Cross-Lingual Hate Speech Detection Using Zero-Shot & 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 across the globe 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. Building on previous research, this paper explores the viability of hate speech detection in low-resource languages by leveraging multilingual transformer-based models enhanced with multi-task and zero-shot learning techniques. 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

This study hypothesizes that in a zero-shot setting for a low-resource language like Hindi, a multilingual language model trained on multiple related languages and related tasks will outperform a model trained on a single related language for the specific task of detecting hate speech. The study further speculates that the model could probably achieve comparable performance, albeit not better than state-of-the-art models explicitly trained on Hindi hate speech data. So, by incorporating related languages and leveraging related tasks, this approach aims to reduce the dependency on extensive datasets for low-resource languages, enabling effective hate speech detection for the same.

2 Related Work

Recent studies (Narayan et al., 2023; Vashistha and Zubiaga, 2021; Velankar et al., 2021; Kumar and Ojha, 2019) have consistently demonstrated that Transformer-based models outperform traditional machine learning techniques in hate speech detection, particularly for low-resource languages like Hindi. Comprehensive reviews (Ramos et al., 2024; Shishah and Fajri, 2022) further reinforce this notion that transformers act as catalysts in advancing hate speech research by addressing challenges inherent to low-resource languages, such as data scarcity and linguistic diversity. Moreover, Ghosh and Senapati (2022) validates the effectiveness of the different pre-trained multilingual transformers on the HASOC2019/20 dataset, demonstrating that even without fine-tuning, contextual embeddings significantly improve classification performance across multiple languages. Collectively, these studies underscore the transformative impact of transformer models in the realm of hate speech detection, particularly for languages with limited annotated resources.

Zero-Shot Learning (ZSL) enables machine learning models to generalize across tasks in languages they weren't explicitly trained on by leveraging multilingual frameworks. In the realm of hate speech detection, ZSL is particularly valuable for low-resource languages such as Hindi. Studies like (Sharma et al., 2021) utilize mBERT to identify hate speech in transliteration of Hindi-English code-switched text without Hindi-specific labels. The paper (Mnassri et al., 2024) combines semi-supervised techniques with multilingual models to enhance detection across Indo-European languages, including Hindi. Additionally, (Kapil and Ekbal, 2024) reveals strong ZSL performance for Hindi and other languages, underscoring ZSL's effectiveness in applying patterns learned from high-resource languages like English to detect hate speech in low-resource contexts.

Research in hate speech detection has increasingly explored Multi-Task Learning (MTL) to improve detection accuracy by incorporating auxiliary tasks. For instance, De la Pena Sarracén and Rosso (2021) used MTL to combine offensive language detection with hate speech detection, finding that shared

learning across related tasks can reduce false negatives in hate speech classification. Similarly, Plaza-Del-Arco et al. (2021) demonstrated that incorporating sentiment analysis and emotion detection tasks alongside hate speech detection could help the model capture emotional cues associated with hate speech, thereby improving performance on Spanish hate speech datasets.

Kapil et al. (2023) took a multilingual approach, leveraging high-resource languages (e.g., English and Urdu) to improve hate speech detection for Hindi, a low-resource language, through MTL. Their study used a transformer-based MTL model that combined shared and task-specific layers, demonstrating a significant performance boost in Hindi hate speech detection when supported by related auxiliary tasks and multilingual training data. These studies underscore the utility of MTL for hate speech detection, especially when auxiliary tasks (e.g., sentiment and offensive language detection) provide relevant contextual signals. By integrating sentiment analysis as an auxiliary task and employing zero-shot learning for cross-lingual transfer, this study aims to further explore the impact of MTL on hate speech detection performance in low-resource languages like Hindi.

3 Methodology

For our study, we employed XLM-RoBERTa along with its pre-trained tokenizer from Hugging Face to perform hate speech detection. We selected XLM-R over mBERT due to its enhanced performance on morphologically rich, low-resource languages (Conneau et al., 2020). As a baseline, we performed minimal data preprocessing and fine-tuned the model using a single-task approach in a zero-shot setting using English and Marathi datasets. The training process utilized a batch size of 16 with a learning rate of 2e-5, and was carried out over 5 epochs. Subsequently, the fine-tuned model was evaluated on Hindi data to assess its zero-shot performance in detecting hate speech within the low-resource language context.

3.1 Multi-Task Learning

This project employs MTL to enhance the performance of hate speech detection by leveraging information from auxiliary tasks. In MTL, multiple related tasks are learned together within a single model, allowing it to develop shared representations that benefit the primary task, in this case, hate speech detection. By simultaneously learning from related tasks, MTL improves the model's ability to generalize and capture nuances that are often critical for distinguishing hate speech.

The primary task in this MTL setup is hate speech detection. Two auxiliary tasks, sentiment analysis and offensive language detection, are included to introduce additional linguistic context and refinement for the main task. The main task, is assigned a higher weight, as it is the target task. Auxiliary tasks are assigned lower weights, allowing them to contribute useful features without overshadowing the primary objective. These weights ensure that the shared learning optimally supports hate speech detection while benefiting from the contextual cues provided by sentiment and offensive language tasks.

The chosen XLM-Roberta is used for this MTL setup. The architecture includes shared layers that capture language features beneficial to all tasks, along with individual output layers for each task, allowing it to specialize where needed. Task-specific weights guide the model's learning process, emphasizing hate speech detection while integrating supplementary insights from the auxiliary tasks. This approach helps the model build a more comprehensive representation, improving its robustness in identifying hate speech.

3.2 Zero-Shot Learning

To implement zero-shot learning for hate speech detection in low-resource languages, we plan to utilize the multilingual model, XLM-RoBERTa, which is pre-trained on high-resource languages with ample labeled data. These models facilitate cross-lingual generalization by learning hate speech-specific semantic, contextual, and syntactic cues that transfer to low-resource languages through shared representations. In our study, Hindi serves as the low-resource, unseen target language, with the model fine-tuned on high-resource hate speech datasets. By leveraging XLM-R for zero-shot learning, we capture relevant features for Hindi without additional language-specific training. We then assess the model's performance on Hindi datasets to evaluate the effectiveness of zero-shot learning in bridging the gap between source and target languages for hate speech detection.

3.3 Datasets

Our project uses a diverse set of multilingual datasets to conduct zero-shot and multi-task evalu-

ations, focusing primarily on hate speech detection. For the test data, we designate Hindi as the target language, using the HASOC-2019 Hindi dataset (approximately 9,000 datapoints) to evaluate model performance and cross-lingual transfer.

For training, we employ datasets across multiple languages and tasks. For hate speech detection, we use the hate speech 18 Stormfront dataset for English, HASOC-2019 for German, the L3-Cube MahaHate two-class dataset for Marathi, and the Bangla Hate Speech dataset from Kaggle. For offensive language detection, we include HASOC-2019 datasets for English and German, the L3-Cube MahaHate four-class dataset for Marathi, and the Bangla Hate Speech dataset from Kaggle. For sentiment analysis, we incorporate the Amazon Reviews dataset (FastText format) for English, the SB10k dataset for German, the L3-Cube MahaSent dataset for Marathi, and the Bengali Sentiment Classification dataset from Kaggle for Bangla.

In total, we use approximately 82,000 datapoints for hate speech detection, 63,000 datapoints for offensive language detection, and 3.03 million datapoints for sentiment analysis.

3.4 Evaluation Methods

To evaluate the model's performance in a cross-lingual, zero-shot learning setup, we employed the following core metrics:

1. **Evaluation Loss:** Uses the Cross Entropy Loss, which measures the prediction error on the test set. Lower loss indicates more accurate alignment with the test data.
2. **Macro F1 Score:** Useful for imbalanced datasets, as it averages F1 scores across classes, providing a balanced view of model performance on hate and non-hate categories.
3. **Accuracy:** Reflects the overall percentage of correct predictions, offering a general sense of the model's effectiveness.

4 Initial Results

In this project, we assess XLM-RoBERTa's performance in a zero-shot learning setting by training the model on one language and testing it on another. Specifically, we conducted two experiments:

1. Training on the English hate speech dataset (Hate Speech 18) and testing on Hindi.
2. Training on a custom Marathi hate speech dataset and testing on the same Hindi dataset.

These experiments allow us to evaluate how a cross-lingual model performs on Hindi when trained on either a linguistically similar language (Marathi) or a relatively more distant one (English). The results from these setups are as follows:

Training Data	Loss	Macro F1 Score	Accuracy
English	2.21	0.5593	0.5949
Marathi	1.31	0.7698	0.7704

Table 1: Zero-Shot Evaluation Results on Hindi Dataset

The results show that the model trained on Marathi outperformed the English-trained model on Hindi, with higher Macro F1 and accuracy, indicating that training on a similar language enhances cross-lingual transfer for hate speech detection.

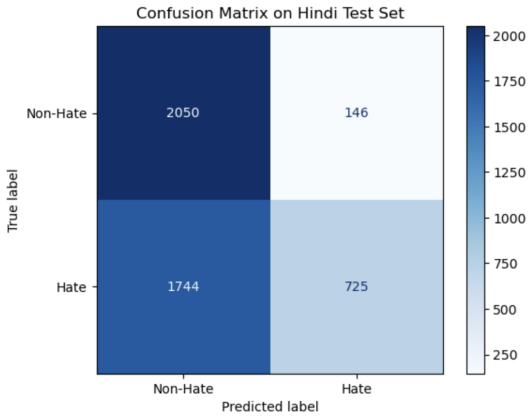

Figure 1: Confusion Matrix for Zero-Shot Evaluation on Hindi Test Set (Trained on English Dataset)

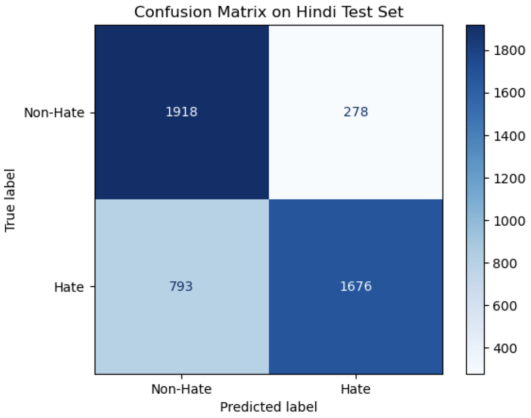

Figure 2: Confusion Matrix for Zero-Shot Evaluation on Hindi Test Set (Trained on Marathi Dataset)

References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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

Gretel Liz De la Pena Sarracén and Paolo Rosso. 2021. Multi-task learning to analyze the influence of offensive language in hate speech detection. In *Multimodal Hate Speech Workshop 2021*, pages 13–18.

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*, 1, pages 853–865, Manila, Philippines. Association for Computational Linguistics.

Prashant Kapil and Asif Ekbal. 2024. Cross-lingual zero-shot and few-shot learning to hate speech detection. *SSRN Electronic Journal*.

Prashant Kapil, Gitanjali Kumari, Asif Ekbal, Santanu Pal, Arindam Chatterjee, and B. N. Vinutha. 2023. Hhsd: Hindi hate speech detection leveraging multitask learning. *IEEE Access*, 11:101460–101473.

Ritesh Kumar and Atul Kr. Ojha. 2019. Kmi-panlingua at hasoc 2019: Svm vs bert for hate speech and offensive content detection. In *Proceedings of FIRE 2019*.

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.

Nikhil Narayan, Mrutyunjay Biswal, Pramod Goyal, and Abhranta Panigrahi. 2023. Hate speech and offensive content detection in indo-aryan languages: A battle of lstm and transformers. *Preprint*, arXiv:2312.05671. ArXiv:2312.05671.

Flor Miriam Plaza-Del-Arco, M. Dolores Molina-González, L. Alfonso Ureña-López, and María Teresa Martín-Valdivia. 2021. A multi-task learning approach to hate speech detection leveraging sentiment analysis. *IEEE Access*, 9:112478–112489.

Gil Ramos, Fernando Batista, Ricardo Ribeiro, Pedro Fialho, Sérgio Moro, António Fonseca, Rita Guerra, Paula Carvalho, Catarina Marques, and Cláudia Silva. 2024. A comprehensive review on automatic hate speech detection in the age of the transformer. *Social Network Analysis and Mining*, 14(1):204.

Arushi Sharma, Anubha Kabra, and Minni Jain. 2021. Ceasing hate withmoh: Hate speech detection in hindi-english code-switched language. *Preprint*, arXiv:2110.09393.

Wesam Shishah and Ricky Maulana Fajri. 2022. Large comparative study of recent computational approaches in automatic hate speech detection. *TEM Journal*, 11(1):82–93.

Neeraj Vashistha and Arkaitz Zubiaga. 2021. Online multilingual hate speech detection: Experimenting with hindi and english social media. *Information*, 12(1):5.

Abhishek Velankar, Hrushikesh Patil, Amol Gore, Shubham Salunke, and Raviraj Joshi. 2021. Hate and offensive speech detection in hindi and marathi. *Preprint*, arXiv:2110.12200. ArXiv:2110.12200.