Hope_for_best@LT-EDI 2025: Detecting Racial Hoaxes in Code-Mixed Hindi-English Social Media Data using a multi-phase fine-tuning strategy

Abhishek Singh Yadav[1] **Deepawali Sharma**[2] **Aakash Singh**[1] **Vivek Kumar Singh**[1]
[1]Department of Computer Science, University of Delhi, India
[2]School of Computer Science Engineering and Technology, Bennett University, Noida, India
meabhishek8965@gmail.com, deepawali21@bhu.ac.in
asingh@cs.du.ac.in, vivek@cs.du.ac.in

Abstract

In the age of digital communication, social media platforms have become a medium for the spread of misinformation, with racial hoaxes posing a particularly insidious threat. These hoaxes falsely associate individuals or communities with crimes or misconduct, perpetuating harmful stereotypes and inflaming societal tensions. This paper describes the team "Hope_for_best" submission that addresses the challenge of detecting racial hoaxes in code-mixed Hindi-English (Hinglish) social media content and secured the 2^{nd} rank in the shared task (Chakravarthi et al., 2025). To address this challenge, the study employs the HoaxMixPlus dataset, developed by LT-EDI 2025, and adopts a multi-phase fine-tuning strategy. Initially, models are sensitized using the THAR dataset—targeted hate speech against religion (Sharma et al., 2024) —to adjust weights toward contextually relevant biases. Further fine-tuning was performed on the HoaxMixPlus dataset. This work employed data balancing sampling strategies to mitigate class imbalance. Among the evaluated models, HingBERT achieved the highest macro F1-score of 73% demonstrating promising capabilities in detecting racially charged misinformation in code-mixed Hindi-English texts.

1 Introduction

In the digital era, social media platforms have revolutionized global communication by enabling individuals to disseminate information across vast and diverse audiences. However, this accessibility has also facilitated the rapid spread of misinformation, including racially charged hoaxes that falsely implicate individuals or communities in criminal or unethical behavior. Such hoaxes are not merely misinformative but are deliberately crafted to reinforce harmful stereotypes, incite hostility, and exacerbate societal divides (Singh et al., 2025).

Platforms like Twitter, Facebook, and YouTube empower users with the ability to express opinions publicly and anonymously. While this democratization of speech has positive implications, it also opens the door to misuse. These platforms, with their anonymous nature and rapid content spread, often intensify hate-fueled narratives. This makes it increasingly important to build automated systems that can identify and curb such content before it leads to real-world consequences (Shanmugavadivel et al., 2022).

In the domain of targeted hate speech detection, Natural Language Processing (NLP) has made significant progress with the advent of deep learning architectures (Sharma et al., 2025b). Early approaches employed Long Short-Term Memory networks (LSTM) (Hochreiter and Schmidhuber, 1997) and Gated Recurrent Units (GRU) (Chung et al., 2014), but the emergence of transformer-based models, particularly BERT (Vaswani et al., 2017), has significantly improved performance across a range of language understanding tasks.

One of the major challenges in this domain is dealing with code-mixed text, especially Hindi-English (Hinglish), which is commonly used in Indian social media discourse. This linguistic mixing complicates tokenization, syntactic parsing, and semantic understanding. While several studies have focused on hate speech detection in Indic languages (Mathew et al., 2021; Patwa et al., 2020), the problem of detecting racial hoaxes in Hinglish remains underexplored.

The present study contributes to this growing field by introducing a transformer-based approach for the classification of racial hoaxes in code-mixed Hindi-English social media content. Building on the HoaxMixPlus dataset introduced by LT-EDI 2025, this study employs a multi-phase fine-tuning strategy to adapt models for the detection of contextually biased misinformation. Initially, pretraining is conducted on the THAR dataset (Sharma et al., 2024), which targets hate speech against religious communities, followed by fine-tuning on

task-specific HoaxMixPlus data. Among the models evaluated, Hing-BERT achieved the best performance, demonstrating its effectiveness in capturing racially hostile content embedded in informal, code-mixed linguistic structures.

This paper is structured as follows: Section 2 provides a comprehensive review of the existing literature on misinformation and hate speech detection, specifically within the context of code-mixed languages. Section 3 outlines the datasets utilized in this study, namely the THAR and HoaxMixPlus datasets, along with a detailed description of their respective features. Section 4 describes the methodology employed, with an emphasis on the multi-phase fine-tuning approach, model architecture, and data preparation techniques. In Section 5, the experimental results are presented. Section 6 offers a detailed discussion and analysis of the model's performance. Finally, Section 7 concludes the paper and highlights potential avenues for future research.

2 Related Work

The detection of misinformation and racially motivated hoaxes in social media has attracted increasing attention in recent years, particularly in multilingual and code-mixed contexts. Prior studies have explored various linguistic and contextual challenges in identifying harmful narratives, such as hate speech, fake news, and racially biased misinformation.

There is a growing need for research addressing harmful and biased content in code-mixed and multilingual social media, supported by the creation of linguistically diverse datasets and model strategies. Studies such as HopeEDI (Chakravarthi, 2020) and the ensemble-based model for hope speech detection in English and Dravidian languages (Sharma et al., 2025a) highlight the effectiveness of such approaches in promoting inclusive and equitable language technologies.

Code-mixed language, especially Hindi-English (Hinglish), presents significant challenges for natural language understanding due to its informal structure and lack of standardized grammar. Recent efforts, such as Patwa et al. (2020), have addressed sentiment analysis and offensive language detection in code-mixed texts through SemEval-2020. Similarly, Barman et al. (2014) provided foundational insights into part-of-speech tagging in Bengali-Hindi-English code-mixed social media content, underscoring the complexity of such texts.

Although racial hoaxes remain an underexplored domain, prior research in hate speech detection provides valuable foundations. Datasets such as THAR (Sharma et al., 2024) target religious hate in multilingual Indian contexts, offering pretraining potential for models tackling similar sociolinguistic phenomena. Other notable works include Mathew et al. (2021), who introduced HateXplain—a benchmark dataset for hate speech detection with multi-perspective annotations including class labels, target communities, and human-provided rationales to improve model explainability and reduce bias, and Vidgen and Yasseri (2020), who developed a multi-class classifier to distinguish between non-Islamophobic, weakly Islamophobic, and strongly Islamophobic content, emphasizing the need for nuanced categorization over binary classification.

Transfer learning through sequential fine-tuning has shown considerable promise in improving task-specific performance for low-resource and domain-specific problems. The use of a multiphase fine-tuning pipeline, where models are initially exposed to related bias-aligned data (e.g., hate speech or religious hostility) and subsequently adapted to the target task (e.g., racial hoaxes), aligns with strategies explored in Gururangan et al. (2020), who demonstrated the efficacy of domain-adaptive pretraining. In the current work, models such as Hing-BERT leverage this approach by first calibrating on THAR before task-specific tuning on HoaxMixPlus.

Pretrained multilingual models like MuRIL (Kakwani et al., 2020) and domain-specific transformers such as Hing-BERT (Kumar et al., 2020) have been specifically optimized for Indian languages and their mixed variations. These models benefit from pretraining on diverse scripts and colloquial structures, making them suitable for nuanced detection tasks in Hinglish texts. Moreover, models like hing-roberta-mixed have demonstrated competitive performance in identifying hate speech in informal, noisy, and multilingual settings.

Given the skewed nature of real-world social media datasets, strategies like data sampling, loss re-weighting, and oversampling are commonly adopted to mitigate bias and improve minority class detection. Approaches documented in Rathpisey and Adji (2022) emphasize the importance of balancing in achieving fairer performance across all classes.

Despite the substantial progress in detecting hate

[1]https://www.aclweb.org/anthology/2020.peoples-1.5/

speech, fake news, and offensive content in code-mixed and multilingual contexts, the specific problem of identifying racially motivated hoaxes remains insufficiently addressed. Most existing studies focus on broad categories of harmful content, often overlooking the nuanced linguistic and contextual markers unique to racial hoaxes, especially in informal, code-mixed languages like Hinglish. This presents a significant research gap, as racially charged misinformation can have far-reaching societal impacts. In this work, we aim to address this gap by proposing a novel multi-phase fine-tuning approach—first sensitizing models on a related hate speech dataset (THAR), then adapting them to the task-specific HoaxMixPlus dataset for racial hoax detection. This strategy enhances model performance in low-resource settings while introducing a focused lens on racial misinformation.

3 Dataset Description

The datasets used in this work were provided by the organizers of the LT-EDI 2025 shared task [1] (Chakravarthi et al., 2025). Two datasets were employed in our multi-phase fine-tuning approach: the THAR dataset (Sharma et al., 2024), which targets religion-based hate speech, and the HoaxMixPlus dataset, a novel resource annotated for racial hoaxes in code-mixed Hindi-English social media content.

3.1 THAR Dataset

The Targeted Hate Against Religion (THAR) dataset comprises social media comments annotated for the presence of religious hate speech. The dataset consists of binary labels, with values Non-AntiReligion and AntiReligion. This dataset was used to contextually sensitize the model toward sociocultural bias before fine-tuning on the target task. Due to limited data in HoaxMixPlus, we first fine-tune the model on the larger, related THAR dataset to help it learn code-mixed hate speech patterns, enhancing its performance on racial hoax detection.

3.2 HoaxMixPlus Dataset

The HoaxMixPlus dataset consists of 5,105 YouTube comment posts written in code-mixed Hindi-English (Hinglish). It is annotated specifically for racial hoaxes, which are a subcategory of misinformation that falsely associates individuals or groups with crimes or controversial events. This dataset represents an important advancement for low-resource language settings, offering a benchmark for racial hoax detection in multilingual social contexts. The dataset(Training and validation) includes two fields: clean_text and labels, and the test set contains three fields: id, clean_text, and labels. The labels are binary, with values non-racial hoax and racial hoax.

The distribution of both datasets is provided in Table 1 and 2.

Table 1: THAR Dataset Distribution for Religious Hate Speech Detection

Dataset	Non-AntiReligion	AntiReligion	Total
THAR	6,095	5,454	11,549

Table 2: HoaxMixPlus Dataset Distribution for Racial Hoax Detection

Dataset	Non-Racial	Racial	Total
HoaxMixPlus (Train)	2,319	741	3,060
HoaxMixPlus (Dev)	774	247	1,021
HoaxMixPlus (Test)	774	247	1,021

4 Methodology

Text classification remains a fundamental task in Natural Language Processing (NLP), particularly when dealing with complex phenomena such as racial hoaxes in multilingual contexts. Our approach addresses the challenge of detecting racial hoaxes in code-mixed Hindi-English social media content through a novel multi-phase sequential fine-tuning architecture using Hing-BERT model.

This paper aims to highlight the importance of detecting racially motivated hoaxes in online discourse and presents a robust methodology that integrates contextual pretraining, class balancing techniques, and model architecture selection tailored for code-mixed inputs.

4.1 Data Preparation and Balancing

The experiment utilizes two distinct datasets: 'Racial Hoaxes dataset', and 'THAR dataset' (Targeted Hate Against Religion dataset). Due to the inherent class imbalance in the racial hoaxes dataset, we implemented an upsampling technique for the minority class (racial hoaxes) to create a balanced training dataset. This process involves randomly

[1] https://codalab.lisn.upsaclay.fr/competitions/21885

sampling with replacement from the minority class until it matches the size of the majority class, followed by shuffling the combined dataset. This approach prevents bias toward the majority class and improves model generalization.

4.2 Model Architecture

Our approach leverages the "l3cube-pune/Hing-BERT" pre-trained model, which is specifically designed for Hindi-English code-mixed text. This model builds upon the BERT architecture but has been pre-trained on a corpus of code-mixed Hindi-English data, making it particularly suitable for our task. We adapted this model for sequence classification with a binary output layer to classify text as either containing racial hoaxes (1) or not (0). The Hing-BERT model maintains the transformer-based architecture with multiple self-attention heads, which allows it to effectively capture contextual relationships in code-mixed text where linguistic patterns differ significantly from monolingual content.

4.3 Multi-Phase Sequential Fine-tuning

The core innovation in our methodology is the multi-phase sequential fine-tuning approach:

1. **First Fine-tuning Phase (Domain Adaptation/Sensitivity Conditioning):** We initially fine-tune the Hing-BERT model on the THAR dataset focused on anti-religious hate speech content. This phase sensitizes the model's weights towards recognizing nuanced and sensitive linguistic cues commonly present in harmful content. The model thereby develops a refined sensitivity to contextually offensive and hate-indicative language patterns, effectively conditioning it for task adaptation in subsequent fine-tuning stages.

2. **Second Fine-tuning Phase (Task Adaptation):** Building on the sensitized weights from the first phase, we perform a second round of fine-tuning using the racial hoaxes dataset. This phase involves relatively minor adjustments to the preconditioned weights, steering them toward the specific task of detecting racially motivated hoaxes while preserving the model's learned sensitivity to harmful content. This sequential approach allows the model to build upon the knowledge acquired in the domain adaptation phase while specializing in the specific characteristics of racial hoaxes.

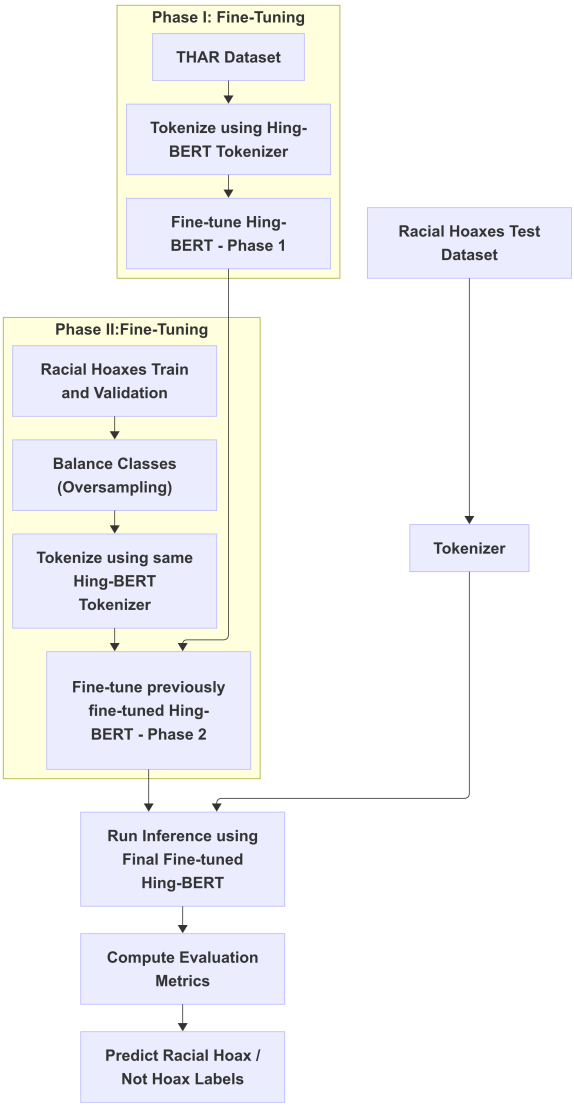

Figure 1: Two-stage fine-tuning and evaluation using THAR and HoaxMixPlus datasets.

This multi-phase approach follows the principle of curriculum learning (Bengio et al., 2009), where the model progressively learns from a broader or similar related domain (religious hate speech) to the specific target domain (racial hoaxes).

4.4 Tokenization and Model Configuration

We employed the specialized tokenizer from "l3cube-pune/Hing-BERT", which effectively handles code-mixed Hindi-English text. Texts were tokenized with a maximum sequence length of 128 tokens, applying padding and truncation as needed. This configuration balances computational efficiency with the need to capture sufficient context from social media posts.

4.5 Training Configuration

The model fine-tuning involves adjusting several standard hyperparameters during the training process, which are set explicitly in the `TrainingArguments` objects. The learning rate is tuned in two phases: Phase I (THAR dataset pre-training) uses a learning rate of 2×10^{-5}, while Phase II (racial hoax dataset fine-tuning) uses a reduced learning rate of 2×10^{-6}. This staged reduction allows the model to converge smoothly and helps prevent catastrophic forgetting after initial domain adaptation. The batch size is set to 16 per device for both training and evaluation. The number of training epochs is set to 4 in both phases. A weight decay of 0.01 is applied to regularize the model and reduce the risk of overfitting. The model saving strategy includes `load_best_model_at_end=True`, ensuring that the best model, based on validation F1-score, is retained at the end of training. All fine-tuning is done by updating the standard transformer layers and the classification head parameters with no weights are frozen. No adapters are used in this model and the fine-tuning directly optimizes the full model without incorporating any additional adapter modules. The code uses `AutoModelForSequenceClassification`, which internally applies cross-entropy loss for binary classification. This is standard and not overridden or custom-defined in the code. The optimization was performed with the AdamW optimizer.

4.6 Inference Pipeline

For deployment and testing, we developed a prediction function that processes new text samples through the following steps:

1. Tokenization of the input text
2. Forward pass through the model
3. Classification based on the output logits
4. Return of human-readable prediction (Racial Hoax detected/Not a Racial Hoax)

Several transformer models were trained using the training and development datasets, and their performance is shown in Table 3. After testing the performance of various transformer models, the top three models with the best performance were selected. The models, Hing-BERT (Nayak and Joshi, 2022), BAAI BGE-M3 (Sun et al., 2024), hing-roberta-mixed (Nayak and Joshi, 2022), MuRIL (Khanuja et al., 2021) were selected. The chosen models were trained using a combined version of the training and validation datasets to make the final predictions.

5 Results

The proposed approach, centered around Hing-BERT and refined using a multi-phase fine-tuning strategy, demonstrated strong effectiveness in identifying racial hoaxes in code-mixed Hindi-English (Hinglish) social media data. After initially adapting the model to socio-religious hate contexts using the THAR dataset, the system was further fine-tuned on the HoaxMixPlus dataset, allowing it to capture task-specific linguistic and contextual cues.

On the test set consisting of 1,021 instances, the model achieved a high overall accuracy of 80%, affirming the robustness of the learned representations. Notably, the model attained a macro-averaged F1-score of 0.73, indicating balanced performance across both hoax and non-hoax classes. The weighted average precision and recall values, both reaching 0.81 and 0.80 respectively, highlight the model's strong capability to make reliable predictions while handling class distribution effectively.

The BAAI BGE-M3 model also performed well, with a macro F1-score of 0.72 and accuracy of 0.79, closely followed by MuRIL and hing-roberta-mixed. Models like Indic-BERT, Hing-BERT LID (Nayak and Joshi, 2022) and roberta-en-hi-codemixed exhibited comparatively lower scores, suggesting they are less effective for this specific task. The results affirm that domain-specific pre-training and code-mixed adaptability significantly enhance model effectiveness for this challenge. The results of all the models evaluated using the training data are presented in Table 3.

The results presented in Table 4 show the per-class F1-scores, precision, and recall for various models in detecting non-racial hoax (Class 0) and racial hoax (Class 1) content.

6 Discussion

Hing-BERT outperforms other models in detecting racial hoaxes within code-mixed Hindi-English social media text due to its alignment with the linguistic characteristics of such data. Unlike models like Indic-BERT, mBERT, or MuRIL, which

Table 3: Model Performance with Multi-Phase Fine-Tuning

Model	Macro F1 Score	Accuracy
Indic-BERT	0.68	0.74
roberta-en-hi-codemixed model	0.67	0.73
BAAI BGE-M3	0.72	0.79
Muril model	0.70	0.75
hing-roberta-mixed	0.70	0.75
Hing-BERT LID	0.69	0.78
Hing-BERT	0.73	0.80

Table 4: Performance Metrics for Each Model (Per-Class F1 Scores)

Model	Non-Hoax			Hoax		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Indic-BERT	0.80	0.87	0.83	0.62	0.50	0.55
roberta-en-hi-codemixed	0.75	0.87	0.81	0.66	0.46	0.54
BAAI BGE-M3	0.87	0.84	0.85	0.56	0.59	0.57
Muril	0.83	0.86	0.85	0.59	0.53	0.56
hing-roberta-mixed	0.76	0.89	0.82	0.70	0.49	0.57
Hing-BERT LID	0.77	0.88	0.69	0.67	0.49	0.56
Hing-BERT	**0.88**	**0.86**	**0.87**	**0.57**	**0.60**	**0.58**

were pre-trained on formal or monolingual corpora, Hing-BERT was pre-trained on large-scale real-world code-mixed data from platforms such as Twitter and YouTube. This exposure enables it to model code-switching patterns, transliteration variants (e.g., *acha*, *accha*, *achha*), and the blending of grammatical structures across languages more effectively.A key strength is its ability to deal with the informal, messy nature of social media, including slang, spelling variations, hashtags, emojis, and subtle code-switch points that may signal sarcasm or misinformation. However, the model has its drawbacks. It tends to favor the majority class due to dataset imbalance and can be sensitive to noisy inputs like excessive emojis or special characters. There's also the risk of hidden biases linked to demographics or dialects in the training data. Finally, the model's decisions are not easily explainable, making it harder to understand or trust why certain posts are flagged as racial hoaxes.

7 Conclusion and Future Work

This study presents an effective multi-phase fine-tuning approach for detecting racial hoaxes in Hinglish social media content, achieving a macro F1-score of 73% using Hing-BERT. By incorporating bias-aware pretraining via the THAR dataset and addressing class imbalance through strategic sampling, the model demonstrates enhanced contextual sensitivity and robustness. Future enhancements may include using contrastive learning (Chen et al., 2020) to better identify subtle forms of hate, incorporating other types of information such as images and hashtags, and expanding the system to support more code-mixed languages spoken in India and fairness audits and bias mitigation to improve reliability. Techniques like adversarial testing (Goodfellow et al., 2015) and explainability (Ribeiro et al., 2016; Lundberg and Lee, 2017) can also help make the model more reliable and easier to understand when used in real-world settings.

8 Source code availability

https://github.com/Abhi-3022/Detecting-Racial-Hoaxes-in-Code-Mixed-Hindi-English-Social-Media-Data

References

Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014. Code-mixing: A challenge for language identification in the language of social media. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 13–23, Doha, Qatar. Association for Computational Linguistics.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, pages 41–48. ACM.

Bharathi Raja Chakravarthi. 2020. Hopeedi: A multilingual hope speech detection dataset for equality, diversity, and inclusion. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 41–53, Barcelona, Spain (Online). Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Prasanna Kumar Kumaresan, Shanu Dhawale, Saranya Rajiakodi, Sajeetha Thavareesan, Subalalitha Chinnaudayar Navaneethakrishnan, and Durairaj Thenmozhi. 2025. Overview of the shared task on detecting racial hoaxes in code-mixed hindi-english social media data. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. *CoRR*, abs/2002.05709.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning*.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *ICLR 2015*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Divyanshu Kakwani, Vedanuj Goswami, Janani Prabhakar, Anoop Kunchukuttan, and Pratyush Kumar. 2020. Indiccorp and muril: Large-scale language models for indian languages. *arXiv preprint arXiv:2005.00085*.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, and 1 others. 2021. Muril: Multilingual representations for indian languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3406–3412.

Raghav Kumar, Kunal Sinha, Saurabh Varshney, and Manish Shrivastava. 2020. Hingbert: A hinglish language model. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 47–51.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *CoRR*, abs/1705.07874.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875.

Ravindra Nayak and Raviraj Joshi. 2022. L3Cube-HingCorpus and HingBERT: A code mixed Hindi-English dataset and BERT language models. In *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 7–12, Marseille, France. European Language Resources Association.

Parth Patwa, Gustavo Aguilar, Sudipta Kar, Suraj Pandey, Srinivas PYKL, Björn Gambäck, Tanmoy Chakraborty, Thamar Solorio, and Amitava Das. 2020. Semeval-2020 task 9: Overview of sentiment analysis of code-mixed tweets. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation (SemEval-2020)*, pages 774–790, Barcelona, Spain (Online). International Committee for Computational Linguistics.

Heng Rathpisey and Teguh Bharata Adji. 2022. Handling imbalance issue in hate speech classification using sampling-based methods. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2583–2592, Gyeongju, Republic of Korea. IEEE.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938.

K. Shanmugavadivel, A. Narayanan, and M. Senthilkumar. 2022. The challenge of detecting online hate speech: A systematic review. *International Journal of Computer Applications*, 184(12):1–8.

D. Sharma, V. Gupta, V. K. Singh, and B. R. Chakravarthi. 2025a. Stop the hate, spread the hope: An ensemble model for hope speech detection in english and dravidian languages. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

D. Sharma, T. Nath, V. Gupta, and V. K. Singh. 2025b. Hate speech detection research in south asian languages: A survey of tasks, datasets and methods. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 24(3):1–44.

Deepawali Sharma, Aakash Singh, and Vivek Kumar Singh. 2024. Thar: Targeted hate speech against religion: A high-quality hindi-english code-mixed dataset with the application of deep learning models for automatic detection. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

A. Singh, D. Sharma, and V. K. Singh. 2025. Misogynistic attitude detection in youtube comments and replies: A high-quality dataset and algorithmic models. *Computer Speech & Language*, 89:101682.

Xiaofei Sun, Xu Han, Hao Sun, Yankai Lin, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2024. Bge-m3: A foundational model for multilingual, multimodal, and multitask retrieval. *Preprint*, arXiv:2403.17818.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Bertie Vidgen and Taha Yasseri. 2020. Detecting weak and strong islamophobic hate speech on social media. *Journal of the Association for Information Science and Technology*, 71(12):1475–1487.