EE559: Group Mini-Project summary

Group number:

7

Title of the Group Mini-Project:

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

Brief summary (max 300 words):

One of the main limiting factors in reaching great results in text-based hate-speech detection for languages other than English is the issue that there is not as much data available in most languages other than English. This limits the extent to which larger models (such as Bert based models) can be trained on these languages. With our project we want to explore how we could leverage text translation to improve hate-speech detection in these low-resource language. We intend to use a pretrained Bert-based model which we then fine tue first using a large English data set and then further finetune it with the english translation of a smaller dataset from a lower resource language. The main difficulty will be to try and account for cultural differences that are present in the different languages and might be lost in translation. Depending on the results from this first approach we might then further explore this idea or also leverage translation in a different way, e.g. by translating English data into a low-resource language and thus getting more data in this language.

Keyword 1:

Translation

Keyword 2:

Transfer Learning

Keyword 3:

Multilingual Learning

Main objective (max 100 words):

Our main objective is to extend the work of Lee et. Al (Key Reference 1) who found significant performance drops when simply training on English data and then applying these models to other languages. Furthermore, we intend to overcome the fundamental cultural differences embedded in languages to create a hate-speech detection pipeline which allows for fostering safer online spaces for languages where annotated data is scarce.

Key reference 1:

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: Association for Computational Linguistics, May 2023, pp. 35–46. [Online]. Available: https://aclanthology.org/2023.c3nlp-1.5

Key reference 2:

I. Bigoulaeva, V. Hangya, I. Gurevych, and A. Fraser, "Addressing the Challenges of Cross-Lingual Hate Speech Detection," Jan. 2022, arXiv:2201.05922 [cs]. [Online]. Available: http://arxiv.org/abs/2201.05922

Key reference 3:

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: Association for Computational Linguistics, Apr. 2021, pp. 15–25. [Online]. Available: https://aclanthology.org/2021.ltedi-1.3

Dataset used 1:

Main English dataset: Learning From the Worst (Dynamically generated hate speech dataset) [https://github.com/bvidgen/Dynamically-Generated-Hate-Speech-Dataset]

Model used 1:

RobertaForSequenceClassification (from HuggingFace: https://huggingface.co/docs/transformers/index)

Other:

We will use more datasets for fine-tuning (in the low-resource languages) but we are currently still finalising our selection process for this part.