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 Introduction

The advent of digital platforms has revolutionized communication, fostering inclusivity and global interconnectivity. However, this digital transformation has also led to the pervasive problem of hate speech, which undermines societal cohesion by targeting individuals or groups based on attributes such as ethnicity, religion, or gender (Vidgen and Derczynski, 2020). Hate speech, when left unchecked, can incite violence, perpetuate discrimination, and disrupt online spaces. The challenge of detecting and mitigating hate speech is compounded by the linguistic diversity of online platforms, where low-resource languages lack adequate tools and datasets to effectively tackle this issue (Ranasinghe and Zampieri, 2020).

Traditional approaches to hate speech detection primarily rely on supervised learning techniques that require extensive labeled datasets. However, for low-resource languages, such datasets are often unavailable or insufficient, making it difficult to build robust detection systems (Waseem et al., 2017). Multilingual language models have emerged as a promising solution, leveraging their ability to transfer learning across languages and adapt to tasks with limited resources (Conneau et al., 2020). These models harness cross-lingual embeddings, enabling effective generalization across diverse linguistic contexts, including morphologically rich and syntactically complex languages.

This study builds on the hypothesis that multilingual models trained on related languages and auxiliary tasks can outperform single-task, single-language models in hate speech detection for low-resource languages. By incorporating Multi-Task Learning (MTL) and Zero-Shot Learning (ZSL) techniques, we aim to enhance the model's generalization capabilities, reducing dependency on large datasets and achieving comparable performance to state-of-the-art models trained explicitly on the target language (Ruder, 2017; Liu et al., 2021)

The integration of auxiliary tasks allows for the extraction of complementary features that improve hate speech detection performance (Benton et al., 2017). Where as leveraging datasets from multiple related languages improves cross-lingual transfer, demonstrating that linguistic diversity can be an asset rather than a limitation in combating hate speech.

The subsequent sections detail the related research and methodology, including the datasets, implementation strategies, and evaluation metrics employed. We discuss the implications of our findings for the broader task of hate speech detection in low-resource languages and outline directions for future research to address the limitations of current approaches.

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

The methodology section outlines the approaches employed to address the challenges of hate speech detection in low-resource languages, leveraging the capabilities of multilingual models and techniques like Zero-Shot Learning (ZSL) and Multi-Task Learning (MTL). This study therefore tries to explore both cross-lingual generalization and multi-task optimization to improve detection capabilities. The section also delves into the datasets used, evaluation metrics employed, and the choice of low-resource language. All of this collectively forms the backbone of this research.

3.1 Zero-Shot Learning

To implement zero-shot learning for hate speech detection in low-resource languages, we utilized XLM-RoBERTa, a multilingual model pre-trained on high-resource languages with ample labeled data (Liu et al., 2021). Zero-shot learning facilitates cross-lingual generalization by enabling the model to learn semantic, contextual, and syntactic cues relevant to hate speech, which can transfer to low-resource languages through shared representations. In this study, Hindi was selected as the low-resource, unseen target language, with the model fine-tuned on high-resource hate speech datasets (Kumar and Ojha, 2019; Mandl et al., 2019). The effectiveness of zero-shot learning was evaluated by testing the model on Hindi datasets to assess its ability to bridge the gap between source and target languages for hate speech detection.

3.2 Multi-Task Learning

This project employs Multi-Task Learning (MTL) to enhance hate speech detection by simultaneously learning from multiple related tasks: hate speech detection, sentiment analysis, and offensive language detection. MTL enables the model to develop shared representations that improve generalization and capture nuanced linguistic features relevant to the primary task—hate speech detection.

Usually, equal weights are assigned to all tasks to ensure a fair contribution of auxiliary tasks without biasing the learning process. XLM-RoBERTa serves as the foundation for the MTL setup, leveraging shared layers to capture general language features while maintaining task-specific output layers to allow specialization for individual tasks. By learning from sentiment analysis and offensive language detection, the model benefits from additional linguistic signals that refine its ability to distinguish hate speech (Caruana, 1997; Ruder, 2017; Plaza-Del-Arco et al., 2021).

This approach builds a more comprehensive understanding of text, enabling improved robustness and generalization in hate speech detection. Shared representations from auxiliary tasks enhance the primary task by providing supplementary context, such as sentiment cues and offensive content patterns, which are critical for distinguishing hate speech from benign content.

3.3 Datasets

To ensure robust generalization, we incorporated datasets spanning multiple related and unrelated languages, alongside auxiliary tasks such as sentiment analysis and offensive language detection, in addition to the primary task of hate speech detection.

3.3.1 Training Data

To develop a comprehensive training setup, we integrated datasets across various languages, emphasizing both task diversity and linguistic variation.

English: We utilized the widely recognized ICWSM 2017 dataset (Davidson et al., 2017) for hate speech and offensive language detection, sampling 5,000 examples for each task. Additionally, 5,000 samples from the Sentiment140 dataset from Kaggle were employed for sentiment analysis.

Marathi: The Marathi model was trained using datasets from the L3Cube-MahaNLP repository, a comprehensive resource for Marathi NLP. The MahaHate dataset supported hate speech and offensive language detection, while the MahaSent dataset facilitated sentiment analysis, with each dataset providing over 20,000 samples.

Bangla: Unlike the previous two languages, the Bangla datasets were curated from multiple sources. For our experiment, we made use of - BD-SHS consisting of approximately 50,000 samples for hate speech detection, 5,000 samples from HASOC 2024 for offensive language detection, and SAIL, a relatively small dataset with approximately 1,000 samples for sentiment analysis.

German: For the German language, training data was compiled from multiple reliable sources to ensure diversity and robustness. The German Hate Speech Corpus, a repository collated from several datasets, included approximately 1,200 samples for each class: hate, non-hate, offensive, and non-offensive. Additionally, for sentiment analysis, the German Sentiment Dataset was utilized. From this large-scale dataset, containing over 5 million samples, approximately 800 examples were sampled for each class, specifically the positive, negative and neutral sentiment categories.

3.3.2 Test Data

Hindi: Hindi designated as the target language for our evaluations. We employed Sub-task A of the HASOC 2019 Hindi dataset, which consists of approximately 8,000 samples. Each sample is annotated into one of two classes: Hate and Offensive (HOF) or Non-Hate and Offensive (NOT), thereby establishing a binary classification task.

Hinglish: The Hinglish test set was created to evaluate the model's performance on code-switched data, where Hindi is written in the Latin script. For this purpose, we used the Hinglish Hate Detection dataset (Sengupta et al., 2021), consisting of approximately 5,000 samples. Each sample is annotated into one of two classes: Hate (yes) or Non-Hate (no), forming a binary classification task. This dataset reflects the prevalent use of code-switched language on social media platforms, where users seamlessly mix Hindi and English for communication, posing additional challenges for traditional multilingual models.

Nepali: To further evaluate the model performance and validate our approach, we have selected a Nepali test dataset, sourced from Kaggle, the Nepali Abusive Language NER and Sentiment Analysis. We have tested our model on approx-

imately 4000 samples, primarily written in the Devanagari script. Each sample is categorized into two classes: Hate (1) and Non-Hate (0), presenting a binary classification task.

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.

3.5 Choice of low-resource language

Selecting Hindi as the focal low-resource language for this project is advantageous for several strategic reasons. Firstly, compared to high-resource languages like English, Hindi still needs extensive computational and linguistic resources, presenting a meaningful challenge that aligns with our research objectives.

However, despite this paucity, a considerable amount of research has been dedicated to Hindi, which can be attributed to the fact that it is one of the most widely spoken languages. This provides a robust repertoire of baselines and upper bounds against which our experiments can be effectively benchmarked.

Additionally, Hindi is part of the Indo-Aryan language family and, therefore, shares similarities with its sister languages. This linguistic similarity facilitates cross-lingual knowledge transfer in our zero-shot experimentation, potentially enhancing the model's performance even with limited direct resources. Moreover, Hindi is prominently included in the pre-training of most multilingual models, eliminating the need to train models from scratch and saving time and computational resources.

Ultimately, the focus of this experiment is to underscore the broader applicability of multi-task learning for zero-shot hate speech detection. The insights derived should be inherently transferable and, therefore, could be applied to other low-resource languages.

4 Experimentation and Results

4.1 Implementation Details

Data Pre-processing: Minimal data preprocessing was employed to retain the linguistic richness of datasets. This involved cleaning the textual data by removing URLs, mentions, hashtags, and extraneous whitespace using regular expressions, thereby enhancing the model's ability to focus on meaningful linguistic patterns. Tokenization was performed using XLM-RoBERTa's pre-trained tokenizer to standardize input for model training (Wolf et al., 2020).

Dataset Balancing: To ensure fairness and consistency across experiments, we constructed nearly balanced datasets for each task and language by undersampling larger classes. We maintained equal or near-equal label counts across all classes within each dataset for the three tasks. Additionally, we computed class weights to retain flexibility for scenarios where class imbalances might occur.

Model Architecture: As stated before, the core of the model architecture consists of the XLM-RoBERTa encoder, renowned for its robust multilingual capabilities, and superior performance on morphologically rich, low-resource languages (Conneau et al., 2020). Training was conducted in two settings: single-task learning and multi-task learning. Furthermore, for multitask learning, the model was supplemented by distinct classification heads tailored for each specific task. This modular design facilitates simultaneous learning across tasks while preserving task-specific nuances.

Training: Training was conducted using the AdamW optimizer with a carefully scheduled learning rate of 2e-5 to promote stable convergence. We employed a batch size of 16 and trained the model for 5 epochs. To address class disparities, loss functions were weighted appropriately. During training, we encountered the issue of exploding gradients, where the model tended to predict a single class. This was effectively mitigated by applying gradient clipping.

4.2 Results

We conducted a series of experiments to evaluate the impact of multilingual and multi-task learning on zero-shot hate speech detection using the XLM-RoBERTa model. The results have been summa-

rized in the Table 1 and Table 2. As a benchmark, a prior study, Ghosh and Senapati (2022), achieved an F1-score of 0.82 on the HASOC2019 dataset using a single-task approach.

Our initial experiment combined two Indo-Aryan languages, Marathi and Bangla, in a multi-task setting, resulting in a comparable F1 score of 0.79. To assess the benefits of multilingual training, we then evaluated each language separately. Marathi, sharing the same script and being more linguistically similar to Hindi, outperformed Bangla by 0.11 points, achieving an F1 score of 0.77. Further, to understand the advantages of multi-task learning, we trained Marathi in a single-task setting, which led to a 9% decrease in performance, underscoring the efficacy of multi-task approaches.

Expanding our investigation to non-Indo-Aryan languages, we combined English and German in a multi-task framework. The performance mirrored that of Bangla's multi-task results. However, individual evaluations revealed a significant disparity: English achieved an F1 score of 0.64, while German lagged with a score of just 0.40, accompanied by a similar 12% drop when switching to a single-task English setup.

Training Languages	Macro F1 Score	Accuracy
English	0.56	0.59
Marathi	0.70	0.72

Table 1: Zero-Shot Single Task Evaluation Results on Hindi Dataset

Training Languages	Macro F1 Score	Accuracy
German	0.40	0.40
English	0.64	0.64
English + German	0.69	0.67
Bangla	0.66	0.67
Marathi	0.77	0.77
Marathi + Bangla	0.79	0.79

Table 2: Zero Shot, Multi Task Learning Evaluation Results on Hindi Dataset

Training Languages	Macro F1 Score	Accuracy
German	0.50	0.55
English	0.57	0.58
English + German	0.65	0.65
Bangla	0.60	0.60
Marathi	0.65	0.65
Marathi + Bangla	0.69	0.71

Table 3: Zero Shot, Multi Task Learning Evaluation Results on Nepali Dataset

From these limited experiments in Table 1 and Table 2, we infer that the incorporation of multiple languages yields minimal benefits, mainly dominated by the performance of the most similar language, leading to diminishing returns with additional languages. Therefore, when training models, it is crucial to balance the availability of related language data, computational costs, and the marginal gains achieved.

In addition, training in languages of similar linguistic origin, such as Marathi, produces more substantial generalization capabilities than distant languages such as German. This aligns with the hypothesis that related languages provide greater contextual and syntactic overlap, thereby improving the model's performance, as verified by the results in Table 2 and Table 3

Training Languages	Macro F1 Score	Accuracy
German	0.28	0.37
English	0.52	0.52
English + German	0.42	0.45
Bangla	0.48	0.52
Marathi	0.51	0.56
Marathi + Bangla	0.51	0.51

Table 4: Zero Shot, Multi Task Learning Evaluation Results on Hinglish Dataset

To further explore the performance of the model, we evaluated it on code-switched Hinglish data, where Hindi is written in Latin script. Table 4 summarizes the results for this experiment. Despite the strong performance of the model on Devanagari-script Hindi in previous evaluations, its performance on Hinglish data degraded significantly,

with the best F1 score of 0.52 achieved when trained on English or Bangla, and the lowest score of 0.28 observed when trained on German. Even combinations of related languages such as Marathi and Bangla failed to improve performance beyond 0.51. These findings highlight the challenges posed by code-switched text, where the absence of explicit Hinglish training data creates a gap in the model's understanding of mixed-language syntax and semantics.

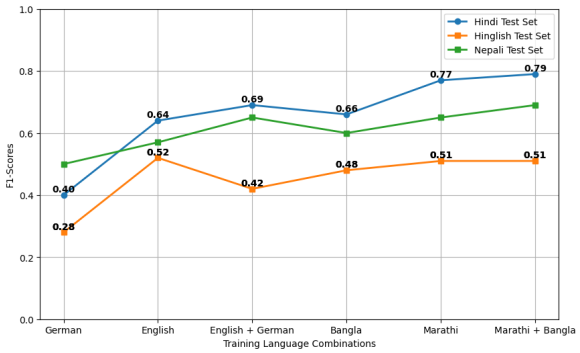

Figure 1: F1 Scores for Hate Speech Detection In Hindi

Conversely, multi-task learning consistently demonstrated substantial improvements, highlighting its importance. Obviously, the selection of appropriate auxiliary tasks and the availability of relevant data in related languages are pivotal factors that influence the effectiveness of multi-task approaches.

5 Conclusion

This study explored cross-lingual hate speech detection for low-resource languages, starting with zero-shot learning. While training on unrelated languages like English provided limited performance, a notable improvement was observed when using culturally and linguistically similar Indo-Aryan languages, such as Marathi, highlighting the value of shared linguistic characteristics. Building on this, multi-task learning—combining related tasks like hate speech detection, offensive language identification, and sentiment analysis—further enhanced performance by leveraging shared knowledge across tasks. Together, these results demonstrate a clear progression: from the lower-bound performance of a simple zero-shot model to increasingly stronger outcomes through the integration of related languages and tasks. While zero-shot learning provides a strong starting point, combining it with linguistically similar data and multi-task objectives proves to be a more promising solution for low-resource settings. This work underscores the importance of exploring such techniques to reduce reliance on annotated datasets and improve hate speech detection for underrepresented languages.

The inclusion of Nepali data allows us to analyze the model's cross-lingual generalization capabilities for another Indo-Aryan language that shares certain syntactic and morphological similarities with Hindi. This evaluation further highlights the challenges and opportunities of applying multilingual models to closely related but resource-constrained languages.

While the results are promising, the model struggles when confronted with more complex linguistic scenarios, such as code-switched text. Code-switching, particularly in the case of Hinglish, poses unique challenges due to the dynamic mixing of Hindi and English across scripts. This blending of languages introduces additional ambiguity, which makes it harder for models to accurately process and classify the text. The decline in performance on Hinglish data underscores the limitations of current multilingual models and highlights the need for adaptations tailored to such hybrid linguistic constructs. To address these challenges, fine-tuning models on specialized code-switched datasets and implementing more robust preprocessing techniques can help improve the model's ability to generalize across mixed-language inputs.

The performance shortfall on Hinglish data can be traced to two critical factors. Firstly, models like XLM-RoBERTa are primarily pre-trained on monolingual datasets, particularly in native scripts, and thus have minimal exposure to transliterated or code-switched content. Secondly, Hinglish introduces distinct syntactic structures and morphological variations arising from the combination of Hindi and English. These complexities create ambiguities that pre-trained multilingual models struggle to resolve without targeted fine-tuning. Bridging this gap will require integrating code-switched corpora into model training and exploring advanced approaches, such as character-level embeddings or hybrid architectures, to better handle the nuances of Latinized Hindi text.

6 Future Works

While this study demonstrates the effectiveness of multilingual models in hate speech detection, several limitations and future directions remain.

First, sentiment analysis in our current framework is limited to broad categories (positive, neutral, and negative). Incorporating fine-grained emotional classifications, such as Ekman's seven universal emotions—happiness, sadness, anger, fear, disgust, surprise, and contempt—could provide richer contextual information to improve hate speech detection (Ekman, 1992).

Second, annotator biases in labeling datasets remain an underexplored issue. Cultural and moral differences among annotators may introduce inconsistencies in labeling, which could impact model fairness. Leveraging frameworks such as the Moral Foundations Theory can help structure annotations and mitigate these biases (Davani et al., 2024).

Our current multi-task and multilingual learning approaches rely on an empirical grouping of related tasks and languages. Future work could explore methods to identify optimal groupings through ablation studies, as alternative configurations may yield improved performance.

Furthermore, equal task weighting was applied in our MTL setup, which may not optimally leverage auxiliary tasks. Implementing adaptive task-weighting strategies, such as uncertainty-based methods or gradient normalization techniques, could help the model prioritize the primary task during training.

Finally, our approach struggles with code-switched text, particularly Hinglish (Hindi written in the Latin script). Since a significant portion of social media hate speech content involves code-switching, future work should focus on fine-tuning multilingual models with code-switched datasets to enhance performance (Sharma et al., 2021; Velankar et al., 2021). Similarly, incorporating hate speech type classification—such as distinguishing between content targeted at individuals, groups, or specific demographics like race, religion, or gender—could provide further granularity and improve detection capabilities (Narayan et al., 2023).

These directions aim to address the limitations of the current work, ensuring more robust, fair, and adaptable solutions for hate speech detection across diverse linguistic and cultural contexts.

References

Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017. Multi-task learning for mental health using social media text. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 152–162.

R. Caruana. 1997. Multitask learning. *Machine Learning*, 28(1):41–75.

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.

Aida Davani, Yu-Cheng Yeh, Pierre-François Bouillon, and François Simard. 2024. Annotator biases and the moral foundations of offensive content. *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, pages 1–12.

T. Davidson, D. Warmsley, M. Macy, and I. Weber. 2017. Automated hate speech detection and the problem of offensive language. *ICWSM*.

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.

Paul Ekman. 1992. *An Argument for Basic Emotions*, volume 6. Cognition and Emotion.

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 multi-task 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*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pretrain, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

T. Mandl, S. Modha, and D. Patel. 2019. Overview of hasoc 2019: Hate speech and offensive content identification in indo-european languages. In *FIRE*.

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.

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.

Tharindu Ranasinghe and Marcos Zampieri. 2020. Multilingual offensive language identification with cross-lingual embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5838–5844, Online. Association for Computational Linguistics.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Ayan Sengupta, Sourabh Kumar Bhattacharjee, Md. Shad Akhtar, and Tanmoy Chakraborty. 2021. Does aggression lead to hate? detecting and reasoning offensive traits in hinglish code-mixed texts. *Neurocomputing*.

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. *arXiv preprint arXiv:2110.12200*.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data: Garbage in, garbage out. *PLOS ONE*, 15(12):e0243300.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84. Association for Computational Linguistics.

T. Wolf, L. Debut, and V. Sanh. 2020. Transformers: State-of-the-art natural language processing. In *EMNLP*.