Transformers at HSD-2Lang 2024: Hate Speech Detection in Arabic and Turkish Tweets Using BERT Based Architectures

Kriti Singhal, Jatin Bedi
Computer Science and Engineering Department
Thapar Institute of Engineering and Technology
kritisinghal711@gmail.com, jatin.bedi@thapar.edu

Abstract

Over the past years, researchers across the globe have made significant efforts to develop systems capable of identifying the presence of hate speech in different languages. This paper describes the team Transformers' submission to the subtasks: Hate Speech Detection in Turkish across Various Contexts and Hate Speech Detection with Limited Data in Arabic, organized by HSD-2Lang in conjunction with CASE at EACL 2024. A BERT based architecture was employed in both the subtasks. We achieved an F1 score of 0.63258 using XLM RoBERTa and 0.48101 using mBERT, hence securing the 6th rank and the 5th rank in the first and the second subtask, respectively.

1 Introduction

Hate Speech is defined as the usage of expressions or phrases which are hostile, offensive, or threatening in nature. Hate Speech is usually targeted against an individual or a group of individuals, highlighting those unique characteristics that distinguish those individuals. Some people use online platforms, such as Twitter, to spread hateful content at the click of a button.

Access to the internet, along with social media platforms such as Twitter, Instagram, and Facebook, can enable anyone, anywhere in the world, to share their ideas with millions of people across the globe within a few milliseconds (Shanmugavadivel et al., 2022).

With the advancing technological age, it is getting easier to spread hateful content thousands of miles across the globe, even without revealing their identity, due to the increased anonymity offered by online platforms. Automated detection of hateful content has become crucial for enhancing content moderation to mitigate societal harm. Social media platforms encourage the users to report any hate speech content that violates the hateful conduct policy so that appropriate action can be taken. However, it is still visible to many users, which necessitates the use of an automated system to detect and curb such content (Abuzayed and Elsayed, 2020).

The task organized by HSD-2Lang[1] at CASE 2024 aimed at identifying the presence of hate speech in Turkish and Arabic languages (Gökçe Uludoğan, 2024). The task was divided into two subtasks, as listed below:

i. Subtask A: Hate Speech Detection in Turkish across Various Contexts

ii. Subtask B: Hate Speech Detection with Limited Data in Arabic

In the past few years, Natural Language Processing (NLP) has experienced major breakthroughs, especially in the Hate Speech identification domain. Some of which are Long Short Term Memory (Hochreiter and Schmidhuber, 1997) and the Gated Recurrent Units (Chung et al., 2014). But, there has been a paradigm shift with the introduction of transformers (Vaswani et al., 2017).

Arabic language is one of the six official languages of the United Nations. Arabic is also a critical and strategically useful language (Ryding, 2013). With 18.55 million users, Turkey had the 7th highest number of Twitter users in 2023 (Statista, 2022). Turkish is also one of the most widely spoken languages of the Turkic language family.

Both Arabic and Turkish are very different from the English language. The orthography of both languages significantly differs from English due to the right-to-left text orientation and the utilization of connecting letters. The presence of word elongation, common ligatures, zero-width diacritics, and allographic variants leads to further complications. The morphology is extremely intricate, showcasing a wealth of morphemes that are used as prefixes,

[1] 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 190–194
March 22, 2024 ©2024 Association for Computational Linguistics

suffixes, or even circumfixes. These elements can denote various grammatical features such as case, number, gender, and definiteness, among others, resulting in a sophisticated morphotactic system (Malmasi and Dras, 2014; Budur et al., 2020).

2 Related Work

Researchers have made multiple efforts in the past to automatically detect the presence of hate speech in Arabic and Turkish, in the past. A variety of different approaches have been used in the past to address this problem.

Abuzayed and Elsayed (2020) compared 15 classical and neural network models to classify Arabic tweets based on the presence of hate speech. To solve the problem efficiently, a “quick and simple” approach was used. The experiments were conducted on a collection of 8,000 tweets, and it was found that neural learning models outperformed the classical ones. The best classifier was a joint architecture of a convolution and recurrent neural network. The classifier used data after pre-processing, in which the punctuation, foreign characters, numbers, repeated characters, and diacritics were removed from the text. The remaining Arabic text was then normalized.

In the approach adopted by Husain (2020), extensive pre-processing was performed. The pre-processing step involved seven different steps. The work showed the improvement that pre-processing the data makes by retaining only the important content and performing dimensionality reduction. The first step in pre-processing was the conversion of emojis and emoticons to a textual label to ensure that the meaning conveyed by them did not suffer due to their removal. Next, since the Arabic dialect exists in various different forms, the variations in the different forms were normalized. The words were then categorized and then letter normalization was performed. This was followed by hashtag segmentation, where the ‘#’ symbol was removed and the text following it was left untouched. After this, the numbers, more than two consecutive spaces, and the occurrence of more than three repetitive characters was removed along with Arabic stop words. Lastly, to address the data imbalance, upsampling was performed.

Neural networks and discourse analysis techniques were used by Hüsünbeyi et al. (2022) to identify the presence of hate speech in Turkish text. A Hierarchical Attention Network and BERT based deep learning models were implemented to apprehend evolving verbal cues and comprehend the contextual nuances within the discourse. Additionally, linguistic features using critical discourse analysis techniques were designed and integrated with neural network models.

Table 1: Dataset Distribution for Subtask A

Dataset	Label	
	Hateful	Non-Hateful
Anti-Refugee sentiment	1447	4477
Israel-Palestine conflict	880	1360
Anti-Greek discourse	555	421

3 Dataset Description

3.1 Subtask A

The dataset provided by the organizers for subtask A comprised of three parts, broadly classified into three categories, namely, refugees, the Israel-Palestine conflict, and Anti-Greek discourse. The dataset contained a total of 9,140 tweets in Turkish language. The detailed data distribution has been shown in Table 1. The text also contained emojis, emoticons, special symbols, numbers, and hyperlinks.

3.2 Subtask B

The dataset provided for subtask B comprised of 1000 Arabic tweets. In this dataset provided by the organizers, 778 tweets did not contain hate speech, whereas the remaining 82 tweets contained hate speech. The text in some tweets had special symbols such as ‘#’, ‘@’, ‘’, and ‘[’. The text also contained links, and some words were written in English.

4 Methodology

In NLP, text classification in languages with limited resources and code-mixed nature, has been a prominent problem. It can be defined as assigning text labels depending upon the content, context and intention of it. Researchers have devised multiple models to tackle this problem. Many of these models followed a transformer based approach and have been pre-trained on large corpora of text and

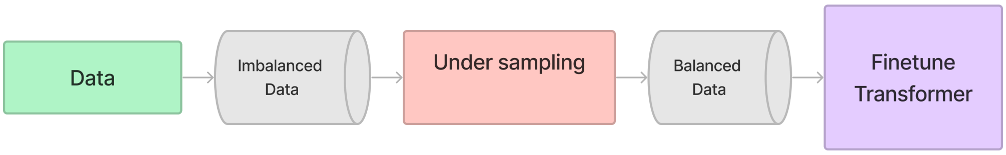

Figure 1: Proposed Methodology

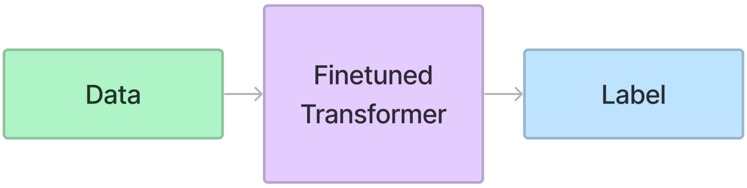

Figure 2: Label Generation for Unseen Data

have been made available for a multitude of solving problems like text classification. However, the corpora of text available to train such models is usually dominated by high-resourced languages like English. This problem is solved by the use of cross-lingual transfer learning.

4.1 Subtask A

The dataset provided for subtask A, as described in Section 3.1, comprised of three subsets. The data from all the three was first concatenated together to form one large dataset. The data also, had a huge data imbalance problem between the two classes of hateful and non-hateful tweets. The number of tweets for categorised as non-hateful were almost three times as many as the ones categorised as hateful. This issue was addressed by performing undersampling on the data, such that the number of tweets for both the classes became equal. The undersampling was performed randomly with no preference given to any particular type of data.

The data was then split such that 80% of the data was used for training and 20% was used for testing. The data was used to finetune the XLM RoBERTa Large (XLMR) model (Conneau et al., 2019). It was observed that the model performed the best when trained for 8 epochs using the weighted Adam optimizer and negative log likelihood loss.

The XLMR model is an unsupervised model trained on data of 100 different languages. This model is derived from the 2019 RoBERTa model launched by Facebook. XLMR is a large multilingual model which has been trained on 2.5TB of filtered data acquired from CommonCrawl. XLMR uses its own tokenizer, known as the XLMRoberta-Tokenizer.

4.2 Subtask B

In the dataset for subtask B, as elaborated in Section 3.2, there was a significant difference between the number of samples with and without hateful content. Hence, the data imbalance was addressed by performing undersampling on the data. After performing undersampling, both the classes had 82 tweets each. The undersampling was performed to randomly select 82 tweets from all the 778 tweets classified as non-hateful.

Next, 70% of the remaining data was used for training, and 30% of the data was used for testing, the multilingual BERT (mBERT) based architecture which was employed in this subtask. The model showed the best performance after training for 13 epochs and using the Adam optimizer and negative log likelihood loss.

The model, mBERT is a self-supervised transformer model pre-trained on a huge multilingual corpus. The corpus comprised of 104 languages, with the largest Wikipedia utilizing a masked language modeling objective. mBERT uses the BertTokenizer to perform tokenization on the data.

5 Results and Discussion

Multilingual transformer models were used to detect the presence of hate speech in Arabic and Turkish tweets. The BERT based architectures were finetuned to improve their performance further.

The data imbalance present in the data of both the subtasks was addressed by performing random undersampling on the data to ensure that the number of tweets for both the classes are equal.

It was found that the model performed better when the unprocessed data was used to train the model. Hence, the text data was used without any pre-processing to train the model.

The methodology followed to finetune the transformers for both the subtasks has been summarized in Figure 1. The label generation for the testing data has been summarised in Figure 2 for both the subtasks.

The models achieved an F1 score of 0.63258 and 0.48101 in subtask A and subtask B, respectively. Overall, the highest F1 score achieved was 0.69644 and 0.68354 in subtask A and subtask B respectively by the teams ranked 1^{st} in the shared task.

6 Conclusion and Future Work

Hate speech detection is the process of classifying text based on the presence or absence of hateful content. The aim of the shared task organized by HSD-2Lang in conjunction with CASE at EACL 2024 was to automatically detect whether the tweet was hateful or not in nature.

In this paper, we discussed our use of two multilingual BERT based transformers in the Hate Speech Detection in Turkish and Arabic Tweets shared task. We achieved an F1 score of 0.63258 in subtask A with XLMR and an F1 score of 0.48101 in subtask B with mBERT with the discussed approaches.

Transformers have shown great potential in the field of NLP and have consistently outperformed the classical models. Hence, combining different transformer models using ensembling techniques can help improve performance. Also, since limited resources are available for both Arabic and Turkish, the performance may be further enhanced by combining multiple datasets.

References

Abeer Abuzayed and Tamer Elsayed. 2020. "Quick and Simple Approach for Detecting Hate Speech in Arabic Tweets". In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 109–114, Marseille, France. European Language Resource Association.

Emrah Budur, Rıza Özçelik, Tunga Gungor, and Christopher Potts. 2020. "Data and Representation for Turkish Natural Language Inference". In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8253–8267, Online. Association for Computational Linguistics.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

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. *arXiv preprint arXiv:1911.02116*.

İnanç Arın Elif Erol Berrin Yanikoglu Arzucan Özgür Gökçe Uludoğan, Somaiyeh Dehghan. 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.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Fatemah Husain. 2020. "OSACT4 Shared Task on Offensive Language Detection: Intensive Preprocessing-Based Approach". In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 53–60, Marseille, France. European Language Resource Association.

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.

Shervin Malmasi and Mark Dras. 2014. "Arabic Native Language Identification". In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 180–186, Doha, Qatar. Association for Computational Linguistics.

Karin C. Ryding. 2013. Teaching Arabic in the United States. In Kassem M Wahba, Zeinab A Taha, and Liz England, editors, *Handbook for Arabic Language Teaching Professionals in the 21st Century*. Routledge.

Kogilavani Shanmugavadivel, Malliga Subramanian, Prasanna Kumar Kumaresan, Bharathi Raja Chakravarthi, B Bharathi, Subalalitha Chinnaudayar Navaneethakrishnan, Lavanya Sambath Kumar, Thomas Mandl, Rahul Ponnusamy, Vasanth Palanikumar, et al. 2022. Overview of the Shared Task on Sentiment Analysis and Homophobia Detection of YouTube Comments in Code-Mixed Dravidian Languages.

Statista. 2022. Number of active Twitter users in selected countries. Accessed: January 24, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.