Evaluation of ChatGPT and BERT-based Models for Turkish Hate Speech Detection

Nur Bengisu Çam
Department of Computer Engineering
Bogazici University
Istanbul, Türkiye
bengisu.cam@boun.edu.tr

Arzucan Özgür
Department of Computer Engineering
Bogazici University
Istanbul, Türkiye
arzucan.ozgur@boun.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/UBMK59864.2023.10286663

***Abstract*—The popularity of large language models (LLMs) is increasing day by day. ChatGPT is one of the most popular LLMs. It is known for its success in many areas of natural language processing (NLP). Most importantly, we have yet to find zero-shot performance on various NLP tasks for low-level languages such as Turkish. Detection of hate speech is among the most important problems in NLP. With the growing social media usage, the prevalence of hate speech has also increased. However, automatic detection of hate speech in Turkish is rare compared to studies conducted in English. In our work, we analyzed the performance of ChatGPT and various fine-tuned BERT-based transformer models in detecting hate speech in Turkish. We found that ChatGPT provides similar results to the BERT-based models in detecting Turkish hate speech; thus, it is promising. In this study, a dataset consisting of 1000 Turkish tweets labeled "hate," "aggressor," and "none" was used.**

Keywords — hate speech; ChatGPT; BERT; transformers; classification; Turkish.

I. Introduction

Nowadays, there are many social media platforms. These platforms allow anonymous sharing of content. This leads to an increase in hateful content. Therefore, automatic hate speech detection is among the most critical topics in natural language processing (NLP) [1]. It has been observed that hate speech is expressed on Twitter, especially in terms of ethnicity, language, religion, and gender [2]. Traditional machine learning based methods have been investigated for detection of hate speech in various languages, including Turkish [3, 4]. With the improvements in deep learning models, GRU and LSTM based models have also been widely used for hate speech detection [5, 6]. In recent works, BERT and its variant models have been studied for hate speech detection in various languages, including Turkish. [7, 8, 9, 10]. After the publication of the ChatGPT [11], its performance in detecting hate speech in English was analyzed [12, 13, 14]. As with many tasks in NLP, there are fewer studies for Turkish than for English. To our knowledge, there is no study that investigates ChatGPT performance on hate speech detection in Turkish.

A. BERT-based Models

The transformer architecture introduced by Vaswani et al. has led to significant advances in numerious NLP tasks [16]. BERT is among the most popular Large Language Models (LLMs) based on the same architectural model proposed by Vaswani et al. and developed by Google [17]. BERT-based and other LLMs are trained on an enormous amount of data. Once trained, they can predict the next word and almost understand the language. These LLMs can be used for various NLP tasks by fine-tuning a small amount of task-specific data.

B. ChatGPT

ChatGPT is an API that serves GPT-based models trained with human feedback reinforcement learning. It was developed by OpenAI. The experiments in this study used the GPT-3.5 version provided by ChatGPT. The parameters of the BERT-based models and ChatGPT (GPT-3.5) are listed in Table I.

TABLE I. ChatGPT And BERT Parameters

Model Details	
Model	***Number of Parameters***
ChatGPT (GPT-3.5)	175 B
BERT-base	110 M

After the release of the ChatGPT API, people began to analyze the performance of the zero-shot setting on many different NLP tasks. Zhu et al. measured ChatGPT performance on stance detection, hate speech, sentiment analysis, and bot detection in English [14]. Their results show that ChatGPT performs the worst in detecting hate speech [14]. Zhong et al. found that the zero-shot results of ChatGPT is comparable to the BERT-based models which are fine-tuned on hate speech detection for English [15].

In the study by Rehana et al. the performance of ChatGPT and GPT-based models in extracting relationships between protein entities in biomedical texts was measured and compared with the fine-tuned BERT-based models [18].

Jiao et al. done their studies in evaluating and comparing the performance of machine translation in different languages. They found that the GPT-4 model provided better results [19]. Gilardi et al. found that the performance of ChatGPT in text classification outperformed that of MTurk taggers while costing

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

Authorized licensed use limited to: University of Plymouth. Downloaded on September 29,2024 at 09:25:17 UTC from IEEE Xplore. Restrictions apply.

much less [20]. Huang et al. measured ChatGPT performance in classifying implicitly made hate speech using 795 tweets in English. They found that ChatGPT correctly classified most of the tweets [12]. Li et al. measured the classification performance of ChatGPT on texts containing hate speech, offensive speech, and toxic content in English [13]. As can be seen, measuring the zero-shot performance of LLMs and ChatGPT is becoming increasingly popular in various NLP subjects. Since no work has been made in Turkish hate speech detection, we aimed to contribute to the literature by comparing the zero-shot setting results of ChatGPT with the fine-tuned BERT-based.

II. Dataset

The Turkish dataset created by Mayda et al. was used [3]. There are 1000 Turkish tweets in this dataset. In this dataset, the keywords for the topics about which the most hate speech was made were determined as "Syrians, Armenians, British, Kurds, Turks, Greeks, Arabs, Alevis, Muslims, Jews, atheists, infidels, Christians, women, perverts, homosexuals" [3]. There are 100 tweets for each of the keywords "Syrian"," "Armenian" and "Jew" and 50 tweets for each of the other keywords [3]. After collecting the tweets related to the keywords, each of them is labeled as "hate," "aggressor," and "none." In Table II, amount of tweets for each label is given.

TABLE II. Number of Tweets In Each Label

Dataset Details	
Label	***Number of Tweets***
Hate	276
Aggression	60
None	664
Total	**1000**

III. Models And Methods

The dataset was divided into five separate folds for k-fold cross-validation. Next, the BERT-based models were trained with the 5-fold cross-validation. After each fold, both training and test results were reported.

Three different experiments were performed in this work. The first two are classifications with two labels and one is a classification with three labels. In the first classification, the "aggression" tweets are counted as "hate." In this way, the performances of the ChatGPT and the BERT-based models were analyzed when "aggressive" tweets were accepted as "hate" containing very similar sentences in terms of sentiment. This experiment is referred to as "Experiment 1".

In the second experiment, the label "aggression" and the associated tweets were removed from the dataset. Thus, we analyze how the classification results would change if the label "aggressor" and the examples were not present at all. This experiment is referred to as "Experiment 2".

In the third experiment, the data set remained unchanged in its original form. Each of the three labels and their tweets were used for classification. This experiment is referred to as "Experiment 3."

For each fold of the 5-fold dataset, the zero-shot evaluation of ChatGPT was performed by prompting five tweets at each query. As with the BERT models, ChatGPT classification was performed for "Experiment 1," "Experiment 2," and "Experiment 3."

The query was chosen similarly to the study by Rehana et al. [18]. To prevent ChatGPT from being memorized, the API page was refreshed after each query. In addition, the experiments were repeated three times for a more accurate analysis. The results of the three trials were averaged.

An example of a query and a ChatGPT response can be seen in Figure 1. Since the tweets were in Turkish, the query was also written in Turkish during the experiments. For better understanding, you can also find the English version of the query and ChatGPT response in Figure 2.

Query (in Turkish)	Verilen tweetlerin hepsi için nefret söylemi durumlarını incele. Her bir tweeti "nefret", "saldırgan" veya "hiçbiri" etiketlerinden biri olarak sınıflandır ve sıra no, etiket şeklinde paylaş. Tweetler: “diyanet avrupa birliğinden hıristiyan birleşik devletlerine kadar yardım alıyor bilen var mıdır.”, “bu kız sapık gibi yaa bir rahat bırak çocuğu sinir olmaya başladım buna”, “sen de arap maymununa benzemişsin”, “arap yağı çok bulunca gerisini siz anladınız”, “ya biz süryaniyiz ya onlar alevi!”
ChatGPT Response	1. Hiçbiri 2. Saldırgan 3. Nefret 4. Nefret 5. Hiçbiri

Fig. 1. Query Example in Turkish

Query (in English)	Review the hate speech statuses for all given tweets. Classify each tweet as one of the hashtags "hate", "aggression" or "none" and share it as a sequence number, label Tweets: . “diyanet avrupa birliğinden hıristiyan birleşik devletlerine kadar yardım alıyor bilen var mıdır.”, “bu kız sapık gibi yaa bir rahat bırak çocuğu sinir olmaya başladım buna”, “sen de arap maymununa benzemişsin”, “arap yağı çok bulunca gerisini siz anladınız”, “ya biz süryaniyiz ya onlar alevi!”
ChatGPT Response	1. None 2. Aggression 3. Hate 4. Hate 5. None

Fig. 2. Query Example in English

Authorized licensed use limited to: University of Plymouth. Downloaded on September 29,2024 at 09:25:17 UTC from IEEE Xplore. Restrictions apply.

IV. EXPERIMENTAL RESULTS

In addition to ChatGPT, four number of fine-tuned BERT models were used. Macro, precision, recall, and F1 values are averaged for "Experiment 1," "Experiment 2," and "Experiment 3." Row-based normalized confusion matrices were created for the ChatGPT and BERT models, which performed best in each experiment after three trials of 5-fold cross-validation. Confusion matrices were then averaged for three trials.

During fine-tuning of the BERT models, 10% of the data set was used for validation. The hyperparameters were selected *considering* the best performance *of the models o*n *the* validation set. These hyperparameters are listed in Table III. Pytorch [21] and Hugging Face [22] libraries were used for training.

The classification result for "Experiment 1" is shown in Table IV. According to the classification results, the model BERT-base-offensive-3 performed the best in all metrics. The ChatGPT achieved the second best result based on the recall score, with a difference of 0.8%.

TABLE III. HYPERPARAMETERS

Hyperparameters	**Values**
Epoch	6
Batch	8
Learning Rate	5e-5
Weight Decay	0.1

On the other hand, the ChatGPT performed similarly to the BERT-base-offensive-1 and BERT-base-offensive-2 models. For "Experiment 1," the confusion matrices of the best BERT-based model, which is BERT-base-offensive-3, and ChatGPT are provided in Table V and Table VI, respectively. The matrices were normalized by setting the column sums for each row to one and rounding the values to the nearest hundredth.

TABLE IV. "EXPERIMENT 1" AVERAGED SCORES ON TEST DATA

	Metrics		
Model	***Precision***	***Recall***	***F1***
BERT-base-128k[1]	0.741	0.706	0.710
BERT-base-offensive-1[2]	0.700	0.692	0.691
BERT-base-offensive-2[3]	0.703	0.686	0.687
BERT-base-offensive-3[4]	**0.770**	**0.724**	**0.735**
ChatGPT (GPT-3.5)	0.699	0.716	0.666

TABLE V. "EXPERIMENT 1" AVERAGED CONFUSION MATRIX FOR BERT-BASE-OFFENSIVE-3

		Predicted	
	Label	***None***	***Hate***
Real	***None***	0.90	0.10
	Hate	0.44	0.56

TABLE VI. "EXPERIMENT 1" AVERAGED CONFUSION MATRIX FOR CHATGPT

		Predicted	
	Label	***None***	***Hate***
Real	***None***	0.59	0.41
	Hate	0.14	0.86

The classification result for "Experiment 2" is shown in Table VII. According to the classification results, the model BERT-base-offensive-3 performed the best in all metrics. The ChatGPT achieved the second best result based on the recall score with a difference of 0.4%. The model BERT-base-offensive-2 scored second best in precision and F1 scores. For "Experiment 1," the best results of the BERT-based model, which is BERT-base-offensive-3 and the ChatGPT confusion matrices are provided in Table VIII and Table IX, respectively. The matrices were normalized by setting the column sums for each row to one and rounding the values to the nearest hundredth.

TABLE VII. "EXPERIMENT 2" AVERAGED SCORES ON TEST DATA

	Metrics		
Model	***Precision***	***Recall***	***F1***
BERT-base-128k[1]	0.639	0.644	0.628
BERT-base-offensive-1[2]	0.690	0.670	0.674
BERT-base-offensive-2[3]	0.709	0.695	0.693
BERT-base-offensive-3[4]	**0.740**	**0.714**	**0.721**
ChatGPT (GPT-3.5)	0.665	0.710	0.642

TABLE VIII. "EXPERIMENT 2" AVERAGED CONFUSION MATRIX FOR BERT-BASE-OFFENSIVE-3

		Predicted	
	Label	***None***	***Hate***
Real	***None***	0.91	0.09
	Hate	0.62	0.38

Authorized licensed use limited to: University of Plymouth. Downloaded on September 29,2024 at 09:25:17 UTC from IEEE Xplore. Restrictions apply.

TABLE IX. "EXPERIMENT 2" AVERAGED CONFUSION MATRIX FOR CHATGPT

		Predicted	
	Label	***None***	***Hate***
Real	***None***	0.59	0.41
	Hate	0.13	0.87

The classification result for "Experiment 3" is shown in Table X. The classification results of three labels show that the model BERT-base-offensive-2 performs best in all metrics. On the other hand, the model BERT-base-offensive-3 achieved the second best results. ChatGPT achieved the third best result based on recall score, with a difference of 1.4%. The BERT-base-128k model achieved the lowest performance. For "Experiment 1," the confusion matrices of the best BERT-based model, which is BERT-base-offensive-2 and ChatGPT are provided in Table XI and Table XII. The matrices were normalized by setting the column sums for each row to one and rounding the values to the nearest hundredth.

TABLE X. "EXPERIMENT 3" AVERAGED SCORES ON TEST DATA

	Metrics		
Model	***Precision***	***Recall***	***F1***
BERT-base-128k[1]	0.472	0.471	0.458
BERT-base-offensive-1[2]	0.576	0.529	0.539
BERT-base-offensive-2[3]	**0.641**	**0.572**	**0.584**
BERT-base-offensive-3[4]	0.616	0.554	0.564
ChatGPT (GPT-3.5)	0.497	0.540	0.467

TABLE XI. "EXPERIMENT 3" AVERAGED CONFUSION MATRIX FOR BERT-BASE-OFFENSIVE-2

		Predicted		
	Label	***None***	***Hate***	***Aggression***
Real	***None***	0.81	0.16	0.03
	Hate	0.38	0.60	0.02
	Aggression	0.23	0.38	0.39

TABLE XII. "EXPERIMENT 3" AVERAGED CONFUSION MATRIX FOR CHATGPT

		Predicted		
	Label	***None***	***Hate***	***Aggression***
Real	***None***	0.59	0.22	0.19
	Hate	0.16	0.58	0.26
	Aggression	0.06	0.41	0.53

V. RESULTS & FUTURE WORK

In this study, the performance of ChatGPT and BERT-based models in classifying Turkish hate speech was measured and compared. It was found that some fine-tuned BERT-based models and ChatGPT performed similarly on Turkish hate speech. Based on the precision, recall, and F1 results of the experiments, it can be seen that Experiment 1 scores are better than the other two.

Due to the similarity of the labels "offensive" and "hate," the models performed better when these labels were treated as a single label. As in Experiment 3, the performance of both ChatGPT and BERT-base models decreased significantly when the task consisted of classifying three labels. The confusion matrices of Experiment 1 and Experiment 2 show that the BERT-based models performed better when classifying the"none" labeled tweets. On the other hand, ChatGPT performed better when classifying the "hate" labeled tweets.The confusion matrix of Experiment 3 shows that the BERT-based models again performed better in classifying tweets labeled "none," while ChatGPT performed better in classifying tweets labeled "hate" and "aggression."

Our experiments show that the zero-shot setting ChatGPT results are comparable to the fine-tuned BERT-based models in Turkish hate speech detection. However, ChatGPT needs to be improved to exceed the fine-tuned performances of the BERT models. In our work, we used the dataset which the performance of the BERT models has not yet been analyzed before. Moreover, this work is the initial study which evaluates ChatGPT performance on Turkish hate speech task.

In the future, we would like to further evaluate ChatGPT and its variants on different datasets for Turkish hate speech detection. The study will pave the way for future ChatGPT evaluations for Turkish hate speech detection.

VI. ACKNOWLEDGEMENT

TUBA-GEBIP Award of the Turkish Science Academy (to A.O.) is gratefully acknowledged.

[1]huggingface.co/dbmdz/bert-base-turkish-128k-uncased

[2]huggingface.co/Overfit-GM/bert-base-turkish-cased-offensive

[3]huggingface.co/Overfit-GM/bert-base-turkish-128k-uncased-offensive

[4]huggingface.co/hemekci/off_detection_turkish

Authorized licensed use limited to: University of Plymouth. Downloaded on September 29,2024 at 09:25:17 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] Zhang, Z., & Luo, L. (2019). Hate speech detection: A solved problem? the challenging case of long tail on twitter. Semantic Web, 10(5), 925-945.

[2] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in Tweets," *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*, pp. 759-760, 2017

[3] İ. MAYDA, B. DİRİ, and T. YILDIZ, "Türkçe Tweetler üzerinde Makine öğrenmesi Ile Nefret söylemi tespiti," *European Journal of Science and Technology*, no. 24, pp. 328-334, 2021.

[4] D. C. Asogwa, C. I. Chukwuneke, C. C. Ngene, and G. N. Anigbogu, "Hate speech classification using SVM and naive Bayes," *arXiv preprint arXiv:2204.07057,* 2022.

[5] Z. Zhang, D. Robinson, and J. Tepper, "Detecting hate speech on Twitter using a convolution-gru based deep neural network," *In The Semantic Web: 15th International Conference, ESWC 2018*, Proceedings 15 (pp. 745-760), 2018.

[6] A. Bisht, A. Singh, H. S. Bhadauria, J. Virmani, and Kriti, "Detection of hate speech and offensive language in Twitter data using LSTM model," *Advances in Intelligent Systems and Computing*, pp. 243–264, 2020.

[7] M. Mozafari, R. Farahbakhsh, and N. Crespi, "A Bert-based transfer learning approach for hate speech detection in online social media," *Complex Networks and Their Applications VIII*, pp. 928–940, 2019.

[8] H. Sohn and H. Lee, "MC-BERT4HATE: Hate speech detection using multi-channel Bert for different languages and translations," *2019 International Conference on Data Mining Workshops (ICDMW)*, pp. 928–940, 2019.

[9] Beyhan, F., Çarık, B., Arın, İ., Terzioğlu, A., Yanikoglu, B., & Yeniterzi, R. (2022, June). A Turkish hate speech dataset and detection system. In Proceedings of the Thirteenth Language Resources and Evaluation Conference (pp. 4177-4185).

[10] Hüsünbeyi, Z. M., Akar, D., & Özgür, A. (2022, June). Identifying Hate Speech Using Neural Networks and Discourse Analysis Techniques. In Proceedings of the First Workshop on Language Technology and Resources for a Fair, Inclusive, and Safe Society within the 13th Language Resources and Evaluation Conference (pp. 32-41).

[11] OpenAI, "CHATGPT: Optimizing language models for dialogue," OpenAI, https://openai.casa/blog/chatgpt/ (accessed Aug. 4, 2023).

[12] F. Huang, H. Kwak, and J. An, "Is Chatgpt better than human annotators ? Potential and limitations of ChatGpt in explaining implicit hate speech," *arXiv preprint arXiv:2302.07736*, 2023.

[13] L. Li, L. Fan, S. Atreja, and L. Hemphill, "" HOT" ChatGPT: The promise of ChatGPT in detecting and discriminating hateful, offensive, and toxic comments on social media," *arXiv preprint arXiv:2304.10619*, 2023.

[14] Y. Zhu, P. Zhang, E. U. Haq, P. Hui, and G. Tyson, "Can chatgpt reproduce human-generated labels? a study of social computing tasks," *arXiv preprint arXiv:2304.10145*, 2023.

[15] Q. Zhong, L. Ding, J. Liu, B. Du, and D. Tao, "Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert," *arXiv preprint arXiv:2302.10198*, 2023.

[16] A. Vaswani et al., "Attention is all you need," *Advances in neural information processing systems*, 30, 2017.

[17] J. Devlin, M. W. Chang, K. Lee, nad K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[18] H. Rehana, N. B. Çam, M. Basmaci, Y. He, A. Özgür, and J. Hur, "Evaluation of GPT and BERT-based models on identifying protein-protein interactions in biomedical text," *arXiv preprint arXiv: 2303.17728*, 2023

[19] W. Jiao, W. Wang, J. T. Huang, X. Wang, and Z. Tu, "Is chatgpt a good translator? yes with gpt-4 as the engine," *arXiv preprint arXiv:2301.08745,* 2023.

[20] F. Gilardi, M. Alizadeh, and M. Kubli, "Chatgpt outperforms crowd-workers for text-annotation tasks," *arXiv preprint arXiv:2303.15056*, 2023.

[21] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, 32, 2019.

[22] T. Wolf et al., "Transformers: State-of-the-art natural language processing," *In Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pp. 38-45, 2020.

Authorized licensed use limited to: University of Plymouth. Downloaded on September 29,2024 at 09:25:17 UTC from IEEE Xplore. Restrictions apply.