Research Article

Comparison of Transformer-Based Turkish Models for Question-Answering Task

Mehmet Arzu and Murat Aydogan

***Abstract*— Question-answering systems facilitate information access processes by providing fast and accurate answers to questions expressed by users in natural language. Today, advances in Natural Language Processing (NLP) techniques increase the effectiveness of such systems and improve the user experience. However, in order for these systems to work effectively, the structural features of the language must be properly understood. Traditional rule-based and knowledge retrieval-based systems cannot analyze the contextual meaning of questions and texts deeply enough and therefore cannot produce satisfactory answers to complex questions. For this reason, Transformer-based models that can better capture the contextual and semantic integrity of the language have been developed. This study aims to evaluate the performance of Transformer-based models on Turkish question-answering tasks. In this context, a new dataset created by combining THQuAD (Turkish Historic Question Answering Dataset) on Turkish Islamic History of Science and Ottoman History with BQuAD (Biology Question Answering Dataset) consisting of topics in biology course was used. On this dataset, the performances of BERTurk, ELECTRA Turkish and DistilBERTurk models for Turkish question-answering tasks were compared by fine-tuning under the same hyperparameters and the results were evaluated. According to the findings, higher Exact Match (EM) and F1 scores were obtained in models with case sensitivity; the best performance was obtained in BERTurk (Cased, 128k) model with 63.99% EM and 80.84% F1 scores. These findings reveal the effectiveness of Transformer-based models in Turkish question-answering tasks and especially the performance superiority of models with case sensitivity.**

Index Terms—**Natural Language Processing, Question-Answering System, Transformer, BERTurk**

I. Introduction

WITH THE rapid advancement of technology, it has become necessary to make natural language understandable and processable by machines. This requirement paved the way for the emergence of the field of Natural Language Processing (NLP). NLP is a sub-discipline of artificial intelligence that enables computers to understand, analyze and process human language and use it in various applications [1]. NLP techniques are widely used especially in areas such as processing, analyzing and interpreting written texts [2]. These techniques are effectively used in different fields such as text analysis, sentiment analysis, machine translation and information extraction [3].

Question-answering systems, one of the application areas of NLP, are systems that aim to produce accurate and fast answers to the questions asked by users. These systems aim to provide appropriate answers to questions by analyzing the information obtained from large data sets [4]. QA systems enable users to find information in texts more quickly and thus offer significant advantages in many fields such as education, health, and scientific research [5], [6]. NLP processes in suffixed languages such as Turkish have some difficulties due to the structural features of the language. The grammatical structure of Turkish, derivation of words with roots and affixes, and syntactic structure can make natural language processing processes complex [7]. This has led to limited research on Turkish compared to research on common languages such as English. However, in recent years, technological advances in the field of NLP have enabled successful results to be obtained on languages such as Turkish. In particular, Transformer [8]-based models offer an important solution for processing languages such as Turkish with the ability to analyze contextual meaning in more depth. In this context, this study compares Transformer-based models on Turkish texts for a question-answer task and analyzes the experimental results.

There are various studies in the literature on the development of question-answering systems with NLP techniques. These studies have been carried out using different methods and techniques due to the grammatical features and complex structure of different languages. Research topics cover a wide range of areas such as text semantic analysis, word and sentence processing techniques, contextual semantic inference, and the use of machine learning and deep learning models. Some of the important and influential works in this field are as follows: Incidelen and Aydogan [9] created a dataset for QA tasks in medical texts for Turkish, a low-resource language, and fine-tuned the BERTurk model. As a result, the BERTurk (cased, 128k) model showed the best performance with an Exact Match (EM) score of 55.121 and F1 score of 77.187. These findings demonstrate that QA tasks can be successfully implemented in low-resource languages and provide an important foundation for Turkish in the field of medical text processing. Ozkurt [10] evaluated the performance of BERT, DistilBERT, RoBERTa and ALBERT models in text-based QA system on the SQuAD

Mehmet Arzu, is with Department of Artificial Intelligence and Data Engineering of Fırat University, Elazig, Turkey, (e-mail: marzu@firat.edu.tr).

https://orcid.org/0000-0001-6610-2788

Murat Aydogan, is with Department of Software Engineering of Fırat University, Elazig, Turkey, (e-mail: m.aydogan@firat.edu.tr).

https://orcid.org/0000-0002-6876-6454

Manuscript received Oct 31, 2024; accepted Dec 11, 2024.
DOI: 10.17694/bajece.1576976

v2 dataset. The ALBERT model achieved the best performance with 86.85% EM and 89.91% F1 score, while BERT, RoBERTa and DistilBERT models showed lower performance. Soygazi et al. [11] presented THQuAD, a Turkish historical question-answer dataset. This dataset consists of passages from Wikipedia articles on Ottoman history and the history of Turkish Islamic Science, and question-answer pairs based on these passages. In the study, they conducted experiments on the dataset using the pre-trained language models BERT, ELECTRA and ALBERT and evaluated the performance of the models with F1 score and EM metrics. According to their findings, they found that the ELECTRA model showed the highest performance with 63.08% EM and 81.55% F1 score. Ugurlu et al. [12] developed a virtual assistant that can answer questions about COVID-19, mainly in the field of health, in order to provide access to reliable information during the COVID-19 pandemic period. As a pioneering work in the Turkish health field, this assistant contributes to the field of society and health with natural language processing techniques. Unlu and Cetin [13] emphasized the impact and importance of deep learning methods on NLP and question answering systems. They underlined that deep learning methods such as text analysis, meaning extraction, keyword extraction significantly affect the performance of question-answering systems. They emphasized the importance of deep learning techniques in reaching the right information in large data sets. With their study, they argued that deep learning techniques are effective in NLP and question-answer systems and provide many advantages in areas such as keyword extraction. Amasyali and Diri [14] developed a question answering system called “BayBilmiş” using NLP techniques. They used NLP techniques such as finding word roots, identifying words according to their types and semantic analysis to understand the questions from users. By searching databases, the most accurate answers were obtained by information extraction according to the content of the questions. The answers obtained were tested with NLP algorithms to ensure reliable answers. A number of techniques were used to make the answers understandable. The system they designed provided high success in question answering processes by using a large database. Gemirter et al. [15] addressed the difficulties in the field of NLP in languages that do not have a simple structure such as Turkish. They developed a question-answering system that can give correct answers on documents in the banking sector. Using large data sets in their system, they used the BERT model and then optimized this model by fine-tuning it. In order to overcome the difficulties encountered due to the complexity of Turkish, they conducted a number of investigations on translated data sets. Mukanova et al. [16] aimed to develop a geographical QA system for the Kazakh language. In this context, they created a dataset of 50,000 question-answer pairs related to Kazakhstan geography and tested the system by conducting experiments on this dataset. The performance of the model was evaluated with BLEU and F1 score metrics, and an average BLEU score of 95.76% and F1 score of 95.8% were obtained. Staš et al. [17] developed a question-answer dataset based on machine translation for the Slovak language. The English SQuAD v2.0 dataset was translated into Slovak and experimental studies were conducted on this dataset by fine-tuning the SlovakBERT and mBERT models. The performance of the models was tested on two different datasets: machine translated and manually labeled. According to the findings, it was observed that the hand-labeled dataset provided higher performance. In particular, the mBERT model achieved EM and F1 scores of 56.02% and 63.02% on the machine-translated dataset, while these scores were 69.48% EM and 78.87% F1 on the hand-labeled dataset, respectively. In another study, Rajpurkar et al. [18] considered the Stanford Question Answering Dataset (SQuAD) v2.0. In addition to the question-answer mappings from SQuAD v1.1, SQuAD v2.0 has approximately 50,000 unanswered questions generated by participants, and when they trained and tested the DocQA and ELMo model on SQuAD v2.0, they obtained an F1 score of 66.3%. As a result of their results, they argue that there is a significant difference between humans and machines in SQuAD v2.0 compared to SQuAD 1.1. They also explained that SQuAD v2.0 is a significantly more challenging dataset for existing models

II. TRANFORMERS ARCHITECTURE

Rapidly developing technologies in the field of artificial intelligence and NLP are continuously improving text processing and comprehension capabilities and require new approaches to tackle more complex tasks. At this point, the so-called “Transformer” architecture has attracted a great deal of attention in a short period of time. Transformer is considered to be a revolutionary innovation in the field of language processing, enabling impressive results in a variety of application areas.

The concept of transformer was first introduced in the article “Attention is all you need” [8]. This paper states that a transformer model is a model that uses self-attention and multi-headed attention methods to find the relationship of a word given as input to other words in both the forward and backward directions. In this way, it can find the relation of a word to other words without the need for models such as RNN, which uses sequentially aligned data, or CNN, which performs convolutional operations in the middle layers.

Transformer is a deep learning model used in NLP. Unlike traditional language models, it uses purely attentional mechanisms to identify word relationships in language, without relying on preceding and following word order. This allows it to capture long-distance connections more efficiently and provide better performance in language processing tasks [19].

Figure 1 shows the overall architecture of the Transformer model. The Transformer model consists of two main components, the encoder and the decoder. The encoder, located on the left side of the figure, processes and encodes the input data and generates meaningful representation vectors from this data. The decoder, on the right-hand side, uses these representation vectors to produce the target outputs. The encoder and decoder components are supported by various structures such as multi-head attention mechanisms, feed-forward layers and positional coding. Working together, these components enable the model to take input data and process it through various processes to generate the final outputs.

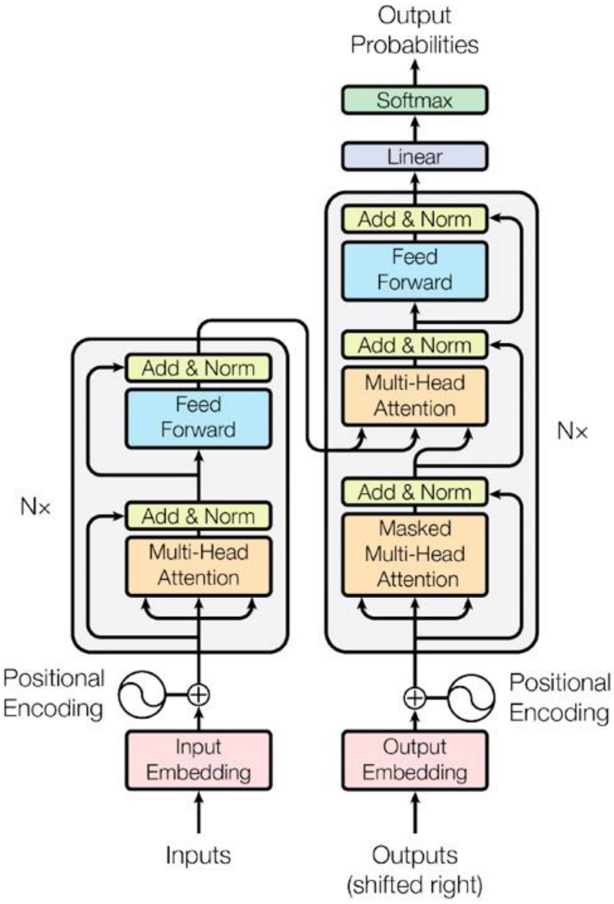

Fig.1. Transformer Architecture [8]

III. MATERIALS AND METHODS

A. Dataset

The dataset used in the experiments is a dataset obtained by combining THQuAD (Turkish Historic Question Answering Dataset) [20] prepared by Kokcu et al. and BQuAD (Biology Question Answering Dataset) [21] prepared by Akyon et al. THQuAD contains texts on the history of Turkish Islamic Science and Ottoman history. BQuAD, on the other hand, contains texts, questions and answers compiled from high school 1st, 2nd, 3rd and 4th grade biology textbooks published by the Turkish Ministry of National Education. BQuAD, which is designed for training biology-based QA, question generation and answer generation models, is organized in the widely used SQuAD [22] format.

If the data sets have different structures, it is inevitable that the structure of the code will change and become more complex after merging. This can lead to errors and complicate the process of training the model. Since THQuAD and BQuAD have different structures, the two datasets were converted to the same data format. By combining THQuAD and BQuAD, it was aimed to obtain a data set consisting of more data. Details about the data set are given in Table I

TABLE I
PARAGRAPH AND QUESTION-ANSWER COUNTS IN TRAINING AND TEST DATASETS

	Paragraph	**Question-Answer Pair**
Training Dataset	2554	14839
Test Dataset	341	1483
Total Data	2895	16322

As shown in Table I, the training dataset contains 2554 paragraphs and 14839 question-answer pairs, while the test dataset has a smaller volume of 341 paragraphs and 1483 question-answer pairs. Overall, this combined dataset contains 2,895 paragraphs and 16,322 question-answer pairs.

The dataset is organized in JSON (JavaScript Object Notation) format, reflecting a typical structure used for question-answering systems. This structure is in SQuAD format, which is widely used for question-answering systems. A sample paragraph and question-answer pairs from the dataset are given in Figure 2.

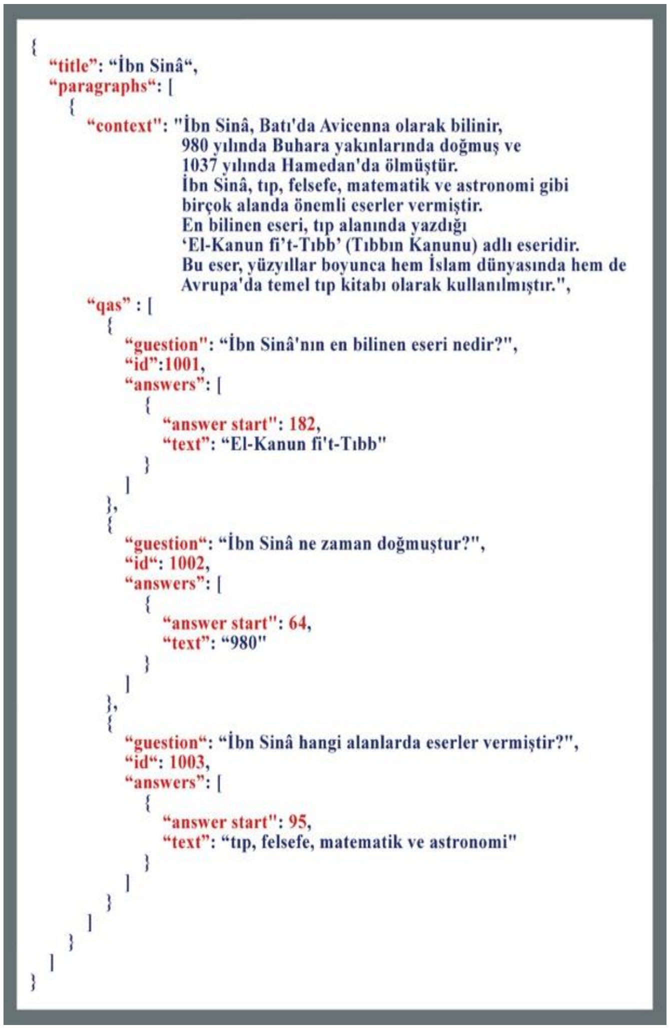
```
{
  “title”: “İbn Sinâ“,
  “paragraphs“: [
    {
      “context": "İbn Sinâ, Batı'da Avicenna olarak bilinir,
                  980 yılında Buhara yakınlarında doğmuş ve
                  1037 yılında Hamedan'da ölmüştür.
                  İbn Sinâ, tıp, felsefe, matematik ve astronomi gibi
                  birçok alanda önemli eserler vermiştir.
                  En bilinen eseri, tıp alanında yazdığı
                  ‘El-Kanun fi’t-Tıbb’ (Tıbbın Kanunu) adlı eseridir.
                  Bu eser, yüzyıllar boyunca hem İslam dünyasında hem de
                  Avrupa'da temel tıp kitabı olarak kullanılmıştır.",
      “qas” : [
        {
          “guestion": “İbn Sinâ'nın en bilinen eseri nedir?",
          “id”:1001,
          “answers”: [
            {
              “answer start": 182,
              “text”: “El-Kanun fi't-Tıbb"
            }
          ]
        },
        {
          “guestion“: “İbn Sinâ ne zaman doğmuştur?",
          “id“: 1002,
          “answers”: [
            {
              “answer start": 64,
              “text”: “980"
            }
          ]
        },
        {
          “guestion“: “İbn Sinâ hangi alanlarda eserler vermiştir?",
          “id“: 1003,
          “answers”: [
            {
              “answer start": 95,
              “text”: “tıp, felsefe, matematik ve astronomi"
            }
          ]
        }
      ]
    }
  ]
}
```

Fig.2. A Sample Paragraph and Question-Answer Pairs from the Dataset

B. Methods

This section introduces the models used in the experiments performed. These models are customized variations of Transformer-based pre-trained language models for Turkish.

1) BERTurk

BERTurk [23] is a model based on the BERT[24] architecture and customized for Turkish language processing. This model was created by Stefan Schweter and is specifically designed to understand and process the rich linguistic structure of Turkish. BERTurk was trained using various data sources such as the Turkish OSCAR corpus, Wikipedia, OPUS corpora and a custom corpus provided by Kemal Oflazer. Variations of the BERTurk model are as follows:

BERTurk (Cased, 32k): This variation performs tokenization by distinguishing between uppercase and lowercase letters. That is, it takes into account uppercase and lowercase differences. For example, “Mehmet” and “mehmet” will have separate tokens. It has a vocabulary of 32 thousand tokens.

BERTurk (Uncased, 32k):This variant performs tokenization on Turkish texts without distinguishing between uppercase and lowercase letters. That is, it does not care about uppercase and lowercase. For example, “Mehmet” and “mehmet” will have the same token. Again, it has a vocabulary of 32 thousand tokens.

BERTurk (Cased, 128k): This variant performs tokenization by distinguishing between uppercase and lowercase letters in Turkish text and has a larger vocabulary. It contains 128 thousand tokens.

BERTurk (Uncased,128k):This model performs tokenization of Turkish text without distinguishing between uppercase and lowercase letters and has a large vocabulary of 128k tokens.

2) ELECTRA Turkish

ELECTRA [25] was developed to accelerate the training process and increase the efficiency of BERT and similar models. The main difference that distinguishes this model from BERT is its approach during the training phase. ELECTRA Turkish [26] is a model of ELECTRA developed for Turkish and is a model created by training on the same data as BERTurk. There are two different variations for Turkish:

ELECTRA Small [27]: This model is a small and fast model because it contains fewer parameters. It performs tokenization by distinguishing between uppercase and lowercase letters. A 35GB data set was used in the training process. Thanks to its small and lightweight structure, it is preferred in systems with limited resources and applications that require fast results.

ELECTRA Base [28]: ELECTRA Base model is a larger and more powerful model because it contains more parameters. This model also performs tokenization by distinguishing between uppercase and lowercase letters. A 35GB dataset was used in the training process. Thanks to its larger size, it is preferred for more complex applications that require high accuracy.

3) DistilBERTurk

DistilBERTurk [29] is a customized version of the DistilBERT [30] model for Turkish. DistilBERTurk is a lightweight and fast natural language processing model for Turkish. This model was trained using the cased version of BERTurk and 7GB of original training data. The training process was carried out by distillation, a technique of building a smaller model from a larger model.

C. Evaluation Metrics

Two different performance metrics, EM and F1score, were used to analyze the performance of the models. These 2 metrics are widely used in the literature for QA tasks.

EM is a widely used evaluation metric in QA systems. It is used to measure the similarity between the answer produced by a model and the correct answer [22]. The EM ratio checks whether the model's answer and the correct answer match exactly, i.e. word for word. If the model's answer is exactly the same as the correct answer, it is considered EM.

EM is important for applications that require precise accuracy because incorrect or missing information can lead to unintended consequences, especially in areas of critical importance (e.g. medical information systems or legal regulations). However, the main limitation of the EM metric is that it does not consider partial accuracies or contextual semantic matches. Therefore, it may not be sufficient to use this metric on its own, as it only provides a precise measure of accuracy; it is often combined with complementary metrics such as the F1 score.

The Number of Exact Matches refers to the sum of cases where the answers generated by the model match the correct answers exactly. The Total Number of Answers represents the total number of answers in the dataset. The ratio of the number of exactly matched answers to the total number of answers gives the EM. The equation for the EM ratio is shown in (1).

$$EM = \frac{Number\ of\ Exactly\ Matching\ Answer}{Total\ Number\ of\ Answer} \tag{1}$$

F1 Score is a widely used metric for evaluating the performance of a model, representing the harmonic mean of the precision and recall metrics.[31]. The F1 score is calculated with the formula given in Equation (2).

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

In this equation, precision and recall metrics are combined to provide a balanced measure that evaluates the overall performance of the model. Precision is the ratio of correctly predicted answers to total predicted answers. This metric is used to evaluate the prediction accuracy of the model by measuring how many of all the answers produced by the model are correct. Sensitivity is the ratio of correctly predicted answers to total correct answers in the dataset and measures the model's ability to find all correct answers.

The F1 score combines precision and sensitivity, expressing a model's performance on these two metrics in a single metric. It is used in this study as the model is expected to perform with both high accuracy and high coverage.

IV. Results and Discussion

Table II shows the performance results of BERTurk, ELECTRA Turkish and DistilBERTurk models with the same hyperparameter settings on the new dataset obtained by combining THQuAD and BQuAD.

Table II
PERFORMANCE RESULTS OF THE MODELS FOR THE SAME HYPER PARAMETERS VALUES

Model Name	**Hyper Parameters**				**Evaluation Metrics**	
	Training Batch Size	Test Batch Size	Epoch	Learning Rate	Exact Match (%)	F1 (%)
BERTurk (Cased, 32k)	16	32	5	3e-5	63.31	80.07
BERTurk (Uncased, 32k)	16	32	5	3e-5	42.21	64.49
BERTurk (Cased, 128k)	16	32	5	3e-5	**63.99**	**80.84**
BERTurk (Uncased, 128k)	16	32	5	3e-5	42.41	64.50
ELECTRA Turkish Base	16	32	5	3e-5	61.15	79.25
ELECTRA Turkish Small	16	32	5	3e-5	34.12	50.70
DistilBERTurk	16	32	5	3e-5	28.65	44.16

In this study, the performance of different Transformer-based language models for Turkish QA tasks was evaluated. The experimental results obtained show that the BERTurk models with casing achieved significantly higher EM and F1 scores than the variation without casing. The BERTurk (Cased, 128k) model achieved the highest performance with EM scores of 63.99% and F1 scores of 80.84%. This supports the conclusion that case sensitivity is semantically critical in a language like Turkish and improves contextual accuracy. This result shows that especially in Turkish, proper nouns, sentence beginnings and important concepts are distinguished by capitalization, which allows the cased models to better understand the context.

Comparisons with ELECTRA models show that the ELECTRA Turkish Base model also performs well, but still lags behind the BERTurk Cased models. While this demonstrates the effectiveness of ELECTRA, it also suggests that the Turkish-specific fine-tuning of the BERTurk models yields superior results. On the other hand, ELECTRA Turkish Small variation performed quite low with EM scores of 34.12% and F1 scores of 50.70%. This finding emphasizes the impact of model size and capacity on performance, and suggests that smaller and simpler models cannot adequately capture complex contextual relationships.

The DistilBERTurk model performed the worst with an EM score of 28.65% and an F1 score of 44.16%. This suggests that reducing the size of this DistilBERT-based model has a negative impact on accuracy and comprehensiveness. Especially in structurally complex languages such as Turkish, lower model capacities and small size models cause information loss and make it difficult to capture the semantic integrity of the text.

These findings emphasize the importance of case sensitivity, model size and vocabulary in model selection in Turkish QA tasks. In particular, large and cased models showed higher performance by better capturing contextual information. In this context, when developing high-performance QA systems in low-resource languages, the preference for cased structures may increase the model's capacity to extract contextual meaning. It was also concluded that the accuracy of low-dimensional and minimized models may be limited and that models trained with larger vocabularies may provide better results, especially in areas with sensitive information. These findings suggest that in Turkish natural language processing research, the preference for large, cased models is critical for effective QA systems.

V. CONCLUSION

This study aims to analyze the experimental results by comparing the performance of Transformer-based language models for QA task on Turkish texts. In this direction, the language models developed for Turkish were fine-tuned with the same hyper parameters using the dataset obtained by merging the THQuAD and BQuAD datasets and the performance of the models was evaluated. Various experiments were conducted with this merged dataset, and the highest EM rate of 63.99% and the highest F1 score of 80.84% were obtained for the BERTurk (Cased, 128k) model.

The reasons why the BERTurk (Cased, 128k) model gives the best results can be explained in terms of the structural features of the Turkish language and the technical advantages of the model. Since Turkish is a case-sensitive language, capitalization of proper nouns, sentence beginnings and important concepts provides a more accurate understanding of the context. This makes it possible for case-sensitive models to better capture contextual meaning. In particular, the BERTurk (Cased, 128k) model analyzes the semantic relations and contexts of words more accurately by taking capitalization into account. Moreover, the model's large vocabulary of 128k allows it to better understand the diversity of word derivations in languages with an agglutinative structure such as Turkish. Since Turkish is a language that is open to word derivation using many different affixes, models with large vocabularies can capture these structural features of the language more effectively. This contributes to increased contextual accuracy.

As a result of the experiments, it was observed that models with case sensitivity generally achieved higher EM and F1 scores. In addition, the number of parameters of the models has a significant impact on performance. Models with more parameters exhibited higher performance. Especially the BERTurk cased models showed superior performance thanks to these features. These findings show that case sensitivity, number of parameters and hyperparameter optimization are critical in the development of Turkish text processing and QA systems.

This study makes valuable contributions to the literature on the development of question-answering systems for Turkish by demonstrating the applicability and effectiveness of question-answering in Turkish. The results obtained will guide future research and applications in this field. The study encourages further research in the field of question-answering for Turkish and shows that high-performance systems can be developed despite the unique challenges of the language.

References

[1] D. Khurana, A. Koli, K. Khatter, and S. Singh, ‘Natural language processing: state of the art, current trends and challenges’, Multimed. Tools Appl., vol. 82, no. 3, pp. 3713–3744, Jan. 2023, doi: 10.1007/s11042-022-13428-4.

[2] K. Crowston, E. E. Allen, and R. Heckman, ‘Using natural language processing technology for qualitative data analysis’, Int. J. Soc. Res. Methodol., vol. 15, no. 6, pp. 523–543, Nov. 2012, doi: 10.1080/13645579.2011.625764.

[3] M. Arzu and M. Aydoğan, ‘Türkçe Duygu Sınıflandırma İçin Transformers Tabanlı Mimarilerin Karşılaştırılmalı Analizi’, Comput. Sci., no. IDAP-2023, pp. 1–6, 2023.

[4] A. Allam and M. Haggag, ‘The Question Answering Systems: A Survey’, Int. J. Res. Rev. Inf. Sci., vol. 2, pp. 211–221, Sep. 2012.

[5] E. Mutabazi, J. Ni, G. Tang, and W. Cao, ‘A Review on Medical Textual Question Answering Systems Based on Deep Learning Approaches’, Appl. Sci., vol. 11, no. 12, Art. no. 12, Jan. 2021, doi: 10.3390/app11125456.

[6] V. Redhu, A. K. Singh, and M. Saravanan, ‘AI-Enhanced Learning Assistant Platform: An Advanced System for Q&A Generation from Provided Content, Answer Evaluation, Identification of Students’ Weak Areas, Recursive Testing for Strengthening Knowledge, Integrated Query Forum, and Expert Chat Support’, in 2024 2nd International Conference on Artificial Intelligence and Machine Learning Applications Theme: Healthcare and Internet of Things (AIMLA), Mar. 2024, pp. 1–6. doi: 10.1109/AIMLA59606.2024.10531533.

[7] K. Tohma and Y. Kutlu, ‘Challenges Encountered in Turkish Natural Language Processing Studies’, Nat. Eng. Sci., vol. 5, no. 3, Art. no. 3, Nov. 2020, doi: 10.28978/nesciences.833188.

[8] A. Vaswani et al., ‘Attention is All you Need’, in Advances in Neural Information Processing Systems, Curran Associates, Inc., 2017. Accessed: May 22, 2024. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[9] M. İncidelen and M. Aydoğan, ‘Developing Question-Answering Models in Low-Resource Languages: A Case Study on Turkish Medical Texts Using Transformer-Based Approaches’, in 2024 8th International Artificial Intelligence and Data Processing Symposium (IDAP), Sep. 2024, pp. 1–4. doi: 10.1109/IDAP64064.2024.10711128.

[10] C. Özkurt, Comparative Analysis of State-of-the-Art Q\&A Models: BERT, RoBERTa, DistilBERT, and ALBERT on SQuAD v2 Dataset. 2024. doi: 10.21203/rs.3.rs-3956898/v1.

[11] F. Soygazi, O. Çiftçi, U. Kök, and S. Cengiz, ‘THQuAD: Turkish Historic Question Answering Dataset for Reading Comprehension’, in 2021 6th International Conference on Computer Science and Engineering (UBMK), Sep. 2021, pp. 215–220. doi: 10.1109/UBMK52708.2021.9559013.

[12] Y. Uğurlu, M. Karabulut, and İ. Mayda, ‘A Smart Virtual Assistant Answering Questions About COVID-19’, in 2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Oct. 2020, pp. 1–6. doi: 10.1109/ISMSIT50672.2020.9254350.

[13] Ö. Ünlü and A. Çetin, ‘A Survey on Keyword and Key Phrase Extraction with Deep Learning’, in 2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Oct. 2019, pp. 1–6. doi: 10.1109/ISMSIT.2019.8932811.

[14] M. F. Amasyalı and B. Diri, ‘Bir Soru Cevaplama Sistemi: BayBilmiş’, Türkiye Bilişim Vakfı Bilgi. Bilim. Ve Mühendisliği Derg., vol. 1, no. 1, Art. no. 1, Jun. 2016.

[15] C. B. Gemirter and D. Goularas, ‘A Turkish Question Answering System Based on Deep Learning Neural Networks’, J. Intell. Syst. Theory Appl., vol. 4, no. 2, Art. no. 2, Sep. 2021, doi: 10.38016/jista.815823.

[16] A. Mukanova, A. Barlybayev, A. Nazyrova, L. Kussepova, B. Matkarimov, and G. Abdikalyk, ‘Development of a Geographical Question- Answering System in the Kazakh Language’, IEEE Access, vol. 12, pp. 105460–105469, 2024, doi: 10.1109/ACCESS.2024.3433426.

[17] J. Staš, D. Hládek, and T. Koctúr, ‘Slovak Question Answering Dataset Based on the Machine Translation of the Squad V2.0’, J. Linguist. Cas., vol. 74, no. 1, pp. 381–390, Jun. 2023, doi: 10.2478/jazcas-2023-0054.

[18] P. Rajpurkar, R. Jia, and P. Liang, ‘Know What You Don’t Know: Unanswerable Questions for SQuAD’, Jun. 11, 2018, arXiv: arXiv:1806.03822. doi: 10.48550/arXiv.1806.03822.

[19] N. Patwardhan, S. Marrone, and C. Sansone, ‘Transformers in the Real World: A Survey on NLP Applications’, Information, vol. 14, no. 4, Art. no. 4, Apr. 2023, doi: 10.3390/info14040242.

[20] Okan, okanvk/Turkish-Reading-Comprehension-Question-Answering-Dataset. (Oct. 26, 2024). Jupyter Notebook. Accessed: Oct. 30, 2024. [Online]. Available: https://github.com/okanvk/Turkish-Reading-Comprehension-Question-Answering-Dataset

[21] ‘TurQuest/turkish-bquad: Türkçe dilinde biyoloji soru/cevap veriseti’. Accessed: Jul. 20, 2024. [Online]. Available: https://github.com/TurQuest/turkish-bquad/tree/main

[22] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, ‘SQuAD: 100,000+ Questions for Machine Comprehension of Text’, presented at the Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Nov. 2016, pp. 2383–2392. doi: 10.18653/v1/D16-1264.

[23] S. Schweter, BERTurk - BERT models for Turkish. (Apr. 27, 2020). Zenodo. doi: 10.5281/zenodo.3770924.

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ‘BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding’, May 24, 2019, arXiv: arXiv:1810.04805. doi: 10.48550/arXiv.1810.04805.

[25] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, ‘ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators’, Mar. 23, 2020, arXiv: arXiv:2003.10555. doi: 10.48550/arXiv.2003.10555.

[26] ‘turkish-bert/electra/README.md at master · stefan-it/turkish-bert’, GitHub. Accessed: Oct. 30, 2024. [Online]. Available: https://github.com/stefan-it/turkish-bert/blob/master/electra/README.md

[27] ‘dbmdz/electra-small-turkish-cased-discriminator · Hugging Face’. Accessed: Oct. 30, 2024. [Online]. Available: https://huggingface.co/dbmdz/electra-small-turkish-cased-discriminator

[28] ‘dbmdz/electra-base-turkish-cased-discriminator · Hugging Face’. Accessed: Oct. 30, 2024. [Online]. Available: https://huggingface.co/dbmdz/electra-base-turkish-cased-discriminator

[29] 'dbmdz/distilbert-base-turkish-cased · Hugging Face'. Accessed: Aug. 09, 2024. [Online]. Available: https://huggingface.co/dbmdz/distilbert-base-turkish-cased

[30] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, 'DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter', Feb. 29, 2020, arXiv: arXiv:1910.01108. doi: 10.48550/arXiv.1910.01108.

[31] P. Flach and M. Kull, 'Precision-Recall-Gain Curves: PR Analysis Done Right', in Advances in Neural Information Processing Systems, Curran Associates, Inc., 2015. Accessed: Aug. 09, 2024. [Online]. Available: https://papers.nips.cc/paper_files/paper/2015/hash/33e8075e9970de0cfea955afd4644bb2-Abstract.html

BIOGRAPHIES

Mehmet Arzu was born in Kırklareli, Turkey, in 1997. He received the B.S. degree in Software Engineering from Fırat University in 2021 and the M.S. degree in Software Engineering from Fırat University in 2024. From 2023 to 2024, he worked as a Research Assistant in the Department of Computer Engineering at Malatya Turgut Ozal University. Currently, he is a Research Assistant in the Department of Artificial Intelligence and Data Engineering at Fırat University

Murat Aydogan received his doctorate. He received the B.S. degree in Electronic and Computer Education, Computer Teaching Program, from Fırat University in 2011. He also received the M.S. degree in Software Engineering from Fırat University in 2014, and the Ph.D. degree in Computer Engineering from Inonu University in 2019. Since 2020, he has been an Assistant Professor in the Software Engineering Department at Fırat University. His research interests include Natural Language Processing, Artificial Intelligence, and Data Mining.