Overview of the Hate Speech Detection in Turkish and Arabic Tweets (HSD-2Lang) Shared Task at CASE 2024

Gökçe Uludoğan[1] and **Somaiyeh Dehghan**[2,3] and **İnanç Arın**[2,3]
Elif Erol[4] and **Berrin Yanikoglu**[2,3] and **Arzucan Özgür**[1]
[1] Department of Computer Engineering, Bogazici University, Turkey 34342
[2] Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul, Turkey 34956
[3] Center of Excellence in Data Analytics (VERIM), Sabanci University, Istanbul, Turkey 34956
[4] Hrant Dink Foundation, Istanbul, Turkey 34373
{gokce.uludogan,arzucan.ozgur}@bogazici.edu.tr, eliferol@hrantdink.org
{somaiyeh.dehghan,inanc.arin,berrin}@sabanciuniv.edu

Abstract

This paper offers an overview of the Hate Speech Detection in Turkish and Arabic Tweets (HSD-2Lang) Shared Task at CASE workshop that was held jointly with EACL 2024. The task was divided into two subtasks: Subtask A, targeting hate speech detection in various Turkish contexts, and Subtask B, addressing hate speech detection in Arabic with limited data. The shared task attracted significant attention with 33 teams that registered and 10 teams that participated in at least one task. In this paper, we provide the details of the tasks and the approaches adopted by the participant along with an analysis of the results obtained from this shared task.

1 Introduction

Hate speech, which targets groups based on characteristics such as ethnicity, nationality, religion, colour, gender, and sexual orientation, is a significant problem on social media platforms. The automated detection of such content is crucial for efficient content moderation and the mitigation of societal harm. Moreover, it can also be instrumental in socio-political event analysis.

The effectiveness of current hate speech detection models is often hampered by issues such as limited data and lack of generalizability. Following the SIU2023-NST competition (Arın et al., 2023), which was organized to benchmark progress in Turkish hate speech detection and classification, we present a new shared task, Hate Speech Detection in Turkish and Arabic Tweets (HSD-2Lang) Shared Task, in conjunction with The 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024). This shared task focuses on tackling the challenge of identifying hate speech in tweets in Turkish and Arabic languages.

2 Tasks

The shared task involves the development of models for hate speech detection in social media, with a specific focus on Turkish and Arabic languages. The task is divided into two distinct subtasks: a) Hate Speech Detection in Turkish across Various Contexts (Subtask A), b) Hate Speech Detection with Limited Data in Arabic (Subtask B).

Both subtasks are formulated as binary classification problems where the objective is to determine whether individual tweets are hateful or non-hateful.

Subtask A: Hate Speech Detection in Turkish across Various Contexts

The objective of this subtask is to develop a model capable of detecting hate speech in Turkish tweets.

Data. The dataset contains Turkish tweets on three topics, each annotated for the presence or absence of hate speech. The topics encompass tweets concerning refugees, the Israel-Palestine conflict, and Anti-Greek discourse. The training set contains a total of 9,140 tweets while the test set comprises 2,295 tweets. The distribution of data with respect to topics, labels, and splits is shown in Table 1.

Evaluation. The performance of the models is evaluated using the F1 metric on the combined test data from all three topics.

Table 1: Statistics for Subtask A data, with respect to topics, labels, and splits.

Topic	Train set		Test set	
	Hateful	Non-hateful	Hateful	Non-hateful
Anti-Refugee	1447	4477	361	1119
Isr-Pal conflict	880	1360	73	498
Anti-Greek	451	555	105	139
Total	2778	6392	539	1756

Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024), pages 229–233

Table 2: Statistics for Subtask B data splits.

Label	Train set	Test set
Hateful	82	52
Non-hateful	778	470
Total	860	522

Subtask B: Hate Speech Detection with Limited Data in Arabic

The goal in this subtask is to build a model for Arabic hate speech detection under data-constrained conditions.

Data. The dataset comprises Arabic tweets, particularly focusing on anti-refugee hate speech. This task is challenging with a smaller data set and high class imbalance. The data statistics are reported in Table 2.

Evaluation. The performance of the models is evaluated using the F1 metric on test data, which includes tweets related to anti-refugee hate speech.

3 System Descriptions

The HSD-2Lang shared task attracted participation from 33 teams associated with various universities and organizations. This task involved developing systems for specific subtasks, detailed in the following subsections.

3.1 Subtask A

A total of 33 teams registered for the subtask, with 10 eventually submitting their results. All systems were based on BERT (Devlin et al., 2019). However, teams employed diverse approaches, including different base models, data processing techniques, and training strategies. The base models varied from monolingual models such as BERTurk (Schweter, 2020) and TurkishBERTweet (Najafi and Varol, 2023), to the multilingual XLM-RoBERTa model (Conneau et al., 2019).

The winner in Subtask A, DetectiveReDASers (Qachfar et al., 2024), utilized the ConvBERTurk model[1] (Schweter, 2020), enhancing it with a novel pooling strategy, cross-lingual data augmentation, and a soft-voting ensemble approach. During preprocessing, they corrected encoding errors and translated emoji characters into corresponding text descriptions in Turkish. For cross-lingual data augmentation, the team translated Arabic tweets from Subtask B using Google Translate. Their pooling strategy combined the standard [CLS] token representation with mean and max pooling of token representations, further refined by additional linear and dropout layers. This approach aimed to improve sequence representation by integrating the [CLS] token with mean and max values from the last hidden layer. For ensembling, they utilized a soft-voting ensemble of five identical ConvBERTurk models, distinguished only by their initializations.

The second and third place teams, ReBERT (Yagci et al., 2024) and VRLLab (Najafi and Varol, 2024), both used TurkishBERTweet[2], which was specifically trained on a large corpus of Turkish tweets. While both utilized LoRA fine-tuning (Hu et al., 2021), they differed in their preprocessing, hyperparameters, and data filtering approaches. Both systems applied the TurkishBERTweet preprocessing pipeline, which transforms Twitter-specific entities into special tags and converts emojis' unicode characters into descriptive words. However, their configurations differed: ReBERT used a smaller batch size (32), a lower learning rate (5e-5), a longer training duration (100 epochs), and polynomial learning rate scheduling with a 10% warm-up steps in the AdamW optimizer (Loshchilov and Hutter, 2017). In contrast, VRLLab chose a batch size of 128, a learning rate of 3e-4, 20 epochs of training, and 6% warm-up steps.

3.2 Subtask B

Compared to Subtask A, this was a smaller dataset and also attracted fewer participants. Altogether, there were 15 teams who registered to the subtask, with 5 eventually submitting results. Among these submissions, all systems used BERT variants (Antoun et al., 2020; Safaya et al., 2020).

The winner of Subtask B, ReBERT (Yagci et al., 2024), finetuned AraBERTv0.2[3] (Antoun et al., 2020), which was pretrained on approximately 60 million Arabic tweets. This version of AraBERT includes emojis and previously omitted common words in its vocabulary, and was used without any preprocessing. Unlike their parameter-efficient approach in Subtask A, the team performed 4 epochs of full supervised fine-tuning using a batch size of

[1] https://huggingface.co/dbmdz/convbert-base-turkish-cased

[2] https://huggingface.co/VRLLab/TurkishBERTweet

[3] https://huggingface.co/aubmindlab/bert-base-arabertv02

Table 3: Scores of top three ranking teams in public and private leaderboards of Subtask A.

Rank	Team	Public			Private		
		Precision	Recall	F1 Score	Precision	Recall	F1 Score
1	DetectiveReDASers (Qachfar et al., 2024)	0.70588	0.76364	0.73362	0.68161	0.71194	**0.69645**
2	ReBERT (Yagci et al., 2024)	0.79167	0.69091	0.73786	0.75989	0.62998	0.68886
3	VRLLab (Najafi and Varol, 2024)	0.71296	0.70000	0.70642	0.66588	0.66276	0.66432

Table 4: Confusion matrices for the top three ranking systems in Subtask A.

Actual Label	**Predictions** (DetectiveReDASers)		**Prediction** (ReBERT)		**Prediction** (VRLLab)	
	Hateful	Non-Hateful	Hateful	Non-Hateful	Hateful	Non-Hateful
Hateful	388	149	345	192	360	177
Non-Hateful	177	1578	105	1650	173	1582

8, a learning rate of 5e-4 with a linear decay, and the AdamW optimizer.

The second place team, Team Curie (Barkhordar et al., 2024), employed a different Arabic BERT model[4] (Safaya et al., 2020), which was pretrained on around 8.2 billion words from the Arabic subset of OSCAR (Suárez et al., 2019) and Arabic Wikipedia. They fine-tuned this model on raw tweets, opting not to preprocess the data based on their findings that preprocessing could negatively impact performance. Their fine-tuning parameters included 5 training epochs, a batch size of 128, a learning rate of 5e-5 with a weight decay regularization parameter of 0.01, and they also used the AdamW optimizer.

In third place, Team Uriel also used the AraBERTv0.2 model trained on tweets, similar to the winning team. However, their approach differed by introducing an additional layer to map BERT representations to a lower dimension before output mapping. This process involved two fully connected layers with ReLU activation functions, distinguishing their method from the standard approach of direct mapping of BERT representations to outputs. The first layer reduced the dimensionality to 100, while the second served as a binary classification output layer.

4 Competition Results

For both tasks, the performance of models is evaluated using the test samples of the corresponding dataset. The test samples are randomly divided into public and private samples. Public samples make up 20% of the test samples and are used by participants for validation during the test phase. Private samples were used to evaluate the model's performance after the test phase has concluded and to generate the final leaderboard of the shared task. Table 3 and 5 display the precision, recall, and F1 scores achieved by the top three systems, officially ranked by F1 score, in the private leaderboard. The confusion matrices for these systems are presented in Table 4 and 6.

5 Results for Subtask A

DetectiveReDASers took first place on the private leaderboard and second on the public leaderboard, with F1 scores of 0.69645 and 0.73362, respectively. This system, leveraging a soft-ensemble of ConvBERTurk models, outperformed the competing systems in recall and F1 scores on both leaderboards. ReBERT and VRLLab, although employing a similar approach using TurkishBERTweet with LoRA fine-tuning, showed a noticeable difference in their F1 scores (ReBERT: 0.73786 public, 0.68886 private; VRLLab: 0.70642 public, 0.66432 private). Notably, ReBERT achieved significantly higher precision scores compared to VRLLab. This variation underscores the critical role of hyperparameter tuning in optimizing model performance.

6 Results for Subtask B

In Subtask B, ReBERT and Team Curie showed very similar performances in the public leaderboard, while their private leaderboard scores varied. ReBERT took the first-place with an F1 score of 0.683532, while Team Curie came second place with an F1 score of 0.65854.

Interestingly, despite using the same Arabic

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

Table 5: Scores of top three ranking teams in public and private leaderboards of Subtask B.

Rank	Team	Public			Private		
		Precision	Recall	F1 Score	Precision	Recall	F1 Score
1	ReBERT (Yagci et al., 2024)	0.76923	0.76923	0.76923	0.67500	0.69231	**0.68354**
2	Team Curie (Barkhordar et al., 2024)	0.76923	0.76923	0.76923	0.62791	0.69231	0.65854
3	Team Uriel	0.66667	0.61538	0.64000	0.57143	0.71795	0.63636

Table 6: Confusion matrices for the top three ranking systems in Subtask B.

Actual Label	**Prediction** (ReBERT)		**Prediction** (Team Curie)		**Prediction** (Team Uriel)	
	Hateful	Non-Hateful	Hateful	Non-Hateful	Hateful	Non-Hateful
Hateful	37	15	37	15	36	16
Non-Hateful	16	454	19	451	25	445

BERT model trained on tweets as the top-ranking system, Team Uriel lagged behind the second place system (Team Curie) that used an Arabic BERT model not specifically pre-trained on tweets. This result highlights the importance of hyperparameter tuning, especially in scenarios with limited data.

The winner of Subtask A, DetectiveReDASers (Qachfar et al., 2024), took 4th place in subtask B, with an F1 score of 0.6 on the private dataset. They reported that they did not use cross-lingual augmentation in this task (i.e. translating Turkish tweets into Arabic) as it degraded the performance, even though this strategy had worked well in Subtask A. This may be due to the relative numbers of the two datasets.

7 Conclusion

This paper presented an overview of the HSD-2Lang shared task that was organized to benchmark models for hate speech detection on social media platforms, in Turkish and Arabic languages. The task consisted of two distinct subtasks, each addressing unique challenges: Subtask A focused on hate speech detection in various contexts in Turkish, and Subtask B addressed the challenge of hate speech detection in Arabic under limited data conditions.

The results from these subtasks provided valuable insights into the efficacy of different models and strategies. All participating systems used BERT-based models, demonstrating their effectiveness. On the other hand, systems using the same model achieved noticeably different results due to hyperparameter choices.

In Subtask A, the top-performing system employed a soft-ensemble of ConvBERTurk models, achieving an F1 score of 0.69645 on the private test set. This shows the effectiveness of ensemble methods in tackling hate speech detection across various contexts. Moreover, the noticeable performance differences among systems using the same method, specifically TurkishBERTweet with LoRA fine-tuning, underscore the importance of hyperparameter tuning in improving model performance.

Subtask B yielded comparable results due to similar model implementations with minor variations. The top system achieved an F1 score of 0.68354 on the private test set. The small performance difference between the two tasks are interesting, considering that the Turkish dataset had three times more data compared to the Arabic set, showing the effectiveness of pretrained models. On the other hand, the difference between the performances of ReBERT and Team Uriel, even though they used the same Arabic BERT model, further highlights the importance of hyperparameter tuning, especially in scenarios with limited data availability.

Acknowledgements

This work was supported by the EU project "Utilizing Digital Technology for Social Cohesion, Positive Messaging and Peace by Boosting Collaboration, Exchange and Solidarity" (EuropeAid/170389/DD/ACT/Multi), carried out by the Hrant Dink Foundation.

References

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

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. IEEE.

Ehsan Barkhordar, Işık S. Topçu, and Ali Hürriyetoğlu. 2024. Team Curie at HSD-2Lang 2024: Hate speech detection in Turkish and Arabic tweets using BERT-based models. In *Proceedings of The Seventh Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Malta. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

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 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Ali Najafi and Onur Varol. 2023. TurkishBERTweet: Fast and reliable large language model for social media analysis. *arXiv preprint arXiv:2311.18063*.

Ali Najafi and Onur Varol. 2024. VRLLab at HSD-2Lang 2024: Turkish hate speech detection online with TurkishBERTweet. In *Proceedings of The Seventh Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Malta. Association for Computational Linguistics.

Fatima Zahra Qachfar, Bryan E. Tuck, and Rakesh M. Verma. 2024. DetectiveReDASers at HSD-2Lang 2024: A new pooling strategy with cross-lingual augmentation and ensembling for hate speech detection in low-resource languages. In *Proceedings of The Seventh Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Malta. Association for Computational Linguistics.

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.

Stefan Schweter. 2020. BERTurk - BERT models for Turkish. https://zenodo.org/records/3770924.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache.

Utku Ugur Yagci, Ahmet Emirhan Kolcak, and Egemen Iscan. 2024. ReBERT at HSD-2Lang 2024: Fine-tuning BERT with AdamW for hate speech detection in Arabic and Turkish. In *Proceedings of The Seventh Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, Malta. Association for Computational Linguistics.