Leveraging translation-based transfer learning to improve hate speech detection in low-resource languages

Marc Schenk [1] **Romino Steiner** [1] **Moritz Widmer** [1]

Abstract

This work investigates translation-based transfer learning to enhance hate speech detection in low-resource languages, focusing on Arabic. The lack of annotated data in these languages limits classifier effectiveness. Our approach fine-tunes pre-trained BERT-based models on a large English dataset and refines it with English translations of a smaller Arabic dataset. This method leverages the abundance of English data while maintaining linguistic and cultural specificity. Results show improved classification performance. The findings underscore the potential of leveraging abundant English data to enhance hate speech detection across diverse linguistic contexts, contributing to more inclusive and culturally sensitive NLP models.

Keywords: Translation, Hate Speech, Transfer Learning, Multilingual NLP.

1. Introduction

Hate speech detection is a critical task in natural language processing (NLP) with significant implications for social media platforms, online communities, and society at large. [1] While significant progress has been made in developing effective hate speech classifiers in English [2], a major limitation persists for other languages: the scarcity of annotated data. [3] This restricts the performance of larger models in low-resource languages. [4] Addressing this limitation is essential for equitable and effective hate speech detection across different linguistic communities, thereby fostering safer online spaces.

A recent study highlights the disparity in hate speech detection research, showing that over half of the analyzed records focused solely on English sources. [2] Multiple works [5, 6] demonstrate the potential of cross-lingual transfer learning for hate speech detection by training classifiers on English and applying them to lower-resource languages. However, existing classifiers often fail to account for cultural differences, making them culturally insensitive in cross-lingual NLP tasks. [7]

This work aims to leverage text translation to enhance hate speech detection in low-resource languages. We propose fine-tuning pre-trained BERT-based models using a large English dataset and subsequently refining the models with English translations of a smaller dataset from a low-resource language. This approach seeks to balance the high availability of English training data with the linguistic and cultural specificity required for accurate hate speech detection in other languages.

2. Related Work

To provide a comprehensive understanding of the current landscape and position our project within the existing research, this section reviews key studies and developments in cross-lingual hate speech detection.

Amplayo *et Al.* [8] suggest using translations as additional context for sentence classification in NLP tasks. Sohn and Lee [9] incorporate this idea in their multi-channel BERT model. The study, however, does not present the detailed effect translations have on the model performance.

Bigoulaeva *et Al.* [5, 6] address the challenge of detecting hate speech in low-resource languages. The researchers demonstrate the effectiveness of using cross-lingual transfer-learning by training classifiers with bilingual word embeddings on English data and then applying these trained models to German data.

As later shown by Lee *et Al.* [7], however, hate speech classifiers fail to account for cultural nuances and linguistic variations across languages. The authors show that, due to their cultural insensitivity, hate speech classifiers trained on monolingual datasets perform poorly when applied to translated data from other languages.

Aluru *et Al.* [10, 11] analyze hate speech detection across datasets in 9 languages, showing that LASER embeddings excel in low-resource scenarios, while BERT-based models perform better with larger datasets. This presents an incisive limitation for lower-resource languages, as they cannot fully

[1]EPFL.

EE-559: *Deep Learning*, 2024.

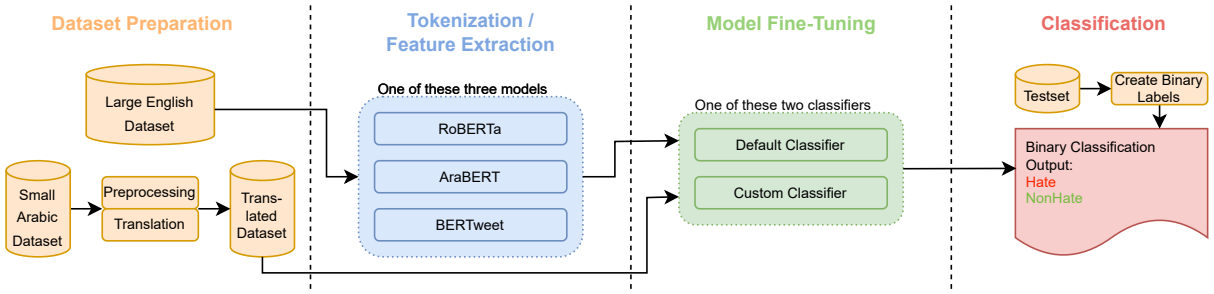

Figure 1. General pipeline for hate speech detection used in this project.

leverage the potential of the deep contextual representations of data-intensive BERT-based language models.

3. Method

To address the scarcity of annotated data in low-resource languages identified in section 2 as the primary limitation in hate speech detection for most languages, we pursue a transfer-learning based approach. This approach also tackles the second challenge: the insensitivity of classifiers trained on monolingual corpora to cultural differences.

3.1. General Pipeline

The general pipeline used can be gathered from Figure 1. It starts with two datasets: a larger English dataset and a smaller Arabic dataset. For the Arabic dataset, adequate preprocessing and translation are applied as detailed in section 3.2 Next, one of the language models RoBERTa [12], BERTweet [13], or AraBERT [14] is used to create the contextual representations for training the classifier. For classification, the default classifier provided by `RobertaForSequenceClassification` is used. For the best-performing models we also implement a custom classifier to further improve performance. Finally, we evaluate the models and classifiers with a validation set or a dedicated test set.

3.2. Dataset Selection and Preprocessing

For pre-training the classifiers, a dynamically generated English dataset containing $40\,000$ samples is used. [15, 16] This dataset offers several advantages: (1) $54\,\%$ of the dataset is hateful, providing a rich source of training data. (2) It was created over four rounds of dynamic data collection, enhancing model robustness. (3) It includes $15\,000$ challenging perturbations, making models more resilient in real-world applications.

To fine-tune the classifier with the English translation of a low-resource language, we use an Arabic dataset from [17]. Arabic was chosen due to its lack of annotated data compared to English, while it still has adequate data available. Additionally, Arabic culture has fewer commonalities with Western cultures, making it ideal for investigating cultural sensitivity. The selected Arabic dataset has the following qualities: (1) It is one of the largest Arabic datasets for offensive content, allowing robust training. (2) Models trained on it show strong generalization capabilities on external datasets.

The Arabic dataset, comprised of around $13\,000$ Tweets, includes Emojis and placeholders for empty lines. To improve translation quality, we remove Emojis, and replace empty line placeholders with full stops. For the translation of the data to English we use LibreTranslate [18], an open source machine translation API that offers state-of-the-art translation and can be self-hosted offline.

Both datasets have detailed labels beyond simple hate or no hate speech. For binary classification, all violent, offensive, or specific hate speech categories are treated as hate speech, and all others as non-hate speech.

3.3. Model Pre-Training

For the classification, we used three different models based on BERT, a pre-trained deep learning model, that generates contextualized word embeddings by processing text bidirectionally, capturing the context of words based on both preceding and succeeding words.

RoBERTa (Robustly optimized BERT approach): This model improves upon BERT by training on larger datasets and with optimized hyperparameters, resulting in enhanced performance on various natural language understanding tasks. **BERTweet**: Specializing in understanding and generating social media text, particularly tweets, BERTweet is used because the Arabic dataset is comprised of Twitter data, which may lead to better performance than RoBERTa. **AraBERT**: Pre-trained on a large corpus of Arabic text, AraBERT is trained and evaluated directly on the non-

translated Arabic data, serving as a baseline for hate speech classification without the extensive data available in English and the transfer-learning from translation.

For RoBERTa and BERTweet, we freeze the model layers to only train the classifier using the English dataset. This results in pre-trained models for hate speech detection which can then be further specialized for detecting hate speech in Arabic. Each model was trained for 35 epochs to ensure adequate training, with the loss not significantly decreasing further while maintaining reasonable resource consumption, running no more than a day on a single GPU.

3.4. Fine-Tuning with Translated Data

After pre-training the models for hate speech detection on English data, we freeze the pre-trained model layers to fine-tune the classifier further. In this second step, training is conducted with the translated samples from the originally Arabic data. This step adapts the model to better understand the linguistic and cultural contexts of the low-resource language.

3.5. Classifier Architecture

For the classification, we employed two different architectures. First, we utilized `RobertaForSequenceClassification`. This classification head consists of a combination of a single pooling layer, a dropout layer, and a fully connected linear layer. In addition, we implement a more complex custom classifier (CC). The CC introduces additional layers and normalization steps to enhance representation and robustness. It adds a second dense layer with Tanh activation and includes residual connections to improve training stability. The CC also normalizes the output after each dense layer with residual addition, and introduces two more dropout layers, providing enhanced regularization compared to the single dropout layer in the standard classification head.

4. Validation

The implemented solutions described in section 3 are evaluated on a separate validation set, with validation loss and validation accuracy being computed. All results are presented in Tabels 1 and 2. The validation loss is calculated using the `CrossEntropyLoss` criterion, which measures the difference between the predicted logits and the true labels. The validation accuracy measures the proportion of correctly predicted labels out of the total number of labels.

The results show that fine-tuning the classifier on translated data improved classification in all three cases. Despite the Arabic dataset consisting of Twitter data, the BERTweet model does not outperform the standard RoBERTa model after fine-tuning. Note that removing Emojis during pre-processing likely deteriorates performance, suggesting that retaining such contextual elements could enhance accuracy.

MODEL	DATASET	ACCURACY	LOSS
ROBERTA	ENGLISH	0.6769	0.0194
BERTWEET	ENGLISH	0.6357	0.0201
ROBERTA CC	ENGLISH	0.6825	0.0183

Table 1. VALIDATION ACCURACY AND LOSS OF MODELS PRE-TRAINED ON A LARGE ENGLISH DATASET.

MODEL	DATASET	ACCURACY	LOSS
ARABERT	ARABIC	0.8024	0.0138
ROBERTA	ARABIC T	0.7135	0.0183
ROBERTA CC	ARABIC T	0.7135	0.0181
BERTWEET	ARABIC T	0.6989	0.0183

Table 2. CLASSIFICATION PERFORMANCES OF PRE-TRAINED MODELS AFTER FINE-TUNING ON ENGLISH TRANSLATIONS OF ARABIC DATA AND PERFORMANCE OF ARABERT.

Our custom classifier (CC) improved accuracy for the pre-trained models, but this gain was offset by the performance improvements from fine-tuning. The AraBERT model, using only low-resource language, Arabic data, achieved higher accuracy and lower loss than all other setups.

5. Conclusion

In this project, we explore translation-based transfer learning to enhance hate speech detection in low-resource languages, focusing on Arabic. By fine-tuning pre-trained BERT-based models on a large English dataset and then refining them with English translations of Arabic data, we significantly improve classification performance.

Our findings demonstrate the value of translation-based fine-tuning, leveraging abundant English data to address the scarcity of annotated data in other languages. This approach partially overcomes the cultural insensitivity of monolingual hate speech classifiers by incorporating linguistic and cultural contexts from translated data. Although the purely Arabic AraBERT model outperforms our translation-based models, it is trained on a considerably larger datasets with significant computational resources, which might not be feasible for other low-resource languages. Our approach offers a resource-efficient alternative, especially valuable for languages with even less annotated data than Arabic.

In conclusion, translation-based fine-tuning is a viable and effective solution for improving hate speech detection in low-resource languages, contributing to the development of more inclusive and culturally sensitive NLP models.

References

[1] M. Mondal, L. A. Silva, and F. Benevenuto, "A Measurement Study of Hate Speech in Social Media," in *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, HT '17, (New York, NY, USA), pp. 85–94, Association for Computing Machinery, July 2017.

[2] M. S. Jahan and M. Oussalah, "A systematic review of hate speech automatic detection using natural language processing," *Neurocomputing*, vol. 546, p. 126232, Aug. 2023.

[3] O. Adams, A. Makarucha, G. Neubig, S. Bird, and T. Cohn, "Cross-Lingual Word Embeddings for Low-Resource Language Modeling," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers* (M. Lapata, P. Blunsom, and A. Koller, eds.), (Valencia, Spain), pp. 937–947, Association for Computational Linguistics, Apr. 2017.

[4] B. Vidgen and L. Derczynski, "Directions in abusive language training data, a systematic review: Garbage in, garbage out," *PLOS ONE*, vol. 15, p. e0243300, Dec. 2020. Publisher: Public Library of Science.

[5] I. Bigoulaeva, V. Hangya, I. Gurevych, and A. Fraser, "Addressing the Challenges of Cross-Lingual Hate Speech Detection," Jan. 2022. arXiv:2201.05922 [cs].

[6] I. Bigoulaeva, V. Hangya, and A. Fraser, "Cross-Lingual Transfer Learning for Hate Speech Detection," in *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion* (B. R. Chakravarthi, J. P. McCrae, M. Zarrouk, K. Bali, and P. Buitelaar, eds.), (Kyiv), pp. 15–25, Association for Computational Linguistics, Apr. 2021.

[7] N. Lee, C. Jung, and A. Oh, "Hate Speech Classifiers are Culturally Insensitive," in *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)* (S. Dev, V. Prabhakaran, D. Adelani, D. Hovy, and L. Benotti, eds.), (Dubrovnik, Croatia), pp. 35–46, Association for Computational Linguistics, May 2023.

[8] R. K. Amplayo, K. Lee, J. Yeo, and S.-w. Hwang, "Translations as Additional Contexts for Sentence Classification," June 2018. arXiv:1806.05516 [cs].

[9] H. Sohn and H. Lee, "MC-BERT4HATE: Hate Speech Detection using Multi-channel BERT for Different Languages and Translations," in *2019 International Conference on Data Mining Workshops (ICDMW)*, pp. 551–559, Nov. 2019. ISSN: 2375-9259.

[10] S. S. Aluru, B. Mathew, P. Saha, and A. Mukherjee, "A Deep Dive into Multilingual Hate Speech Classification," in *MACHINE LEARNING AND KNOWLEDGE DISCOVERY IN DATABASES: APPLIED DATA SCIENCE AND DEMO TRACK, ECML PKDD 2020, PT V* (Y. Dong, G. Ifrim, D. Mladenic, C. Saunders, and S. VanHoecke, eds.), vol. 12461, (Cham), pp. 423–439, Springer International Publishing Ag, 2021. ISSN: 0302-9743, 1611-3349 Num Pages: 17 Series Title: Lecture Notes in Artificial Intelligence Web of Science ID: WOS:000716884800026.

[11] S. S. Aluru, B. Mathew, P. Saha, and A. Mukherjee, "Deep Learning Models for Multilingual Hate Speech Detection," Dec. 2020. arXiv:2004.06465 [cs].

[12] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," July 2019.

[13] D. Q. Nguyen, T. Vu, and A. Tuan Nguyen, "BERTweet: A pre-trained language model for English Tweets," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (Q. Liu and D. Schlangen, eds.), (Online), pp. 9–14, Association for Computational Linguistics, Oct. 2020.

[14] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based Model for Arabic Language Understanding," in *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection* (H. Al-Khalifa, W. Magdy, K. Darwish, T. Elsayed, and H. Mubarak, eds.), (Marseille, France), pp. 9–15, European Language Resource Association, May 2020.

[15] B. Vidgen, T. Thrush, Z. Waseem, and D. Kiela, "Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (C. Zong, F. Xia, W. Li, and R. Navigli, eds.), (Online), pp. 1667–1682, Association for Computational Linguistics, Aug. 2021.

[16] B. Vidgen, "bvidgen/Dynamically-Generated-Hate-Speech-Dataset," Feb. 2024. original-date: 2020-12-31T10:10:46Z.

[17] H. Mubarak, S. Hassan, and S. A. Chowdhury, "Emojis as Anchors to Detect Arabic Offensive Language and Hate Speech," Jan. 2022.

[18] "LibreTranslate/LibreTranslate," May 2024. original-date: 2020-12-19T19:19:34Z.