Sentiment Analysis from Turkish News Texts with BERT-Based Language Models and Machine Learning Algorithms

Engin Demir
Department of Computer Engineering
Bursa Uludağ University
Bursa Türkiye
512231006@ogr.uludag.edu.tr

Metin Bilgin
Department of Computer Engineering
Bursa Uludağ University
Bursa Türkiye
metinbilgin@uludag.edu.tr

2023 8th International Conference on Computer Science and Engineering (UBMK) | 979-8-3503-4081-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/UBMK59864.2023.10286719

***Abstract*- Sentiment analysis is defined as text analysis and is defined as identifying the class that the text wants to express emotionally. In this study, sentiment analysis was performed with BERT-based language models and machine learning algorithms on the data obtained from Turkish news texts. ALBERT, DistilBERT, and RoBERTa were used as BERT-based language models, and Naive Bayes, Support Vector Machine, and Random Forest methods were used as machine learning algorithms. Our dataset contains 5000 two-class (positive-negative) sentences, with 90% of the data used for training and 10% for testing. When the results of the experimental studies are examined, the accuracy values of the studies performed with language models have reached higher values than machine learning algorithms. The success rates of the language models are DistilBERT, RoBERTa, and ALBERT and the values obtained are 80%, 80%, and 77% respectively. The ranking of machine learning algorithms is Naive Bayes, Support Vector Machine, and Random Forest and the values obtained are 71%, 68%, and 68%.**

Keywords*- *Language Models, Machine Learning, BERT, Sentiment Analysis.

I. INTRODUCTION

Today, natural language processing (NLP) technology has become an important tool for developing artificial intelligence systems that help humans understand texts written in natural language. Sentiment analysis is a field of artificial intelligence used to recognize and understand emotional content from text, audio, visual or other types of data. It is also called emotion recognition or emotion detection. Sentiment analysis combines techniques from linguistics, psychology, and artificial intelligence to understand people's emotional states.

In the study by Güven, Turkish sentiment analysis was conducted on product reviews obtained from an e-commerce site. The study compared the success rates between language models and machine learning models. Among machine learning models, the Naive Bayes model achieved the highest success rate with 89.95%, and among language models, ELECTRA achieved the highest success rate with 92.54% [1]. In his study, Tayşi aims to develop a Turkish-specific system for emotional state detection using audio data. In the study, a dataset consisting of 2194 recordings of anger, sadness, excitement, happiness, helplessness, and neutral emotions was created and various classification algorithms were used. When the experimental results of the study are analyzed, the Random Forest algorithm obtained the best results [2]. Maşrifoğlu et al. conducted evaluations using BERT-based models and machine learning algorithms on Turkish data collected through Net Promoter Score (NPS) surveys of customers using banking services. It was stated that the Turkish BERT model achieved the highest score with 91% F1-Score and machine learning models also achieved a close performance [3]. Wandhavan and Aggarwal compared CNN, LSTM, and Bi-LSTM with BERT-based language models using a dataset of 150,000 Hindi-English tweets. In the results of the study, the BERT language model reached the highest score with 71.43% [4]. Dani et al. applied K-Nearest Neighbor and Decision Trees machine learning techniques on the Toronto Emotional Speech Set dataset for 7 emotions. In the results of the study, they stated that the K-Nearest Neighbor technique provided 98% accuracy, Decision Trees 92%, and Extra-Tree classifier 99% [5]. Jahan et al. analyzed suicidal texts for the prediction of suicidal thoughts using BiLSTM and BERT, ALBERT, RoBERTa, and XLNET language models. The RoBERTa language model achieved the highest result with 95.21% accuracy [6].

In this study, BERT-based language models and machine learning algorithms are used to compare the accuracy, Precision, Recall, and F1-Score metrics on a dataset of Turkish news texts. Thus, it is aimed to contribute to Turkish natural language processing studies.

II. METHODS AND IMPLEMENTATION

In this section, information about the dataset, the methods used, and the experimental studies are presented.

A. Data Set

The dataset used in the study consists of 5000 positive and negative data including Turkish news texts. Of the data labeled as positive and negative, 2950 are negative and 2050 are positive [7].

The data were preprocessed before processing with machine learning models and BERT-based language models. These preprocesses are: lowercase conversion, punctuation removal, number removal and stop words removal. After this preprocessing, the data reached a certain standard.

B. Machine Learning Algorithms

In this section, information about the machine learning algorithms used in the study will be presented.

1. Naïve Bayes

Naive Bayes is a widely used classification algorithm in machine learning. Basically, it calculates the probabilities that features in a dataset belong to a class and classifies new instances using these probabilities. The basic assumption of this algorithm is that all features are independent of each other,

979-8-3503-4081-5/23/$31.00 ©2023 IEEE

i.e. one feature does not affect the other. This is why it is called "naive". The naive Bayes algorithm is used in many applications such as spam filtering, sentiment analysis, and feature selection as well as classification problems. It is also widely used in industry and research fields due to its easy applicability and high performance [8].

2. *Support Vector Machine*

Support Vector Machine (SVM) is a widely used classification and regression algorithm in the field of machine learning. SVM uses a hyperplane to classify data points into classes, and this hyperplane best passes through data points with a marginal distance separating the classes. SVM improves classification accuracy by maximizing this marginal distance. SVM also allows data points to be classified in high-dimensional spaces using kernel functions. SVM is widely used to provide high accuracy on high-dimensional and complex datasets and is used in many application areas such as image processing, bioinformatics, and web page classification. The advantages of SVM include its high accuracy, low data requirement, and scalability [8].

3. *Random Forest*

Random Forest is an ensemble-based classification and regression algorithm widely used in machine learning. This algorithm is built by combining multiple decision trees and each tree is trained with a different set of sub-features and data samples. This allows it to better adapt to variability and noise in the data set and produce more consistent results. Random Forest is widely used in classification and regression problems and is used in many application areas such as image processing, bioinformatics, financial analysis, and marketing. Moreover, the Random Forest algorithm can be used on large datasets due to its fast training and high performance [8].

C. *BERT-Based Language Models*

Language models are artificial intelligence models used in natural language processing to understand, create or process text-based data. These models are typically trained with a large dataset and then used to perform text-based tasks. Among the areas where language models are most widely used are natural language processing, automatic text generation and translation, and speech recognition.

In this section, information about the BERT language models used in the study will be presented.

1. *DistilBERT*

DistilBERT is a compressed version of the BERT language model. BERT is a deep learning-based language model that has revolutionized the field of natural language processing. DistilBERT is designed to achieve a faster and more efficient language model by reducing the size and computational cost of the BERT model. For this purpose, DistilBERT removes some layers of the BERT model and makes the remaining layers smaller. This reduces the size of the model by up to 40%, while drastically reducing its speed and memory usageDistilBERT provides faster training and inference with a slight compromise on the performance of the BERT model and maybe a particularly suitable option for mobile and embedded systems due to its lower hardware requirements. DistilBERT has been successfully used in many natural language processing tasks such as natural language understanding, text classification, and sentiment analysis [9].

2. *RoBERTa*

RoBERTa (Robustly Optimized BERT Pretraining Approach) is an improved version of the BERT language model. The BERT model is a deep learning-based language model that has revolutionized the field of natural language processing. RoBERTa is designed to obtain a better language model by optimizing the training phase of BERT. To this end, RoBERTa modifies some of the assumptions used in the learning process of the BERT model and is trained longer on a larger datasetThanks to these changes, RoBERTa builds a better language model and achieves higher accuracy in natural language understanding tasks. RoBERTa has achieved state-of-the-art results in the field of language models and has been successfully used in many natural language processing tasks such as natural language understanding, text classification and sentiment analysis [10].

3. *ALBERT*

The ALBERT (A Lite BERT) language model is a lighter and faster version of the BERT language model. The BERT model is a deep learning-based language model that has made a huge impact in the field of natural language processing. However, training the BERT model is quite costly and time-consuming, and therefore a lighter and faster version of the BERT model is needed. ALBERT uses a more efficient training method to improve the sharing of parameters and learning efficiency in the BERT model. This makes the ALBERT model lighter and faster than the BERT model, while at the same time matching the performance of the BERT model. ALBERT is used in many natural language processing tasks such as natural language understanding, text classification and sentiment analysis [11].

D. *Experimental Study Results*

This study was conducted in Google's Colab environment. ALBERT, DistilBERT, and RoBERTa language models were used as BERT-based models. In addition, machine learning algorithms such as Naive Bayes, Support Vector Machine (SVM), and Random Forest (RF) methods are also applied and compared with language models. All models were trained on training data and then their performance was measured on test data.

For the implementation of the study, the training-test data ratio was applied as 90% to 10%. 4500 training data and 500 test data were used.

In order to compare the performance of BERT-based language models and machine learning algorithms, the accuracy metric, sensitivity metric, nominal value, and F1-Score metrics were used. Definitions and expressions for these metrics are given below.

Accuracy: Accuracy is the summation part of classification. That is, it can also be characterized as truths (negative and positive) / total data.

$$Accuracy = \frac{TN+TP}{Total\ Data} \qquad (1)$$

Precision: It is a performance criterion term that shows how close the measurements are to each other as a result of the tests. This value is calculated by dividing the number of classified positive samples by the total number of positive samples. The range of this value is between 0 and 1.

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

Recall: It is known as the target attainment rate. In other words, it is a value that shows to what extent the information that needs to be reached has been reached. This value is calculated by dividing the number of correctly classified relevant positive samples by the total number of relevant documents.

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

F1-Score: It is obtained by calculating the harmonic averages of sensitivity and rated values. This value is between 0 and 1. It is the value that expresses the accuracy of the test performed.

$$F1 - Score = \frac{2*Precision*Recall}{Precision+Recall} \qquad (4)$$

Accuracy metric is used to compare the performance of BERT-based language models and machine learning algorithms. The accuracy metric represents the percentage of correct classification of the model.

According to the results obtained, BERT-based models DistilBERT, RoBERTa and ALBERT achieved accuracy rates of 80%, 80% and 77%, respectively.

The machine learning algorithms Naive Bayes, SVM and RF methods achieved accuracy rates of 71%, 68% and 68% respectively. The results of the accuracy metric are given in Figure 1.

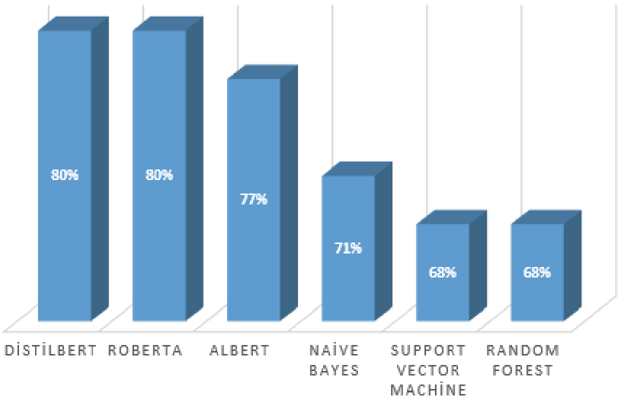

Fig.1. Results of the study

The sensitivity value, nominal value, and F1-Score metrics obtained with BERT-based language models and machine learning algorithms are given in Table 1.

TBALE 1. PRECISION, RECALL, and F1-SCORE RESULTS OBTAINED with BERT-based LANGUAGE MODELS and ACHINE LEARNING ALGORITHMS

Language Models and Algorithms	Precision	Recall	F1-Score
ALBERT	0.77	0.86	0.75
RoBERTa	0.82	0.80	0.81
DistilBERT	0.78	0.77	0.78
NB	0.7	0.68	0.70
SVM	0.66	0.66	0.66
RF	0.67	0.65	0.65

The results show that RoBERTa and DistilBERT perform better than ALBERT in the evaluation of language models. In machine learning algorithms, Naïve Bayes algorithm performed better than SVM and RF algorithms.

In the overall application, language models achieved higher accuracy rates and emerged as a more effective option for sentiment analysis.

III. Discussion

Sentiment analysis was performed on the data obtained from Turkish news texts. Naïve Bayes, Support Vector Machine, and Random Forest algorithms from machine learning models, ALBERT, DistilBERT, and RoBERTa language models from BERT-based language models were used for these analyses. These algorithms and language models were chosen because they are frequently used in the literature and are prominent in sentiment analysis studies.

In the analysis of BERT-based language models, DistilBERT and RoBERTa language models obtained the best results. Looking at both model structures separately, DistilBERT's structure, with fewer layers and reduced memory usage, has been effective in achieving better results. The RoBERTa language model, on the other hand, obtained a better result in this analysis because the size of the datasets used in its structure and the training time contributed to better learning of the model compared to other models.

In machine learning algorithms, the Naïve Bayes algorithm achieved the best result. The fact that the Naïve Bayes algorithm considers the data as independent variables and uses this more effectively in sentiment analysis with two labels has enabled it to achieve the best accuracy metric result.

When we compare our study with similar studies in the literature, we observe that language models are more successful than machine learning models and the RoBERTa language model achieves better results than other language models.

IV. Results

Sentiment analysis was performed using BERT-based models and classical machine learning algorithms on the data obtained from Turkish news texts. ALBERT, DistilBERT, and RoBERTa language models and classical machine learning algorithms such as Naive Bayes, Support Vector Machine, and

Random Forest are used in this analysis. Out of 5000 two-class (positive-negative) data, 90% is allocated for training and 10% for testing.

According to the results of the study, sentiment analysis studies using language models achieved higher accuracy rates compared to classical machine learning algorithms. DistilBERT, RoBERTa, and ALBERT language models achieved 80%, 80%, and 77% accuracy rates, respectively. These results show that BERT-based models are an effective option for sentiment analysis in Turkish news texts. On the other hand, classical machine learning algorithms such as Naive Bayes, Support Vector Machine, and Random Forest achieved accuracy rates of 71%, 68%, and 68% respectively. These results show that language models outperform classical methods in sentiment analysis.

This study emphasizes the importance of sentiment analysis studies on the data obtained from Turkish news texts. Language models have been able to achieve higher accuracy rates in sentiment analysis using large, pre-trained language representations. These results demonstrate the potential of language models in natural language processing.

In the future, the results of studies using more comprehensive data sets and different language models can be further analyzed. Furthermore, by exploring new techniques and methods in sentiment analysis, it will be possible to achieve higher accuracy rates.

REFERENCES

[1] ZA Guven, Türkçe Ürün Yorumları için BERT, ELECTRA ve ALBERT Dil Modellerinin Duygu Analizine Etkisi,2021.

[2] F. D. Tayşi, Konuşma Verisinden Duygu Durum Tespiti, 2019.

[3] M.Masrifoglu, U. Tıgrak, S. Hakyemez, G. Gul, E. Bozan, AH. Buyuklu, A. Ozgur, Bankacılık Alanında Müşteri Yorumlarının BERT Tabanlı Yaklaşımlar ile Duygu Analizi, 2021.

[4] A. Wandhavan, A. Aggarwal, Towards Emotion in Hindi-English Code Mixed Data: A Tranformer Based Approach, 2021.

[5] Mande AA, Dani S, Telang S, Shao Z. EMOTION DETECTION USING AUDIO DATA SAMPLES. International 2019.

[6] RU. Nur, S. Jahan, Z. Mahmud, FM. Shah, A Transformer Based Approach to Detect Suicidial Ideation Using Pre-Trained Language Models, 2020.

[7] https://www.kaggle.com/datasets/engindmr/seskayitverileri

[8] E. Demir, A. Tepecik, Türkçe Ses Kayıt Verilerinin CountVectorizer ve TFIDFVectorizer Modelleri Olarak Google Colab Platformunda ve RapidMiner'da Makine Öğrenmesi Algoritmalarıyla Analizi, 2022.

[9] V. Sanh, L. Debut, J. Chaumond, T. Wolf, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, 2020.

[10] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa : A Robustly Optimized BERT Pretraining Approach, 2019.

[11] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations, 2020.