Black Sea Journal of Engineering and Science
doi: 10.34248/bsengineering.1596832

Open Access Journal
e-ISSN: 2619 – 8991

Research Article
Volume 8 - Issue 2: 323-329 / March 2025

PERFORMANCE EVALUATION OF TRANSFORMER-BASED PRE-TRAINED LANGUAGE MODELS FOR TURKISH QUESTION-ANSWERING

Mert İNCİDELEN[1], Murat AYDOĞAN[2]*

[1]Fırat University, Faculty of Engineering, Department of Artificial Intelligence and Data Engineering, 23200, Elazığ, Türkiye
[2]Fırat University, Faculty of Technology, Department of Software Engineering, 23200, Elazığ, Türkiye

Abstract: Natural language processing (NLP) has made significant progress with the introduction of Transformer-based architectures that have revolutionized tasks such as question-answering (QA). While English is a primary focus of NLP research due to its high resource datasets, low-resource languages such as Turkish present unique challenges such as linguistic complexity and limited data availability. This study evaluates the performance of Transformer-based pre-trained language models on QA tasks and provides insights into their strengths and limitations for future improvements. In this study, variations of the mBERT, BERTurk, ConvBERTurk, DistilBERTurk, and ELECTRA Turkish pre-trained models were fine-tuned using the SQuAD-TR dataset, which is the machine-translated Turkish version of the SQuAD 2.0 dataset. The performance of these fine-tuned models was tested using the XQuAD-TR dataset. The models were evaluated using Exact Match (EM) and F1 Score metrics. Among the tested models, the ConvBERTurk Base (cased) model performed the best, achieving an EM of 57.82% and an F1 Score of 71.59%. In contrast, the DistilBERTurk Base (cased) and ELECTRA TR Small (cased) models performed poorly due to their smaller size and fewer parameters. The results indicate that case-sensitive models generally perform better than case-insensitive models. The ability of case-sensitive models to discriminate proper names and abbreviations more effectively improved their performance. Moreover, models specifically adapted for Turkish performed better on QA tasks compared to the multilingual mBERT model.

Keywords: Natural language processing, Question-answering, Transformers, BERT, ELECTRA

***Corresponding author:** Fırat University, Faculty of Technology, Department of Software Engineering, 23200, Elazığ, Türkiye
E mail: m.aydogan@firat.edu.tr (M.AYDOĞAN)
Mert İNCİDELEN https://orcid.org/0009-0002-1975-8332
Murat AYDOĞAN https://orcid.org/0000-0002-6876-6454

Received: December 05, 2024
Accepted: January 15, 2025
Published: March 15, 2025

Cite as: İncidelen M, Aydoğan M. 2025. Performance evaluation of transformer-based pre-trained language models for turkish question-answering. BSJ Eng Sci, 8(2): 323-329.

1. Introduction

With the onset of the digital age and the rise of artificial intelligence, there have been significant developments in the ability of machines to understand and use natural languages like humans. These developments have enabled the growth of the field of natural language processing (NLP) (Locke et al., 2021). NLP has emerged as an interdisciplinary field of study where computer science and linguistics intersect. NLP focuses on machines understanding and generate human language. Many tasks such as machine translation, sentiment analysis, text classification, and question-answering (QA) can be performed in the field of NLP (Khurana et al., 2023). There has been a significant increase in the amount of text data created by online sources such as social media, blogs and many other text sources (Hassani et al., 2020). Thus, a wide variety of resources have emerged for use in NLP tasks for different languages. With the increase in text data, NLP studies have also gained importance. The announcement of the Transformer architecture (Vaswani et al., 2017) and the emergence of Transformer-based pre-trained language models have led to the beginning of a new era in the field of NLP. Many pre-trained language models can be used for NLP tasks. These models are pre-trained on large-scale text data and then fine-tuned for specific NLP tasks. One of these tasks is answer extraction from text. Answer extraction is the process of extracting information about a query text from the content of texts obtained from information sources. Unlike traditional information search methods, the answer extraction method aims to obtain clear answers directly from sources instead of directing to sources where the information is available (Allam and Haggag, 2012; Yiğit and Amasyalı, 2021). Thus, effective QA systems can be developed. English datasets such as SQuAD (Stanford Question-Answering Dataset) have been a turning point for the development of QA systems. The SQuAD dataset, which is compiled from Wikipedia texts and contains more than 100,000 question-answer pairs, has become a standard for open-domain QA systems (Rajpurkar et al., 2016). SQuAD 2.0 expanded this dataset by adding unanswered questions, making it possible to evaluate the ability of QA models to handle more complex situations (Rajpurkar et al., 2018). In a high-resource language such as English, large datasets support the training, testing and development processes of QA models. This has made English the most frequently studied language in the field of NLP and has led to the formation of a large literature on tasks such as QA. English QA systems have been developed with SQuAD and its derivative datasets, and the effectiveness of Transformer-based models has been proven in different fields (Raza et al., 2022; Alzubi et al., 2023; Zhu et al., 2023). NLP studies on low-resource languages such as Turkish are limited. Although pre-trained models for Turkish are widely used in different NLP tasks (Çelikten and Bulut, 2021; Türkmen et al., 2023; Arzu and Aydoğan, 2023), the lack of resources for Turkish QA tasks has limited work on this task.

The limitations in the literature on QA studies in Turkish, combined with the structural and linguistic difficulties arising from this language, make the development of QA systems more complex. In this context, there are a limited number of QA studies and datasets for Turkish. Gemirter and Goularas (2021) aimed to develop a deep learning-based Turkish QA system in their work. This capability was developed using the BERT model to develop a question and answer system. The system is designed to provide precise and concise answers to banking-related questions by using large datasets and advanced deep learning techniques. Wikipedia, news corpora, and banking sector-specific data were utilized in the study. Soygazi et al. (2021) constructed a Turkish QA dataset called THQuAD for Turkish reading comprehension evaluations. The study utilized texts consisting of Turkish Wikipedia articles and Ottoman history and Turkish-Islamic history of science datasets. Analyses were conducted on this dataset by fine-tuning language models such as BERT, ELECTRA and ALBERT. The study demonstrated how these models perform in a morphologically rich language such as Turkish. The results showed that especially ELECTRA models achieved the highest performance compared to other models, emphasizing the potential of language models in Turkish reading comprehension tasks. Akyon et al. (2021) investigated the mT5 model in multitask settings to perform automatic question generation and question answering tasks from Turkish texts. In the study, the performance of the model was evaluated on datasets such as TQuADv1, TQuADv2 and XQuAD and meaningful question-answer pairs were generated with the proposed methods. In particular, the focus was on the multitask learning approach that performs answer extraction, question generation and answering tasks simultaneously. In addition, sentence segmentation algorithms specific to the Turkish language were developed and the morphological complexities of the language were taken into account in this process. The obtained results showed that the mT5 model exhibited better performance compared to existing methods. The study focused on the applications of transformer-based models in morphologically rich languages such as Turkish. Budur et al. (2024) proposed a method to develop an efficient and effective Open Domain QA (OpenQA) system for low-resource languages. In the study, focusing on Turkish, SQuAD 2.0 dataset was translated into Turkish using machine translation approach, resulting in the creation of the SQuAD TR dataset. Models such as ColBERT-QA and DPR were customized and adapted to Turkish sources. The results indicate that these systems achieved 24-32 percent and 22-29 percent improvement in Exact Match (EM) and F1 scores, respectively.

In this study, the performance of Transformer based pre-trained language models for QA tasks is analyzed. mBERT, BERTurk Base (cased), BERTurk Base (uncased), BERTurk 128k (cased), BERTurk 128k (uncased), ConvBERTurk Base (cased), ConvBERTurk mC4 (cased), ConvBERTurk mC4 (uncased), DistilBERTurk Base (cased), ELECTRA TR Base (cased), ELECTRA TR Small (cased), ELECTRA TR mC4 (cased) and ELECTRA TR mC4 (uncased) pre-trained language models are fine-tuned using a comprehensive dataset, SQuAD-TR. The performance of the fine-tuned models on QA tasks is compared using XQuAD-TR dataset by evaluating them with metrics such as EM and F1 score. The findings of this study contribute to the growing body of research on Turkish QA systems by providing a comprehensive evaluation of Transformer-based pre-trained language models. In particular, adapting pre-trained language models based on Transformer architecture to Turkish poses many challenges due to both morphological richness and semantic parsing difficulties. In order to overcome these challenges, this paper comprehensively evaluates and performs performance analyses of different configurations of Turkish-specific models. This analysis not only identifies the most effective models for Turkish QA but also underscores the potential of Transformer architectures in addressing linguistic challenges unique to Turkish. The insights gained from this study pave the way for future research aimed at enhancing NLP tools for Turkish and other low-resource languages, ultimately advancing the development of robust QA systems.

2. Materials and Methods

2.1. Datasets

In this study, SQuAD-TR dataset is used for qa task. SQuAD-TR dataset is created by Budur et al. by translating original SQuAD 2.0 dataset from English to Turkish using machine translation (Budur et al., 2024). In addition, the performances of fine-tuned models are compared using XQuAD-TR dataset. This dataset is a Turkish QA dataset included in XQuAD (Cross-lingual Question-Answering Dataset) which was created to evaluate QA performance across different languages (Artetxe et al., 2019). Table 1 shows sample data from SQuAD-TR and XQuAD-TR datasets.

Datasets contain paragraphs from various articles and questions and answers related to these paragraphs. The answers to the questions consist of a part of the text in the paragraph. The starting positions of these answers within the paragraph are included in the datasets.

Table 1. Sample data from QA datasets

	SQuAD-TR	XQuAD-TR
Paragraph	“Beyoncé Giselle Knowles-Carter (d. 4 Eylül 1981), ABD'li şarkıcı, söz yazarı, prodüktör ve aktris. Houston, Teksas'ta doğup büyüdü, çocukken çeşitli şarkı ve dans yarışmalarında sahne aldı ve 1990'ların sonlarında R&B kız grubu Destiny's Child'ın solisti olarak ün kazandı. …”	“230.000$ bütçeyle, Apollo 11'den kalan orijinal ay yayın verileri, Nafzgerand tarafından derlendi ve restorasyon için Lowry Digital görevlendirildi. Video, tarihi meşruiyeti bozmadan, rastgele gürültüyü ve kamera sarsıntısını gidermek için işlendi. …”
Question	*“Beyonce ne zaman popüler olmaya başladı?”*	*“Kalan orijinal Apollo 11 iniş verilerini kim derledi?”*
Answer	*“1990'ların sonlarında”*	*“Nafzgerand”*

The SQuAD-TR dataset contains 113,082 questions, 63,639 answerable and 49,443 unanswerable, related to 19,980 Turkish paragraphs. Preprocessing steps were implemented to ensure the quality and suitability of the datasets for experiments focusing only on answerable questions. First, cases where the answers provided in the dataset did not match the content of the relevant paragraphs were identified and removed. Additionally, unanswerable questions were excluded to limit the scope of the analysis to answerable scenarios. These preprocessing steps aimed to eliminate inconsistencies and increase the overall reliability of the datasets for evaluating the QA model. The size and scope of the dataset provides a solid basis for comparing and evaluating the performance of models on QA tasks. Furthermore, the XQuAD-TR dataset contains 1,190 questions from 240 Turkish paragraphs. The XQuAD-TR dataset contains 1,190 questions from 240 Turkish paragraphs. After the preprocessing steps, the numbers of paragraphs and questions in the SQuAD-TR dataset and XQuAD-TR dataset are shown in Table 2.

Table 2. Datasets

	Paragraph	Question
SQuAD-TR	18273	60797
XQuAD-TR	240	1190

2.2. Methods

In the study, Transformer-based pre-trained language models were used. Transformer architecture has led to significant developments in NLP and is used for many NLP tasks (Acheampong et al., 2021). The architecture consists of encoder and decoder blocks. The encoder block takes text data and creates a representation of the text, while the decoder block uses the representations to create the target text. Positional coding takes into account the position of each word in the sentence. The architecture draws its power from the attention mechanism. The attention mechanism focuses on each word in the text and tries to understand the relationships between these words. Softmax is used to normalize the outputs (Vaswani et al., 2017). By using Transformer architecture, effective models can be created for NLP tasks by utilizing large text corpora in different languages. In this study, Transformer-based pre-trained language models developed for NLP tasks were used.

2.2.1. BERT

BERT is a pre-trained language model based on Transformer. BooksCorpus and English Wikipedia corpora were used in the pre-training phase. The model uses Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) techniques. The MLM technique randomly masks a section of the text. An attempt is made to predict the masked words correctly, and the semantic relationships of these words with the words before and after them are analyzed. For the NSP technique, a sentence and its direct continuation or random sentences from the corpus are selected. Then, it is predicted whether the selected sentences are consecutive or not. This technique is used to understand the relationships between sentences in the text (Devlin et al., 2018). The NSP and the MLM techniques of the BERT model are shown in figure 1.

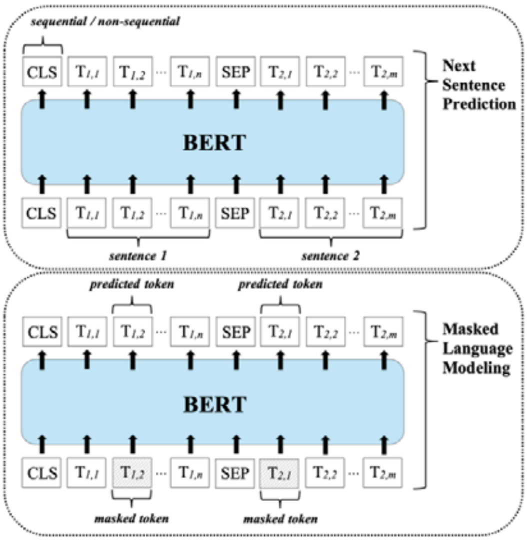

Figure 1. NSP and MLM techniques of the BERT (Devlin et al., 2018).

Thanks to its advanced natural language capabilities, the model can work effectively on various NLP tasks such as QA. For these tasks, the model is fine-tuned using relevant datasets, allowing it to be optimized for different purposes. Multilingual and Turkish customized versions of the BERT model are used in this study.

mBERT: The mBERT model is a multilingual version of the BERT model. Multilingual models can transfer learning from one language to another. Thus, tasks in low-resource languages such as Turkish can be performed using learning from different languages. The model has been trained using texts in 104 languages and can be used for NLP tasks in different languages, including Turkish.

BERTurk: BERTurk variations are BERT models specifically adapted for Turkish NLP tasks. The BERTurk model was trained on 35 GB of Turkish text data from the Wikipedia, OPUS and OSCAR corpus. There are case-sensitive "cased" and case-insensitive "uncased" versions of the models that can handle 32,000 and 128,000 unique words (Schweter, 2020).

ConvBERTurk: ConvBERT model structure aims to improve the performance and efficiency of BERT model by combining the capabilities of convolutional layers with the attention mechanism (Jiang et al., 2020). There are variations of the model specially adapted for Turkish. ConvBERTurk (cased) is a case-sensitive model trained using Turkish texts. In addition, ConvBERTurk mC4 (cased) and ConvBERTurk mC4 (uncased) variations were trained using Turkish texts from multilingual C4 corpus containing texts in many languages.

DistilBERTurk: DistilBERTurk is a faster and lighter variation of the BERT model, specifically adapted for the Turkish language. Using the distillation method, the learned knowledge from the base model is transferred from a larger teacher model to a smaller student model (Sanh et al., 2019). Thus, the model provides faster and more efficient results than the original model.

2.2.2. ELECTRA

ELECTRA is a Transformer based model that uses an approach to detect randomly changed words in the text. The architecture uses a Generator and a Discriminator. Random words are selected from the text and masked. For each masked word, the Generator generates the sentence with meaningful possible values according to the structure of the sentence. The Discriminator evaluates whether the words in the sentence generated by the Generator are the words in the original sentence (Clark et al., 2020). The structure of the language is understood in depth by using the Generator and the Discriminator. Figure 2 shows the structure of ELECTRA.

ELECTRA TR: There are different variations of the ELECTRA model specifically adapted for Turkish NLP tasks. ELECTRA Turkish Base (cased) is a case-sensitive model trained on Turkish texts, while ELECTRA TR Small (cased) is a more efficient version with fewer parameters. Both models were trained using the same data as the BERTurk model (Savci and Das, 2023). Additionally, there are both cased and uncased variations of the model trained on the Turkish part of the multilingual C4 corpus

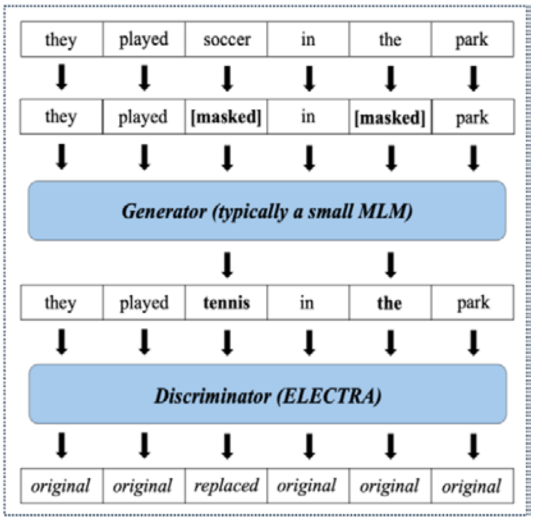

Figure 2. The ELECTRA architecture (Clark et al., 2020).

2.3. Performance Metrics

EM and F1 Score metrics were used to evaluate the performance of the models. EM is a metric based on the proportion of the model's predictions that exactly match the actual responses in the test dataset. This metric is an important and strict metric for measuring the model's ability to give correct answers that are exact matches. It has been frequently used in the literature to analyze the performance of models in QA tasks. The formula used to calculate the EM metric is given in equation 1.

$$EM\ Rate = \frac{Number\ of\ Correctly\ Answered\ Questions}{Total\ Number\ of\ Questions} \tag{1}$$

The F1 score is another commonly used metric for evaluating performance in QA tasks. It provides a balance between precision and recall, offering a single metric that reflects the accuracy of the predicted answers and their coverage of the correct answers. To calculate the F1 score, Precision and Recall results of the models are required. For these metrics, it is necessary to know the true positive (TP), false positive (FP) and false negative (FN) predictions of the model. TP refers to the number of tokens common between the correct answer and the model's prediction. FP refers to the number of tokens included in the model's prediction but not in the correct answer. FN refers to the number of tokens included in the correct answer but not in the model's prediction. The formula for the Precision metric is given in equation 2 and the formula for the Recall metric is given in equation 3.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP\ + FN} \tag{3}$$

In the study, the F1 Score used to evaluate the performance of QA models was calculated using the formula given in equation 4.

$$F1\ Score\ =\ 2\ x\ \frac{Precision\ x\ Recall}{Precision\ +\ Recall} \tag{4}$$

3. Results and Discussion

In this study, mBERT Base (cased), BERTurk Base (cased), BERTurk Base (uncased), BERTurk 128k (cased), BERTurk 128k (uncased), ConvBERTurk Base (cased), ConvBERTurk mC4 (cased), ConvBERTurk mC4 (uncased), DistilBERTurk Base (cased), ELECTRA TR Base (cased), ELECTRA TR Small (cased), ELECTRA TR mC4 (cased) and ELECTRA TR mC4 (uncased) pre-trained language models were fine-tuned using the SQuAD-TR dataset. The same parameter values were used in the fine-tuning process to ensure that the models could be evaluated under the same conditions. The parameter values used for training the models are given in Table 3.

Table 3. Fine-tuning parameters

Parameter	Value
Train Batch Size	16
Learning Rate	3e^{-5}
Maximum Length	512
Stride	256
Epoch	3

The training was performed with an NVIDIA Tesla V100 GPU using the Python programming language and the Transformers library. The fine-tuned models were then evaluated with the XQuAD-TR dataset. The performance of the models was evaluated using the EM and F1 Score. In the performance evaluations performed after finetuning the models, the ConvBERTurk Base (cased) model achieved the highest performance among all models with 57.82% EM and 71.59% F1 Score. As a result of the fine-tuning process, the mBERT Base (cased) model achieved 52.61% EM and 66.84% F1 Score. BERTurk 128k (cased) model showed the best performance among BERTurk variations with 57.06% EM and 71.17% F1 Score. DistilBERTurk Base (cased) model showed poor performance in Turkish QA tasks with 41.26% EM and 55.75% F1 Score. Among the ELECTRA Turkish models, the ELECTRA TR Base (cased) model achieved the best results with 57.31% EM and 71.39% F1 Score.

Among all models, the ConvBERTurk Base (cased) model performed the best, while the ELECTRA TR Small (cased) model showed the worst performance. This demonstrates that model architecture and parameter configuration significantly influence the effectiveness of pre-trained language models, particularly in complex QA tasks. Additionally, cased variations of the models were generally found to be more successful than their uncased variants. This can be attributed to the ability of cased models to better capture the semantic and syntactic intricacies of Turkish, where capitalization often conveys critical meaning.

The performance results of the fine-tuned models are presented in Table 4.

Table 4. Results of the fine-tuned models

Model	EM (%)	F1 Score (%)
mBERT Base (cased)	52.61	66.84
BERTurk Base (cased)	55.29	70.07
BERTurk Base (uncased)	52.44	67.68
BERTurk 128k (cased)	57.06	71.17
BERTurk 128k (uncased)	53.70	69.09
ConvBERTurk Base (cased)	57.82	71.59
ConvBERTurk mC4 (cased)	56.39	69.56
ConvBERTurk mC4 (uncased)	55.71	70.79
DistilBERTurk Base (cased)	41.26	55.75
ELECTRA TR Base (cased)	57.31	71.39
ELECTRA TR Small (cased)	41.18	55.19
ELECTRA TR mC4 (cased)	56.64	70.63
ELECTRA TR mC4 (uncased)	55.21	69.43

Cased models were found to be more successful than uncased models because they were able to effectively discriminate important language elements, such as proper names and abbreviations, by better managing the case sensitivity of the language. This advantage is particularly significant in Turkish, a language where orthographic case often serves as a marker of named entities, formal terms, and other contextually important elements. In particular, the ConvBERTurk Base (cased) model appears to perform best by successfully combining attentional mechanisms with convolutional layers. This architecture not only enhances the model’s ability to focus on relevant portions of text but also allows it to extract features hierarchically, improving contextual understanding. In contrast, smaller models with smaller dimensions and fewer parameters showed lower performance in language comprehension and correct response extraction. Such limitations highlight the trade-off between computational efficiency and the ability to handle linguistically rich and context-dependent tasks. The fact that the mBERT model adapted to multilingual NLP tasks was not as successful as the models adapted specifically for Turkish Language, emphasizes the importance of language-specific adaptations. This finding underscores the critical need to design NLP models tailored to the unique linguistic and syntactic properties of target languages, especially for low-resource contexts. Performance comparisons of all fine-tuned models are given in Figure 3.

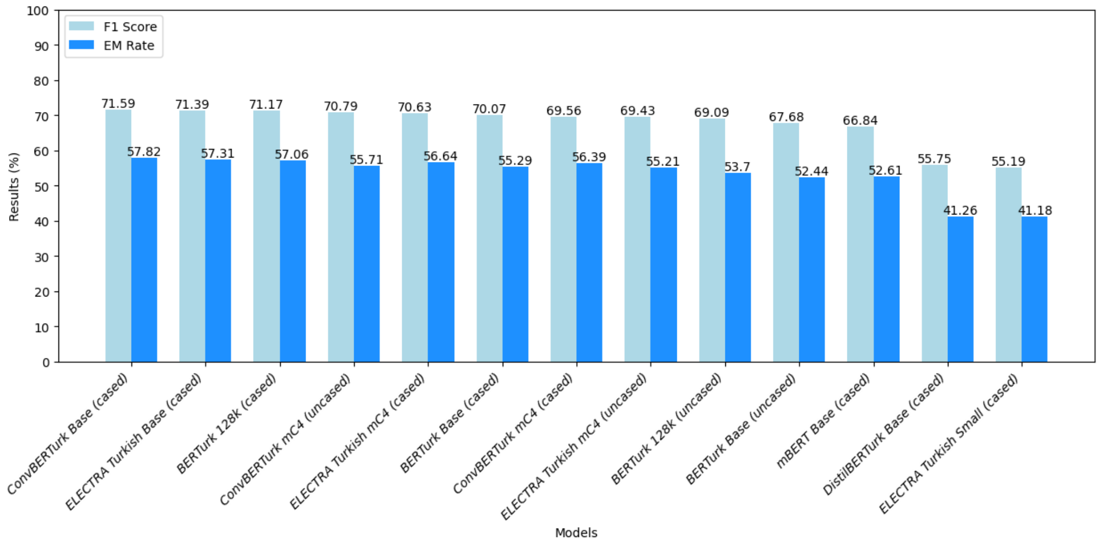

Figure 3. Performance comparisons of the fine-tuned models

4. Conclusion

This study investigated the performances of Transformer-based pre-trained language models for Turkish QA tasks. Different variations of the pre-trained language models were fine-tuned for answer extraction from Turkish texts using the SQuAD-TR dataset. As a result of the fine-tuning process, the performances of the models were tested and evaluated on the XQuAD-TR dataset. Among all models, the ConvBERTurk Base (cased) model performed the best, while the ELECTRA TR Small (cased) model showed the worst performance. Additionally, cased variations of the models were generally found to be more successful than their uncased variants. The results of this study clearly demonstrate the performance differences of Transformer-based models in Turkish QA tasks. Cased models were found to be more successful than uncased models because they were able to effectively discriminate important language elements, such as proper names and abbreviations, by better managing the case sensitivity of the language. In particular, the ConvBERTurk Base (cased) model appears to perform best by successfully combining attentional mechanisms with convolutional layers. In contrast, smaller models with smaller dimensions and fewer parameters showed lower performance in language comprehension and correct response extraction. The fact that the mBERT model adapted to multilingual NLP tasks was not as successful as the models adapted specifically for Turkish emphasizes the importance of language-specific adaptations. This research contributes to the development of efficient and accurate QA systems for low-resource languages like Turkish, paving the way for further advancements in NLP for diverse languages.

Author Contributions

The contribution percentages of the authors are given below. All authors have reviewed and approved the manuscript.

	M.İ.	M.A.
C	40	60
D	40	60
S	40	60
DCP	80	20
DAI	80	20
LS	80	20
W	60	40
CR	20	80
SR	40	60

C= concept, D= design, S= supervision, DCP= data collection, and/or processing, DAI= data analysis and/or interpretation, LS= literature search, W= writing, CR= critical review, SR= submission and revision.

Conflict of Interest

The authors declared that there is no conflict of interest.

Ethical Approval Declaration

Ethics committee approval was not required for this study because of there was no study on animals or humans.

References

Acheampong FA, Nunoo-Mensah H, Chen W. 2021. Transformer models for text-based emotion detection: a review of BERT-based approaches. Artif Intell Rev, 54(8): 5789-5829.

Akyon FC, Cavusoglu D, Cengiz C, Altinuc SO, Temizel A. 2021. Automated question generation and question answering from Turkish texts. arXiv preprint arXiv:2111.06476.

Allam AMN, Haggag MH. 2012. The question answering systems: a survey. Int J Res Rev Inf Sci (IJRRIS), 2(3).

Alzubi JA, Jain R, Singh A, Parwekar P, Gupta M. 2023. COBERT: COVID-19 question answering system using BERT. Arab J Sci

Eng, 48(8): 11003-11013.

Artetxe M, Ruder S, Yogatama D. 2019. On the cross-lingual transferability of monolingual representations. arXiv preprint arXiv:1910.11856.

Arzu M, Aydoğan M. 2023. Türkçe duygu sınıflandırma için transformers tabanlı mimarilerin karşılaştırmalı analizi. Comput Sci, 2023: 1-6.

Budur E, Özçelik R, Soylu D, Khattab O, Güngör T, Potts C. 2024. Building efficient and effective OpenQA systems for low-resource languages. arXiv preprint arXiv:2401.03590.

Clark K, Luong MT, Le QV, Manning CD. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555.

Çelikten A, Bulut H. 2021. Turkish medical text classification using BERT. 29th Signal Processing and Communications Applications Conference (SIU), June 9-11, İstanbul, Türkiye, pp: 1-4.

Devlin J, Chang MW, Lee K, Toutanova K. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Gemirter CB, Goularas D. 2021. A Turkish question answering system based on deep learning neural networks. J Intell Syst Theory Appl, 4(2): 65-75.

Hassani H, Beneki C, Unger S, Mazinani MT, Yeganegi MR. 2020. Text mining in big data analytics. Big Data Cogn Comput, 4(1): 1.

Jiang Z, Yu W, Zhou D, Chen Y, Feng J, Yan S. 2020. ConvBERT: improving BERT with span-based dynamic convolution. Adv Neural Inf Process Syst, 33: 12837-12848.

Khurana D, Koli A, Khatter K, Singh S. 2023. Natural language processing: state of the art, current trends and challenges. Multimed Tools Appl, 82(3): 3713-3744.

Locke S, Bashall A, Al-Adely S, Moore J, Wilson A, Kitchen GB. 2021. Natural language processing in medicine: a review. Trends Anaesth Crit Care, 38: 4-9.

Rajpurkar P, Jia R, Liang P. 2018. Know what you don't know: unanswerable questions for SQuAD. arXiv preprint arXiv:1806.03822.

Rajpurkar P, Zhang J, Lopyrev K, Liang P. 2016. SQuAD: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250.

Raza S, Schwartz B, Rosella LC. 2022. CoQUAD: A COVID-19 question answering dataset system, facilitating research, benchmarking, and practice. BMC Bioinfo, 23(1): 210.

Sanh V, Debut L, Chaumond J, Wolf T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01106.

Savci P, Das B. 2023. Comparison of pre-trained language models in terms of carbon emissions, time and accuracy in multi-label text classification using AutoML. Heliyon, 9(5).

Schweter S. 2020. BERTurk-BERT models for Turkish. Zenodo, 2020: 3770924.

Soygazi F, Çiftçi O, Kök U, Cengiz S. 2021. THQuAD: Turkish historic question answering dataset for reading comprehension. 6th International Conference on Computer Science and Engineering (UBMK), September 15-17, Ankara, Türkiye, pp: 215-220.

Türkmen H, Dikenelli O, Eraslan C, Callı MC, Özbek SS. 2023. BioBERTurk: exploring Turkish biomedical language model development strategies in low-resource setting. J Healthc Inform Res, 7(4): 433-446.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. 2017. Attention is all you need. Adv Neural Inf Process Syst, 2017: 30.

Yiğit G, Amasyalı F. 2021. Soru cevaplama sistemleri üzerine detaylı bir çalışma: veri kümeleri, yöntemler ve açık araştırma alanları. Bilisim Teknol Derg, 14(3): 239-254.

Zhu P, Yuan Y, Chen L. 2023. ELECTRA-based graph network model for multi-hop question answering. J Intell Inf Syst, 61(3): 819-834.