Leveraging Translation-Based Transfer Learning to Improve Hate Speech Detection in Low-Resource Languages

Marc Schenk, Romino Steiner, Moritz Widmer

Group 7, EPFL

Problem definition

Hate speech detection in low-resource languages is limited by the scarcity of annotated data, which hinders the effectiveness of classifiers. This project aims to enhance hate speech detection in Arabic, a low-resource language, by leveraging translation-based transfer learning. The goal is to improve classification performance by fine-tuning pre-trained BERT-based models on a large English dataset and then finetuning them with English translations of an Arabic dataset.

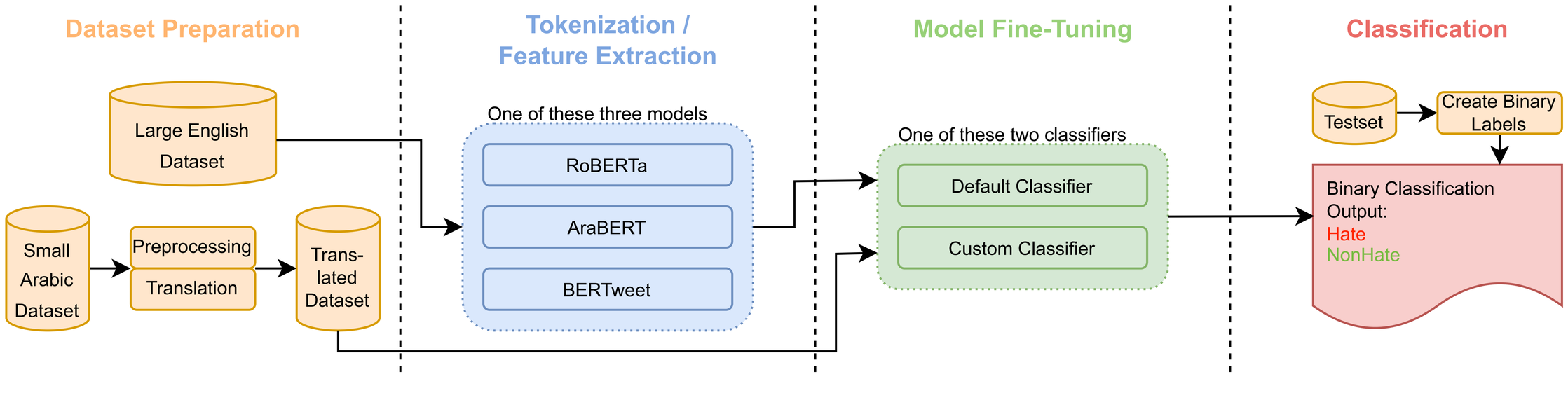
Key Related Works

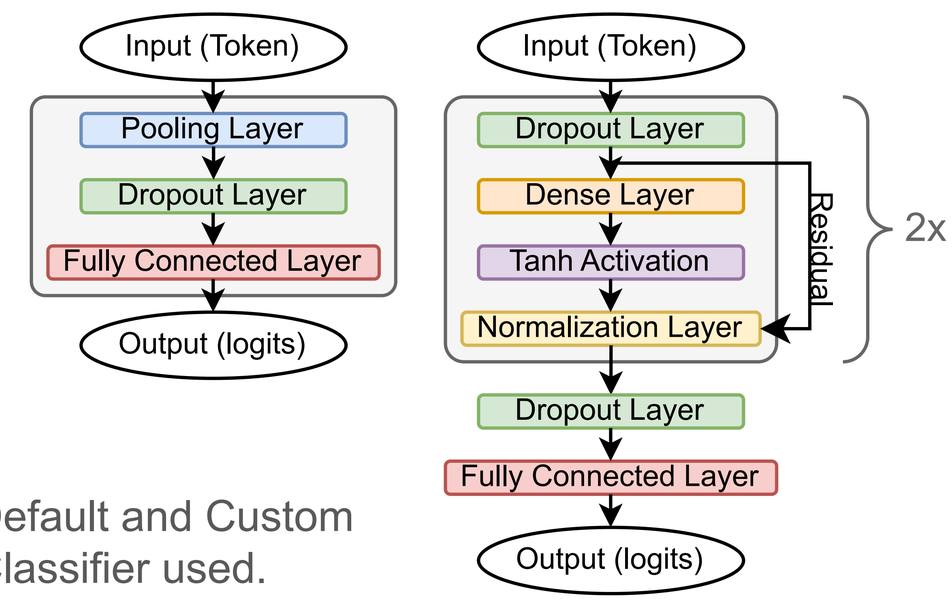
- **Bigoulaeva et Al. (2022)** [1] – Demonstrated the effectiveness of cross-lingual transfer learning by training classifiers with bilingual word embeddings.
- **Lee et Al. (2023)** [2] – Showed that hate speech classifiers often fail to account for cultural nuances and linguistic variations across languages.
- **Aluru et Al. (2021)** [3] – Analysed hate speech detection across multiple languages, showing BERT-based models perform worse on low-resource languages due to lack of training data.

Dataset(s)

- **Large English dataset**: 40,000 samples with 54% labelled as hate speech.
- **Arabic dataset**: ~13,000 Tweets used for fine-tuning on translated data after translation to English.

Method





Default and Custom Classifier used.

Validation

The performance of the models was evaluated using a separate validation set. Validation accuracy and loss (cross-entropy loss) were computed, showing that fine-tuning on translated data improved classification performance.

Model	Dataset	Accuracy	Loss
RoBERTa	English	0.6769	0.0194
BERTweet	English	0.6357	0.0201
RoBERTa CC	English	0.6828	0.0183

Model	Dataset	Accuracy	Loss
AraBERT	Arabic	0.8024	0.0138
RoBERTa	Arabic T	0.7135	0.0183
BERTweet	Arabic T	0.6989	0.0183
RoBERTa CC	Arabic T	0.7135	0.0181

Limitations

- Performance is influenced by the quality of translations.
- The removal of Emojis in preprocessing might deteriorate performance, suggesting the need for retaining such contextual elements.
- If enough data is available, directly training in the low-resource language achieves better results.

Conclusion

Our translation-based transfer learning approach effectively enhances hate speech detection in low-resource languages like Arabic by leveraging abundant English data and refining it with translated data. While AraBERT outperforms our models using purely Arabic data, our method offers a resource-efficient alternative suitable for languages with even less annotated data than Arabic, contributing to more inclusive and culturally sensitive NLP models.

References

[1] 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., Cham: Springer, 2021, pp. 423-439. doi: 10.1007/978-3-030-67670-4_26.

[2] I. Bigoulaeva, V. Hangya, I. Gurevych, and A. Fraser, "Addressing the Challenges of Cross-Lingual Hate Speech Detection," *arXiv*, Jan. 2022. doi: 10.48550/arXiv.2201.05922.

[3] H. Mubarak, S. Hassan, and S. A. Chowdhury, "Emojis as Anchors to Detect Arabic Offensive Language and Hate Speech," *arXiv*, accessed May 25, 2024. [Online]. Available: https://arxiv.org/abs/2201.06723v2