Team Curie at HSD-2Lang 2024: Hate Speech Detection in Turkish and Arabic Tweets using BERT-based models

Ehsan Barkhordar
Koç University
İstanbul, Turkey
ebarkhordar23@ku.edu.tr

Işık S. Topçu
Koç University
İstanbul, Turkey
itopcu21@ku.edu.tr

Ali Hürriyetoğlu
Wageningen
Food Safety Research (WFSR)
Wageningen, the Netherlands
ali.hurriyetoglu@wur.nl

Abstract

This study focuses on hate speech detection in Turkish and Arabic tweets using advanced BERT-based models. Performance metrics demonstrate the models' effectiveness, with the Turkish variant achieving a 71.8% F1 score and the Arabic model a 76.9% F1 score, ranking them fourth and third, respectively, in a competitive leaderboard. Performance enhancements were realized through targeted preprocessing, including emoji translation and user mention exclusion, and thoughtful data balancing approaches. Future directions include refining model accuracy and broadening language support. Our reproducible approach and detailed findings are accessible on GitHub[1].

1 Introduction

Social media platforms like Twitter, Facebook, and YouTube have become pivotal for expressing opinions and sharing information. However, hate speech—targeting ethnic, religious, gender, or other societal groups—poses a significant challenge to social harmony. The need for efficient detection mechanisms is amplified by the global reach of such content, yet languages like Turkish and Arabic present specific hurdles due to their intricate linguistic features and scarce annotated datasets (Beyhan et al., 2022).

The *Hate Speech Detection in Turkish and Arabic Tweets (HSD-2Lang)* shared task[2], part of CASE @ EACL 2024 Uludoğan et al. (2024), builds on the SIU2023-NST competition's groundwork in Turkish to include Arabic. This expansion highlights the need for language-specific solutions capable of accurately identifying hate speech in varied contexts.

Our contribution to Subtask A and Subtask B of this shared task underscores our commitment to advancing hate speech detection in Turkish and Arabic. Through our methodologies, we aim to contribute to the development of safer digital environments.

2 Related Work

The detection of hate speech, especially in linguistically complex languages like Turkish, has garnered significant attention in natural language processing research. Beyhan et al. (2022) presented a BERTurk-based approach at LREC 2022, highlighting the effectiveness of context-specific training with domain-specific datasets, achieving notable accuracies on the Istanbul Convention and Refugees datasets.

Toraman et al. (2022) advanced the field by creating large-scale, human-labeled tweet datasets, demonstrating the superiority of Transformer-based models over traditional methods. In the context of detecting homophobic and related hate comments in Turkish social media, Karayiğit et al. (2022) successfully employed a pre-trained Multilingual Bidirectional Encoder Representations from Transformers (M-BERT) model. Their approach yielded an impressive average F1-score of 90.15% on the Homophobic-Abusive Turkish Comments (HATC) dataset.

Hüsünbeyi et al. (2022) explored the integration of BERT models with linguistic features, showing their potential in surpassing traditional and CNN-based models in hate speech detection. Çam and Özgür (2023) examined the efficacy of ChatGPT and BERT variants in identifying Turkish hate speech, contributing to the evolving landscape of automated detection systems.

The SIU2023-NST Hate Speech Detection Contest, reported by Arın et al. (2023), emphasized the dominance of transformer-based and LightGBM models, with the leading entries achieving significant Macro F1 scores in both binary and multi-class hate speech detection tasks.

[1] https://github.com/politusanalytics/team-curie-case-2024-hsd-2lang

[2] https://github.com/boun-tabi/case-2024-hsd-2lang

Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024), pages 215–220
March 22, 2024 ©2024 Association for Computational Linguistics

Epoch	Training Loss	Validation Loss	Validation Performance		
			F1 Score	Accuracy	Recall
1	0.5561	0.5600	0.7151	0.7250	0.7250
2	0.3997	0.5845	0.7486	0.7556	0.7556
3	0.3167	0.4701	0.8022	0.8028	0.8028

Table 1: Training and Validation Results for Subtask A over Epochs

3 System Architecture and Training

This section details the system architecture and training processes for each distinct subtask.

3.1 Subtask A: Turkish Hate Speech Detection

Our goal in Subtask A was to develop a model capable of accurately detecting hate speech in Turkish tweets, encompassing data handling, preprocessing, model tuning, and a strategic training approach.

3.1.1 Data Preparation and Preprocessing

Social media data is inherently noisy, containing informal language, slang, misspellings, and unique language usage. To address this, a thorough preprocessing pipeline is essential for cleaning and standardizing text data for model analysis. In our preprocessing for Subtask A, we employ the emoji library[3] to convert emojis into their English textual descriptions, preserving their semantic value. Newline characters are replaced with spaces, and extra spaces are trimmed to streamline the text. URLs, user mentions, and standalone '@' symbols are removed to reduce non-essential information. Hashtags are also removed; this step not only reduces the word count but also aids in better tokenization by eliminating characters that could disrupt the model's ability to understand the context. The entire text is then converted to lowercase to ensure consistency across the dataset.

3.1.2 Train-Test Split

The division of our dataset into training and testing subsets is crucial for the unbiased development and evaluation of our model. We employ a stratified sampling strategy to ensure a balanced representation of label-topic combinations across both subsets.

For the validation set, we use a specific configuration to determine the number of samples for each label-topic combination, as outlined in the Table 2. The allocation of more samples for certain topics, such as Anti-Refugee, is informed by their proportion in the training data, ensuring a representative and balanced validation set.

Topic	Not Hateful	Hateful
Anti-Refugee	70	70
Israel-Palestine	60	60
Turkey-Greece	50	50

Table 2: Numbers of Validation Samples for Each Label-Topic Combination

This structured approach ensures that the validation set accurately reflects the diversity and distribution of the original dataset. The remaining data, after allocating the specified samples to the validation set, is used for training purposes.

3.1.3 Model Architecture

Our model architecture for detecting hate speech in Turkish tweets is based on the `dbmdz/bert-base-turkish-128k-uncased`[4] model, a pre-trained BERT variant optimized for Turkish text. We utilize the same tokenizer provided with this model to ensure consistency in text processing. The model is fine-tuned for binary classification, focusing on distinguishing between hateful and non-hateful content within various topics relevant to the subtask. Input sequences are processed with a maximum length of 128 tokens, aligning with the model's specifications.

3.1.4 Training Regime

The training regime for Subtask A is meticulously designed to balance representativeness and efficiency. We employ stratified sampling for the creation of training and validation sets and use the AdamW optimizer with a learning rate of 5×10^{-5} and a batch size of 128. The weight decay for the optimizer is set to 0.01 to prevent overfitting. The model is iterated over the dataset for 3 epochs,

[3] https://github.com/carpedm20/emoji/

[4] https://huggingface.co/dbmdz/bert-base-turkish-128k-cased

Epoch	Training Loss	Validation Loss	Validation Performance		
			F1 Score	Accuracy	Recall
1	0.3279	0.2201	0.8627	0.9070	0.9070
2	0.1957	0.1475	0.9207	0.9186	0.9186
3	0.1109	0.1573	0.9207	0.9186	0.9186
4	0.0569	0.1576	0.9070	0.9070	0.9070
5	0.0164	0.2253	0.9242	0.9186	0.9186

Table 3: Updated Training and Validation Results for Subtask B over Epochs

with careful monitoring of performance metrics to ensure optimal model tuning, as detailed in Table 1.

3.2 Subtask B: Hate Speech Detection with Limited Data in Arabic

This subsection outlines our strategy for detecting hate speech in Arabic tweets, a task challenged by the scarcity of comprehensive training data.

3.2.1 Data Preparation and Preprocessing

In addressing Subtask B—hate speech detection in Arabic tweets—we divided the dataset into training and validation sets. Initial preprocessing aimed to clean and standardize Arabic texts, typically involving noise reduction and format normalization for NLP tasks.

However, initial findings revealed that preprocessing diminished performance, suggesting that raw data, with its inherent linguistic nuances, might be more effective for this task. This led us to minimize preprocessing to preserve the original tweets' contextual and linguistic integrity, enhancing hate speech detection accuracy in Arabic.

3.2.2 Model Architecture

For Arabic hate speech detection, we utilized the `asafaya/bert-base-arabic`[5] model, a BERT variant optimized for Arabic (Safaya et al., 2020). This model was fine-tuned for binary classification to identify hateful versus non-hateful content. Data management was streamlined through a custom PyTorch `Dataset` class and `DataLoader` instances for efficient training and validation.

3.2.3 Training Regime

The training of the model for Subtask B was meticulously executed over the course of 5 epochs, employing a batch size of 128 for each iteration. We opted for the AdamW optimizer, configuring it with a learning rate set at 5×10^{-5} and incorporating a weight decay parameter of 0.01 to mitigate overfitting risks. Throughout the training process, we diligently monitored the model's loss metrics and subjected its performance to rigorous evaluation against the validation set upon the completion of each epoch. Please refer to Table 3 for more details.

4 Experimental Results

In this section, we summarize the performance of our models for each subtask. Our models were evaluated on a test dataset provided by the shared task organizers on Kaggle[6][7].

4.1 Performance Terminology Clarification

In this section, we clarify the terms used in Tables 4 and 6 to describe our model's performance and its comparison with other submissions within the competition.

Competition Best refers to the highest F1-score achieved by any team or participant in the official competition leaderboard. This score represents the best performance recorded during the competition period, under the contest's constraints and evaluation protocols.

Our Peak Performance denotes the highest F1-score our team achieved through late submissions, after the official competition period ended. These late submissions allowed us to further refine and test our models without the daily submission limits imposed during the competition. Thus, "Our Peak Performance" reflects our model's optimal performance obtained without the constraints of the competition's submission cap.

Official Submission represents the F1-score of our model that was officially submitted during the competition period, adhering to the contest's rules,

[5] `https://huggingface.co/asafaya/bert-base-arabic`

[6] `https://www.kaggle.com/competitions/hate-speech-detection-in-turkish/leaderboard?tab=public`

[7] `https://www.kaggle.com/competitions/hate-speech-detection-with-limited-data-in-arabic/leaderboard?tab=public`

including the limitation of three test evaluations per day. This score is what was officially recorded and considered in the competition's final rankings.

It is important to note that the methodologies and system architectures described in the sections for Subtask A and Subtask B were instrumental in achieving "Our Peak Performance". The results and insights derived from these sections are based on the models and approaches that contributed to our highest achieved scores, post-competition. This distinction is crucial for understanding the potential of our proposed solutions when not limited by the competition's constraints on model submissions and evaluations.

4.2 Subtask A: Hate Speech Detection in Turkish across Various Contexts

The performance of our model for Subtask A is summarized in Table 4. It is important to note that these results were obtained through a late submission, and as such, they might not appear on the official leaderboard. Despite this, our model's code is fully reproducible, allowing other researchers to verify our results and use them as a foundation for future work.

Metric	F1-Score	
	Public	Private
Competition Best	0.74876	0.69644
Our Peak Performance	0.71889	0.66129
Official Submission	0.71365	0.60790

Table 4: F1-Score Comparison in Subtask A

Furthermore, the confusion matrix depicted in Figure 1 offers valuable insights into the model's performance on the validation set.

4.3 Subtask B: Hate Speech Detection with Limited Data in Arabic

The performance of our model in Subtask B was rigorously evaluated over 5 training epochs, demonstrating the model's capability in accurately identifying hate speech within Arabic tweets, even with the constraints of limited data.

Metric	F1-Score	
	Public	Private
Competition Best	0.88888	0.68354
Our Peak Performance	0.76923	0.65853
Official Submission	0.76923	0.65853

Table 6: F1-Score Comparison in Subtask B

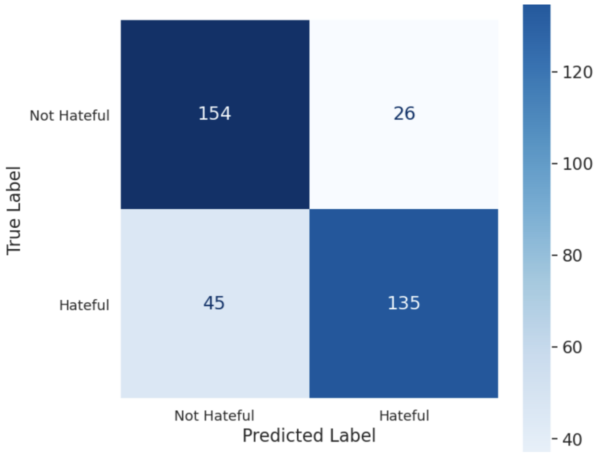

Figure 1: Confusion Matrix of the Model on the Validation Set for Subtask A

For a comparison of our model's F1-Score with the top scores in the task, see Table 6, which contrasts our results against the competition's best on both public and private leaderboards.

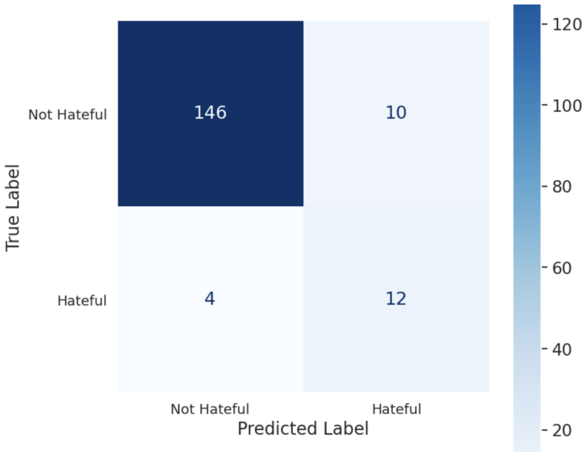

Figure 2: Confusion Matrix of the Model on the Validation Set for Subtask B

Additionally, the confusion matrix provided in Figure 2 further elucidates the model's classification prowess.

5 Ablation Study for Subtask A

In our ablation study for Subtask A, we systematically evaluated the impact of various preprocessing steps and data balancing techniques on the model's F1 score. This involved selectively omitting individual preprocessing steps—such as newline and extra space removal, URL removal, emoji conversion to text, mention and symbol removal, and hashtag processing—to assess their contribution to the model's overall performance. Additionally, we explored the effects of label and topic balancing, both

Experiment	**Public F1 Score**	**Private F1 Score**
Our Peak Performance	**0.71889**	**0.66129**
Preprocessing		
Without Newline/Extra Space Removal	0.71889	0.66129
Without URL Removal	0.71171	0.64947
Without Emoji Conversion	0.69868	0.64391
Without Mention/Symbol Removal	0.71544	0.63705
Without Hashtag Processing	0.67868	0.62391
Data Balancing		
With Label Balancing	0.70646	0.64332
With Topic Balancing	0.63917	0.60550
Data Balancing (1 Epoch Training)		
With Label Balancing	0.70769	0.64024
With Topic Balancing	0.64000	0.62585

Table 5: Effects of Preprocessing and Data Balancing on F1 Scores for Subtask A

with the standard training duration and a shortened training span of just one epoch.

Data Balancing Techniques: In our study, we employed two distinct data balancing strategies to mitigate class imbalance and enhance model performance:

- **Label Balancing:** We addressed class imbalance by equalizing the representation of labels in the training data. Specifically, we resampled the minority class (hateful content, labeled as '1') to match the quantity of the majority class (non-hateful content, labeled as '0'). This technique ensures that both classes contribute equally to the training process, preventing model bias toward the more prevalent class.

- **Topic Balancing:** Recognizing the importance of thematic representation, we also balanced the dataset based on topics. This involved resampling tweets within specific topics (e.g., Anti-Refugee, Israel-Palestine, Turkey-Greece) to ensure that hateful and non-hateful contents within each topic were equally represented. This approach acknowledges the contextual nuances of hate speech and aims for a model that is sensitive to topic-specific expressions of hate.

The findings from this study, as detailed in Table 5, are instrumental in elucidating the significance of each preprocessing step and data balancing strategy. For instance, the removal of hashtag processing exhibited a notable decrease in F1 scores, highlighting its critical role in the model's ability to accurately classify tweets. Similarly, the impact of data balancing techniques provides valuable insights into optimizing the training process for enhanced model performance.

Conclusion

Our participation in the HSD-2Lang 2024 contest underscored the effectiveness of BERT-based models in hate speech detection for Turkish and Arabic tweets. Leveraging innovative techniques and sophisticated architectures, we achieved notable F1 scores of 71.8% and 76.9% for Turkish and Arabic, respectively. These results highlight our system's proficiency in handling linguistic complexities and its contribution to improving online safety.

Acknowledgments

This work is supported by the European Research Council Politus Project (ID:101082050) and European Union's HORIZON projects EFRA (ID: 101093026) and ECO-Ready (ID: 101084201).

References

İnanç Arın, Zeynep Işık, Seçilay Kutal, Somaiyeh Dehghan, Arzucan Özgür, and Berrin Yanikoğlu. 2023. Siu2023-nst - hate speech detection contest. In *2023 31st Signal Processing and Communications Applications Conference (SIU)*, pages 1–4.

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 Thirteenth Language Resources and Evaluation Conference*, pages 4177–4185, Marseille, France. European Language Resources Association.

Nur Bengisu Çam and Arzucan Özgür. 2023. Evaluation of chatgpt and bert-based models for turkish hate speech detection. In *2023 8th International Conference on Computer Science and Engineering (UBMK)*, pages 229–233. IEEE.

Zehra Melce Hüsünbeyi, Didar Akar, and Arzucan Özgür. 2022. 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*, pages 32–41, Marseille, France. European Language Resources Association.

H. Karayiğit, A. Akdagli, and Ç. İ. Aci. 2022. Homophobic and hate speech detection using multilingual-bert model on turkish social media. *Information Technology and Control*, 51(2):356–375.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059, Barcelona (online). International Committee for Computational Linguistics.

Cagri Toraman, Furkan Şahinuç, and Eyup Yilmaz. 2022. Large-scale hate speech detection with cross-domain transfer. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2215–2225, Marseille, France. European Language Resources Association.

Gökçe Uludoğan, Somaiyeh Dehghan, İnanç Arın, Elif Erol, Berrin Yanikoglu, and Arzucan Özgür. 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)"*, Malta. Association for Computational Linguistics.