, multilingual BERT, and an LSTM+attention-based baseline model. 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 BERTurk model by using only radiology Ph.D. theses to test the effect of the task-related text. This model slightly outperforms all models on the impressions dataset and

showed that using only radiology-related data for continual pretraining could be effective. The third model continues to pretrain by adding radiology theses to biomedical corpus but does not show a statistically meaningful difference. The final model combines radiology and biomedicine corpora with the corpus of BERTurk and pretrained a BERT model from scratch. This model is the worst performed model of the BioBERT family, even worse than BERTurk and multilingual BERT.

Keywords: biomedicine, pretrained language model, transformer, radiology reports

1 Introduction

After the impressive performance of BERT[1] in several downstream NLP tasks, the usage of pretrained language models became the standard engineering approach for NLP systems. These models are trained on public domain corpora, like Wikipedia and Book Corpus, to make them general enough. One natural following research question is whether the usage of domain text corpora improves the performance of these models in domain tasks. Biomedicine is one of the most likely domain since resources like Pubmed and MIMIC III provides ready-to-use, high volume, and quality data for generating such models. Hence, a recent survey found 13 models based on Pubmed, 12 based on MIMIC, and 16 different ones using private data in other languages [2].

Two engineering decisions seem critical when analyzing the proposed pretrained language models for the biomedicine domain: pretraining approach and corpus selection for pretraining.

Continual pretraining is the first tried pretraining approach in the literature to create domain-specific models. The new model is initialized from an existing one like BERT and continued pretraining using the domain-specific corpus in continual pretraining. BioBERT [3] is the first model that shows the effectiveness of continual pretraining. It was initialized from the general BERT version and continued training on Pubmed abstracts and full-text articles. Including Pubmed data with continuous pretraning improved performance over BERT for all tasks (Named entity recognition, relation extraction, and question answering) in 15 open biomedical datasets. Clinical BERT [4] is another work that evaluates continual pretraining in different settings. The authors used all MIMIC notes and only discharge summaries and continued to pretrain them initializing from the general BERT and BioBERT. Results showed that versions initiated from BioBERT perform better in 3 of 5 clinical tasks than BERT and BioBERT and are very similar to the rest of the tasks. In other words, using the MIMIC data via continual pretraining improves the performance in clinical tasks. An alternative approach to continual pretraining is pretraining from scratch. PubMedBERT [5] evaluated this approach by pretraining a BERT model and creating the vocabulary from scratch using

Pubmed abstracts. They created a new benchmark that includes a set of biomedical NLP tasks from publicly available datasets for evaluation. Results were close but slightly better than BioBERT, and significantly better than ClinicalBERT. But this does not mean that creating a model from scratch always performs better. For example, FS-BERT is a BERT model built from scratch using 3.8 million unstructured radiology reports in German [6]. But it performs worse than RAD-BERT, which is initialized from general German BERT and continued pretraining using the corpus of FS-BERT. So, it seems that how comprehensive the domain data is also critical for pretraining from scratch.

Although the selection of pretraining approach can be critical, corpus selection is crucial for the success of the new domain-specific model. The primary strategy for corpus selection is to mix general domain knowledge with in-domain knowledge. Continual pretraining applies this by transferring model weights. BioBERT, ClinicalBERT, and BlueBERT [7] are examples of mixing in-domain corpus with the general one via continual pretraining. They all perform better than their baseline model. But the relevance of added in-domain corpus with the task domain seems to affect the performance. For example, adding MIMIC data for pretraining performs better in clinical tasks, as shown in ClinicalBERT and BlueBERT. Using only in-domain data is the other alternative for corpus selection. PubMedBERT proves that if you have large size, comprehensive, and quality data like PubMed in a domain, using only it to generate vocabulary and model can be effective. So, suppose you have a small in-domain corpus, which is the case for low resource languages. In that case, the only alternative is to mix it with general domain corpus via continual pretraning. BioBERTpt [8] evaluates this situation in Portuguese using a small corpus that includes clinical notes and abstracts of scientific papers. It performed slightly better than multilingual BERT and Portuguese BERT in two NER tasks. They also observed the effect of only clinical data and only abstracts. Both cases slightly improved the performance. ABioNER [9] showed similar results for Arabic, which is initialized from general Arabic BERT and pretrained with a small biomedical corpus.

There is only one Turkish biomedical text classification study in the literature [10]. In this study, the authors used existing Turkish BERT (BERTurk) [11] and multilingual BERT (mBERT)[1] to classify Turkish medical abstracts into disease categories. Our purpose of creating pretrained language models for biomedicine is totally different and can be used by any biomedical task to improve performance. This work introduces four pretrained language models for the biomedicine domain in the Turkish language. These models explore the effects of different corpus selection and pretraining strategies in the Turkish biomedicine domain. We also created a labeled dataset for classifying head CT radiology reports to evaluate the models. The main contributions can be listed as:

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

- We create two in-domain corpora by collecting open full-text scientific papers in biomedicine and theses on radiology. Then, we built four domain-specific pretrained language models using these corpora and made both corpora and models public for the first time in Turkish.
- We create a text classification task for head CT radiology reports in Turkish for the first time and evaluate two different parts of reports, *impressions*, and *findings*. As far as we know, this work is the first to evaluate the performance of pretrained language models in a Turkish clinical text.
- The positive effect of task-related corpus on model performance has been shown in the literature. Turkish biomedical corpus showed similar results. We also evaluate the effect of pretraining with theses in radiology corpus and pretraining from scratch approach on radiology report classification task for the first time.

2 Materials and Method

This section introduces the details of four pretrained language models developed in this work and the characteristics of the domain corpora used to generate these models. The first model, BioBERTurk$_{con}$(+trM), uses only Turkish biomedical text and applies the continual training approach, initializing weights from available general Turkish BERTurk [11]. It tests the hypothesis that using biomedical corpus via continual pretraining improves the performance of biomedical and clinical tasks, which is still valid for Turkish. The second model named BioBERTurk$_{con}$(trR) uses only radiology theses corpus for continuing pretraining to understand the task-related corpus's impact better. In addition, The third model adds a corpus built by the radiology theses to Turkish biomedical text and evaluates task-related corpus with Turkish biomedical text usage in continual pretraining. This one is called BioBERTurk$_{con}$(+trM+trR). In naming trM and trR indicate biomedical and radiology theses corpus, respectively. Finally, we trained a BERT model from scratch to evaluate the pretraining from the scratch approach in a low resource setting. It is called BioBERTurk$_{sc}$(+trW+trM+trR), which uses a mixed corpus composed of collected Turkish biomedical and radiology theses corpora and general domain corpus. We used the general domain corpus that BERTurk was trained for pretraining from scratch to be fairly comparable. We release the models and Turkish biomedical corpora, which can be accessed on the Github repository.

2.1 Building Domain-Specific Corpora

To develop BioBERTurk$_{con}$(+trM), the first step is to collect text in the biomedicine domain. Turkish abstracts in Pubmed are very limited, and we need other resource(s) to build a meaningful size corpus. We used Dergipark[2]

[2] www.dergipark.com.tr

to build the corpus. It has been developed and managed by Ulakbim[3] (Turkish Academic Network and Information Center), a gateway to access periodic refereed journals. We wrote a boot to visit all biomedical journals under the Dergipark and collected all full-text pdf articles published in those journals. We then scraped the collected pdf documents based on some heuristics rules similar to ABioNER [9]. For example, a rule defines the parts of articles to get necessary data for the domain like the beginning part should be "özet" (abstract) and the ending part should be "referanslar" (references) in Turkish articles. It is challenging and time-consuming to define all rules for extracting text from unstructured pdf. After retrieving the required text by this way, we applied a cleaning pipeline with custom steps to the raw text data. First, we combined all data in one large text file with one sentence per line. We then aggressively processed the file using language detection and hand-written heuristics rules. Written rules identify suspicious patterns like too high ratio of digits or punctuations, non-Turkish alphabet characters, or low average token numbers. Finally, to avoid repetitive content, the remaining corpora was deduplicated.

The second corpus aims to evaluate the effect of the task-related text. Since we use a classification task for head CT radiology reports, we searched open-domain text on radiology. Turkish Council of Higher Education provides a website[4] to search and access all open Ph.D. theses. We filtered all theses conducted in radiology departments of medical schools. We combined all collected theses again and applied the cleaning pipeline, building the corpus on radiology. As far as we know, this is the first attempt to use Ph.D. theses as task related domain corpus in pretrainig. The statistics of the final pretraining data produced in the cleaning steps are summarized in Table 1

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

2.2 Analyzing domain similarity

Before performing pretraining of our BERT models, we aim to calculate the similarity between our BERT's domain and target task domain. We evaluated the similarity of domains by calculating the intersection ratio of their domain vocabularies. Underlying assumption of this approach, the number of vocab

[3]https://ulakbim.tubitak.gov.tr/
[4]www.tez.yok.gov.tr/

words shared between domains should show how similar they are [12]. We consider domain vocabularies containing the most frequently used 10k unigram after removing stopwords, punctuations and numbers. We also used 100k sentences from random samples of documents in each BERT domain's corpus to generate vocabularies. For task vocabulary, we used 50k radiology report's impression since they are much shorter. Figure 1 shows the shared vocabulary ratio between domains. The measures calculated show that Turkish medical articles domain has strong vocabulary overlap with Turkish radiology theses. Although articles domain is more comprehensive than theses, they are similar because they have a similar tenor. We also observed that the target domain is the most similar to the theses domain (%22.37) on account of the radiology field and the least similar to the general domain (%5.67). So, the collected corpora seem appropriate to observe the effect of task-related corpus usage in pretrained language model generation.

Fig. 1: Vocabulary overlap ratio (%) between domains

2.3 Data Preprocessing

Data preprocessing is the transformation of raw textual data into BERT supported inputs. In order to do this, we first tokenized text and then encoded into integer numbers to feed the BERT model. The original BERT uses the Wordpiece tokenizer to implement this transformation for English text. However there are several studies using different tokenizers trained in languages other than English [13]. Turkish is a morphologically rich language and has special characteristic features due to its agglutinative structure. The rich morphology of Turkish provides generating words in many different meanings from a given root. From the NLP perspective, this linguistic feature leads to a high rate

of out-of-vocabulary (OOV) problem and reduces the performance of training accuracy. Wordpiece tokenization is a powerful approach to mitigate the challenging OOV problem and has been proven to have the highest performance in several Turkish NLP tasks [13]. In the light of these informations, we inherited Wordpiece vocabulary from BERTurk to preprocessing inputs of pretraining and finetuning of BioBERTurk$_{con}$. On the other hand, we constructed a new Wordpiece vocabulary for preprocessing of BioBERTurk$_{sc}$. We also used the tokenizer library from HuggingFace[5] to build uncased vocabulary and set the vocabulary size 32k to adjust the size defined in the BERTurk configuration file. We then utilized the official *create_pretraining_data.py* script provided by Google AI Research team to convert the all raw BERT input into structured tensorflow examples.

2.4 Pretraining Process

We conducted a set of experiments on two pretraining approaches for our models. BioBERTurk$_{sc}$ was pretrained from scratch using mixed corpora while BioBERTurk$_{con}$ variants were initialized with a tensorflow version of BERTurk checkpoints to continue pretraining. For training of our BERT variants, we followed the same procedure of BERTurk training. Each model was trained for 1M steps, with a max sequence length of 512 and a batch size of 128. We set Adam with a learning rate of 1e-4 warming up for 10K steps. We trained all models with open-source training scripts available in the official BERT Github repository using V3 TPUs with 8 cores from Google Cloud Compute Services.

2.5 Model Baseline

Following [1], we implement a fully-connected layer on the top of the BERT for classification tasks. We also establish a baseline model as presented in [14] to show comparative classification performance. This model was used to classify radiology reports of head CT exams in a non-English language (in Hebrew), the same task used in our experiments. We took the best-performed model with 90.8 classification accuracy and performed significantly better than Logical Regression and Gradient Boosting. The model has an LSTM layer stacked with an attention layer and, on top of that, a fully connected layer. It takes as input the word2vec embedding induced from our Turkish biomedical corpora. We referred to this baseline model as LSTM-attn-wvc.

3 Experiments

3.1 Classification of radiology reports

Turkish is a low resource language that lacks a labeled clinical dataset to construct NLP tasks. To evaluate our models at a text classification, we created two datasets based on different sections of radiology reports, one containing

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

findings and the other containing impressions.
In curation radiology datasets, we used an in-house corpus of 45,304 de-identified Turkish CT head radiology examinations produced at Ege University Hospital, Turkey. The reports cover the period from 2016 to the present. We also used the same individual's report to separate findings and impressions datasets. Before data analysis, we filtered out texts with less than 300 characters and removed newlines and domain-specific encodings. After the cleaning process, the final dataset had 5514 reports.
The annotation process was conducted by three radiologists (C.E, M.C.Ç, and S.S.O) with years of experience in radiology reporting. Two annotators (C.E, M.C.Ç) first independently labeled all reports. Then, the annotations were checked by the third annotator, and the consensus of all radiologists solved the conflicted ones. The annotation schema included three classes, "Presence of Intracranial Pathology (Abnormal)," "No Intracranial Pathology (Abnormal)" and "Out of Series", respectively, to indicate the presence or absence of intracranial pathology.
Finally, we divided the final annotated dataset into two datasets, *findings*, and *impressions*, to evaluate them separately. Our findings dataset has 28704 sentences and 13892 tokens; on the other hand, our *impressions* dataset has 78939 sentences and 17348 tokens. The impression part of a report usually includes longer text that lets us observe the performance of models with longer text. The annotation datasets were then randomly split into test (%10), validation (%10), and training (%80) set for fine-tuning. The two datasets' class distributions are the same, and this distribution is shown in Figure2. As seen in the Figure2, the datasets have an unbalanced distribution, a common condition for text processing in the radiology domain [15].

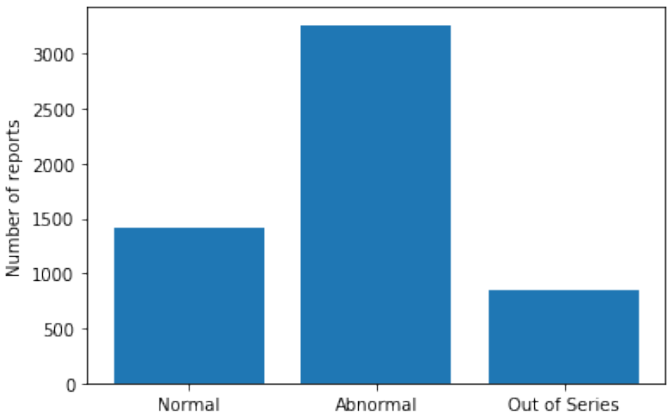

Fig. 2: Dataset class distribution

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 the 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 gensim framework using the CBOW architecture [16]. We evaluated a range of parameter combination for our baseline model, selecting the maximum length for 128, batch size for 16 and training for 25 epochs.

3.3 Evaluation criteria

The results of the models were evaluated using precision, recall and F1-score. For detailed analysis, the performance of each class was evaluated separately as well. Besides precision, recall and F1-score, t test [17] was conducted to determine whether there were statistical differences between models. We used 0.05 as the threshold to consider that results are statistically significant.

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

Model	Precision	Recall	F1-Score
BERTurk *+trW(c)*[1]	91.88%	91.87%	91.86%
BioBERTurk$_{con}$ *+trM(c)*	93.00%	93.02%	92.99%
BioBERTurk$_{con}$ *+(trM+trR)(c)*	92.74%	92.77%	92.75%
BioBERTurk$_{con}$ *+trR(c)*	**93.13%**	**93.14%**	**93.13%**
BioBERTurk$_{sc}$ *+(trW+trM+trR)(u)*[2]	89.52%	89.51%	89.48%
mBERT(c)	91.45%	91.43%	91.42%
LSTM-attn-wvc	80.80%	82.00%	80.72%

The best scores are in bold.

[1]refers cased model

[2]refers uncased model

4 Experimental Results

We conducted our experiments on the *impressions* and *findings* datasets separately. All scores are given in the best hyperparameter settings for each model. Table2 presents the average F1-scores over ten runs for *impressions* dataset. According to results, all BERT variants significantly outperformed the

baseline model (lstm-attn-word2vec). Furthermore, our in domain model BioBERTurk$_{con}$+(trR) achieved statistically higher F1-score than BERTurk(c) (P value 2.46e-05), BioBERTurk$_{sc}$ (P value 1.77e-11) and multilingual BERT (mBERT) (P value 9.11e-07). Although BioBERTurk$_{con}$+(trR) performed better than BioBERTurk$_{con}$+(trM), there is no statistical difference between these models (P value 0.59). We also compare all Turkish BERT models with mBERT to measure the effect of language on radiology report text classification. Although some studies have shown that mBERT performs more robustly than monolingual BERT models for some tasks [8, 18], our study shows that BioBERTurk$_{con}$ variants and BERTurk more accurately determine the class of Turkish radiology reports. For detailed analysis, per class, F1-scores are also reported in Table 3. While the highest F1-score for class "Normal" and "Out of Series" obtained by BioBERTurk$_{con}$+(trR) and for class "Abnormal" obtained by BioBERTurk$_{con}$+(trM). We also observed the winning model, BioBERTurk$_{con}$+(trR), obtained higher precision and recall than the others (Table 2).

Table 4 shows the average F1-scores over ten runs for *findings* dataset. The first clear observation of these experiments is that all model performs worse in findings data. Similar results were observed in English [19] and it is expected since findings are longer and less informative in terms of classification than impressions. In the findings dataset, BioBERTurk$_{con}$+(trM) performed well with the highest F1-score of 89.97% followed by BioBERTurk$_{con}$+(trM+trR) (P value 0.02) with no statistically meaningful difference and all BERT variants significantly outperformed our baseline model. When we examine the other metrics from Table 5, BioBERTurk$_{con}$+(trM) model performed very effectively for the Normal class but surprisingly not in Abnormal and Out of Series classes.

Table 3: per-label F1-score on *impressions* test set.

Model	Normal	Abnormal	Out of Series
BERTurk *+trW(c)*[1]	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)*	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.

5 Discussions

When we evaluate the experiments in general, we can draw the following conclusions from our study. First, our results show that all BioBERTurk$_{con}$ variants give better results in both datasets than the existing generic BERT

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

Model	Precision	Recall	F1-Score
BERTurk *+trW(c)*	89.00%	88.55%	88.60%
BioBERTurk$_{con}$ *+(trM)*	**90.34%**	**89.98%**	**89.97%**
BioBERTurk$_{con}$ *+(trM+trR)(c)*	88.93%	89.35%	89.38%
BioBERTurk$_{con}$ *+trR(c)*	88.61%	88.76%	88.75%
LSTM-attn-wvc	82.49%	83.01%	82.61%

The best scores are in bold.

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.

model and traditional model. Similar results are observed in English, where in-domain models outperform generic models [3, 4]. But, in our case continuing pretraining with a very small-sized domain corpus compared to the generic corpus is still very effective in classifying the clinical task. We observed similar results with medical articles and the radiology theses corpora. Although these corpora include longer and more noisy data than the Pubmed abstracts.

Another critical observation is the effect of the theses corpus in continuous pretraining. The theses corpus is very small compared to the medical article's corpus (0,11 GB vs 0,48 GB). BioBERTurk$_{con}$+(trR) model that is continuously trained only with the theses corpus performed statistically similar with other models. These results show that a corpus similar to the task domain can be very effective even with small size and noisy and long text data. This model outperformed other models to classify "Out of Series" label in the impression dataset. When the theses corpus was combined with medical articles, the outcome model (BioBERTurk$_{con}$+(trM+trR)) performed effectively, especially to classify "Abnormal" and "Out of Series" labels of the findings dataset. Thus, we can conclude that continual pretraining with small task-related data led to better accuracy for low-frequency label (Out of Series) in the classification of Turkish radiology reports.

Lastly, we compare the result of BioBERTurk$_{sc}$ with other models to investigate the pretraining technique. Our model, BioBERTurk$_{sc}$, gives bad classification accuracy for both datasets except our baseline model. So, combining very small domain data with large generic data is not an effective approach, at least in Turkish domain-oriented pretrained model generation from scratch.

In light of these results, we have demonstrated that if the target domain is dramatically different from general domain (the similarity of the Turkish general domain and Turkish clinical domain is %9), using task-related data for continual pertaining can boost the classification performance.
There has also been previous study that applied mBERT, which has significant zero-shot cross-lingual transfer abilities for low-resource language [20]. When we compare the F1score of mBERT with other models, our monolingual models developed using continual pretraining give better results in the Turkish clinical task. In summary, the success of our in-domain models presents that continual pretraining of biomedical articles can improve model performance on a clinical task in Turkish even when the available language resources are restricted.
Finally and most importantly, we have introduced the first Turkish biomedical resources and made them available to the NLP community.
Our study also 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. Second, we have reported longer than 512 character size, which is the limit of input size required by the BERT model.

6 Conclusion

In this study, we introduced the BioBERTurk family, four pretrained biomedical language models and evaluated them for the classification of Turkish radiology reports. Our work shows that further pretrained model with a small-scale radiology corpus, our domain-specific BioBERTurk$_{con}$ variant, achieved better performance than out-of-box BERT embeddings in classifying Turkish radiology reports. In the future work, we would like to investigate different pretraining and fine-tuning approaches on low resource settings for clinical Turkish domains and evaluate our model for different tasks in clinical NLP.

Acknowledgments. We would like to acknowledge the TPU Research Cloud program (TRC) [6] and the Google's CURe program in providing access to TPUv3 units and GCP credits, respectively.

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

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

Conflict of Interest. The authors declare no competing interests.

Authors' contributions. 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.

Availability of data and materials. 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/TurkRADBERT

References

[1] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

[2] Kalyan, K.S., Rajasekharan, A., Sangeetha, S.: Ammu: A survey of transformer-based biomedical pretrained language models. Journal of biomedical informatics, 103982 (2021)

[3] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020)

[4] Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., McDermott, M.: Publicly available clinical bert embeddings. arXiv preprint arXiv:1904.03323 (2019)

[5] Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H.: Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH) **3**(1), 1–23 (2021)

[6] Bressem, K.K., Adams, L.C., Gaudin, R.A., Tröltzsch, D., Hamm, B., Makowski, M.R., Schüle, C.-Y., Vahldiek, J.L., Niehues, S.M.: 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 (2020)

[7] Peng, Y., Yan, S., Lu, Z.: Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets. arXiv preprint arXiv:1906.05474 (2019)

[8] Schneider, E.T.R., de Souza, J.V.A., Knafou, J., e Oliveira, L.E.S., Copara, J., Gumiel, Y.B., de Oliveira, L.F.A., Paraiso, E.C., Teodoro, D., Barra, C.M.C.M.: Biobertpt-a portuguese neural language model for clinical named entity recognition. In: Proceedings of the 3rd Clinical Natural Language Processing Workshop, pp. 65–72 (2020)

[9] Boudjellal, N., Zhang, H., Khan, A., Ahmad, A., Naseem, R., Shang, J., Dai, L.: Abioner: a bert-based model for arabic biomedical named-entity recognition. Complexity **2021** (2021)

[10] Çelıkten, A., Bulut, H.: Turkish medical text classification using bert. In: 2021 29th Signal Processing and Communications Applications Conference (SIU), pp. 1–4 (2021). IEEE

[11] Schweter, S.: BERTurk - BERT models for Turkish. Zenodo (2020). https://doi.org/10.5281/zenodo.3770924. https://doi.org/10.5281/zenodo.3770924

[12] Dai, X., Karimi, S., Hachey, B., Paris, C.: Using similarity measures to select pretraining data for ner. arXiv preprint arXiv:1904.00585 (2019)

[13] Toraman, C., Yilmaz, E.H., Şahinuç, F., Ozcelik, O.: Impact of tokenization on language models: An analysis for turkish. arXiv preprint arXiv:2204.08832 (2022)

[14] Barash, Y., Guralnik, G., Tau, N., Soffer, S., Levy, T., Shimon, O., Zimlichman, E., Konen, E., Klang, E.: Comparison of deep learning models for natural language processing-based classification of non-english head ct reports. Neuroradiology **62**(10), 1247–1256 (2020)

[15] Qu, W., Balki, I., Mendez, M., Valen, J., Levman, J., Tyrrell, P.N.: Assessing and mitigating the effects of class imbalance in machine learning with application to x-ray imaging. International journal of computer assisted radiology and surgery **15**(12), 2041–2048 (2020)

[16] Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)

[17] Dietterich, T.G.: Approximate statistical tests for comparing supervised classification learning algorithms. Neural computation **10**(7), 1895–1923 (1998)

[18] Muller, B., Elazar, Y., Sagot, B., Seddah, D.: First align, then predict: Understanding the cross-lingual ability of multilingual bert. arXiv preprint arXiv:2101.11109 (2021)

[19] Gundogdu, B., Pamuksuz, U., Chung, J.H., Telleria, J.M., Liu, P., Khan,

F., Chang, P.J.: Customized impression prediction from radiology reports using bert and lstms. IEEE Transactions on Artificial Intelligence (2021)

[20] Wu, S., Dredze, M.: Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. arXiv preprint arXiv:1904.09077 (2019)