European Journal of Technique

journal homepage: https://dergipark.org.tr/en/pub/ejt

Vol.15, No.1, 2025

Research Article

Sentiment Analysis in Turkish Using Language Models: A Comparative Study

Mert Incidelen[1*], Murat Aydogan[2]

[1*]Firat University, Artificial Intelligence and Data Engineering Department, Elazig, Turkey. (e-mail: mincidelen@firat.edu.tr).
[2] Firat University, Software Engineering Department, Elazig, Turkey. (e-mail: m.aydogan@firat.edu.tr).

ARTICLE INFO

Received: Jun., 23. 2025
Revised: Jun., 8. 2025
Accepted: Nov, 27. 2024

Keywords:
Turkish Sentiment Analysis
Language Models (LMs)
Natural Language Processing (NLP)

Corresponding author: *Mert Incidelen*

ISSN: *2536-5010* | e-ISSN: *2536-5134*

DOI: https://doi.org/10.36222/ejt.1592448

ABSTRACT

Sentiment analysis is a natural language processing (NLP) task that aims to automatically identify positive, negative and neutral emotions in texts. Agglutinative languages such as Turkish pose challenges for sentiment analysis due to their complex morphological structure. Traditional methods are inadequate for detecting sentiment in texts. Language models (LMs), on the other hand, achieve successful results in sentiment analysis as well as in many other NLP tasks thanks to their ability to learn context and structural features of the language. In this study, XLM-RoBERTa, mBERT, BERTurk 32k, BERTurk 128k, ELECTRA Turkish Small and ELECTRA Turkish Base models were fine-tuned using the Turkish Sentiment Analysis – Version 1 (TRSAv1) dataset and the performances of the models were compared. The dataset consists of 150,000 texts containing user comments on e-commerce platforms. The classes have a balanced distribution for positive, negative and neutral classes. The fine-tuned models are evaluated using the test set with metrics such as accuracy, precision, recall and F1 score. The findings show that models customized for the Turkish language exhibit better performance in emotion detection compared to multilingual models. The BERTurk 32k model achieved strong results with an accuracy of 83.69% and an F1 score of 83.65%, while the BERTurk 128k model followed closely with an accuracy of 83.68% and an F1 score of 83.66%. On the other hand, the XLM-RoBERTa model, a multilingual model, delivered competitive performance with an accuracy of 83.27% and an F1 score of 83.22%.

1. INTRODUCTION

In recent years, with the rise of internet usage, there has been a significant increase in text data [1]. A significant portion of this data consists of social media posts, customer comments on e-commerce sites, news sites, and content on similar sources. The need to interpret and analyze such a large amount of text data has increased the importance of natural language processing (NLP) techniques and the interest in this field. Today, social media platforms are commonly used to share personal thoughts and feelings, while e-commerce sites allow customers to express product-related experiences [2,3]. Analysis of data on these platforms has the potential to provide valuable insights into understanding the needs of individuals and societies, as well as predicting future trends. Sentiment analysis, an NLP task that focuses on detecting positive, negative, and neutral emotions in texts, is a crucial area of research in this context. Sentiment analysis is widely accepted as an effective method for addressing important issues, such as determining marketing strategies and making strategic decisions, by governments. In recent years, advancements in sentiment analysis methods have significantly increased interest in this field [4,5].

Before language models (LMs), sentiment analysis methods were divided into machine learning and lexical-based approaches. Machine learning-based approaches consist of various techniques such as Naive Bayes, Support Vector Machines (SVM), and Decision Trees. In these methods, machine learning algorithms are trained on a given labeled dataset and perform classification on texts [6]. Lexical-based methods, on the other hand, analyze the emotional words in the text and assign scores. These approaches use a pre-created dictionary to determine whether the words in the text contain emotions. Thus, a sentiment score is calculated according to the number of words [7]. Ahmad et al. [8] used the SVM algorithm for sentiment analysis with tweets and found that the performance varies depending on the characteristics of the dataset, and it is especially successful in neutral sentiment classification. Dhaoui et al. [9] compared the performance of lexical-based and machine learning-based approaches using social media comments. They observed that combining both methods improved the performance in sentiment analysis. Onan [10] performed sentiment analysis on Turkish tweets using Naive Bayes, SVM, and Logistic Regression methods. In the study, it was revealed that Naive Bayes method gave more successful results than other machine learning approaches.

Traditional approaches are limited in terms of context understanding and vocabulary, as they cannot go beyond the words in the predefined list. Machine learning and lexical based methods may exhibit limited performance depending on the size and diversity of the data. The emergence of LMs with the introduction of the Transformer architecture has led to significant improvements for NLP tasks. LMs can learn the context and relationships between words in texts thanks to the large corpora and self-attention mechanism used during their pre-training [11]. This contextual information enables the accurate analysis of sentence structure. The representations learned by the models can be fine-tuned for development purposes for specific domains and tasks. Thus, LMs can be adapted for tasks in various domains. These pre-trained models have been successfully applied to a wide range of NLP tasks such as text classification [12], machine translation [13], text summarization [14] and question-answer systems [15]. Unlike traditional methods, LMs have the potential to exhibit high performance with fine-tuning, even with imbalanced or complex datasets. This is particularly advantageous for low-resource languages. The contextual capabilities of LMs increase their performance in NLP tasks. Tan et al. [16] proposed a model that combines Transformer architecture with recurrent neural network architecture for sentiment analysis. This model has a hybrid structure with Robustly Optimized Bidirectional Encoder Representations from Transformers Pretraining Approach (RoBERTa) and Long Short-Term Memory (LSTM) approaches that have Bidirectional Encoder Representations from Transformers (BERT) structure. The proposed approach exhibited superior performance on IMDb, Twitter US Airline Sentiment and Sentiment140 datasets by combining the strong contextual ability of RoBERTa with the ability of LSTM to capture long dependencies. Arroni et al. [17] developed a simple model that uses the Transformer architecture's self-attention mechanism to analyze the sentiment of tweets about hotels. The model aims to classify tweets about hotels according to their sentiments. The study revealed that Transformer-based LMs are more effective than traditional methods. Khan et al. [18] fine-tuned the Multilingual BERT (mBERT) model for Urdu, a low-resource language, by comparing traditional methods with deep learning methods. The findings from the study revealed that the contextual capabilities of the BERT model outperformed other traditional methods. mBERT stood out as an effective model, especially for low-resource languages. Yürütücü and Demir [19] performed sentiment analysis using tweets about COVID-19. In the study, LMs and Naive Bayes method were compared for sentiment analysis. Higher accuracy sentiment detection was achieved with the BERT-based model compared to the Naive Bayes method. The results of the study showed that the BERT-based model performed successfully in capturing contextual information compared to traditional methods. Köksal and Özgür [20] created an original dataset consisting of tweets for Turkish sentiment analysis. Using this dataset, BERTurk, mBERT and XLM-RoBERTa models were fine-tuned, and the performances of the models were compared. The best performance among the models was obtained with BERTurk.

The increasing importance of NLP tasks and sentiment analysis has necessitated the development of the capabilities of LMs and the evaluation of their performance. The success of LMs in NLP tasks and sentiment analysis varies depending on many factors. These factors include the structural complexity of the model used, the size and quality of the corpora used in pre-training, grammatical features, and syntactic structure. There are challenges for NLP tasks and sentiment analysis in low-resource languages compared to models in resource-rich languages, such as English [21]. Turkish is also among the low-resource languages. The agglutinative structure of Turkish causes roots and suffixes to combine in different ways, providing a flexible structure. In particular, the fact that the subject and predicate can be in different places in the sentence makes the language complex in terms of syntax. Therefore, in the analysis of the performance of Turkish LMs, this unique structure of the language must be considered. Considering the agglutinative structure and contextual diversity of Turkish, using models to grasp the structure of the language and to perform tasks such as sentiment analysis is a difficult but necessary goal. Models have the potential to exhibit superior performance even in a low-resource language such as Turkish, provided that the size and diversity of the corpora are provided in the pre-training. The performance of the models depends on their ability to grasp the structural and grammatical features of the language.

This study examines the performance of LMs for sentiment analysis. The limited contextual representation and inadequacy of traditional methods highlight the potential of LMs in this field. Evaluating the performance of these models in low-resource languages plays a critical role in the development of NLP studies in these languages. The aim of this study is to examine the performance of XLM-RoBERTa, mBERT, BERTurk 32k, BERTurk 128k, Efficiently Learning an Encoder that Classifies Token Replacements Accurately (ELECTRA) Turkish Small and ELECTRA Turkish Base models in the Turkish sentiment analysis task and to contribute to the development of methods that will provide higher performance for Turkish. In addition, it is aimed to obtain important findings on how to make fine-tuning processes more effective in accordance with the language structure of Turkish and to provide guidance on how these models can be adapted sensitively to the linguistic features of Turkish. The key contributions of this study are as follows:

- A comparative evaluation of LMs on Turkish sentiment analysis.
- Performance benchmarking on a balanced and representative Turkish dataset.
- Analysis of monolingual vs. multilingual model performance.
- Interpretation of model outputs across sentiment classes.
- Recommendations for fine-tuning strategies in morphologically rich, low-resource languages like Turkish.

This study consists of five main sections. In the first section, the background of the study, previous studies, motivation and objectives are presented. In the second section, details about the dataset used in the study and details of the LMs compared are given. In the third section, the experimental setup is explained, information about the fine-tuning process and the metrics used to evaluate the model performances are given. In the fourth section, the findings obtained from the experiments, the comparison of the models and the detailed evaluation of their performance are included. In the fifth section, a general evaluation of the study is made, the main findings are summarized and suggestions for future studies are presented.

2. MATERIAL AND METHOD

This section outlines the dataset and models utilized in the study. First, the TRSAv1 dataset is introduced in detail, followed by descriptions of the LMs and their fine-tuning procedures.

2.1. Dataset

Turkish Sentiment Analysis-Version 1 (TRSAv1) was used in this study [22]. The dataset serves as a comprehensive resource for evaluating sentiment analysis in Turkish. The TRSAv1 dataset contains 150,000 e-commerce reviews of products in the Turkish market from real users. These reviews consist of authentic comments expressing users' opinions about their shopping experiences and product quality. The comments in the dataset are categorized into three classes as positive, negative, and neutral. The number of comments in the classes shows a balanced distribution. Each class includes exactly 50,000 reviews, which were manually labeled using a custom-developed annotation tool. The dataset includes nearly 2 million words and more than 80,000 unique terms. Specifically, the positive class contains 717,674 words, the negative class 613,737, and the neutral class 583,541. This balanced dataset structure allows the models to measure their performance for different sentiment classes. The balanced distribution prevents models from overfitting or underfitting. Examples of the data in the classes in the dataset are given in Table 1.

TABLE I
EXAMPLES OF SENTIMENT CLASSES IN THE TRSAV1 DATASET

Sentiment Class	Example
Positive	"Saçlarda dökülmeyi belirgin derecede azaltıyor. Hem de yumuşacık yapıyor. Kendi yakın arkadaşlarıma dahi önerdiğim bir ürün"
Neutral	"Annem için aldım bakalım memnun kalacak mı boyutu normal"
Negative	"Çok kötü oyuncak gibi sakın almayın bir işe yaramaz"

A data preprocessing pipeline was applied to the dataset to ensure a cleaner input for the models and improve sentiment classification performance. In the data preprocessing phase, elements that could affect the accuracy of the models in the dataset were cleaned. Comments were cleaned by removing symbols such as emojis, hashtags, user mentions, and hyperlinks, which are known to introduce noise in sentiment prediction. As a result, the dataset was structured to be compatible with sentiment classification tasks.

2.2. Language Models

LMs are models trained with large text data that focus on understanding and reproducing the complex structure of human language in the field of NLP. These models learn the contextual structure of language in depth by using the self-attention mechanism in the Transformer architecture [23]. Thus, these models are versatile and applicable to a wide range of NLP tasks. These models are typically trained in two stages. First, they undergo pretraining on large-scale corpora to learn general language representations. Then, they are fine-tuned on task-specific datasets to adapt to NLP applications and domains [24].

LMs can be pre-trained in multiple languages or customized in a single language. In this study, multilingual models, including Turkish and customized models for Turkish, were used. The models used in the study, their parameters and types are given in Table 2.

TABLE II
LANGUAGE MODELS

Model Name	Parameter Count	Language Type
XLM-RoBERTa	270M	Multilingual
mBERT	110M	Multilingual
BERTurk 32k	110M	Monolingual
BERTurk 128k	110M	Monolingual
ELECTRA Turkish Small	14M	Monolingual
ELECTRA Turkish Base	110M	Monolingual

XLM-RoBERTa: A multilingual model developed by Facebook AI and pre-trained with text data in more than 100 languages [25]. It is based on the RoBERTa architecture, an extended version of BERT [26]. XLM-Roberta's multilingual structure, including Turkish, provides successful results for different NLP tasks in different languages. In this study, this multilingual model is fine-tuned for Turkish sentiment analysis and its performance is analyzed.

mBERT: mBERT is a multilingual model based on the BERT architecture. The BERT model is based on Masked Language Modeling and Next Sentence Prediction techniques [24]. The multilingual version developed by Google can be widely used in text classification, sentiment analysis and other NLP tasks in different languages. In the pre-training process of the model, it was trained with text data in 104 different languages, including Turkish. mBERT's multilingual structure, including Turkish, allows it to be used in Turkish NLP tasks.

BERTurk: BERTurk is a BERT-based model trained on Turkish corpora, including news articles, Wikipedia, and OSCAR datasets. The BERTurk model has been specially developed for high performance on Turkish language tasks. The model is sensitive to the linguistic features and semantic details of Turkish. There are several versions of the BERTurk model with different vocabularies [27]. The 32k and 128k case versions of the BERTurk model, which have shown successful results in Turkish text-based tasks, are used in this study for fine-tuning the sentiment analysis task.

ELECTRA Turkish: ELECTRA Turkish is a customized version of the ELECTRA base model for Turkish. This model has been pre-trained with large Turkish texts and has a good command of the general structure of the Turkish language. ELECTRA is based on the Transformer architecture and uses the Replaced Token Detection technique. This method offers a pre-training process with a different approach than the masked

language model [28]. The ELECTRA model adapted for Turkish has the potential to exhibit high performance in Turkish NLP tasks. This model was fine-tuned for the sentiment analysis task in this study.

3. EXPERIMENTAL SETUP

In this section, the fine-tuning stage and evaluation metrics for the experimental setup of the study are explained. LMs pre-trained on large corpora were fine-tuned with the TRSAv1 dataset for sentiment analysis, and the results were evaluated. Figure 1 shows the pre-training, fine-tuning, and evaluation processes of the LMs.

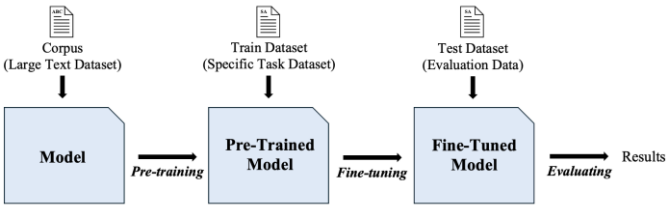

Figure 1. The Process of Pre-training, Fine-tuning, and Evaluation of LMs

3.1. Fine-Tuning

XLM-RoBERTa, mBERT, BERTurk 32k, BERTurk 128k, ELECTRA Turkish Small and ELECTRA Turkish Base models were fine-tuned with TRSAv1 dataset for Turkish sentiment analysis task. Fine-tuning is the process of customizing pre-trained models for specific NLP tasks. 80% of the dataset was allocated for training and 20% for testing. During data splitting, balanced representation of all sentiment classes was ensured to avoid class imbalance issues. Identical hyperparameter values were applied during fine-tuning to ensure fair comparison across models. The fine-tuning parameters, including the number of epochs, learning rate, and batch size, are presented in Table 3.

TABLE III
HYPERPARAMETER SETTINGS FOR FINE-TUNING

Hyperparameter	Assigned Value
Epoch	3
Batch Size	32
Learning Rate	3e-5

The fine-tuning of the models was conducted utilizing the Hugging Face Transformers library, a widely recognized toolset for NLP tasks, implemented in Python. An NVIDIA A100 graphics processing unit, with its high memory bandwidth and computational power, was utilized to accelerate the fine-tuning process. After the models were fine-tuned, their overall and class-based performances were evaluated with various metrics using the test dataset.

3.2. Evaluation Metrics

The commonly used metrics of accuracy, precision, recall and F1 score are used to evaluate fine-tuned models for sentiment analysis tasks. These metrics provide the opportunity to evaluate different performance aspects of the model in detail for each class in a classification problem. Thus, the performance of the model for the problem can be analyzed in depth.

Accuracy: The ratio of samples correctly classified by the model to the total number of samples. The formula for the Accuracy metric is given in Equation 1.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Precision: The ratio of instances that the model correctly classifies as positive to the total number of instances that the model correctly classifies as positive. The formula for the metric is given in Equation 2.

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Recall: The ratio of samples correctly classified as positive by the model to the total number of true positive samples. The formula for the Recall metric is given in Equation 3.

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

F1 Score: It is expressed as the harmonic mean of Precision and Recall metrics. Its formula is given in Equation 4.

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

4. RESULTS and DISCUSSION

In this study, we comprehensively evaluate and compare the performance of fine-tuned LMs on a Turkish sentiment analysis task. This evaluation, which includes multilingual models designed in accordance with the linguistic features of Turkish, provides important insights into which model may be more effective in NLP tasks such as sentiment analysis. Table 4 shows the performance of the fine-tuned models in sentiment analysis.

TABLE IV
PERFORMANCE METRICS OF FINE-TUNED MODELS FOR TURKISH SENTIMENT ANALYSIS

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
XLM-RoBERTa	83.27	83.30	83.27	83.22
mBERT	81.86	82.00	81.86	81.89
BERTurk 32k	83.69	83.68	83.69	83.65
BERTurk 128k	83.68	83.69	83.68	83.66
ELECTRA Turkish Small	81.84	81.87	81.84	81.80
ELECTRA Turkish Base	83.64	83.64	83.64	83.58

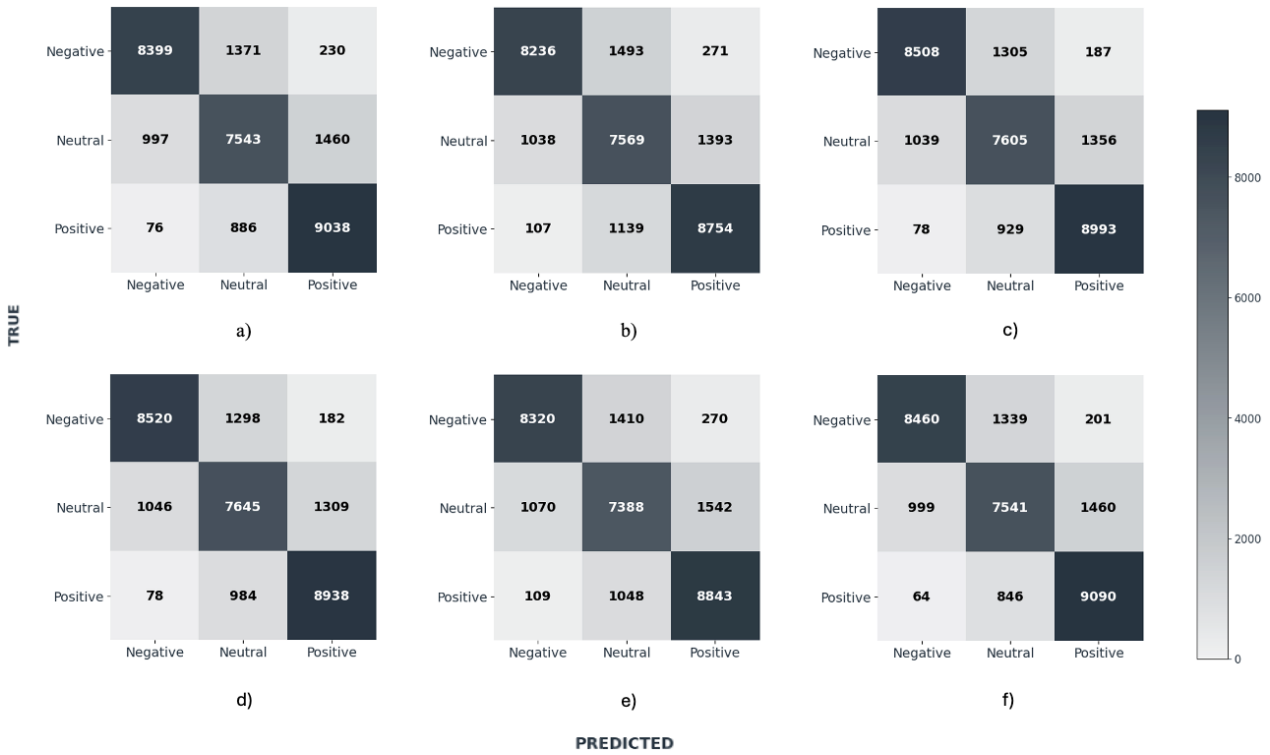

Figure 2. Confusion Matrices of Fine-Tuned Models: **a)** XLM-RoBERTa, **b)** mBERT, **c)** BERTurk 32k, **d)** BERTurk 128k, **e)** ELECTRA Turkish Small, **f)** ELECTRA Turkish Base

Among the models evaluated for Turkish sentiment analysis, BERTurk models were the most effective models. The BERTurk 32k model achieved an accuracy of 83.69% and an F1 score of 83.65%. Similarly, the BERTurk 128k model followed closely, with 83.68% accuracy and an F1 score of 83.66%. Pre-training on Turkish-specific data, which enables the models to effectively grasp the agglutinative and morphological nuances of the language, seems to be effective in Turkish NLP tasks. The ELECTRA Turkish Base model also performed commendably with an accuracy of 83.64% and an F1 score of 83.58%. These results demonstrate the robustness of monolingual models trained on Turkish data. However, the ELECTRA Turkish Small model with only 14 million parameters struggled to capture the rich contextual information of Turkish. Its lower performance with 81.84% accuracy and 81.80% F1 score can be attributed to its limited capacity to learn from data. Therefore, the model is less suitable for tasks involving complex language structures. The importance of model size and pre-training in achieving high performance in morphologically rich languages such as Turkish emerges.

Among the multilingual models, XLM-RoBERTta showed strong results with an accuracy of 83.27% and an F1 score of 83.22%. It performed particularly well in the positive and negative sentiment classes, demonstrating its ability to understand Turkish sentiment in these categories. However, as can be seen in Table 5, its performance in the neutral class was significantly weaker compared to the monolingual BERTurk model. This limitation suggests that despite its multilingual capabilities, XLM-RoBERTa struggles to fully adapt to Turkish-specific grammatical structures and nuanced expression of neutral sentiments. On the other hand, mBERT performed less effectively than the other models, achieving an accuracy of 81.86% and an F1 score of 81.89%. This suggests that its generalized multilingual architecture does not fully address the unique linguistic features of Turkish. Its lower performance compared to monolingual models suggests that pre-trained LMs, especially for Turkish, have an advantage in understanding the articulatory and context-sensitive nature of the language. Table 5 shows the results of the fine-tuned models for different sentiment classes according to the precision, recall and F1 score metrics.

TABLE V
PERFORMANCE OF LMS ACROSS SENTIMENT CLASSES

Model	Class	Precision (%)	Recall (%)	F1 Score (%)
XLM-RoBERTa	Negative	88.67	83.99	86.27
	Neutral	76.97	75.43	76.19
	Positive	84.25	90.38	87.21
mBERT	Negative	87.79	82.36	84.99
	Neutral	74.20	75.69	74.94
	Positive	84.03	87.54	85.75
BERTurk 32k	Negative	88.39	85.08	86.71
	Neutral	77.29	76.05	76.67
	Positive	85.35	89.93	87.58
BERTurk 128k	Negative	88.35	85.20	86.74
	Neutral	77.01	76.45	76.73
	Positive	85.70	89.38	87.50
ELECTRA Turkish Small	Negative	87.59	83.20	85.34
	Neutral	75.04	73.88	74.45
	Positive	82.99	88.43	85.63
ELECTRA Turkish Base	Negative	88.84	84.60	86.67
	Neutral	77.53	75.41	76.46
	Positive	84.55	90.90	87.61

As seen in Figure 2, the confusion matrices show a high number of misclassifications in the neutral sentiment class, especially among multilingual models. A significant number of neutral examples were incorrectly predicted as positive. This indicates that the models have difficulty distinguishing subtle contextual cues that separate neutral from positive sentiment. Correct classification of neutral expressions in the Turkish sentiment analysis task is one of the biggest challenges. Neutral expressions can be perceived as positive or negative depending on the context, which can affect the classification performance of LMs. BERTurk models performed better in the neutral class compared to other models. This suggests that these monolingual models may have better learned the contextual diversity and grammatical features of Turkish. The other multilingual models performed inconsistently in the neutral class. This shows that multilingual models are generally less sensitive to context in languages with a suffixal structure such as Turkish. The suffixal structure of Turkish, the fact that word roots combine with different affixes to acquire new meanings, and the flexibility of sentence structure pose a significant challenge for LMs. In this context, the monolingual models analyzed are more successful than multilingual models in learning this complex structure. BERTurk models clearly show the advantage of being specially trained with Turkish data.

5. CONCLUSION

This study investigates the performance of LMs on Turkish sentiment analysis tasks. Different variations of the BERTurk and ELECTRA Turkish models, as well as the mBERT and XLM-RoBERTa multilingual models, were compared for the sentiment analysis task. The findings revealed that Turkish customized models such as BERTurk 32k, BERTurk 128k and ELECTRA Turkish Base were significantly more effective at capturing Turkish linguistic context. BERTurk variations have proven to be a strong choice for Turkish sentiment analysis tasks due to their consistent performance. While sentiment analysis is challenging with traditional methods and multilingual models due to the complex linguistic structure of Turkish, monolingual models managed to overcome these challenges and achieved high accuracy and F1 scores. These findings demonstrate the importance of models trained on language-specific data. The TRSAv1 dataset, featuring a balanced class structure and real user comments, was used in this study. These characteristics of the dataset enabled more accurate evaluation of the models. In this context, BERTurk models demonstrated better performance in the neutral class compared to other models. Although the multilingual models showed acceptable performance for the affirmative and negative classes, they were insufficient in capturing the context for the neutral class. This suggests that monolingual models may be more suitable for low-source and agglutinative languages such as Turkish. By contrast, ELECTRA Turkish Small, with fewer parameters, was inadequate for handling the complex linguistic features of Turkish due to its limited capacity. Nevertheless, the results indicate that this model can be a viable option for non-critical applications due to its low hardware requirements and high speed. Conversely, the ELECTRA Turkish Base model performs close to the BERTurk models and has considerable success for Turkish sentiment analysis.

This study highlights the potential and limitations of LMs for low-resource and structurally ambiguous languages such as Turkish. The success of models such as BERTurk has demonstrated the power of monolingual models. The success of these Turkish-specific models demonstrates their capabilities in NLP tasks such as language understanding and Turkish sentiment analysis. Future studies can improve model performance in Turkish and similar low-resource languages by investigating the effect of larger and more diverse datasets such as domain-specific texts. In this context, the findings of this study provide important findings and a roadmap for Turkish NLP tasks.

REFERENCES

[1] I. Yaqoob, I. A. T. Hashem, A. Gani, S. Mokhtar, E. Ahmed, N. B. Anuar, and A. V. Vasilakos, "Big data: From beginning to future," Int. J. Inf. Manage., vol. 36, no. 6, pp. 1231–1247, 2016.

[2] S. Mittal, A. Goel, and R. Jain, "Sentiment analysis of E-commerce and social networking sites," in Proc. 3rd Int. Conf. Comput. Sustainable Global Develop. (INDIACom), 2016, pp. 2300–2305.

[3] M. Rodríguez-Ibáñez, A. Casánez-Ventura, F. Castejón-Mateos, and P.-M. Cuenca-Jiménez, "A review on sentiment analysis from social media platforms," Expert Syst. Appl., vol. 223, p. 119862, 2023.

[4] M. Marong, N. K. Batcha, and R. Mafas, "Sentiment analysis in e-commerce: A review on the techniques and algorithms," J. Appl. Technol. Innov., vol. 4, no. 1, p. 6, 2020.

[5] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," Artif. Intell. Rev., vol. 55, no. 7, pp. 5731–5780, 2022.

[6] A. P. Jain and P. Dandannavar, "Application of machine learning techniques to sentiment analysis," in Proc. 2nd Int. Conf. Appl. Theor. Comput. Commun. Technol. (iCATccT), 2016, pp. 628–632.

[7] C. S. G. Khoo and S. B. Johnkhan, "Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons," J. Inf. Sci., vol. 44, no. 4, pp. 491–511, 2018.

[8] M. Ahmad, S. Aftab, and I. Ali, "Sentiment analysis of tweets using SVM," Int. J. Comput. Appl., vol. 177, no. 5, pp. 25–29, 2017.

[9] C. Dhaoui, C. M. Webster, and L. P. Tan, "Social media sentiment analysis: Lexicon versus machine learning," J. Consum. Market., vol. 34, no. 6, pp. 480–488, 2017.

[10] A. Onan, "Sentiment analysis on Twitter messages based on machine learning methods," Yönetim Bilişim Sistemleri Dergisi, vol. 3, no. 2, pp. 1–14, 2017.

[11] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., "Transformers: State-of-the-art natural language processing," in Proc. 2020 Conf. Empirical Methods Natural Lang. Process.: System Demonstrations, 2020, pp. 38–45.

[12] R. Qasim, W. H. Bangyal, M. A. Alqarni, and A. A. Almazroi, "A fine-tuned BERT-based transfer learning approach for text classification," J. Healthcare Eng., vol. 2022, no. 1, p. 3498123, 2022.

[13] X. Zhang, N. Rajabi, K. Duh, and P. Koehn, "Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with QLoRA," in Proc. Eighth Conf. Mach. Transl., 2023, pp. 468–481.

[14] H. Chouikhi and M. Alsuhaibani, "Deep transformer language models for Arabic text summarization: A comparison study," Appl. Sci., vol. 12, no. 23, p. 11944, 2022.

[15] S. Butt, N. Ashraf, M. H. F. Siddiqui, G. Sidorov, and A. Gelbukh, "Transformer-based extractive social media question answering on TweetQA," Computación y Sistemas, vol. 25, no. 1, pp. 23–32, 2021.

[16] K. L. Tan, C. P. Lee, K. S. M. Anbananthen, and K. M. Lim, "RoBERTa-LSTM: A hybrid model for sentiment analysis with transformer and recurrent neural network," IEEE Access, vol. 10, pp. 21517–21525, 2022.

[17] S. Arroni, Y. Galán, X. M. Guzmán Guzmán, E. R. Núñez Valdéz, A. Gómez Gómez, et al., "Sentiment analysis and classification of hotel opinions in Twitter with the transformer architecture," Int. J. Interact. Multimedia Artif. Intell., 2023.

[18] L. Khan, A. Amjad, N. Ashraf, and H.-T. Chang, "Multi-class sentiment analysis of Urdu text using multilingual BERT," Sci. Rep., vol. 12, no. 1, p. 5436, 2022.

[19] Ö. Y. Yürütücü and Ş. Demir, "Ön eğitimli dil modelleriyle duygu analizi," İstanbul Sabahattin Zaim Üniversitesi Fen Bilimleri Enstitüsü Dergisi, vol. 5, no. 1, pp. 46–53, 2023.

[20] A. Köksal and A. Özgür, "Twitter dataset and evaluation of transformers for Turkish sentiment analysis," in Proc. 29th Signal Process. Commun. Appl. Conf. (SIU), 2021, pp. 1–4.

[21] S. Joshi, M. S. Khan, A. Dafe, K. Singh, V. Zope, and T. Jhamtani, "Fine tuning LLMs for low resource languages," in Proc. 5th Int. Conf. Image Process. Capsule Netw. (ICIPCN), 2024, pp. 511–519.

[22] M. Aydoğan and V. Kocaman, "TRSAv1: A new benchmark dataset for classifying user reviews on Turkish e-commerce websites," J. Inf. Sci., vol. 49, no. 6, pp. 1711–1725, 2023.

[23] A. Vaswani, "Attention is all you need," Adv. Neural Inf. Process. Syst., 2017.

[24] J. Devlin, "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[25] A. Conneau, "Unsupervised cross-lingual representation learning at scale," arXiv preprint arXiv:1911.02116, 2019.

[26] Y. Liu, "RoBERTa: A robustly optimized BERT pretraining approach," arXiv preprint arXiv:1907.11692, vol. 364, 2019.

[27] S. Schweter, BERTurk - BERT models for Turkish, version 1.0.0, Zenodo, Apr. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3770924. DOI: 10.5281/zenodo.3770924.

[28] K. Clark, "ELECTRA: Pre-training text encoders as discriminators rather than generators," arXiv preprint arXiv:2003.10555, 2020.

Mert Incidelen was born in Elazığ, Turkey, in 1997. He completed his Bachelor's degree in Department of Computer Engineering at Fırat University in 2021. Currently, he is pursuing a Master's degree in Software Engineering at Fırat University. He served as a Research Assistant in the Department of Computer Engineering at Igdir University between 2023 and 2024. Presently, he is a Research Assistant at Fırat University in the Department of Artificial Intelligence and Data Engineering. His research interests include Natural Language Processing and Artificial Intelligence.

Murat Aydogan obtained his Bachelor's degree from the Computer Education Program in the Department of Electronics and Computer Education at Fırat University in 2011. He earned his Master's degree in Software Engineering from Fırat University in 2014 and completed his Ph.D. in Computer Engineering at İnönü University in 2019. Since 2020, he has been working as an Assistant Professor in the Department of Software Engineering at Fırat University. His research interests include Natural Language Processing, Artificial Intelligence, and Data Mining.