Evaluating ChatGPT's Ability to Detect Hate Speech in Turkish Tweets

Somaiyeh Dehghan[1,2] and **Berrin Yanikoglu**[1,2]
[1]Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul, Turkey 34956
[2]Center of Excellence in Data Analytics (VERIM), Sabanci University, Istanbul, Turkey 34956
{somaiyeh.dehghan, berrin}@sabanciuniv.edu

Abstract

ChatGPT, developed by OpenAI, has made a significant impact on the world, mainly on how people interact with technology. In this study, we evaluate ChatGPT's ability to detect hate speech in Turkish tweets and measure its strength using zero- and few-shot paradigms and compare the results to the supervised fine-tuning BERT model. On evaluations with the SIU2023-NST dataset, ChatGPT achieved 65.81% accuracy in detecting hate speech for the few-shot setting, while BERT with supervised fine-tuning achieved 82.22% accuracy. This results supports previous findings that show that, despite its much smaller size, BERT is more suitable for natural language classifications tasks such as hate speech detection.

1 Introduction

ChatGPT, developed by OpenAI (OpenAI.), has revolutionized the way people interact with technology. As a state-of-the-art language model, ChatGPT leverages the power of deep learning to understand and generate human-like text, enabling natural and coherent conversations. Its applications range from question answering in various domains, to generating creative content like writing, poetry, and more. Thanks to its tremendous success as a large language model, there has been interest to test its abilities in various natural language understanding problems, such as sentiment analysis and hate speech detection.

Hate speech refers to any form of communication, in speech, writing, or behavior, that offends, threatens, or insults individuals or groups based on attributes such as race, ethnicity, religion, sexual orientation, disability, or gender (Beyhan et al., 2022). Hate speech detection, followed by potential measures such as blocking or counter-speech, is aimed to create safer digital spaces. Detecting hate speech is a challenging problem, since hate speech is subjective, context-dependent, and the language of tweets show high variability with the use of contractions, emojis, and typos.

The performances of hate speech detection systems show a lot of variation in the literature, as researchers often report results on proprietary or different datasets. However, state-of-art methods often use transformer based models, such as BERT (Devlin et al., 2019) or ChatGPT (Brown and et al., 2020).

BERT (Devlin et al., 2019), a pre-trained contextual language model, is widely used to detect hate speech. BERT is a transformer-based model designed for various natural language processing tasks, such as sentiment analysis, named entity recognition, and hate speech detection. It was trained in an unsupervised manner by predicting masked words in a sentence.

ChatGPT (Brown and et al., 2020), on the other hand is also based on the transformer architecture, but is specifically designed for generating coherent and contextually relevant text given an input prompt. It is trained using a language modeling objective, where it learns to predict the next word in a sentence given the context of preceding words.

Related to the problem at hand, BERT uses a bidirectional context, which helps capture complex relationships and dependencies within the text. It is also free, open-source and much smaller (110 million parameters) compared to ChatGPT which has 175 billion parameters. Nonetheless, ChatGPT was selected in this work due to the interest it receives and relatively low cost[1].

In this study, we contribute to the body of work assessing ChatGPT's ability to detect implicit or explicit hate speech in Turkish tweets, as well as its estimation of the strength of hate speech. Its performance is compared to that of fine-tuned BERTurk classifier and regressor models.

[1]Its online use is free and API is cheaper than that of GPT-4s

Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024), pages 54–59

The rest of the paper is organized as follows: in Section 2, we provide a summary about related works; in Section 3, the dataset used to train and test our models is defined; in Section 4, the methodology is presented. Experiments are provided in Section 5. Finally, conclusions and future work are presented in Section 6.

2 Related Work

Many studies have been conducted to evaluate ChatGPT in detection of hate speech in English, each of which used different dataset, but similar studies are rare for the Turkish language. Studies show the importance of the prompts when using ChatGPT.

Among the recent works, Chiu et al. (2022) used ChatGPT to classify English text as sexist or racist. They used zero-, one-, and few-shot learning paradigms. For zero- and one-shot learning, they achieved an average accuracy between 55% and 67% depending on the category of text and type of learning. For few-shot learning, they used a different example set in prompt and they found that with few-shot learning, the model's accuracy could be as high as 85%.

Han and Tang (2022) used ChatGPT to detect hate speech and investigated designing effective prompts for better performance. They demonstrated that numbers of training examples in the prompt matters. Additionally, they discovered that giving the model clear instructions works better than other approaches for incorporating our past knowledge into the model and enhancing its functionality. They achieved accuracy of 86% and macro-F1 of 85% for English comments from YouTube and Reddit.

Huang et al. (2023) examined whether ChatGPT can be used for providing natural language explanations (NLEs) for implicit hateful speech detection. They reported that ChatGPT correctly identifies 80% of the implicit hateful tweets in their experiment setting. Additionally, they discovered that ChatGPT-generated NLEs tend to be interpreted as clearer than NLEs created by humans and can reinforce human perception. This does, however, underline the need for more caution when utilizing ChatGPT as a tool to aid in data annotation because, in the event that it makes a mistake, it may mislead lay people

Li et al. (2023) aimed to use the potential power of ChatGPT to detect harmful content in English.They evaluated ChatGPT in comprehending hateful, offensive, and toxic concepts. They showed that ChatGPT can achieve an accuracy of approximately 80% when compared to Amazon MTurker[2] annotations.

Das et al. (2023) evaluated ChatGPT's performance for multilingual and emoji-based hate speech detection for 11 languages. They achieved highest macro-F1 score (89.2%) for English language and lowest macro-F1 score for Hindi language (67.3%).

Similar to our study, Çam and Ozgur (2023) compared ChatGPT to BERT on a Turkish dataset containing 1,000 tweets against ethnic groups, with three labels (None, Aggressor, Hate). They conducted three different experiments: aggressor tweets was counted as hate, aggressor tweets was removed, and multi classification with these three labels. They also used different pretrained versions of Turkish BERT (BERTurk-base and BERTurk-offensive). In all three experiments, BERTurk-offensive (previously fine tuned with 31,277 Turkish twitter data) showed better performance than ChatGPT. They achieved highest F1 score of 66.6% for ChatGPT in their first experiment (aggressor tweets was counted as hate).

3 Dataset Overview

We use the extended version of the publicly available SIU2023-NST dataset[3] towards immigrants and refugees (İnanç Arın et al., 2023). This dataset is labelled in two ways: binary classification (based on presence or absence of hate speech) and multi-class classification, ranging from insult to threat of harm, assumed to correspond to the strength of hate speech. Tables 1 and 2 show the statistics for the train- and test-split of this dataset.

Table 1: Statistics for the train-split of the Refugee dataset.

Strength	Size	Category	Size
0: Not hate speech	4,477	0: No hate speech	4,477
1: Insult	864	1: Hate speech	1,447
2: Exclusion	410		
3: Wishing harm	137		
4: Threatening harm	36		
Total	5,924	-	5,924

[2] https://www.mturk.com/

[3] https://github.com/verimsu/hate-speech-detection-dataset

Zero-shot prompt	You are an AI language model trained to analyze and detect hate speech. Given a tweet, decide whether the post contains implicitly or explicitly hate speech and also measure the strength of hate speech on a scale of [0-4]. Give your prediction as a yes or no in lowercase for using the following JSON format: {"prediction": prediction, "strength": strength} **Tweet:** benim cebime bu ay 50 lira az giricekse 10 milyon suriyeli ve afgan ölebilir. tayyip çıkıp 40 milyar dolar harcadık bi 40 daha harcarız diyor. lan benim paramı harcıyon?
ChatGPT Response	**prediction:** yes, **strength: 3**

Figure 1: Our zero-shot prompt and ChatGPT response for an hate speech post towards refugees

Few-shot prompt	You are an AI language model trained to analyze and detect hate speech. Given a tweet, decide whether the post contains implicitly or explicitly hate speech and also measure the strength of hate speech on a scale of [0-4]. For example: **Tweet:** 'Depremden hemen sonra asker sahaya insin, milletin canını malını korusun dedim.. Yağmacılar çoğalıyor bunun sonu iyi değil..' **label:** no, **strength: 0** ; **Tweet:** 'Yağmacılar ile yağmacı olmayanlar birbirine karıştı, ceza keseceğim diye toplanan gruplar, söylem,duyum ile hareket ediyorlar,daha kötüye doğru gidiyoruz. yağmacılar' **label:** no, **strength: 0** ; **Tweet:** 'ne kadar sığınmacı varsa istisnasız sınır dışı edilmeli ve ülkemiz bunlardan temizlenmelidir. ne tedirginlikler ne göç politikası hakkında bilinçlendirilmeye ihtiyacımız yok, biz her şeyin farkındayız.' **label:** no, **strength: 0** ; **Tweet:** 'gerçekler ağır gelir suri tarih kitapları sizi isyancı terörist olarak yazacak gelecek nesillerdeki suriyeli çocuklar sizi böyle anacak, devlete kim ihanet ederse teröristtir bunun lamı cimi yoktur .' **label:** yes, **strength: 1** ; **Tweet:** 'Tırları yağmalayanları tokat manyağı yapan bir abimiz... Analar aslan doğurmuş helal olsun hırsızlara mallarımızı çaldırmayın ,! suriyeliler Deprem Yağmacılar' **label:** yes, **strength: 2** ; **Tweet:** 'suriyeli çetelerin evlilik vaadiyle kandırıp binlerce tl dolandırılan cahillere zerre kadar üzülmüyorum ...türkiye'de kadın kalmadı de mi? beter olun... 15 ocak çarşamba' **label:** yes, **strength: 3** ; **Tweet:** 'yagmacilar deprem HalukLevent şimdi bunların yagmacidan ne farkı kaldı vatan hainleri hırsızlar bunlar gibiler olduğu sürece daha başımıza çok işler gelir bizim Allah'ım sen kurunun yanında yasida yakma ama bunları cehennemin en dibine....' **label:** yes, **strength:** 4 ; Give your prediction as a yes or no in lowercase for using the following JSON format: {"prediction": label, "strength": strength} ; **Tweet:** Hocam bu yağmacılar gitsin artık ülkemdemülteciistemiyorum ültecilersınırdışıedilsin suriyelileriistemiyoruz SuriyelilerSehirlerdenCıkartın SuriyeliYağmacılar suriyelikatiller
ChatGPT Response	**prediction:** yes, **strength:** 4

Figure 2: Our few-shot prompt and ChatGPT response for an hate speech post towards refugees

Table 2: Statistics for the test-split of the Refugee dataset.

Strength	Size	Category	Size
0: Not hate speech 1: Insult 2: Exclusion 3: Wishing harm 4: Threatening harm	1,119 216 103 34 8	0: No hate speech 1: Hate speech	1,119 361
Total	1,480	-	1,480

4 Methodology

We evaluate two approaches , namely BERT and ChatGPT, to detect hate speech and measure the strength of hate speech. The two problems are formulated as a binary-classification problem and a regression problem respectively.

In the first approach, we fine-tune the BERTurk model in the Huggingface Transformer package[4], using a classification or regression head that consists of a linear layer on top of the pooled output. The input to both models are preprocessed to remove usernames, URLs and the # signs, while keeping the text of the hashtags.

For the classification problem, we use cross-entropy (CE) loss to fine-tune BERT:

$$L_{CE} = -\sum_{i=1}^{N} y_i log(\hat{y}_i) \tag{1}$$

where y_i is the target value for the ith input and $\hat{y}_i$ is the prediction.

For the regression problem, we used mean squared error (MSE) loss to fine-tune BERT:

$$L_{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2}$$

where y_i and $\hat{y}_i$ are desired and predicted values, respectively.

For the second approach, we use the ChatGPT with zero- and few-shot learning paradigms. For zero- and few-shot learning, we design two prompts to interact with ChatGPT as shown in Figure 1 and 2. Our few-shot prompt contains seven examples from train-split of the Refugee dataset, three of which are examples with non-hate label and four examples with hate labels ranging strength from 1 to 4.

[4]https://huggingface.co/docs/transformers

5 Experiments

We conduct two experiments: Experiment-1: binary classification problem (hateful and non-hateful); Experiment-2: regression problem for predicting strength of hate speech.

Using the transfer learning approach, we fine-tune BERTurk[5] model. We use the cross-entropy loss and mean-squared error (MSE) loss for the classification and regression problems respectively, using stratified 10-fold cross validation.

For zero- and few-shot learning, we use "ChatGPT-text-davinci-003" model as it is one of the most powerful versions of the GPT language model developed by OpenAI. It is trained on a larger and more diverse dataset and designed to generate high-quality natural language responses to a wide range of tasks, including language translation, summarization, question-answering, and more.

Tables 3 and 4 show the results for Experiment-1 and Experiment-2, respectively. Moreover, confusion matrices for these three models are show in Figure 3.

Classification Results: As show in Table 3, supervised BERTurk-CE achieved better performance (82.22% accuracy) compared to ChatGPT (70.81% with zero-shot and 65.81% with few-shot learning) in accuracy, macro-F1, precision, and recall values.

In the case of ChatGPT (zero-shot) and ChatGPT (few-shot), we see that although the accuracy of ChatGPT (zero-shot) is higher, ChatGPT (few-shot) has higher macro-F1, precision and recall values compared to it.

While we give accuracy along with the macro-F1 scores so that our results are comparable to those in the literature, we pay importance to macro-F1 score for ranking the systems since our data is imbalanced. Indeed, the confusion matrices shown in Figure 3 show that ChatGPT (few-shot) is able to correctly identify more positives (higher recall) and avoid more false positives (higher precision) compared to ChatGPT (zero-shot).

Regression Results: The mean squared errors are shown in Table 4. We observe that the BERTurk-MSE regressor has significantly lower MSE (0.46) compared to ChatGPT, with either paradigm (zero- or few-shot). In fact, we can say that without any dedicated training, ChatGPT is not able to predict the strength of hate speech, as its mean-squared

[5]https://huggingface.co/dbmdz/bert-base-turkish-uncased

Table 3: Classification results on Refugee dataset in Experiment-1 for detecting hate speech

	Refugee Dataset			
	Accuracy	Macro-F1	Precision	Recall
BERTurk-CE (supervised transfer learning)	**82.22**	**74.86**	**76.12**	**73.89**
ChatGPT-text-davinci-003 (zero-shot learning)	70.81	58.50	59.04	58.17
ChatGPT-text-davinci-003 (few-shot learning)	65.81	60.19	60.27	63.12

BERTurk-CE (supervised)

Predicted Label \ True Label	0	1
0	1009	110
1	153	208

ChatGPT (zero-shot)

Predicted Label \ True Label	0	1
0	927	192
1	240	121

ChatGPT (few-shot)

Predicted Label \ True Label	0	1
0	765	354
1	152	209

Figure 3: Confusion matrix for BERTurk-CE (supervised), ChatGPT (zero-shot), and ChatGPT (few-shot) models for binary classification in Experiment-1

Table 4: Regression results on Refugee dataset in Experiment-2 for estimating strength of hate speech

	Refugee Dataset
	Mean squared error
BERTurk-MSE (supervised transfer learning)	0.46
ChatGPT-text-davinci-003 (zero-shot learning)	2.49
ChatGPT-text-davinci-003 (few-shot learning)	3.10

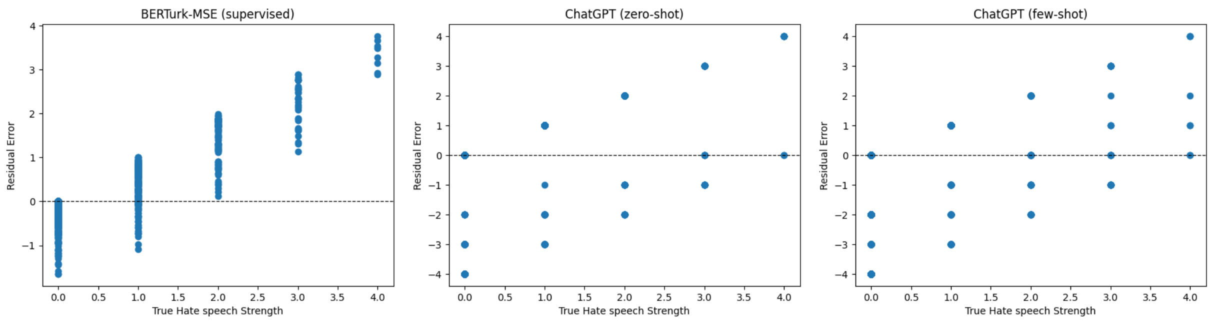

Figure 4: Residual error value for BERTurk-MSE (supervised), ChatGPT (zero-shot), ChatGPT (few-shot)

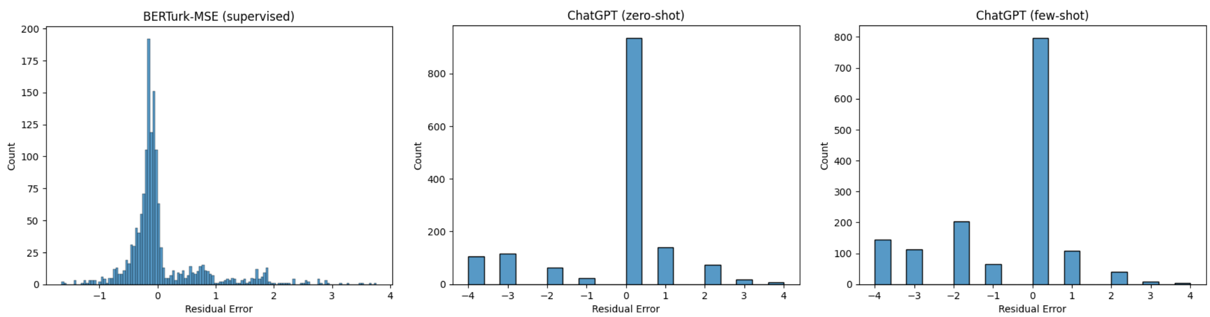

Figure 5: Residual value's histogram for BERTurk-MSE (supervised), ChatGPT (zero-shot), ChatGPT (few-shot)

error is 2.49 for zero-shot and 3.10 for few-shot cases.

The histogram of the residual errors of these approaches are shown in Figure 4 and Figure 5, respectively. Here, we see that the zero-shot paradigm outperforms the few shot with a slight margin.

6 Conclusions and Future Work

In this paper, we evaluate ChatGPT's ability for hate speech detection and measuring strength of hate speech in Turkish tweets. Our experimental results on the extended SIU2023-NST dataset show that fine-tuning the pre-trained BERTurk performs quite well for the challenging problem of hate speech detection. It achieves an accuracy of 82.22% and macro-F1 score of 74.86 in detecting hate speech and a mean square error of 0.46 in estimating the strength of the hate speech. These results are also significantly better than those obtained with ChatGPT, whether in zero- or few-shot paradigm.

Our experience with ChatGPT parallels previous results in the literature, showing that the performance depends strongly on the prompt. Possibly related to this, the relative results of ChatGPT with the zero- or few-shot paradigms are mixed: Zero-shot is better in terms of accuracy and MSE, while the few-shot is better in terms of precision, recall and macro-F1. On the other hand, the performance of the few-shot increased by increasing samples (from 3 to 7), as expected.

As a result, we suggest that ChatGPT may be used as an auxiliary tool in big data annotation. However, care must be taken in the design of prompt that the instructions are simple and clear and the number of samples is appropriate.

As future work direction, we aim to evaluate the explaining ability of ChatGPT in detecting hate speech.

7 Acknowledgements

This work was supported by the EU project "Combating Hate Speech and Discrimination Using Digital Technologies" (EuropeAid/170389/DD/ACT/Multi), carried out by the Hrant Dink Foundation.

References

Fatih Beyhan, Buse Çarık, İnanç Arın, Ayşecan Terzioğlu, Berrin Yanikoglu, and Reyyan Yeniterzi. 2022. A Turkish hate speech dataset and detection system. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 4177–4185.

Tom B. Brown and et al. 2020. Language models are few-shot learners. *ArXiv:2005.14165*.

Ke-Li Chiu, Annie Collins, and Rohan Alexander. 2022. Detecting hate speech with GPT-3. *arXiv:2103.12407*.

Mithun Das, Saurabh Kumar Pandey, and Animesh Mukherje. 2023. Evaluating ChatGPT's performance for multilingual and emoji-based hate speech detection. *arXiv:2305.13276*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 4171–4186.

Lawrence Han and Hao Tang. 2022. Designing of prompts for hate speech recognition with in-context learning. In *2022 International Conference on Computational Science and Computational Intelligence (CSCI)*.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is ChatGPT better than human annotators? potential and limitations of ChatGPT in explaining implicit hate speech. *arXiv:2302.07736*.

Lingyao Li, Lizhou Fan, Shubham Atreja, and Libby Hemphill. 2023. "HOT" ChatGPT: The promise of ChatGPT in detecting and discriminating hateful, offensive, and toxic comments on social media. *arXiv:2304.10619*.

OpenAI. *ChatGPT: Optimizing Language Model for Dialogue*.

Nur Bengisu Çam and Arzucan Ozgur. 2023. Evaluation of ChatGPT and BERT-based models for Turkish hate speech detection. In *Proceedings of the International Conference on Computer Science and Engineering (UBMK)*.

İnanç Arın, Zeynep Işık, Seçilay Kutal, Somaiyeh Dehghan, Arzucan Özgür, and Berrin Yanikoğlu. 2023. SIU2023-NST - hate speech detection contest. In *Proceedings of the 31. IEEE Conference on Signal Processing and Communications Applications, Istanbul*.