Cross-Lingual Hate Speech Detection Using Zero-Shot Learning & Multi-Task Learning

Mitash Shah[1], Aayush Kharwal[1], Shritej Patil[1],
Kasyap Sai Chakkirala[1], Nirmal Malavalli Venkataraman[1],

[1]University of Southern California

Abstract

The rapid proliferation of the internet in India has transformed online communication, introducing the significant diversity of our daily speech into the online discourse. This variety has also, unfortunately, led to an increase in hate speech - defined as communication that disparages individuals or groups based on characteristics such as race, gender, religion, or more. Detecting hate speech in such a cross-linguistic landscape is both crucial and challenging, particularly for low-resource languages like Hindi. Building on previous research, this paper explores hate speech detection in Hindi by leveraging multilingual transformer-based models enhanced with zero-shot learning technique. We hope that this study not only contributes to the task of combating hate speech but also underscores the importance of continuing to develop better solutions in the future.

1 Hypothesis

A cross-lingual model that uses pre-trained transformers such as mBERT and XLM-R can effectively generalize hate speech detection capabilities from high-resource languages like English to low-resource languages like Hindi, demonstrating the possibility of effective zero-shot learning in low-resource settings.

2 Related Work

Recent advancements in Hindi hate speech detection have utilized both traditional machine learning methods and advanced transformer-based models to improve accuracy and address multilingual challenges. Vashistha et al. (2020) demonstrated that transformer models like BERT, when fine-tuned, significantly outperform traditional techniques in identifying hate speech in Hindi text. Pani et al. (2022) further validated the effectiveness of the different pre-trained multilingual transformers on HASOC2019/20 datasets, showing that even without fine-tuning, contextual embeddings alone enhance classification accuracy across multiple languages. To tackle data scarcity, Mnassri et al. (2024) combine Generative Adversarial Networks (GANs) with mBERT, outperforming baseline mBERT in detecting hate speech in resource-limited settings. Additionally, HateCheckHIn (2022) developed functional tests and a template-based Hindi tweet dataset, revealing that existing multilingual BERT models struggle with multilingual scenarios and exhibit biases toward specific target communities, underscoring the need for more research in cross and multi-lingual hate speech detection systems.

3 Methodology

3.1 Baseline Models

The project plans to use pre-trained multilingual transformer models such as mBERT and XLM-R. Thanks to their cross-lingual word embeddings (CLWEs), these models are particularly effective for tasks that span different languages. Fine-tuning them specifically for our hate speech detection task can further enhance their cross-lingual capabilities.

3.2 Training Methods

3.2.1 Zero-Shot Learning

Using zero-shot learning, the model can apply knowledge from resource-rich languages to detect hate speech in under-represented languages. This approach removes the dependency on labeled data in the target language by utilizing cross-lingual transfer, where multilingual embeddings allow for a unified semantic understanding across language boundaries. So, although the fine-tuning will primarily focus on high-resource languages like English and Spanish, the models' understanding of both source and target languages should enable them to generalize hate speech detection across diverse linguistic contexts.

3.2.2 Multi-Task Learning

Multi-task learning will boost the model's generalization by training it on related tasks: hate speech detection as the primary task, with offensive language detection and sentiment analysis as auxiliary tasks.

3.3 Datasets

We will use a combination of datasets annotated for hate speech and offensive language. Below are the datasets that will be used:

- HatEval 2019: Multilingual dataset (English, Spanish) against Immigrants and Women
- HASOC: Collection of social media posts in different Indian Languages
- Stormfront: Dataset comprising of English posts from the Stormfront community, providing rich examples of extremist and derogatory language.

3.4 Evaluation Methods

To determine how well the model generalizes to low-resource languages the model's performance will be based on the following metrics and methodologies:

- F1-Scores: To address data imbalance, we will use macro F1-scores.
- Ablation Study: To determine the contributions of specific model components towards the target hate speech detection.

4 Input/Output

The model will take Hindi sentences unseen during training and use zero-shot learning and multi-task learning for cross-lingual processing. It will classify sentences as hate speech (1) or non-hate speech (0).

References

Mithun Das, Punyajoy Saha, Binny Mathew, and Animesh Mukherjee. 2022. Hatecheckhin: Evaluating hindi hate speech detection models. *Preprint*, arXiv:2205.00328.

Koyel Ghosh and Dr. Apurbalal Senapati. 2022. Hate speech detection: a comparison of mono and multilingual transformer model with cross-language evaluation. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*, pages 853–865, Manila, Philippines. Association for Computational Linguistics.

Khouloud Mnassri, Reza Farahbakhsh, and Noel Crespi. 2024. Multilingual hate speech detection using semi-supervised generative adversarial network. In *Complex Networks & Their Applications XII*, pages 192–204, Cham. Springer Nature Switzerland.

Neeraj Vashistha and Arkaitz Zubiaga. 2020. Online multilingual hate speech detection: Experimenting with hindi and english social media. *Inf.*, 12:5.