# Harnessing the Power of BERT in the Turkish Clinical Domain: Pretraining Approaches for Limited Data Scenarios

**Hazal Türkmen, Oğuz Dikenelli**
Department of Computer Engineering, Ege University, 35100 Izmir, Türkiye
{hazal.turkmen, oguz.dikenelli}@ege.edu.tr

**Cenk Eraslan, Mehmet Cem Çallı, Süha Süreyya Özbek**
Department of Radiology, Ege University, 35100 Izmir, Türkiye
{cenk.eraslan, cem.calli, sureyya.ozbek}@ege.edu.tr

## Abstract

Recent advancements in natural language processing (NLP) have been driven by large language models (LLMs), thereby revolutionizing the field. Our study investigates the impact of diverse pre-training strategies on the performance of Turkish clinical language models in a multi-label classification task involving radiology reports, with a focus on overcoming language resource limitations. Additionally, for the first time, we evaluated the simultaneous pre-training approach by utilizing limited clinical task data. We developed four models: TurkRadBERT-task v1, TurkRadBERT-task v2, TurkRadBERT-sim v1, and TurkRadBERT-sim v2. Our results revealed superior performance from BERTurk and TurkRadBERT-task v1, both of which leverage a broad general-domain corpus. Although task-adaptive pre-training is capable of identifying domain-specific patterns, it may be prone to overfitting because of the constraints of the task-specific corpus. Our findings highlight the importance of domain-specific vocabulary during pre-training to improve performance. They also affirmed that a combination of general domain knowledge and task-specific fine-tuning is crucial for optimal performance across various categories. This study offers key insights for future research on pre-training techniques in the clinical domain, particularly for low-resource languages.

## 1 Introduction

Language models have undergone a significant transformation in the field of natural language processing, demonstrating exceptional capabilities in executing tasks with minimal guidance. This shift can be attributed to pivotal milestones, such as word2vec (Mikolov et al., 2013), which replaced feature engineering methods with deep learning-based representation learning. Furthermore, the emergence of contextualized word embeddings with ELMo has led to the development of (Peters et al., 1802) pre-trained transformer-based models such as BERT (Devlin et al., 2018), GPT (Radford et al., 2018), T5 (Raffel et al., 2020), and BART (Lewis et al., 2019).

Recent advancements in large language models (LLMs) have led to the development of models with parameter sizes exceeding hundred billion, including the GPT (Generative Pre-trained Transformer) series (Radford et al., 2018, 2019b,a; Ouyang et al., 2022), such as ChatGPT and GPT-4 (OpenAI, 2023), which are pre-trained on massive datasets. However, research focusing on LLMs architectures within specialized domains characterized by limited resources is scarce. A range of approaches for developing language models exists to address the issue of limited language resources, including simultaneous pretraining with in-domain data (Wada et al., 2020) and domain-adaptive pretraining by fine-tuning an existing generic language model with in-domain data (Gururangan et al., 2020). The choice of pre-training technique depends on the specific task data and available resources, but determining the optimal utilization of limited clinical task data in pretraining and selecting the most suitable data for pretraining methods remain open questions. This study aimed to assess and contrast different techniques using a limited task corpus for pretraining BERT models in the Turkish clinical domain, a low-resource setting. We introduce two pre-trained language model families, TurkRadBERT-sim and TurkRadBERT-task, each comprising two models for the clinical domain in the Turkish language. These models explore the effects of different corpus selections that combine small task-related corpora and pretraining strategies in the Turkish clinical domain. The TurkRadBERT-sim pre-trained model family, developed via simultaneous pre-training (Wada et al., 2020), involves a balanced combination of two distinct corpora: one general and one limited task-specific. Both corpora were upsampled to create pretraining instances, resulting in robust neural language models.



*Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 161–170

The TurkRadBERT-task pretrained model family, developed via task-adaptive pre-training, involves an additional pretraining stage where the model is adaptively pretrained on a smaller, task-specific dataset following the initial pretraining. We also created a labeled dataset for multi-label document classification using head CT radiology reports to evaluate the models. The main contributions can be listed as follows:

- While simultaneous pretraining has previously been explored with limited biomedical data in the work of (Wada et al., 2020), our study shifts the focus towards applying this approach to limited clinical Turkish radiology data for the first time. We conducted an evaluation of simultaneous pretraining, incorporating limited clinical task radiology data, and compared it with task-adaptive pretraining through continual pre-training. This novel comparison provides valuable insights into the efficacy of these methods in the context of limited clinical radiology data, highlighting their potential in specialized domains.

- We created small task-related corpora, including Turkish head CT radiology reports by Ege University Hospital. Then, we built four pretrained clinical language models, for the first time, using Turkish head CT radiology reports, Turkish general corpus, and Turkish biomedical corpara, including Turkish medical articles (Türkmen et al., 2022) and Turkish radiology theses (Türkmen et al., 2022).

- We developed a multi-label document classification task aimed at identifying the presence or absence of 12 clinically significant observations, as well as a "no findings" label indicating no observations, within head CT radiology reports for the purpose of evaluating language models. To the best of our knowledge, there are no existing multi-label document classification studies in the Turkish clinical domain.

## 2 Related Work

To optimize natural language processing models for specialized domains, various studies have explored different approaches to adapt general BERT models for the biomedical domain. BioBERT (Lee et al., 2020), an early attempt to adapt general BERT models to the biomedical domain, employed continual pretraining to enhance performance. Initialized from the general BERT model, BioBERT was further trained on PubMed abstracts and full-text articles, yielding an improved performance for tasks such as named entity recognition, relation extraction, and question answering. Similarly, ClinicalBERT (Alsentzer et al., 2019), a domain-specific language model, was created using continual pretraining with MIMIC data, demonstrating its effectiveness in improving clinical task performance.
Other studies have explored continual pretraining for biomedical language models, such as SciBERT (Beltagy et al., 2019) and BlueBERT (Beltagy et al., 2019), which were pretrained on a mix of biomedical and general domain corpora. An alternative approach, pretraining from scratch, focuses exclusively on in-domain data, without relying on a generic language model. This method has been effective in creating models, such as PubMedBERT (Gu et al., 2021), which is pretrained solely on PubMed abstracts. Comparisons between the two pretraining methods reveal that continual pretraining often leads to more successful transfers from general to specialized domains. For example, one study proposed four BERT models (Bressem et al., 2020), two pretrained on German radiology free-text reports (FS-BERT and RAD-BERT), and two based on open-source models (MULTI-BERT and GER-BERT). The FS-BERT model, which used the pretraining from scratch approach, performed poorly compared to the other models, suggesting that domain-specific corpora alone might be insufficient for learning proper embeddings. Another study developed RadBERT (Yan et al., 2022), a set of six transformer-based language models pretrained on radiology reports with various language models for initialization, to explore their performance in radiology NLP applications.
Although pretraining BERT models can improve performance across various biomedical NLP tasks, they require significant domain-specific data. Biomedical text data are often limited and scattered across various sources, and few publicly available medical databases are written in languages other than English. This creates a high demand for effective techniques that can work well even with limited resources. One solution to this problem is the simultaneous pre-training technique proposed in (Wada et al., 2020), which up-samples a limited domain-specific corpus and uses it for pre-training

| Corpus | Size (GB) | N tokens | Domain |
|---|---|---|---|
| General Turkish Corpus | 35 | 4,404,976,662 | General |
| Turkish Biomedical Corpus | 0,48 | 60,318,554 | Biomedical |
| Turkish Electronic Radiology Theses | 0,11 | 15,268,779 | Radiology |
| Head CT Reports | 0.036 | 4,177,140 | Clinical Radiology |

Table 1: Corpora statistics

in a balanced manner with a larger corpus. Using small Japanese medical article abstracts and Japanese Wikipedia texts, the authors created a simultaneous pretrained BERT model, ouBioBERT. The study confirmed that their Japanese medical BERT model performed better than conventional baselines and other BERT models in a medical Japanese document classification task. However, they did not focus on applying the simultaneous pre-training approach to limited clinical task radiology data. Building upon this work, our study shifts the focus towards applying the simultaneous pre-training approach to limited clinical task data for the first time. To overcome the limitations of the limited resources problem, many researchers have explored the benefits of continued pretraining on a smaller corpus drawn from the task distribution as task-adaptive pre-training (Gururangan et al., 2020; Schneider et al., 2020). In addition, (Turkmen et al., 2022) previously demonstrated that their biomedical BERT models, the BioBERTurk family, which were continuously pre-trained on a limited Turkish radiology thesis corpus, exhibited improved performance in clinical tasks. However, the authors also highlighted the potential ineffectiveness of domain incompatibility when evaluating Turkish language models, emphasizing the need for a closer alignment between domain-specific data and evaluation tasks.

## 3 Materials and Methods

In this section, we provide a concise overview of the pre-training methods employed for the development of Turkish clinical language models and the characteristics of the corpora used in this process. We developed four Turkish clinical language models, leveraging the BERT-base architecture and constrained language resources by employing two pre-training strategies: simultaneous pre-training and continual pre-training, referred to as task-adaptive pretraining. Two models, referred to as the TurkRadBERT-sim family, were developed by employing simultaneous pre-training techniques that combined general, biomedical, and clinical task corpora, while utilizing distinct vocabularies. In contrast, two models, the TurkRadBERT-task family, were developed by employing task-adaptive pretraining using the task corpus. To construct these clinical models, we employed four distinct corpora: the Turkish biomedical corpus compiled from open-source medical articles (Türkmen et al., 2022), Turkish electronic radiology theses corpus (Türkmen et al., 2022), Turkish web corpus (Schweter, 2020), and newly created Turkish radiology report corpus, which is a limited task corpus. While all corpora were utilized in simultaneous pre-training, only Turkish radiology reports were used in task-adaptive pre-training. Subsequently, the clinical language models were fine-tuned on a downstream NLP task within the Turkish clinical domain. Finally, the clinical language models were compared to the general Turkish domain BERT model, BERTurk (Schweter, 2020), and the BioBERTurk variant (Turkmen et al., 2022), which was continually pretrained on Turkish radiology theses.

### 3.1 Pre-training Strategies

The BERT framework (Devlin et al., 2018) consists of two phases: pretraining and fine-tuning. During pre-training, BERT is trained on large-scale plain text corpora, such as Wikipedia, whereas in the fine-tuning phase, it is initialized with the same pre-trained weights and then fine-tuned using task-specific labeled data, such as sentence pair classification. BERT employs two unsupervised tasks during the pre-training phase: Masked Language Model (MLM) and Next Sentence Prediction (NSP). In MLM tasks, a certain percentage of input tokens is randomly masked, and the model predicts the masked tokens in a sentence, as described in the Cloze task (Taylor, 1953). For the NSP, the model predicts whether the second sentence follows a consecutive sentence in the dataset.
In our study, we implemented several modifications to the BERT architecture for simultaneous pretrain-

ing (Wada et al., 2020), our first technique. This pre-training approach posits that training the BERT model using both large and small corpora can prevent overfitting issues caused by limited medical data. To accurately feed inputs into the model, we followed a procedure from the same study (Wada et al., 2020). We divided the small medical corpus and large general corpus into smaller documents of equal size, and combined them to create structured inputs. This approach mitigates potential overfitting resulting from the limited data size by increasing the frequency of pre-training for MLM instances containing small medical data. In accordance with the same study by (Wada et al., 2020), we utilized domain-specific generated text and the Wordpiece algorithm to generate a domain-specific vocabulary, which is referred to as amplified vocabulary in their research. Thus, we examined the impact of the domain-specific vocabulary
Simultaneous pretraining enables the model to learn language representations by training on large-scale texts. However, this approach is expensive owing to the extensive amount of data involved. Finally, we implemented the task-adaptive pretraining method (Gururangan et al., 2020) using only a small amount of clinical task data. This technique is less resource-intensive than the others. In contrast to the aforementioned pre-training methods, we developed different BERT models based on model initialization for task-adaptive pre-training, using the existing BERT vocabulary instead of creating a new one.

### 3.2 Data Sources for Model Development

In the development of various language models, multiple corpora were utilized to ensure that the models were well suited to the specific domain and task at hand. The selection of appropriate corpora is crucial to the performance of language models, as it directly influences their understanding of domain-specific language patterns, structures, and vocabularies. The corpora used are summarized in Table 1 and listed below:

**Head CT Reports:** We collected 40,306 verified Turkish radiology reports pertaining to computed tomography (CT) examinations for patients aged 8 years and above from the neurology and emergency departments at Ege University Hospital between January 2016 and June 2018. Prior to data analysis, reports containing fewer than 100 characters were excluded, and newline characters and radiology-specific encodings were removed for consistency. All text data underwent de-identification and duplicate removal. Following preprocessing, 2,000 reports were randomly selected for the head CT annotation task, and the remaining data (approximately 36 MB) was reserved for pre-training techniques.

**General Turkish Corpus:** This corpus, which was used in the development of the BERTurk model, contains a large collection of Turkish text data (approximately 35 GB). This serves as a foundation for training language models to understand Turkish language patterns.

**Turkish Biomedical Corpus:** A domain-specific corpus (Türkmen et al., 2022) consisting of full-text articles collected from Dergipark, a platform hosting periodically refereed biomedical journals in Turkey.

**Turkish Electronic Radiology Theses:** A unique corpus of open-domain Ph.D. theses (Türkmen et al., 2022) conducted in radiology departments of medical schools obtained from the Turkish Council of Higher Education's website.

### 3.3 Data preparation

The first phase after data understanding is transforming the text to the BERT-supported inputs, namely tokenization. All engineering processes to be fed into BERT were designed for Google Cloud TPUs and implemented using CPU core i8. Furthermore, Wordpiece algorithm was used to generate vocabulary for tokenization in both pre-training methods due to the success in morphologic-rich languages such as Turkish (Toraman et al., 2023). Each vocabulary config file is the same as BERTurk for a fair comparison. We implemented the tokenizer library from Huggingface [1] to build BERT's vocabulary in simultaneous pre-training. For continual pre-training, we used existing BERT's vocabulary for continual pre-training instead of creating a new one. After this process, we used *create_pretraining_data.py* script provided by the Google AI Research team [2] to convert all documents into TensorFlow examples compatible with TPU devices.

### 3.4 Pretraining setup

We followed BERT-base architecture consisting of 12 layers of transformer blocks, 12 attention heads,

[1] https://huggingface.co/docs/tokenizers/python/latest/
[2] https://github.com/google-research/bert

and 110 million parameters for all pre-training strategies. All models were also generated using the same hyperparameters (see Appendix B, Table 5 ) and were trained with open-source training scripts available in the official BERT GitHub repository using V3 TPUs with 32 cores from Google Cloud Compute Services [3].

### 3.5 Developed Language Models

The simultaneous pre-training technique is the first pre-training method we implemented to utilize a small in-domain corpus. Moreover, the first step in simultaneous pre-training is choosing data for small and large corpus data. We produced different TurkRadBERT-sim models according to vocabulary usage. The distinction between the two models lies in their vocabulary use; the first model leverages an amplified, domain-specific vocabulary, whereas the latter adopts the BERTurk vocabulary.

**TurkRadBERT-sim v1** employed a large Turkish general corpus (35 GB) used for developing BERTurk, alongside a mixed Turkish biomedical corpus, Turkish Electronic Radiology Theses, and Turkish Head CT Reports as smaller counterparts. Excluding the data utilized for labeling (approximately 6 MB), the head CT reports were not used as a standalone small corpus for pre-training due to their limited size (30 MB) compared to other corpora. Furthermore, experimental results suggested that simultaneous training with such a data size did not yield significant outcomes in radiology report classification. To address this, we combined the small-sized corpus to match the large one, creating pre-training instances. The model also employed an amplified vocabulary, built from the generated corpus, for simultaneous pre-training.

**TurkRadBERT-sim v2** was also based on the BERT-base architecture and was pre-trained simultaneously. The model used the same corpus as v1 during pretraining. The difference was that the general domain vocabulary was used to observe the effect of the domain-specific vocabulary.

The last pre-training method is task-adaptive pre-training on radiology reports (30 MB). We developed two different BERT models according to the model initialization.

**TurkRadBERT-task v1** used a general domain language model for Turkish, BERTurk for model initialization and then carried out continual pre-training as a task-adaptive pre-training method. Vocabulary was also inherited from BERTurk.

**TurkRadBERT-task v2** used a Turkish biomedical BERT model, BioBERTurk variant(Turkmen et al., 2022), which was further pre-trained on Turkish electronic theses for model initialization. This Turkish biomedical BERT was chosen because it achieved the best score in classification radiology reports (Turkmen et al., 2022). For tokenization, the model again inherited from the general domain.

## 4 Supervision Task

### 4.1 Multi-label CT radiology reports classification

We developed a multi-label document classification task using 2000 Turkish head CT reports mentioned in Section 3.2. This was necessary as there was no shared task for clinical documents in Turkish. Our dataset has 20618 sentences and 249072 tokens. The objective of the document level classification task is to identify the existence of clinically significant observations in a radiology report that is presented in free-text format. These are 'Intraventricular' ,'Gliosis', 'Epidural', 'Hydrocephalus', 'Encephalomalacia', 'Chronic ischemic changes', 'Lacuna', 'Leukoaraiosis', 'Mega cisterna magna' ,'Meningioma', 'Subarachnoid Bleeding', 'Subdural', 'No Findings'. The classification process involves reviewing sentences within the report and categorizing them into one of two classes: positive or negative. The 13th observation, "No Findings", indicates the absence of any findings. Those 12 labels were selected to indicate major and relatively common clinical pathologies possible to be detected in a pre-contrast cranial computerized tomography (CT) examination. Moreover, the 12 labels used in the study also are not vague radiologic findings, but definite clinical pathologies. Therefore, no hedging was performed regarding these categories radiology experts labeled the dataset at document level according to this annotation schema. The annotation process unfolded in three stages, involving three experienced radiologists (C.E, M.C.C, and S.S.O). In each stage, two annotators (C.E, M.C.C) independently labeled a portion of the reports. Subsequently, the third annotator examined these annotations to detect any discrepancies. At the conclusion of each stage, all three annotators reached a consensus by generating mutually agreed-upon annotations. A spreadsheet file was utilized to facilitate the annotation task for

[3] https://cloud.google.com/

| Model | Precision | Recall | F1 Score |
|---|---|---|---|
| BERTurk | **0.9738** | 0.9456 | **0.9562 (± 0.0077)** |
| TurkRadBERT-task v1 | 0.9736 | **0.9462** | 0.9556 (± 0.0057) |
| BioBERTurk | 0.9731 | 0.9440 | 0.9535 (± 0.0068) |
| TurkRadBERT-task v2 | 0.9643 | 0.9352 | 0.9470 (± 0.0068) |
| TurkRadBERT-sim v1 | 0.8613 | 0.7969 | 0.8149 (± 0.0214) |
| TurkRadBERT-sim v2 | 0.8170 | 0.7863 | 0.7879 (± 0.0135) |

Table 2: Average Precision, recall, and F1 Score for each model. We performed ten separate runs with different random seeds and present both the average and standard deviation.

| Category | BERTurk | TurkRadBERT-task v1 |
|---|---|---|
| Intraventricular | **0.4815 (± 0.4475)** | 0.4000 (± 0.3266) |
| Gliosis | **0.8580 (± 0.0577)** | 0.8155 (± 0.1024) |
| Epidural | **0.9012 (± 0.0349)** | 0.9000 (± 0.0333) |
| Hydrocephalus | 0.9458 (± 0.0327) | **0.9673 (± 0.0459)** |
| Encephalomalacia | 0.9622 (± 0.0173) | **0.9633 (± 0.0081)** |
| Chronic ischemic changes | 0.9918 (± 0.0044) | **0.9921 (± 0.0026)** |
| Lacuna | **0.9655 (± 0.0000)** | **0.9655 (± 0.0000)** |
| Leukoaraiosis | **0.8995 (± 0.1063)** | 0.8762 (± 0.1227) |
| Mega cisterna magna | **0.6000 (± 0.1500)** | 0.4500 (± 0.0577) |
| Meningioma | **1.0000 (± 0.0000)** | **1.0000 (± 0.0000)** |
| Subarachnoid Bleeding | 0.9281 (± 0.0183) | **0.9544 (± 0.0118)** |
| Subdural | 0.9666 (± 0.0119) | **0.9757 (± 0.0081)** |
| No Findings | **0.9455 (± 0.0145)** | 0.9311 (± 0.0167) |

Table 3: Average F1 scores for each label in the TurkRadBERT-task v1 and BERTurk models. In each experiment, we carried out ten distinct runs using different random seeds, from which we determine and report the average and standard deviation.

the annotators The annotated datasets were subsequently divided randomly into test (10%), validation (10%), and training (80%) sets for fine-tuning. The class distributions, as illustrated in Appendix A, demonstrate the varying prevalence of different categories in the datasets. The datasets exhibit an imbalanced distribution, which is a typical characteristic of text processing in the radiology domain (Qu et al., 2020).

### 4.2 Fine-tuning Setup

The fine-tuning of all pretrained models was conducted independently utilizing identical architecture and optimization methods as previously employed in the study (Devlin et al., 2018). In the process of fine-tuning, the objective is not to surpass the current state-of-the-art performance on the downstream tasks, but rather to assess and compare pretraining techniques for developing Turkish clinical language models. So, an exhaustive exploration of hyperparameters was not utilized. Consequently, the optimal parameters identified from a limited hyperparameter search are employed, working under the assumption that the fairness of model evaluations and comparisons isn't compromised by the potential presence of more optimal hyperparameters. Hyperparameter searches were conducted for each model, examining learning rate values $\epsilon$ from the set {2e-4, 3e-5, 5e-5}, maximum sequence lengths $\epsilon$ from the set {128, 256, 512}, batch sizes $\epsilon$ from the set {16, 32}, and the number of training epochs $\epsilon$ from the set {15, 20}. Due to memory constraints, a batch size of 64 was not considered. The configurations employed for the TurkRadBERT-sim and TurkRadBERT-task models are displayed in Table 6 and Table 7 in Appendix B respectively.

The effectiveness of distinct pre-trained BERT models on the clinical multilabel classification task was evaluated by computing average precision, recall, and F1 score across ten runs, utilizing the most suitable hyperparameter settings.

## 5 Experimental Results

In this study, we evaluated the performance of five different models, including BERTurk, TurkRadBERT-task v1, TurkRadBERT-task v2, TurkRadBERT-sim v1, and TurkRadBERT-sim v2, for Turkish clinical multi-label classification. We compared their performance over ten runs in terms of average precision, recall, and F1 score. Additionally, we analyzed the performance of wining two model (BERTurk, TurkRadBERT-task v1) on individual categories using their respective F1 scores. The results are presented in Tables 2 and 3.

Table 2 shows that BERTurk achieves an F1 score of 0.9562, with a precision of 0.9738 and recall of 0.9456. TurkRadBERT-task v1 has a slightly lower F1 score of 0.9556 but with comparable precision (0.9736) and recall (0.9462). Both models demonstrate strong performance on the classification task, with BERTurk slightly outperforming TurkRadBERT-task v1 in terms of the overall F1 score. While BERTurk performed better than TurkRadBERT-task v1, there are no statistical differences between these models (P value 0,255). Additionally, BERTurk has also outperformed BioBERTurk. Other models, such as TurkRadBERT-task v2, TurkRadBERT-sim v1, and TurkRadBERT-sim v2, show lower overall performance compared to BERTurk, TurkRadBERT-task v1 and BioBERTurk.

However, it is essential to evaluate the models' performance for each label, as this offers a deeper understanding of their strengths and weaknesses. Table 3 presents the F1 scores for each category for BERTurk and TurkRadBERT-task v1. The results reveal that the performance of the models varies across categories, with some labels showing a noticeable difference in F1 scores between the two models. BERTurk performs better than TurkRadBERT-task v1 in the following categories: Intraventricular, Gliosis, Epidural, Leukoaraiosis, Mega cisterna magna, and No Findings. In contrast, TurkRadBERT-task v1 outperforms BERTurk in the categories of Hydrocephalus, Encephalomalacia, Chronic ischemic changes, Subarachnoid Bleeding, and Subdural. The F1 scores for Lacuna and Meningioma are identical for both models.

## 6 Discussion

By assessing the experiments as a whole, we derived the following conclusions. When comparing simultaneous pre-training and task-adaptive pre-training, it is observed that, owing to the size difference between the task data and the general data, the limited domain-specific data may be overshadowed by the large general-domain data. This causes the model to focus more on learning general rather than task-specific features. This phenomenon highlights the importance of carefully balancing general and domain-specific data during the pretraining process to ensure that the model effectively captures the nuances of the specialized domain.

The performances of the BERTurk and TurkRadBERT-task v1 models are quite similar because both models leverage the knowledge gained from the large general-domain corpus during pre-training. BERTurk is directly pre-trained on this large corpus, while TurkRadBERT-task v1 is initialized with BERTurk's weights and then fine-tuned using task-adaptive pre-training on a smaller clinical corpus. This fine-tuning enables TurkRadBERT-task v1 to capture domain-specific patterns, structures, and terminologies absent in the general-domain corpus.

However, the small task-specific corpus used in task-adaptive pretraining may limit the model's learning of domain-specific knowledge. Consequently, despite the benefits of task-adaptive pre-training, TurkRadBERT-task v1, which utilized this approach, had a slightly lower performance than BERTurk. In limited data scenarios, the task-adaptive pre-training approach may be prone to overfitting, especially when pre-trained on a small task-specific corpus. The model may become overly specialized in training data and fail to generalize well to unseen examples (Zhang et al., 2022).

In terms of performance, TurkRadBERT-task v1 has a slightly higher F1 score (0.9556) than BioBERTurk (0.9535) and TurkRadBERT-task v2 (0.9470). This suggests that despite the more specialized biomedical knowledge in BioBERTurk, the general-domain BERTurk model provides a more robust foundation for task-adaptive pre-training in this specific clinical task.

Another conclusion reached in this study is that comparison between TurkRadBERT-sim v1 and v2 offers insights into the impact of domain-specific vocabulary on model performance. TurkRadBERT-sim v1, which used an amplified vocabulary built from the generated corpus, outperformed TurkRadBERT-sim v2 that employed

the general domain vocabulary. This finding indicates that using a domain-specific vocabulary during pre-training can enhance the ability of the model to capture and understand domain-specific language patterns, ultimately leading to improved performance on clinical NLP tasks.

Examining the F1 scores for each label in Table 3 provides a more detailed perspective of the performance of the two most successful models. First, the optimal performance on specific labels, such as meningioma and chronic ischemic changes, might be attributed to the use of precise, standard reporting terminology to define these pathologies, a factor that likely provides high results, regardless of the classifier employed. BERTurk outperforms TurkRadBERT-task v1 in certain labels, such as Intraventricular, Gliosis, Epidural, Leukoaraiosis, Mega cisterna magna, with No Findings. The higher performance of BERTurk on certain labels could be attributed to the general domain knowledge acquired during direct pre-training (different from other pre-training methods), which may provide better coverage for specific categories, particularly those with a lower frequency in the task-specific corpus. BERTurk's broader pre-training data exposure could potentially give it an advantage over models like TurkRadBERT-task v1 when dealing with specific labels that have lower representation in the task-specific corpus, even though TurkRadBERT-task v1 is initialized with BERTurk. This suggests that a combination of general domain knowledge and task-specific fine-tuning may be critical for optimal performance across diverse categories. On the other hand, TurkRadBERT-task v1 exhibits superior performance for labels like Hydrocephalus, Encephalomalacia, Subarachnoid Bleeding, and Subdural. This suggests that task-adaptive pre-training can offer a performance boost in some instances by fine-tuning the model based on domain-specific information. However, it is worth noting that the overall performance differences between the two models are relatively small, highlighting the importance of leveraging both general-domain and task-specific knowledge in these models.

## 7 Conclusion

This study provides a comprehensive comparison of the performance of various models, including BERTurk, TurkRadBERT-task v1, TurkRadBERT-task v2, TurkRadBERT-sim v1, and TurkRadBERT-sim v2, on a radiology report classification task. Our findings demonstrate that the BERTurk model achieved the best overall performance, closely followed by the TurkRadBERT-task v1 model. This highlights the importance of leveraging both general domain knowledge acquired during pre-training and task-specific knowledge through fine-tuning to achieve optimal performance on complex tasks.

We also observed that the performance of these models varies across different labels, with BERTurk performing better on certain categories, particularly those with lower representation in the task-specific corpus. This finding suggests that a combination of general domain knowledge and task-specific fine-tuning may be critical for achieving optimal performance across diverse categories. Additionally, it is essential to consider label frequencies when interpreting results because performance on rare labels may be more susceptible to noise and overfitting.

The simultaneous pre-training models TurkRadBERT-sim v1 and v2 exhibit lower performance compared to their task-adaptive counterparts, indicating that task-adaptive pre-training is more effective in capturing domain-specific knowledge. Nevertheless, further investigation of alternative pre-training and fine-tuning strategies could help enhance the performance of these models.

Future research could focus on expanding the task-specific corpus to improve domain-specific knowledge and performance on rare labels as well as explore alternative pre-training and fine-tuning strategies to further enhance model performance. Moreover, investigating the factors contributing to the performance differences between the models for each label could provide valuable insights for developing more effective models in the field of medical natural language processing.

## Acknowledgements

The study was approved by the Ege University Ethical Committee under study number UH150040389 and conducted in accordance with the Declaration of Helsinki. We would also like to express our gratitude to the TPU Research Cloud program (TRC) [4] and Google's CURe program for granting us access to TPUv3 units and GCP credits, respectively.

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

## References


Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Keno K Bressem, Lisa C Adams, Robert A Gaudin, Daniel Tröltzsch, Bernd Hamm, Marcus R Makowski, Chan-Yong Schüle, Janis L Vahldiek, and Stefan M Niehues. 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 1802. Deep contextualized word representations. corr abs/1802.05365 (2018). *arXiv preprint arXiv:1802.05365*.

Wendi Qu, Indranil Balki, Mauro Mendez, John Valen, Jacob Levman, and Pascal N Tyrrell. 2020. 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*, 15:2041–2048.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Dario Amodei, Daniela Amodei, Jack Clark, Miles Brundage, and Ilya Sutskever. 2019a. Better language models and their implications. *OpenAI Blog https://openai.com/blog/better-language-models*, 1(2).

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019b. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Elisa Terumi Rubel Schneider, Joao Vitor Andrioli de Souza, Julien Knafou, Lucas Emanuel Silva e Oliveira, Jenny Copara, Yohan Bonescki Gumiel, Lucas Ferro Antunes de Oliveira, Emerson Cabrera Paraiso, Douglas Teodoro, and Cláudia Maria Cabral Moro Barra. 2020. Biobertpt-a portuguese neural language model for clinical named entity recognition. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 65–72.

Stefan Schweter. 2020. Berturk - bert models for turkish.

Wilson L Taylor. 1953. "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433.

Cagri Toraman, Eyup Halit Yilmaz, Furkan Şahinuç, and Oguzhan Ozcelik. 2023. Impact of tokenization on language models: An analysis for turkish. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(4):1–21.

Hazal Turkmen, Oguz Dikenelli, Cenk Eraslan, and Mehmet Cem Callı. 2022. Bioberturk: Exploring turkish biomedical language model development strategies in low resource setting.

Hazal Türkmen, Oğuz Dikenelli, Cenk Eraslan, Mehmet Cem Çalli, and Suha Sureyya Ozbek. 2022. Developing pretrained language models for turkish

biomedical domain. In *2022 IEEE 10th International Conference on Healthcare Informatics (ICHI)*, pages 597–598. IEEE.

Shoya Wada, Toshihiro Takeda, Shiro Manabe, Shozo Konishi, Jun Kamohara, and Yasushi Matsumura. 2020. Pre-training technique to localize medical bert and enhance biomedical bert. *arXiv preprint arXiv:2005.07202*.

An Yan, Julian McAuley, Xing Lu, Jiang Du, Eric Y Chang, Amilcare Gentili, and Chun-Nan Hsu. 2022. Radbert: Adapting transformer-based language models to radiology. *Radiology: Artificial Intelligence*, 4(4):e210258.

Angela Zhang, Lei Xing, James Zou, and Joseph C Wu. 2022. Shifting machine learning for healthcare from development to deployment and from models to data. *Nature Biomedical Engineering*, pages 1–16.

# A Additional dataset information

| Category | Positive | Negative |
|---|---|---|
| Intraventricular | 22 (%1.1) | 1978 (%98.9) |
| Gliosis | 54 (%2.7) | 1946 (%97.3) |
| Epidural | 51 (%2.55) | 1949 (%97.45) |
| Hydrocephalus | 70 (%3.5) | 1930 (%96.5) |
| Encephalomalacia | 177 (%8.85) | 1823 (%91.15) |
| Chronic ischemic changes | 951 (%47.55) | 1049 (%52.45) |
| Lacuna | 138 (%6.9) | 1862 (%93.1) |
| Leukoaraiosis | 49 (%2.45) | 1951 (%97.55) |
| Mega cisterna magna | 15 (%0.75) | 1985 (%99.25) |
| Meningioma | 39 (%1.95) | 1961 (%98.05) |
| Subarachnoid Bleeding | 209 (%10.45) | 1791 (%89.55) |
| Subdural | 227 (%11.35) | 1773 (%88.65) |
| No Findings | 299 (%14.95) | 1701 (%85.05) |

Table 4: Distribution of frequencies for each label's positive and negative radiology documents in the dataset.

# B Pre-training and fine-tuning hyperparameters

| Hyperparameters | Values |
|---|---|
| Learning rate | 1e-4 |
| Batch size | 256 |
| Optimizer | Adam |
| $\beta_1$ | 0.9 |
| $\beta_2$ | 0.999 |
| Warmp up steps | 10000 |
| Max sequence length | 512 |
| Max prediction per seq | 76 |
| Masked MLM probability | 0.15 |
| epoch | 1000000 |

Table 5: Pre-training configuration for BERT models.

| Parameters | Value |
|---|---|
| Learning rate | 5e-5 |
| Batch size | 32 |
| Optimizer | Adam |
| Max sequence length | 512 |
| epoch | 20 |

Table 6: Best fine-tuning configuration for TurkRadBERT-sim family

| Parameters | Value |
|---|---|
| Learning rate | 3e-5 |
| Batch size | 32 |
| Optimizer | Adam |
| Max sequence length | 512 |
| epoch | 15 |

Table 7: Best fine-tuning configuration for BERTurk, BioBERTurk and TurkRadBERT-task family