VRLLab at HSD-2Lang 2024: Turkish Hate Speech Detection Online with TurkishBERTweet

Ali Najafi[1], **Onur Varol**[1,2,*]
[1]Faculty of Engineering and Natural Sciences, Sabanci University
[2]Center of Excellence in Data Analytics, Sabanci University
[*]Corresponding author
{ali.najafi, onur.varol}@sabanciuniv.edu

Abstract

Social media platforms like Twitter - recently rebranded as X - produce nearly half a billion tweets daily and host a significant number of users that can be affected by content that is not properly moderated. In this work, we present an approach that ranked third at the `HSD-2Lang 2024` competition's subtask-A, along with additional methodology developed for this task and evaluation of different approaches. We utilize three different models, and the best-performing approach uses the publicly available TurkishBERTweet model with low-rank adaptation (LoRA) for fine-tuning. We also experiment with another publicly available model and a novel methodology to ensemble different hand-crafted features and outcomes of different models. Finally, we report the experimental results, competition scores, and discussion to improve this effort further.

1 Introduction

Despite the significant opportunities presented with the use of social media, these platforms are shifting towards more hostile environments, especially for marginalized groups. Social networks have been used to access information efficiently (Aral et al., 2009; Wang et al., 2022), participate important societal events (Bas et al., 2022; Ogan and Varol, 2017), and discuss political issues online (Varol et al., 2014; Tufekci, 2017; Jackson et al., 2020).

The increasing popularity of social networks and the opportunities presented to reach millions of individuals simultaneously made these platforms vulnerable to manipulation of discourse by bad actors who utilize automated accounts (Ferrara et al., 2016; Varol et al., 2017), spread disinformation (Mosleh and Rand, 2022; Keller et al., 2020), and coordinate targeted attacks (Shao et al., 2018; Varol and Uluturk, 2020). These targeted attacks can be coordinated or organic, and mostly, the target is minority and vulnerable groups. To prevent vulnerable groups and improve their experience in the online sphere, researchers develop systems to automatically identify these activities, and platforms build systems to moderate content and accounts.

Hate speech detection is a task to identify hateful content aimed towards groups such as refugees and individuals with certain beliefs or ethnicities (Waseem and Hovy, 2016; Zhang and Luo, 2019; MacAvaney et al., 2019). In this work, we demonstrate our approach as part of the `HSD-2Lang 2024` challenge to detect hate speech from textual information presented in social media posts.

2 Data

This challenge is organized in collaboration with the Hrant Dink Foundation for their ongoing project about "Media Watch on Hate Speech." Collaborative efforts of computational and social scientists defined hate speech on social media and carried out a detailed procedure to annotate posts around specific topics and keywords. The provided dataset in this competition contains 9,140 tweets in the context of Israel-Palestine and Turkish-Greek conflicts and content produced against refugees and immigration (Uludogan et al., 2024).

We preprocessed the dataset by removing samples with inconsistent ground truth information (exact text with different labels), and we applied deduplication, resulting in 8,805 tweets. Figure 1 shows word and character length distributions. When the ground-truth labels are considered, we measure that 30.5% of the dataset contains hate speech, suggesting an imbalance between the two classes. Since the dataset only contains the textual information presented in each tweet, we further processed them to take into account platform-specific features.

Removal of hyperlinks and mentions of other accounts in the tweets. This information could be valuable if we had a chance to process real-time data by scraping external web content or using profile information of accounts from Twitter's API since these fields are omitted in the dataset. Since

Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024), pages 185–189

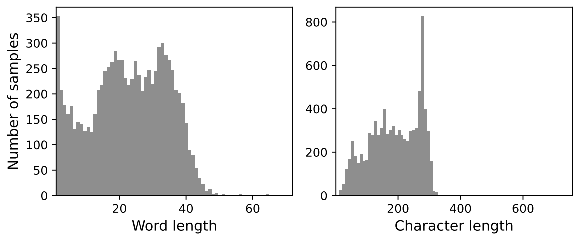

Figure 1: **Tweet statistics.** Distributions for word count (left) and character length (right) presented for the dataset. Character limits exhibit Twitter specific limitations while some tweets may contain fewer words possibly consist of hashtags.

we do not incorporate them into our analysis, we omit them from the dataset.

Preprocessing pipeline for TurkishBERTweet model. We consider different special tags for Twitter-specific entities and translated the Unicode characters of emojis to words describing the meaning using the preprocessor created for the TurkishBERTweet project (Najafi and Varol, 2023).

3 Methodologies

In this challenge, we built different approaches. We considered not only the textual data to fine-tune models but also incorporated additional signals obtained from text and blacklisted word dictionaries. Here, we present the language models used as the foundation and additional features we extracted to improve the model's performance. For the competition, we submitted the model with the best public leaderboard score; however, one of our approaches achieved an even higher score in the private evaluation. We presented all approaches and their respective performances in the results section.

TurkishBERTweet[1] is a new language model that was specifically trained on nearly 894M Turkish tweets and the model offers a special tokenizer that takes social media entities such as hashtags and emojis into account. This model utilized LoRA (Hu et al., 2021), which is a novel way of fine-tuning LLMs in an efficient way, and recent research reports state-of-the-art performance and generalizability capabilities (Najafi and Varol, 2023).

BERTurk[2] is a pre-trained model that utilizes large-scale corpus from various sources. It is a well-known model among the Turkish NLP community (Schweter, 2020).

[1]https://huggingface.co/VRLLab/TurkishBERTweet
[2]https://huggingface.co/dbmdz/bert-base-turkish-128k-uncased

Ensemble of models (EoM) approach combines outputs of aforementioned Hate Speech models along with custom features extracted for this task. These additional features consist of i) logits scores retrieved from an emotion classifier based on a bert-base model fine-tuned model for emotion analysis,[3] ii) logit scores of a sentiment classifier using TurkishBERTweet sentiment analysis model, iii) collection of Turkish blacked-list words[4] used for token level features such as binary exact match feature, Levenshtein distance, hashtag exact match, and hashtag Levenshtein distance. These features are concatenated, resulting in 16 features for the RandomForest classifier with 100 estimators trained to optimize gini-impurity. Since the outputs of ensemble models for imbalanced datasets can be biased, we calibrated the outputs of the model using Platt's scaling for interpreting output scores as probabilities (Niculescu-Mizil and Caruana, 2005).

4 Results

This section presents the experimental evaluation of approaches we tested within the dataset using stratified 5-fold cross-validation. We also report the performance of models we submitted to challenge for comparison. As Table 1 demonstrates, the `Ensemble of models` (EoM) gets the best performance compared to other approaches when all models are evaluated with 5-fold cross-validation. `TurkishBERTweet+Lora` model achieved the best private score, which led us to the third-best rank, although we observed a lower performance than the EoM model in cross-validated experiments. `BERTurk+Lora` model performed similarly to the TurkishBERTweet model using a 5-fold setting; however, it led to a lower private score. We suspect that the BERTurk model with standard or LoRA finetuning models was used by other teams, considering the popularity and availability of that model.

Considering the performance differences between public and private leaderboards, the EoM demonstrates less variability than the other two approaches. Even though it is not our best-performing model in both settings, we may consider it for our research projects since both cross-validated scores point to better performance, and the leaderboard score differences are negligible and can be due to

[3]https://huggingface.co/maymuni/bert-base-turkish-cased-emotion-analysis
[4]https://github.com/ooguz/turkce-kufur-karaliste

Table 1: **Model comparisons.** Weighted F1-score of the models in a 5-fold cross-validation setting. Best scores are presented in bold font, and more than one model is highlighted when the difference is not significant.

Model	F1-Weighted	Public Score	Private Score
TurkishBERTweet+LoRA	0.8137 ± 0.0059	$\mathbf{0.70697}$	$\mathbf{0.66431}$
BERTurk+LoRA	0.8132 ± 0.0054	$\mathbf{0.70476}$	0.64944
Ensemble of Models	$\mathbf{0.8941 \pm 0.0073}$	0.68544	$\mathbf{0.66103}$

noise in the test set of the competition.

We also conduct an error analysis to identify misclassifications that our model is making. This effort can reveal additional features we can implement and issues observed in the labeled dataset. Table 2 shows example tweets classified wrong. We first focus on false negatives since we can learn from these mistakes to improve our model. For instance, we could split hashtags into words to handle cases like #ülkemdemülteciistemiyorum (Turkish for #wedontwantrefugees) or handle popular hashtags differently. Regarding false positives, we noticed that our model correctly classifies tweets as hate speech based on our own judgment. We suspect the existence of mistakes in ground truth labels considering the examples we presented in Table 2. We highlight the words within the tweets that we suspect are mislabeling.

5 Discussion

In the provided dataset, we noticed tweets written in languages other than Turkish, such as Arabic and Hebrew. This could be an artifact of the data collection process, and one can consider i) language-level features, ii) filtering them, or iii) obtaining representation from LLMs. Furthermore, a study about the annotator's influence on the annotation quality for HateSpeech datasets shows that the expertise of annotators positively influences the data quality (Waseem, 2016). Considering the annotators' influence, applying impurity analysis by randomly or strategically changing the annotations and monitoring the Hate Speech system's performance could be a good practice.

Moreover, in this competition, we are only considering the text data to detect the existence of hate speech. Infusing the account information into these systems could help them be more accurate and reliable, such as the number of followers, number of followings, account creation date, etc.

Another approach for improving the performance of the systems is to expose pre-trained models with hateful content by further masked-language modeling on the hate speech dataset, like Caselli et al. (2020) presented in their recent work and improved the system's performance.

Multilingual models could also be utilized for this challenge since Turkish is a low-resource language, and the model can benefit from the other languages' hate speech datasets to infuse the broader knowledge of hate speech and then obtain a better performance (Röttger et al., 2022).

Recently, commercial models like ChatGPT have been used in various challenges. Huang et al. (2023) suggest that the ChatGPT demonstrates high accuracy and can be considered an alternative to human annotators in detecting implicit hate speech (Gilardi et al., 2023). Other work also investigated the performance of LLMs for hate-speech or offensive language detection tasks in English (Guo et al., 2024), Portuguese (Oliveira et al., 2023), and Turkish (Çam and Özgür, 2023). However, we want to raise a concern about the adversarial use of these models to attack vulnerable groups and bypass the detection systems. Additional information about accounts, network structure, and temporal activities should be incorporated into detection systems to address the mentioned risk.

6 Conclusion

In this challenge, the collective effort of research teams points to best practices and demonstrates the capabilities of the state-of-the-art models. Here, we demonstrated different approaches and their respective performances in detecting online hate speech toward three different groups. We obtained the third rank in the final leaderboard of the competition with the `TurkishBERT+Lora` model.

We hope language models like TurkishBERTweet will be used in different downstream tasks on Turkish social media. Research efforts especially need to assess the online participation of minority groups. There is a significant need for publicly available models since the quality of content moderation and use of automated accounts on platforms like X is questionable after the acquisition

Table 2: **Misclassification analysis.** We explored the errors of our model to improve further our approach (studying false negatives) and investigate issues with the ground-truth dataset (pointing to false positives). Here, we select instances where our model produces the correct outcome, but the annotation process suggests otherwise. We color the text in red that we believe suggests hate speech.

False positive Model predicts as HS Labeled no HS	• #Katilİsrail [URL] • Hükümet Cumhurbaşkanı Erdoğan Şerefsiz Suriyeliler Yağma Sizler şu an hem suç hem cinayet işliyorsunuz. İnsanlar Twitter ı kullanmak için VPN kullanıyor ve VPN mobil cihazların şarj süresini oldukça azaltıyor. Tarihe böyle geçeceksiniz. • onursuz ırkıcılar kökünüz kurusun lanet olsun size evet kürdüz türküz ermeniyiz afgan'ız arabız ırkcı itler geberin lan bu ülke hepimizin # #hepimizkürdüz • İnsanlık yapıp ülkeye alıyorsun hainlik,bu zor günde yağmacılık yapıyorlar.Bazı şeref yoksunu suriyeliler yüzünden masum olan insanlar arada kaynıyor.Açıkçası #ülkemdemülteciistemiyorum ! Allah herkesin yardımcısı olsun yardıma ihtiyacı olana koşulsun ama ülkemi terketsinler. [URL]
False negative Model predicts no HS Labeled as HS	• #UELKEMDEMUELTECİİSTEMİYORUM [URL] • Heryerde bilim uzmanı ve yer bilimci prof hocalar. Gerçeği açıklıyor. Sonra unutulup , açgözlü, rantçı,yağmacı yöneticiler soyguna devam eder. 3 yıllık bina yıkılmış, 3 yıl. #deprem #earthquake #Yağmacılar. • sayıları 8 milyon olan suriyeli, afgan, irak ne varsa çok acil ülkelerine geri gönderilmeli. *güvenlik tehdidi oluşturuyorlar. *işsizlik sorunu oluşturuyorlar. bill gates #billgates #sedatpeker10

of Twitter (Varol, 2023a; Hickey et al., 2023). Publicly available models will help researchers monitor these platforms more closely and even help them develop models to protect vulnerable groups.

Pre-trained models available online or developed through challenges can be easily adapted for other projects. Publicly available datasets like *#Secim2023* can be used to study political discourse (Pasquetto et al., 2020; Najafi et al., 2022; Varol, 2023b), and models can be utilized to study these datasets. The TurkishBERTweet that we used approach is publicly available on the HuggingFace platform along with the LoRA adapters for different tasks (Najafi and Varol, 2023).

Open source models: TurkishBERTweet model used in this challenge is available online at the HuggingFace platform. `https://huggingface.co/VRLLab/TurkishBERTweet`

Acknowledgements: We thank Hasan Kemik for discussing and supporting the challenge. We thank TUBITAK (121C220 and 222N311) for partially funding this project. The TurkishBERTweet model was trained and made publicly available thanks to the Google Cloud Research Credits program with the award GCP19980904.

References

Sinan Aral, Lev Muchnik, and Arun Sundararajan. 2009. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51):21544–21549.

Ozen Bas, Christine L Ogan, and Onur Varol. 2022. The role of legacy media and social media in increasing public engagement about violence against women in Turkey. *Social Media+ Society*, 8(4):20563051221138939.

Nur Bengisu Çam and Arzucan Özgür. 2023. Evaluation of chatgpt and bert-based models for Turkish hate speech detection. In *Intl. Conf. on Computer Science and Engineering*, pages 229–233. IEEE.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2020. Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472*.

Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. The rise of social bots. *Communications of the ACM*, 59(7):96–104.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*.

Keyan Guo, Alexander Hu, Jaden Mu, Ziheng Shi, Ziming Zhao, Nishant Vishwamitra, and Hongxin Hu. 2024. An Investigation of Large Language Models for Real-World Hate Speech Detection. *arXiv preprint arXiv:2401.03346*.

Daniel Hickey, Matheus Schmitz, Daniel Fessler, Paul E Smaldino, Goran Muric, and Keith Burghardt. 2023. Auditing elon musk's impact on hate speech and bots. In *Proc. of the Intl. AAAI Conf. on Web and Social Media*, volume 17, pages 1133–1137.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

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 preprint arXiv:2302.07736*.

Sarah J Jackson, Moya Bailey, and Brooke Foucault Welles. 2020. *# HashtagActivism: Networks of race and gender justice*. MIT Press.

Franziska B Keller, David Schoch, Sebastian Stier, and JungHwan Yang. 2020. Political astroturfing on twitter: How to coordinate a disinformation campaign. *Political Communication*, 37(2):256–280.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS One*, 14(8):e0221152.

Mohsen Mosleh and David G Rand. 2022. Measuring exposure to misinformation from political elites on Twitter. *Nature Communications*, 13(1):7144.

Ali Najafi, Nihat Mugurtay, Ege Demirci, Serhat Demirkiran, Huseyin Alper Karadeniz, and Onur Varol. 2022. # Secim2023: First Public Dataset for Studying Turkish General Election. *arXiv preprint arXiv:2211.13121*.

Ali Najafi and Onur Varol. 2023. TurkishBERTweet: Fast and Reliable Large Language Model for Social Media Analysis. *arXiv preprint arXiv:2311.18063*.

Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting good probabilities with supervised learning. In *Proc. of the Intl. Conf. on Machine Learning*, pages 625–632.

Christine Ogan and Onur Varol. 2017. What is gained and what is left to be done when content analysis is added to network analysis in the study of a social movement: Twitter use during Gezi Park. *Information, Communication & Society*, 20(8):1220–1238.

Amanda S Oliveira, Thiago C Cecote, Pedro HL Silva, Jadson C Gertrudes, Vander LS Freitas, and Eduardo JS Luz. 2023. How Good Is ChatGPT For Detecting Hate Speech In Portuguese? In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 94–103. SBC.

Irene V Pasquetto, Briony Swire-Thompson, Michelle A Amazeen, Fabrício Benevenuto, Nadia M Brashier, Robert M Bond, Lia C Bozarth, Ceren Budak, Ullrich KH Ecker, Lisa K Fazio, et al. 2020. Tackling misinformation: What researchers could do with social media data. *The Harvard Kennedy School Misinformation Review*.

Paul Röttger, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. 2022. MULTILINGUAL HATECHECK: Functional Tests for Multilingual Hate Speech Detection Models. *arXiv preprint arXiv:2206.09917*.

Stefan Schweter. 2020. BERTurk - BERT models for Turkish.

Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. 2018. The spread of low-credibility content by social bots. *Nature Communications*, 9(1):1–9.

Zeynep Tufekci. 2017. *Twitter and tear gas: The power and fragility of networked protest*. Yale U. Press.

Gokce Uludogan, Somaiyeh Dehghan, Inanc Arin, Elif Erol, Berrin Yanıkoglu, and Arzucan Ozgur. 2024. Overview of the Hate Speech Detection in Turkish and Arabic Tweets (HSD-2Lang) Shared Task at CASE 2024. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*.

Onur Varol. 2023a. Should we agree to disagree about Twitter's bot problem? *Online Social Networks and Media*, 37:100263.

Onur Varol. 2023b. Who Follows Turkish Presidential Candidates in 2023 Elections? In *Signal Processing and Communications Applications Conference*, pages 1–4. IEEE.

Onur Varol, Emilio Ferrara, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2017. Online human-bot interactions: Detection, estimation, and characterization. In *Proc. of the Intl. AAAI Conf. on Web and Social Media*, volume 11, pages 280–289.

Onur Varol, Emilio Ferrara, Christine L Ogan, Filippo Menczer, and Alessandro Flammini. 2014. Evolution of online user behavior during a social upheaval. In *Proc. of the ACM Conf. on Web Science*, pages 81–90.

Onur Varol and Ismail Uluturk. 2020. Journalists on Twitter: self-branding, audiences, and involvement of bots. *Journal of Computational Social Science*, 3(1):83–101.

Xindi Wang, Onur Varol, and Tina Eliassi-Rad. 2022. Information access equality on generative models of complex networks. *Applied Network Science*, 7(1).

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proc. of the first Workshop on NLP and Computational Social Science*, pages 138–142.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proc. of the NAACL Student Research Workshop, pages=88–93*.

Ziqi Zhang and Lei Luo. 2019. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web*, 10(5):925–945.