SAKARYA UNIVERSITY JOURNAL OF COMPUTER AND INFORMATION SCIENCES
http://saucis.sakarya.edu.tr/
e-ISSN: 2636-8129
Publisher: Sakarya University
Vol. 8, No. 1, 76-88, 2025
DOI: 10.35377/saucis...1564138
SAUCIS
SAKARYA UNIVERSITY

RESEARCH ARTICLE

Turkish Stance Detection on Social Media Using BERT Models: A Case Study of Stray Animals Law

Selma Alav[1], Kristin S. Benli[2*]

[1] Software Engineering, Institute of Science, Fırat University, Elazığ, Türkiye, ror.org/05teb7b63
[2]Software Engineering, Faculty of Engineering and Natural Sciences, Üsküdar University, İstanbul, Türkiye, ror.org/02dzjmc73

Corresponding author:
Kristin S. Benli, Üsküdar University, Software Engineering, Üsküdar, Türkiye
kristin.benli@uskudar.edu.tr

Article History:
Received: 09.10.2024
Revised: 12.01.2025
Accepted: 05.03.2025
Published Online: 27.03.2025

ABSTRACT
Recently, social media has transformed into an essential platform for information dissemination, allowing individuals to articulate their opinions and apprehensions on a wide array of subjects. Stance detection, which refers to the automated examination of text to ascertain the author's perspective regarding a specific proposition or subject, has emerged as a significant area of research. Within the scope of this study, a Turkish-labeled dataset was created to determine the stances of social media users regarding the Stray Animals Law and various pre-trained BERT models were fine-tuned on this dataset, four of which were Turkish (BERTurk 32k and 128k, ConvBERTurk and ConvBERTurk mC4), one multilingual (mBERT) and one base (BERT-Base). The BERTurk 128k model outperformed other BERT models by achieving a remarkable accuracy rate of 87.10%, along with 87.11% precision, 87.10% recall, and 87.10% F1 score. In conclusion, this study has accomplished a contribution in the limited field of Turkish stance detection research by comparing various BERT models in the context of Turkish texts that has not been previously undertaken to our knowledge. The promising results that were obtained from this and similar studies could contribute to the automatic extraction of public opinions, thereby assisting policymakers in formulating efficient policies.

Keywords: Stance detection, BERT, Text mining, Social media analysis, Turkish dataset

1. Introduction

The use of digital platforms like social media, news portals, and discussion forums has grown with the expansion of the Internet. People use these platforms to stay updated on topics such as politics, economy, sports, and social issues and share their thoughts and concerns. As a result, these platforms have become a crucial source of information. These data sources allow decision-makers to understand public opinion about the goals of interest. For example, those in decision-making positions seek to assess the reactions of the populace—whether in support or opposition to trending issues—and can utilize the information gathered from these platforms. However, manually analyzing the enormous amount of data is time-consuming and costly. This is where a growing interest in research on processing and analyzing textual data comes into play. Various fields focusing on automated content analysis include sentiment analysis, emotion recognition, sarcasm/irony detection, rumor detection, and fake news detection [1]. Stance detection, which is closely associated with sentiment analysis, has also gained attention as a recent area of research.

Stance detection is the automatic analysis of text to determine whether the author's perspective on a particular proposition or target is in favor, against, or neutral. The target of this analysis may include a variety of topics, such as individuals, organizations, government policies, movements, or products [2]. Based on application scenarios, stance detection can be divided into four subcategories: single-target, multi-target, cross-target and zero-shot [3].

In this study, we will focus on single target stance which can be expressed as Equation 1:

$$S_i = f(d_i, T) \ \ where \ \ S_i = \{Favor, Against, None\} \tag{1}$$

where T stands for the specified target, d_i represents the comment, which is made by user U_i, and S_i indicates the stance [3].

Research in Stance Detection has predominantly focused on English texts. In contrast, investigations into stance detection in Turkish are relatively scarce, and there is a lack of sufficient Turkish stance datasets for researchers interested in this topic.

Cite as: S. Alav and KS. Benli, "Turkish stance detection on social media using BERT models: A case study of stray animals law," *Sakarya University Journal of Computer and Information Sciences*, vol. 8, no. 1, pp. 76-88, 2025, doi: 10.35377/saucis...1564138

This study employs various Bidirectional Encoder Representations from Transformers (BERT) models, which are based on the transformer architecture and are recognized for their exceptional performance across various natural language processing applications, to analyze stances based on user comments concerning the Stray Animals Law written in Turkish. The lack of comparable studies that examine different BERT models, specifically in Turkish stance detection, increases the original value of our research. The study presents the following contributions:

- Development of a Turkish-Labeled Dataset: A novel and manually labeled Turkish dataset was developed, focusing on the Stray Animals Law as a target.
- Assessment of BERT Models' Performance: The study examines how BERT models (four Turkish, one multilingual and one base model) perform on the recently established Turkish dataset.
- Analysis of models: Examining the words and phrases most significantly contributes to models' decision-making using the LIME (Local Interpretable Model-Independent Explanations) method.

The remaining sections of this paper are organized in the following manner. Section 2 provides an overview of related studies, while Section 3 details the materials and methods that are employed in the research. The study's findings are presented in Section 4, and qualitative analysis is given in Section 5. Finally, Section 6 evaluates the results and discusses potential directions for future research.

2. Related Work

The following paragraphs present recent research studies that employed the BERT model for stance detection and the limited number of studies on the Turkish language.

Küçük and Can [4] conducted a study on stance detection from Turkish tweets about two prominent sports clubs in Turkey, Galatasaray (target-1) and Fenerbahçe (target-2). Researchers created three versions of the Turkish tweet dataset and labeled the stance information as either Favor or Against. They evaluated the performance of the Support Vector Machine (SVM) classifier using different feature sets such as unigrams, bigrams, hashtags, external links, emoticons and named entities. The results showed that a combination of unigrams, hashtags, and named entities performed better than other combinations of features. The study also revealed that named entities benefited from the Turkish stance detection analysis.

Ghosh et al. [5] examined seven stance detection models. They successfully implemented six of them and employed the publicly available code for the remaining model. Their initial focus was on assessing the reproducibility of these models. They applied them to two distinct datasets: 1) the SemEval dataset, which contains microblog data about topics such as atheism, climate change, the feminist movement, Hillary Clinton, and the legalization of abortion, and 2) the Multi Perspective Consumer Health Query (MPCHI) Data, which addresses five specific claims: MMR vaccination may lead to autism, E-cigarettes are less harmful than traditional cigarettes, Hormone Replacement Therapy is advisable for women after menopause, Vitamin C is effective in preventing the common cold, and exposure to sunlight can result in skin cancer. In addition to exploring current stance detection techniques, they utilized a pre-trained BERT (Large-Uncased) model. The results indicated that the performance metrics of the BERT model significantly surpassed those of other competing models. Furthermore, it was noted that the Convolutional Neural Network (CNN) model demonstrated effective performance with shorter tweets, specifically those containing 5 to 10 words, whereas BERT excelled with longer tweets.

Cotfas et al. [6] analyzed public stances towards COVID-19 vaccination using social media posts. Initially, they gathered a dataset of English tweets expressing various stances on COVID-19 vaccination. A subset of this dataset was manually labeled as Neutral, in Favor, or Against vaccination for training the stance classification model. Four different approaches were explored for text representation and classification: 1) Bag-of-Words with classical machine learning, 2) Word embeddings with classical machine learning, 3) Word embeddings with deep learning, and 4) BERT. The BERT model demonstrated better performance compared to other models.

Grimminger and Klinger [7] conducted a study focused on stance detection in political tweets, specifically examining whether supporters of Trump and Biden, who were the candidates in the 2020 US Presidential Elections, engaged in hateful and offensive speeches in their online communications. They developed an annotation task that combined the detection of hateful or offensive speech and stance detection. In addition to the established categories of Favorable and Against opinions, the analysis incorporated Mixed and Neutral positions and instances where a candidate was referenced without any expressed opinion. A pre-trained BERT base model was employed, revealing that the model successfully identified support for a candidate; however, determining an individual's opposition to a candidate proved more challenging.

Polat et al. [8] developed a Turkish stance dataset that contained comments related to different targets such as working from home, mask, e-book, vegan, e-cigarette, and vaccine. Comments were collected from the Ekşi Sözlük online forum, which was also used in this study, and tagged with Favor, Against and None stance classes. They conducted stance detection experiments using various machine learning methods, ensemble learning methods, and CNNs. Texts were represented using Bag-of-Words and Term Frequency-Inverse Document Frequency (TF-IDF) models for machine learning and ensemble learning techniques. For the deep learning approach, Word embedding was utilized for text representation. The results of the experiments showed that, despite target-based variations, the highest performances were observed with XGBoost and CNN

models.

Küçük and Arıcı [9] introduced Turkish datasets related to COVID-19 vaccination for sentiment analysis and stance detection purposes. They gathered tweets from two distinct time frames (December and July). They categorized them into Favor, Against, and None stance classes—the feature set comprised unigrams, hashtag use, and emoticon use. Two different machine learning methods, SVM and Random Forest (RF), were utilized for training and testing, and the evaluation was conducted using a 10-fold cross-validation approach. The process resulted in relatively lower performance rates. The results showed that the SVM outperformed the RF in sentiment analysis and stance detection tasks. In their subsequent research, Küçük and Arıcı [10] assessed the capabilities of BERTurk and ChatGPT utilizing an enhanced version of the same dataset. The findings indicated that ChatGPT demonstrated better performance in stance detection, whereas BERTurk was more successful in sentiment analysis.

Zengin et al. [11] conducted a study on Turkish stance detection, examining how the performance of a fine-tuned BERT model was influenced by training data that was cross-target, cross-domain, and cross-lingual. They developed datasets encompassing football, health, economics, and politics. BERTurk was utilized to process Turkish data, while M-BERT was employed to process English data and cross-lingual experiments. The researchers reached multiple conclusions, notably that the integration of data for different targets within the same domain led to higher performance, manually annotated datasets outperformed automatically assessed datasets, the presence of training data that was aligned with the domain of the test data was a vital element in attaining higher classification performance and training exclusively on Turkish data produced better outcomes than training with a combination of Turkish and English data.

Arslan and Fırat [12] created a labeled dataset in Turkish to analyze user stances on the Russia-Ukraine conflict through social media posts. They categorized the tweets as either Favor or Against and experimented with machine-learning techniques using GloVe and FastText word embeddings. Additionally, they employed the 128K uncased BERT for the Turkish (BERTurk) model. They utilized both undersampling and oversampling techniques to address the imbalance in the dataset. The findings revealed that BERT-based models surpassed all other approaches, with LSTM and GRU yielding comparable results.

3. Materials and Methods

The phases of the study are illustrated in Figure 1. A dataset was developed comprising comments related to the "Stray Animals Law," which sparked considerable debate and commentary in Turkey, using the Ekşi Sözlük platform. The comments were scraped using BeautifulSoup, and the initially dirty data was cleaned. Subsequently, the data was manually categorized into "Favor" and "Against" labels. Next, six pre-trained BERT models were fine-tuned to detect stance in Turkish comments. Experiments were conducted on Google Colab and Drive. The models were assessed using four common metrics: accuracy, precision, recall, and F1 score. Furthermore, the words that had a considerable impact on the predictions made by the models were examined with LIME.

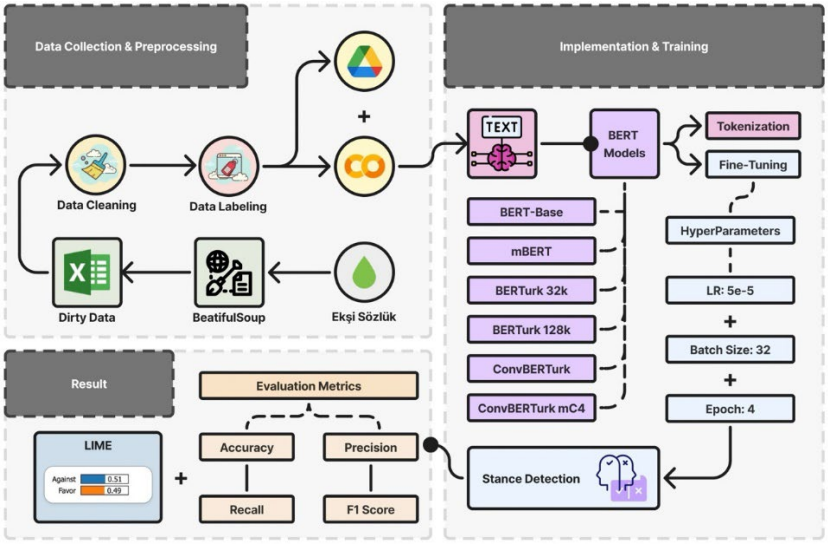

Figure 1. Workflow of the study

3.1. Data Collection and Pre-processing

While developing the dataset, comments on Ekşi Sözlük (a Turkish blog platform) were used. User comments were collected from 29 open headings regarding the "Stray Animals Law." The specific subject headings that were utilized in this study are listed below.

- "29 temmuz 2024 sokak hayvanları yasası değişikliği" (29 July 2024 stray animals law amendment)
- "sokak hayvanları yasasının komisyondan geçmesi" (passing the stray animals law from the commission)
- "sahipsiz hayvanlara yönelik kanun teklifi" (law proposal for homeless animals)
- "sokak hayvanları uyutulacak" (stray animals will be put to sleep).
- "yasayı geri çek" (withdraw the law)

The dataset contained comments where the authors explicitly expressed their viewpoints. The comments were extracted using the BeautifulSoup web scraping library and urllib.request package. In the data cleaning phase, all characters were converted to lowercase, while URLs, symbols, and punctuation marks were removed, and any unnecessary whitespaces were deleted. These operations were performed using string, re and nltk libraries. For the target "Stray Animals Law," each comment was manually labeled as either "Favor" or "Against." Table 1 presents sample target-comment pairs from the Turkish stance-tagged dataset.

Table 1. Target-Comment Pairs from the Turkish Stance Tagged Dataset

Target	Comment	Stance
Stray Animals Law	**(TR[a])** yetkili mercilerden resmi kurumlardan belediyelerden veterinerlerden avukatlardan bunu nasıl engelleyebileceğimize dair acil ve etkili bir yönlendirme bekliyoruz yalvarıyorum [13] **(EN[b])** we expect urgent and effective guidance from competent authorities, official institutions, municipalities, veterinarians, and lawyers on how we can prevent this. I beg	Against
Stray Animals Law	**(TR[a])** başıboş sokak köpeklerinin toplanması gerekiyor gereğinin yapılmasını bekliyoruz [14] **(EN[b])** stray dogs need to be collected; we expect the necessary action to be taken	Favor

[a]Turkish, [b]English

Upon completing the labeling process, there were 5000 comments in the dataset concerning the Stray Animals Law, with 2500 in Favor and 2500 Against. Table 2 presents an in-depth examination of the word count metrics for the comments, such as the shortest, average, and longest word counts within the comments related to the target.

Table 2. Word Count Statistics for the Target "Stray Animals Law"

Stance	Word Count		
	Min	***Max***	***Average***
Against	1	163	41
Favor	1	170	36

The word cloud technique was employed to represent the words in the dataset visually. The visuals were generated utilizing the matplotlib, pandas, and wordcloud modules. Parts (a) and (b) of Figure 2 illustrate the prominent words in the against and favor classes, respectively. The font size of each word in the cloud indicates its frequency or significance within the text. Generally, a word that occurs more often in the text will be displayed larger in the word cloud. While the word clouds associated with the two labels exhibited a general similarity, a closer examination of the smaller font sizes revealed distinct differences. The word cloud for the Against class featured terms such as "katliam" (massacre), "öldürmek" (killing), "masum" (innocent) and "karşı" (against). In contrast, the Favor class included words like "sahiplenin" (adopt), "destekliyorum" (support), "kısırlaştırma" (neuter), "zarar" (harm), "kuduz" (rabies) and "saldırgan" (aggressive).

3.2. BERT Models

BERT, which was created by Google [15] for the field of natural language processing (NLP), represents a significant advancement in language modeling. The architecture of BERT is almost identical to a multilayer bidirectional transformer encoder, as found in research by Vaswani et al. [16]. Unlike its predecessors, this model analyzes text bi-directionally, allowing it to grasp the context more effectively by considering both the preceding and following text.

(a) (b)

Figure 2. Word Cloud Layouts (a) Against and (b) Favor

BERT has two pre-training tasks: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). The task of MLM is to predict the next word based on a given sequence of words. In each sequence, 15% of the words are randomly masked. The words "masked" are not always substituted with the actual [MASK] token. For instance, the i-th token

- replaced with the [MASK] token 80% of the time
- replaced with a random token 10% of the time
- remained unchanged 10% of the time

The model tries to predict the actual values of the masked words with the help of the remaining unmasked words in the sequence.

The NSP task involves understanding the relationship between two sentences or predicting the next sentence in a pair. In this task, the input typically comprises two sentences, A and B. In the %50 of the time, second sentence B directly follows the first sentence A. So, they are related to each other. In the other half, the second sentence, B, is randomly chosen from the dataset, and there is no connection with the first sentence, A.

This research involves experiments with six distinct BERT case models: four Turkish, one multilingual, and one base model. A brief introduction to them will be given in the following lines.

- ***BERT-Base:*** BERT-Base is the foundational model for all BERT variants, having undergone pre-training with an English dataset. It features a vocabulary comprising 30,000 tokens [15]. However, it is limited by a relatively small training dataset and a masking process that occurs only once, which can lead to errors since the masked data remains unchanged across all training epochs [17]. Consequently, it is designed to be fine-tuned and adaptable for further enhancements. Many new models have emerged from this pre-trained version [18].
- ***BERT Multilingual (mBERT):*** BERT includes two multilingual models: one that is cased and another that is uncased. The cased model has been trained in 104 languages, while the uncased model has been trained in 102 languages.
- ***BERTurk:*** BERTurk operates similarly to BERT, being a variant that has been pre-trained on a corpus of 35GB, which includes the Oscar Corpus, Opus Corpora, and a Wikipedia dump. There are variations of BERTurk models that differ in vocabulary size, offering options of 32k and 128k, with both cased and uncased versions available.
- ***ConvBERTurk:*** The ConvBERT Base model has been improved with additions to facilitate using BERT in less challenging tasks [19]. This model requires fewer parameters for operation. ConvBERTurk, on the other hand, is pre-trained in the Turkish language for over 1 million steps with a sequence length of 512, employing a methodology that differs from the conventional approach [20]. ConvBERTurk mC4, a variant of the ConvBERTurk model, is developed utilizing the C4 dataset.

3.3. Tokenization

BERT's pre-trained models utilize a Tokenizer, and each model may require a specific one. For optimal performance of the BERT model, it is essential to divide the text into tokens, which are defined as small text segments. BERT models accept a maximum input of 512 tokens by default [21]. This number also includes special tokens such as [CLS], which is prepended for classification, and [SEP], which indicates where the token belongs [22, 23]. In this study, the text lengths within the dataset were adjusted to not exceed 512 tokens in the BERT models, and the same dataset was utilized across all models. Table 3 presents each model's maximum, minimum, and average token counts.

Table 3. Token Counts of the Dataset as Per the Model Tokenization

Method	Stance	Token Count		
		Min	***Max***	***Average***
BERT-Base	Against	3	505	134
	Favor	3	497	119
mBERT	Against	2	350	94
	Favor	2	368	84
BERTurk 32k	Against	1	242	66
	Favor	2	268	58
BERTurk 128k	Against	1	210	55
	Favor	1	225	48
ConvBERTurk	Against	1	213	56
	Favor	1	268	50
ConvBERTurk mC4	Against	1	213	56
	Favor	1	268	50

Table 4 illustrates the tokenization and token count of the BERT models for the sample comment, “başıboş sokak köpeklerinin toplanması gerekiyor gereğinin yapılmasını bekliyoruz [24]” (stray dogs need to be collected we expect the necessary to be done). This comment comprises eight words and is categorized under the “Favor” class in the dataset.

Table 4. Tokenization of Bert Models in Detail

Method	Tokenization	Token Count
BERT-Base	['b', '##as', '##ı', '##bos', 'so', '##ka', '##k', 'k', '##ope', '##kle', '##rini', '##n', 'top', '##lan', '##mas', '##ı', 'g', '##ere', '##ki', '##yo', '##r', 'g', '##ere', '##gin', '##in', 'ya', '##p', '##ı', '##lma', '##s', '##ı', '##n', '##ı', 'be', '##k', '##li', '##yo', '##ru', '##z']	39
mBERT	['bas', '##ıb', '##os', 'sok', '##ak', 'kop', '##ek', '##lerinin', 'top', '##lanması', 'ger', '##eki', '##yor', 'ger', '##egi', '##nin', 'ya', '##pı', '##lması', '##nı', 'be', '##kli', '##yor', '##uz']	24
BERTurk 32k	['bası', '##bos', 'sokak', 'kop', '##ekler', '##inin', 'toplanması', 'gerekiyor', 'ger', '##eg', '##inin', 'yapılmasını', 'bekliyoruz']	13
BERTurk 128k	['bası', '##bos', 'sokak', 'kopek', '##lerinin', 'toplanması', 'gerekiyor', 'geregi', '##nin', 'yapılmasını', 'bekliyoruz']	11
ConvBERTurk	['başı', '##bo', '##ş', 'sokak', 'köpekler', '##inin', 'toplanması', 'gerekiyor', 'gereğini', '##n', 'yapılmasını', 'bekliyoruz']	12
ConvBERTurk mC4	['başı', '##bo', '##ş', 'sokak', 'köpekler', '##inin', 'toplanması', 'gerekiyor', 'gereğini', '##n', 'yapılmasını', 'bekliyoruz']	12

The “##” symbols represent the splitting of words into smaller pieces of the words. Each tokenization has a specific splitting strategy based on the natural language in which it is pre-trained. The BERT-Base model performed with the highest number of tokenizations, with the mBERT model following closely behind. The Turkish BERT models also demonstrated considerable effectiveness in segmenting sentences into tokens. Also, it was observed that ConvBERTurk and ConvBERTurk mC4 models tokenized the given samples in the same way. In their studies, Kaya and Tantuğ [25] stated that a tokenizer working in English made 2.5 times more word splitting when tokenizing in Turkish.

3.4. Evaluation Metrics

Performance metrics are derived from a confusion matrix, as outlined in Table 5. The prediction results can lead to one of four possible outcomes defined by the confusion matrix: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). TP and TN represent the regions where the model makes accurate predictions, whereas FP and FN denote the regions where the model's predictions are inaccurate.

Table 5. Confusion Matrix

Actual Label	**Predicted Label**		
		Positive (Favor)	Negative (Against)
	Positive (Favor)	True Positive (TP)	False Negative (FN)
	Negative (Against)	False Positive (FP)	True Negative (TN)

Metrics, including accuracy rate, precision, recall, and F1 score, were employed to assess the performances of the models. The formulas for these performance metrics are presented in Equations 2 through 5.

$$Accuracy\ Rate = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{2}$$

$$Precision = \frac{TP}{(TP + FP)} \tag{3}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{4}$$

$$F1\ score = \frac{2 * Recall * Precision}{Recall + Precision} \tag{5}$$

4. Experimental Results

The experiments were conducted on a Tesla T4 GPU via Google Colab. Key hyperparameters, including learning rate, batch size, number of epochs, and learning optimizer, were set the same across all models. All models were trained for four epochs at a batch size of 32 and a learning rate of 5e-5 with the AdamW optimizer. A total of six different BERT case models, four Turkish, one multilingual and one base, were fine-tuned. As Grimminger and Klinger [7] did in their study, the dataset was partitioned into 80% for training and 20% for testing purposes. The TensorFlow and Transformers libraries were employed during the language model training process. All models were trained in approximately 24 minutes. The results are presented in Table 6, including accuracy, precision, recall, and F1 score.

It was observed that every model, except for BERT-Base, attained a success rate exceeding 80%. The best performance on the stance detection problem was obtained through the BERTurk 128k model with 87.10% accuracy, 87.11% precision, 87.10% recall and 87.10% F1 score. BERTurk32k emerged as the second-best model, attaining an accuracy rate of 86.50%.

The findings also revealed that the models accurately classified "Favor" examples more than "Against" examples, except for mBERT and ConvBERTurk. This result aligned with the research conducted by Grimminger and Klinger [7], who observed that the BERT-Base classifier achieved higher accuracy in identifying the "Favor" class. In contrast, detecting the "Against" class proved more challenging. Additionally, in the "Against" class, the BERT-Base model achieved the lowest accuracy rate of 68.80%, while the BERTurk 128k model attained the highest accuracy rate of 86.20%. In the "Favor" class, the mBERT model recorded the lowest accuracy rate at 83%, whereas the BERTurk32k model achieved the highest accuracy rate of 88.60%. Another notable finding is that the BERT base model outperformed the ConvBERTurk and ConvBERTurk mC4 models by achieving a success rate of 88% in classifying the Favor class.

Table 6. Stance Classification Results

Method	Stance	Accuracy	Precision	Recall	F1 score
BERT-Base	Against	68.80	85.15	68.80	76.11
	Favor	88.00	73.83	88.00	80.29
	Weighted Avg.	78.40	79.49	78.40	78.20
mBERT	Against	83.40	83.07	83.40	83.23
	Favor	83.00	83.33	83.00	83.17
	Weighted Avg.	83.20	83.20	83.20	83.20
BERTurk 32k	Against	84.40	88.10	84.40	86.21
	Favor	88.60	85.03	88.60	86.78
	Weighted Avg.	86.50	86.57	86.50	86.49
BERTurk 128k	Against	86.20	87.78	86.20	86.98
	Favor	88.00	86.44	88.00	87.22
	Weighted Avg.	**87.10**	**87.11**	**87.10**	**87.10**
ConvBERTurk	Against	84.80	84.13	84.80	84.46
	Favor	84.00	84.68	84.00	84.34
	Weighted Avg.	84.40	84.40	84.40	84.40
ConvBERTurk mC4	Against	82.60	85.16	82.60	83.86
	Favor	85.60	83.11	85.60	84.34
	Weighted Avg.	84.10	84.13	84.10	84.10

Research on Turkish stance detection has been carried out utilizing datasets from diverse fields, including sports, war, and vaccination. The studies incorporating the BERTurk model are presented alongside the current research in Table 7.

The summary table indicates that Küçük and Arıcı [10] attained an F1 score of 69.93% when focusing on Covid-19 vaccination, while Zengin et al. [11] recorded the highest F1 score of 68.90%, with Trabzonspor target under experimental conditions where the same target was utilized in both the training and testing datasets. Also, the F1 score values Arslan and Fırat [12] reported were 78.7% for the Russia target and 87.0% for the Ukraine target. In our study, both BERTurk 32k and 128k models attained notable levels of success in F1 scores and obtained values of 86.49% and 87.10%, respectively. Additionally, Arslan and Fırat [12] reported a classification accuracy of 78.4% when the target was Russia and 87.2% when the target was Ukraine in their research. The highest accuracy observed in the present study was 87.10%, which closely aligned with the findings of Arslan and Fırat.

5. Qualitative Analysis

LIME [26] was employed to gain insights into the predictive mechanisms of the models and to determine which words were most influential for the predicted labels. Table 8 illustrates the prediction results of each model for a sample text belonging to the "Favor" class. The text that is used for the assessment of this case is: "desteklediğim karar artık sokaklar güvenli olacak herkes için [27]" (the decision I support will now make the streets safe for everyone). Words highlighted in orange signify support (Favor) for the corresponding predicted label, while those in blue indicate opposition (Against). A darker color denotes a greater level of impact.

All models accurately classified the text that was designated as Favor. Moreover, it was observed that the word "desteklediğim" (I support), which indicated that the commenter supported this law, significantly influenced the decision-making processes of the models. Additionally, the word "güvenli" (safe) was another prominent word in decision-making processes and was associated with the "Favor" class.

Table 9 presents the prediction results of each model for a sample text categorized under the "Against" class. The text that is referenced for the evaluation of this case is: "başka çareler mümkünken en basit yola başvurulması üzücü [28]" (it is heartbreaking that the simplest method was used when other solutions were possible).

Table 7. Comparison with Similar Studies

<table>
<tr><th>Study</th><th>Dataset Size</th><th>Method</th><th>Results</th></tr>
<tr><td>Küçük and Arıcı [10]</td><td>830</td><td>BERTurk 32k</td><td><table><tr><th>Target</th><th>Accuracy</th><th>F1 score</th></tr><tr><td>Covid-19 Vaccination</td><td>-</td><td>69.93%</td></tr></table></td></tr>
<tr><td>Zengin et al. [11]</td><td>830

Galatasaray : 353
Fenerbahçe : 276
Beşiktaş : 100
Trabzonspor : 101</td><td>BERTurk 32k</td><td><table><tr><th>Target</th><th>Accuracy</th><th>F1 score</th></tr><tr><td>Galatasaray</td><td>-</td><td>63.30%</td></tr><tr><td>Fenerbahçe</td><td>-</td><td>58.90%</td></tr><tr><td>Beşiktaş</td><td>-</td><td>57.60%</td></tr><tr><td>Trabzonspor</td><td>-</td><td>68.90%</td></tr></table>*Single target, same target in the train and test set results</td></tr>
<tr><td>Arslan and Fırat [12]</td><td>8215

Ukraine: 3264
Russia : 4951</td><td>BERTurk 128k</td><td><table><tr><th>Target</th><th>Accuracy</th><th>F1 score</th></tr><tr><td>Russia</td><td>78.4%</td><td>78.7%</td></tr><tr><td>Ukraine</td><td>87.2%</td><td>87.0%</td></tr></table></td></tr>
<tr><td rowspan="2">Current Study</td><td rowspan="2">5000</td><td>BERTurk 32k</td><td><table><tr><th>Target</th><th>Accuracy</th><th>F1 score</th></tr><tr><td>Stance Animals Law</td><td>86.50%</td><td>86.49%</td></tr></table></td></tr>
<tr><td>BERTurk 128k</td><td><table><tr><th>Target</th><th>Accuracy</th><th>F1 score</th></tr><tr><td>Stance Animals Law</td><td>87.10%</td><td>87.10%</td></tr></table></td></tr>
</table>

Among the models, only the BERTurk 128k model correctly associated the sample text with the Against class. The word “üzücü” (it is sad), which indicated the commentator’s disagreement with this law, was understood by all models as a sign of support for the “Against” label. Nevertheless, this interpretation did not provide the models with enough information to make a correct final decision. Moreover, the word “çareler” (remedies) had a considerable impact on the models that performed “Favor” class prediction.

In addition, the decision-making processes of the BERTurk 128k model, which achieved the highest performance on the dataset, were analyzed under various cases, including TP, TN, FP, and FN, using LIME. The findings are illustrated in Figures 2-5.

- **Case 1: True Positive**

Figure 3 illustrates a case where both the actual and predicted class of the text is Favor. The sample text that is used during the examination of this case is: “sonuna kadar desteklediğim uygulamadır [29]” (this is an application that I fully support).

Table 8. LIME Results of Various Fine-Tuned BERT Models for Favor Class Sample Text

Method	Sample Text	Prediction Probabilities
BERT-Base	desteklediğim karar artık sokaklar güvenli olacak herkes için	Against 0.03 Favor 0.97
mBERT	desteklediğim karar artık sokaklar güvenli olacak herkes için	Against 0.03 Favor 0.97
BERTurk 32k	desteklediğim karar artık sokaklar güvenli olacak herkes için	Against 0.00 Favor 1.00
BERTurk 128k	desteklediğim karar artık sokaklar güvenli olacak herkes için	Against 0.00 Favor 1.00
ConvBERTurk	desteklediğim karar artık sokaklar güvenli olacak herkes için	Against 0.01 Favor 0.99
ConvBERTurk mC4	desteklediğim karar artık sokaklar güvenli olacak herkes için	Against 0.01 Favor 0.99

Table 9. LIME Results of Various Fine-Tuned BERT Models Against Class Sample Text

Method	Sample Text	Prediction Probabilities
BERT-Base	başka çareler mümkünken en basit yola başvurulması üzücü	Against 0.06 Favor 0.94
mBERT	başka çareler mümkünken en basit yola başvurulması üzücü	Against 0.34 Favor 0.66
BERTurk 32k	başka çareler mümkünken en basit yola başvurulması üzücü	Against 0.44 Favor 0.56
BERTurk 128k	başka çareler mümkünken en basit yola başvurulması üzücü	Against 1.00 Favor 0.00
ConvBERTurk	başka çareler mümkünken en basit yola başvurulması üzücü	Against 0.26 Favor 0.74
ConvBERTurk mC4	başka çareler mümkünken en basit yola başvurulması üzücü	Against 0.15 Favor 0.85

The word “desteklediğim” (I support), which reflected a supportive attitude, contributed to the classification of this text as Favor.

Figure 3. Result of Using LIME on Sample Text where the Actual Label and Predicted Label are Favor

- **Case 2: True Negative**

Figure 4 presents a case where both the actual and predicted class of the text is Against. The text that is utilized as a sample in the evaluation of this case is: "tamam saldırganlar için çözüm gerek ama o kadar masumları da var ki nasıl kıyacaksınız o garibanlara çok üzücü [30]" (ok a solution is needed for the aggressive ones but there are so many innocent ones how can you kill those poor ones it is very sad.). The words "kıyacaksınız" (you will kill), "üzücü" (sad) and "masumları" (innocents) were crucial in guiding the classifier towards an Against prediction.

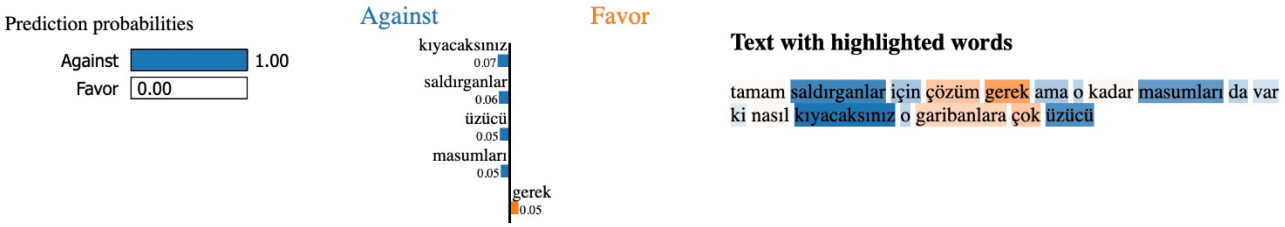

Figure 4. Result of Using LIME on Sample Text where the Actual Label and Predicted Label are Against

- **Case 3: False Positive**

Figure 5 shows a case where the actual class of the sample text is Against, while the predicted class is Favor. The text that is employed for the analysis of this case is: "inşallah uygulanmayacak olan öneridir köpeklerin insanlara zararı yoktur pek [31]" (I hope this is a suggestion that will not be applied dogs do not harm people much). The words "zararı" (harm) and "yoktur" (there is no) played significant roles in leading the classifier to make a Favor prediction.

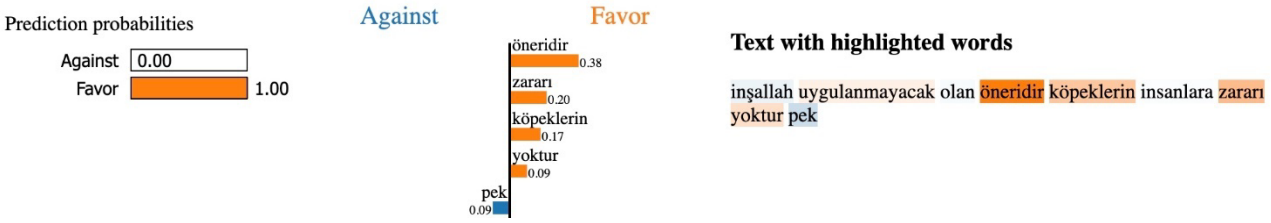

Figure 5. Result of Using LIME on Sample Text where the Actual Label is Against but the Predicted Label is Favor

- **Case 4: False Negative**

Figure 6 depicts a case where the actual class of the sample text is Favor, whereas the predicted class is Against. The text that is referenced during the assessment of this case is: "hadi inşallah korkudan sokağa çıkamaz oldu çoluk çocuk [32]" (hope so children cannot go out on the street because of fear). The words "oldu" (happened) and "korkudan" (fear) were significant in determining the classification of this text as Against.

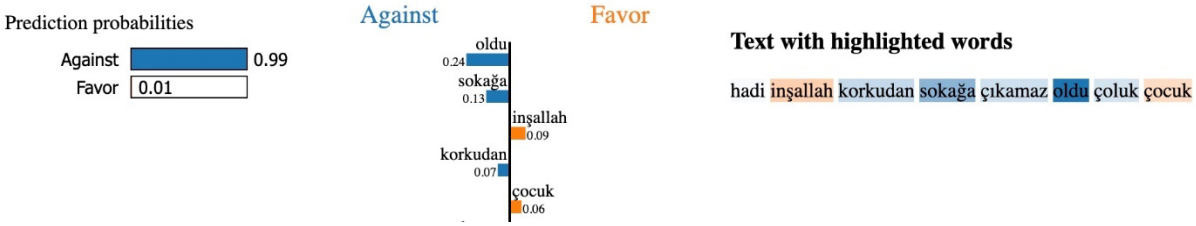

Figure 6. Result of Using LIME on Sample Text where the Actual Label is Favor but the Predicted Label is Against

6. Conclusion

In recent times, social media has evolved into a crucial medium for the distribution of information, enabling individuals to express their views and apprehensions on various subjects. Stance detection studies, which involve the automated examination of text to ascertain the author's stance on a specific proposition or topic, have attracted the attention of researchers and have become a significant focus of research.

This study aims to identify the most effective BERT model for the Turkish stance detection task. To our knowledge, this subject has not been addressed in existing literature. In this context, user comments concerning the Stray Animals Law, which has generated significant discussions and commentaries in Turkey, were collected from Ekşi Sözlük. A total of 5000 comments were manually classified, with 2500 labeled as "Favor" and 2500 as "Against." The performances of fine-tuned language models, including BERT-Base, mBERT, BERTurk 32k, BERTurk 128k, ConvBERTurk and ConvBERTurk mC4, were evaluated in terms of accuracy, precision, recall and F1 score. The experimental findings indicated that BERTurk 128k outperformed all other BERT models. It achieved an accuracy of 87.10%, precision of 87.11%, recall of 87.10% and F1 score of 87.10%. Additionally, most models were more successful in correctly predicting comments labeled as "Favor" than "Against."

This research presents a novel instance within the scarce Turkish stance detection studies. It highlights the effectiveness of BERT models, particularly those tailored for the Turkish language, in stance detection.

Research of this nature and others like it could play a crucial role in the automated extraction of public opinions, enabling governments to formulate policies on animal rights, vaccination, and climate change efficiently and cost-effectively.

As a future work, this study could be expanded to use alternative tokenization techniques and various strategies to improve the performance of BERT models on larger Turkish texts.

References

[1] D. Küçük and F. Can, "Stance detection: A survey," ACM Computing Surveys (CSUR), vol. 53, no.1, pp. 1-37, 2020.
[2] S. Mohammad et al., "Semeval-2016 task 6: Detecting stance in tweets," Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016). 2016.
[3] R. Cao et al., "Stance detection for online public opinion awareness: An overview," International Journal of Intelligent Systems, vol. 37 pp. 11944-11965, 2022.
[4] D. Küçük and F. Can, "Stance detection on tweets: An svm-based approach," arXiv preprint arXiv:1803.08910, 2018.
[5] S. Ghosh et al., "Stance detection in web and social media: a comparative study," Experimental IR Meets Multilinguality, Multimodality, and Interaction: 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9–12, 2019, Proceedings 10. Springer International Publishing, 2019.
[6] L-A. Cotfas et al., "The longest month: analyzing COVID-19 vaccination opinions dynamics from tweets in the month following the first vaccine announcement," Ieee Access, vol. 9, pp. 33203-33223, 2021.
[7] L. Grimminger and R. Klinger, "Hate towards the political opponent: A Twitter corpus study of the 2020 US elections on the basis of offensive speech and stance detection," arXiv preprint arXiv:2103.01664, 2021.
[8] K.K. Polat, N. G. Bayazıt, and O. T. Yıldız, "Türkçe duruş tespit analizi," Avrupa Bilim ve Teknoloji Dergisi vol. 23, pp.99-107, 2021
[9] D. Küçük, and N. Arıcı, "Sentiment analysis and stance detection in Turkish tweets about COVID-19 vaccination," Handbook of research on opinion mining and text analytics on literary works and social media. IGI Global, 371-387, 2022.
[10] D. Küçük, and N. Arıcı, "Deep learning-based sentiment and stance analysis of Tweets about Vaccination," International Journal on Semantic Web and Information Systems (IJSWIS), vol. 19, no.1, pp.1-18, 2023.
[11] M. S. Zengin, B. U. Yenisey, and M. Kutlu, "Exploring the impact of training datasets on Turkish stance detection," Turkish Journal of Electrical Engineering and Computer Sciences, vol. 31, no.7, pp.1206-1222, 2023.
[12] S. Arslan and E. Fırat, "Stance Detection on Short Turkish Text: A Case Study of Russia-Ukraine War," Afyon Kocatepe Üniversitesi Fen Ve Mühendislik Bilimleri Dergisi, vol. 24, no. 3, pp. 602-619, 2024.
[13] Ekşi Sözlük, "29 temmuz 2024 sokak hayvanları yasası değişikliği," Available at https://eksisozluk.com/entry/166760652 (Accessed Date: 08.08.2024)
[14] Ekşi Sözlük, "29 temmuz 2024 sokak hayvanları yasası değişikliği," Available at https://eksisozluk.com/entry/166761830 (Accessed Date: 08.08.2024)
[15] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.
[16] A. Vaswani et al., "Attention is all you need." Advances in neural information processing systems, 30, 2017.
[17] P. Savci and B. Das, "Comparison of pre-trained language models in terms of carbon emissions, time and accuracy in multi-label text classification using AutoML," Heliyon, vol. 9, issue. 5, 2023.
[18] Hugging Face, "Models," Available at https://huggingface.co/models?other=bert (Accessed Date: 12.08.2024)

[19] Z-H. Jiang et al., "Convbert: Improving bert with span-based dynamic convolution," Advances in Neural Information Processing Systems, vol. 33, pp. 12837-12848, 2020.
[20] Hugging Face, "dbmdz Turkish ConvBERT model," Available at https://huggingface.co/dbmdz/convbert-base-turkish-cased (Accessed Date: 12.08.2024)
[21] X. Chen, P. Cong and S. Lv, "A Long-Text Classification Method of Chinese News Based on BERT and CNN," Ieee Access, vol. 10, pp. 34046-34057, 2022.
[22] Medium, "Handle Long Text Corpus for Bert Model," Available at https://medium.com/@priyatoshanand/handle-long-text-corpus-for-bert-model-3c85248214aa (Accessed Date: 13.08.2024)
[23] Medium, "Fine-tuning BERT model for arbitrarily long texts, Part 1," Available at https://medium.com/mim-solutions-blog/fine-tuning-bert-model-for-arbitrarily-long-texts-part-1-299f1533b976 (Accessed Date: 13.08.2024)
[24] Ekşi Sözlük, "29 temmuz 2024 sokak hayvanları yasası değişikliği", Available at https://eksisozluk.com/entry/166761830 (Accessed Date: 29.09.2024)
[25] Y. B. Kaya and A. C. Tantuğ, "Effect of Tokenization Granularity for Turkish Large Language Models," Journal of Intelligent Systems with Applications, vol. 21, 2024.
[26] M. T. Ribeiro, S. Singh, and C. Guestrin, " Why should I trust you?" Explaining the predictions of any classifier, In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135-1144, 2016.
[27] Ekşi Sözlük, "sahipsiz hayvanlara yönelik kanun teklifi," Available at https://eksisozluk.com/entry/166275172 (Accessed Date: 23.09.2024)
[28] Ekşi Sözlük, "sokak hayvanları uyutulacak," Available at https://eksisozluk.com/entry/164628545 (Accessed Date: 23.09.2024)
[29] Ekşi Sözlük, "sokak hayvanları uyutulacak," Available at https://eksisozluk.com/entry/164641295 (Accessed Date: 11.01.2025)
[30] Ekşi Sözlük, "sahipsiz hayvanlara yönelik kanun teklifi," Available at https://eksisozluk.com/entry/166318051 (Accessed Date: 11.01.2025)
[31] Ekşi Sözlük, "14 günde sahiplenilmeyen köpeklerin uyutulması," Available at https://eksisozluk.com/entry/159777438 (Accessed Date: 11.01.2025)
[32] Ekşi Sözlük, "sokak hayvanları uyutulacak", Available at https://eksisozluk.com/entry/164651341 (Accessed Date: 11.01.2025)

Author(s) Contributions
Selma Alav: Data curation, Investigation, Software, Visualization, Writing-Original draft preparation
Kristin S. Benli: Conceptualization, Methodology, Investigation, Software, Visualization, Writing-Original draft preparation.

Conflict of Interest Notice
Authors declare that there is no conflict of interest regarding the publication of this paper.

Ethical Approval and Informed Consent
It is declared that during the preparation process of this study, scientific and ethical principles were followed, and all the studies benefited from are stated in the bibliography.

Availability of data and material
The data that support the findings of this study are available from the authors upon reasonable request.

Plagiarism Statement
This article has been scanned by iThenticate ™.