CUET's_White_Walkers@LT-EDI 2025: Racial Hoax Detection in Code-Mixed on Social Media Data

Md Mizanur Rahman, Jidan Al Abrar, Md Siddikul Imam Kawser, Ariful Islam, Md. Mubasshir Naib, Hasan Murad
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
u1904{116, 080, 081, 129, 089}@student.cuet.ac.bd, hasanmurad@cuet.ac.bd

Abstract

False narratives that manipulate racial tensions are increasingly prevalent on social media, often blending languages and cultural references to enhance reach and believability. Among them, racial hoaxes produce unique harm by fabricating events targeting specific communities, social division and fueling misinformation. This paper presents a novel approach to detecting racial hoaxes in code-mixed Hindi-English social media data. Using a carefully constructed training pipeline, we have fine-tuned the XLM-RoBERTa-base multilingual transformer for training the shared task data. Our approach has incorporated task-specific preprocessing, clear methodology, and extensive hyperparameter tuning. After developing our model, we tested and evaluated it on the LT-EDI@LDK 2025 shared task dataset. Our system achieved the highest performance among all the international participants with an F1-score of 0.75, ranking 1st on the official leaderboard.

1 Introduction

Racial hoaxes on social networks have continuously emerged as a significant concern, which may lead to increased ethnic tension and social unrest. In this study, we define Hoax Speech as intentional, deceptive linguistic content that mimics the tone or structure of hate speech or propaganda, yet lacks genuine hateful intent. It often uses irony, satire, or fabricated narratives to incite confusion or mask malicious undertones for a criticise race. Racial hoaxes refer to fabricated statements that falsely accuse specific racial groups with malicious intent. These intents are designed to manipulate public opinion and provoke communal or ethnic tension.

According to Amnesty International in The News Minute, Racial hoaxes and hate speech have become prevalent in South Asia day by day, leading to violence and social turmoil. For instance, between 2015 and 2019, India witnessed 902 alleged hate crimes resulting in 303 deaths, with Muslims constituting the majority of victims [1]. Social media can amplify such crimes, leading to hatred, misinformation for the minority group or community, especially. A significant number of previous studies have been conducted on spreading misinformation, fake news, and hate speech detection (Barker and Jurasz, 2021), (Arellano et al., 2022), but the task of racial hoax detection has not been explored too much.

The primary objective of this paper is to detect Hindi-English code-mixed racially deceptive text on social media platforms. We have used a transformer-based model, HinG-RoBERTa, which is pre-trained on Hindi-English code-mixed data, to train our model. The core contributions of our research work are as follows:-

1. We have developed an effective model to detect racial deception in Hindi-English code-mixed text.
2. We have conducted a series of experiments on the dataset and comprehensively analyzed their performance outcomes.

The implementation details have been provided in the following GitHub repository:- https://github.com/Mizan116/LT-EDI-LDK-2025/Racial Hoax.

2 Related Study

Hate speech detection identifies degrading information, especially on social media. Hate speech, sexism, homophobia, racism, bullying, and other verbal abuse are detected. Prior work has studied the online dissemination of racially based stereotypes

[1] 2019 sees steepest rise in hate crimes since 2016, finds Amnesty tracker

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 63–67

and disinformation. However, very few studies examine racial hoaxes through multilingual lenses (Bourgeade et al., 2023).

Initially, the discipline was dominated by classical machine learning algorithms such as Support Vector Machines (SVMs) and Naïve Bayes and Random Forests for text mining techniques. The authors of Afroz et al. (2012) used machine learning techniques to detect hoaxes in the English language. Rule-based machine learning methods have also been used in Chopra et al. (2020). They have been used for detecting hate speech in Hindi-English code-mixed text. Research across multiple languages shows how racial hoaxes and stereotypes circulate in social media conversations in Bourgeade et al. (2023). In a related shared task on Dravidian languages, the authors of Rahman et al. (2025) employed transformer models like XLM-R and MuRIL, demonstrating high performance in abusive language detection. The authors of Ahmed et al. (2022) have detected hateful users more accurately and fairly, including social network context. On the other hand, (Papapicco et al., 2022) researched how adolescents show confidence in spotting fake news, but often fail to detect or remember racial hoaxes.

According to Ahmed et al. (2022), the integration of social network data contributes to improved performance and equity in classification systems. The dearth of developed critical thinking skills exposes teenagers to greater deception. Adolescents often feel immune to fake news despite being unable to identify or remember it (Papapicco et al., 2022). The Biradar et al. (2024b) introduced a novel dataset with Hindi-English code-mixed dataset of hateful comments, aiming to explore the link between fake narratives and hate severity. It is problematic to identify hate speech in code-switched Languages such as Hinglish because of its intricacy. Some social network analyses have applied a focus on features of social media such as usual name-calling but study of bias elimination and diversity linguistics is scant (Chopra et al., 2020).

3 Dataset & Task Overview

We have utilized the abusive detection dataset from the LT-EDI@LDK 2025 shared task (Chakravarthi et al., 2025). This research makes use of a code-mixed Hindi-English dataset meant for detecting racial hoaxes in social media posts. The dataset is publicly available in Chakravarthi (2020) research. We used their dataset to train, test and evaluate our model. The dataset is divided into three parts: training, development, and test. The test set does not have labels. Our models predicted the labels for that set. The dataset has two 'Non-Hoax' (0) and 'Racial Hoax' (1) annotations per sample. We have training (3060 samples), validation (1021 samples), and test (1021 samples). Around each text sample, there is a word count of 29-30. The dataset suffers from class imbalance (76% - 24%) as the Non-Hoax class dominates strongly over the Racial Hoax class.

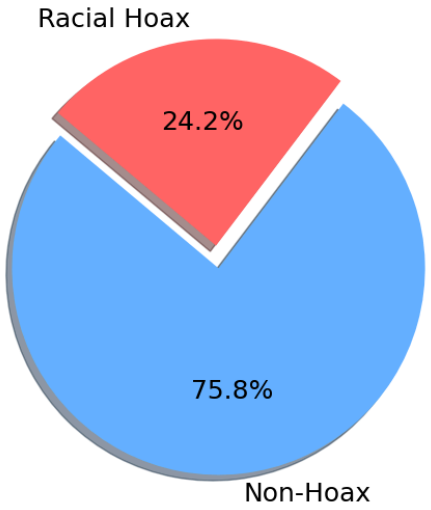

Figure 1: Training Data Distribution

Split	Non-Hoax (0)	Racial Hoax (1)
Train	2319	741
Validation	774	247
Test	774	247

Table 1: Class-wise distribution across dataset splits.

The distribution has been demonstrated in Table 1. To handling the class imbalance problem, we used different preprocessing techniques. The details procedures have been demonstrated in the Methodology section.

4 Methodology

This section provides an overview of the methodology and approach that have been used to build the system using the previously described dataset and transformer model. The methodology of our work

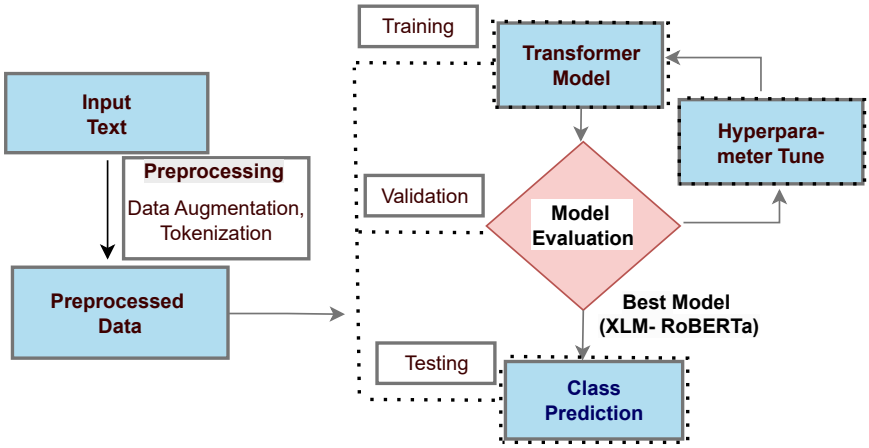

Figure 2: Methodology of our work

is shown in Figure 2.

4.1 Preprocessing

In our approach, preprocessing has focused on handling the unique challenges of code-mixed Hindi-English. The dataset is evaluated to determine its distribution and structure. The labels are encoded as containing Racial Hoax: 1, Not Racial Hoax: 0. The provided dataset was label encoded. We began by the tokenization and encoding the text using the AutoTokenizer from the HuggingFace library. For this task we used hing-roberta for Hindi-English code-mixed language. Text augmentation was applied during training, with a 10% random word masking. Additionally, the dataset was split into training and validation sets with stratified sampling to ensure balanced class distribution. The dataset is divided into training and validation sets using an 80%- 20% ratio.

4.2 Model Architecture and Training

For model selection, we have chosen XLM-RoBERTa, a multilingual transformer-based model, to capture the diverse nature of Hindi-English code-mixed data. We utilized XLM-R as the base model due to its proven effectiveness in multilingual tasks and its strong performance in low-resource languages. Prior work in deception detection, like Biradar et al. (2024a), has used similar transformer-based models successfully. Moreover, XLM-R's ability to capture cross-lingual semantic nuances makes it well-suited for hoax speech detection, which often relies on code-mixing.

Then, the provided dataset is converted into the Hugging Face Dataset format. The model architecture was improved with a custom classifier head that features dropout, layer normalization, and ReLU activation. Initially, all layers of the model were frozen. During the training period, we applied gradual unfreezing, starting with the last two layers. We optimized the model using AdamW with differential learning rates. Early stopping is implemented to prevent overfitting by monitoring validation loss.

4.3 Evaluation and Testing

During model evaluation, we assessed performance using the new dataset for development to fine-tune hyperparameters and ensure optimal performance. Once the model had achieved satisfactory results, we proceeded with the test dataset for the final classification. We have utilized the test dataset that has been provided by the shared task competition, which contains unlabeled comments, to classify racial hoax and non-racial hoax comments. The trained model predicts the labels, distinguishing between racial hoax and non-racial hoax content. This ensured the model's ability to generalize effectively to unseen data.

5 Result and Error Analysis

In this section, we have compared the results and analyzed the different transformer's performance based on the evaluation metrics. The macro F1-score measures the supremacy of the models. Table-3 shows the evaluation metrics for our best model.

5.1 Parameter Setting

We have tuned different hyperparameters to find the corresponding transformer's best model. The

Hyperparameter	Value
Learning Rate	5e-4 (Max)
Batch Size	16
Epochs	10
Dropout	0.3
Weight Decay	0.01
Masking Prob.	0.1
Optimizer	Adam
Scheduler	OneCycleLR
Early Stopping	Patience= 03 epochs

Table 2: Hyperparameters of the model

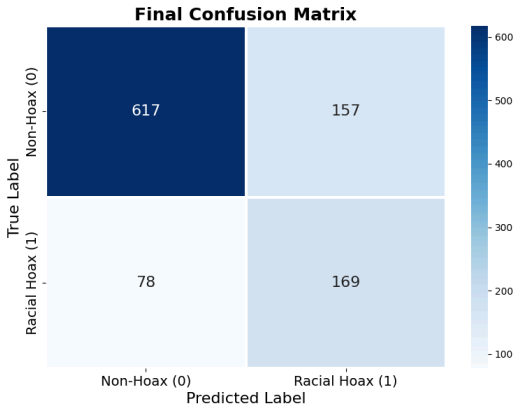

Figure 3: Confusion matrix of our transformer model

hyperparameters that are used in our model are shown in Table 2.

5.2 Metrics Evaluation

The performance of different models are evaluated by various metrics such as F1-score, Accuracy, Precision, and Recall on the test (Provided dev dataset) set. This ensured the model's ability to generalize effectively to unseen data. The evaluation metrics of our best model (XLM- RoBERTa) are shown in Table- 3. Figure- 3 demonstrates the confusion matrics of our model. Both of Table-3 and Figure 3 are based on the validation dataset. So we have an f1 score of 0.81 for the validation dataset, where the test data (unseen data) has an f1 score of 0.75.

5.3 Error Analysis

An analysis of the dataset splits has revealed a consistent class imbalance across the training, validation, and test sets. In each split, approximately 75.8% of the samples belong to the Non-Hoax class, while only 24.2% correspond to the Racial Hoax

Evaluations	Value
F1-Score	0.81
Val. Loss	0.183
Accuracy	85.57%
Precision	0.87
Recall	0.75

Table 3: Evaluation Metrics on the validation set

class. That's why the result is skewed to the non-racial hoax due to class imbalance problem. After analyzing the error, we found that the racial hoax containing text is misclassified as a non-racial hoax in some cases.

The confusion metrics show them well. These are due to the language morphology and lexical ambiguity, sarcasm, and irony when the context of the sentence is ambiguous. Rare words or dialects may also be another reason for these misclassifications. The evaluation metrics of our best model for corresponding languages are shown in Table 3. Incorporating additional context using hierarchical models could help in better understanding the context. Fine-tuning multilingual transformers in domain-specific corpora may also improve performance.

6 Conclusion

In this study, we proposed a transformer-based classification pipeline for detecting racial hoaxes in code-mixed Hindi-English social media content. Our approach incorporated robust preprocessing, Hinglish-specific tokenization, and fine-tuned multilingual models such as XLM-RoBERTa. Due to the problem of class imbalance in the dataset, we placed additional emphasis on preprocessing, incorporating techniques such as oversampling, data augmentation to mitigate the skew and enhance model generalization. That is why the experimental results demonstrated the effectiveness of our methodology, achieving strong performance on the LT-EDI@LDK-2025 shared task. Among all international participants, our system secured first place in the shared task with a satisfactory F1-score of 0.75, demonstrating the effectiveness of our method in addressing racially manipulative content in multilingual contexts. These findings highlight the importance of addressing racial disinformation in multilingual online spaces using deep learning.

Limitations

While our approach demonstrates better performance, it has certain limitations also. First of all, the provided dataset is quite small and class imbalance problem. The impact of the dataset on model development is visible in the result and error analysis section. The class imbalance problem skews the expected output to non-racial hoax class. Improving the dataset volume and more sample for 'Racial Hoax' class, better output can be expected. Secondly, our model shows limitations in capturing the sarcasm, irony, or implicit abusive content. As these are low resources languages and due to their native morphology, capturing the context is challenging.

References

Sadia Afroz, Michael Brennan, and Rachel Greenstadt. 2012. Detecting hoaxes, frauds, and deception in writing style online. In *2012 IEEE symposium on security and privacy*, pages 461–475. IEEE.

Zo Ahmed, Bertie Vidgen, and Scott A Hale. 2022. Tackling racial bias in automated online hate detection: Towards fair and accurate detection of hateful users with geometric deep learning. *EPJ Data Science*, 11(1):8.

Luis Joaquín Arellano, Hugo Jair Escalante, Luis Villaseñor Pineda, Manuel Montes y Gómez, and Fernando Sanchez-Vega. 2022. Overview of da-vincis at iberlef 2022: Detection of aggressive and violent incidents from social media in spanish.

Kim Barker and Olga Jurasz. 2021. Text-based (sexual) abuse and online violence against women: Toward law reform? In *The Emerald International Handbook of Technology-Facilitated Violence and Abuse*, pages 247–264. Emerald Publishing Limited.

Shankar Biradar, Kasu Sai Kartheek Reddy, Sunil Saumya, and Md. Shad Akhtar. 2024a. Proceedings of the 21st international conference on natural language processing (ICON): Shared task on decoding fake narratives in spreading hateful stories (faux-hate). In *Proceedings of the 21st International Conference on Natural Language Processing (ICON): Shared Task on Decoding Fake Narratives in Spreading Hateful Stories (Faux-Hate)*, pages 1–5, AU-KBC Research Centre, Chennai, India. NLP Association of India (NLPAI).

Shankar Biradar, Sunil Saumya, and Arun Chauhan. 2024b. Faux hate: unravelling the web of fake narratives in spreading hateful stories: a multi-label and multi-class dataset in cross-lingual hindi-english code-mixed text. *Language Resources and Evaluation*, pages 1–32.

Tom Bourgeade, Alessandra Teresa Cignarella, Simona Frenda, Mario Laurent, Wolfgang Schmeisser-Nieto, Farah Benamara, Cristina Bosco, Véronique Moriceau, Viviana Patti, and Mariona Taulé. 2023. A multilingual dataset of racial stereotypes in social media conversational threads. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 686–696.

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).

Shivang Chopra, Ramit Sawhney, Puneet Mathur, and Rajiv Ratn Shah. 2020. Hindi-english hate speech detection: Author profiling, debiasing, and practical perspectives. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 386–393.

Concetta Papapicco, Isabella Lamanna, and Francesca D'Errico. 2022. Adolescents' vulnerability to fake news and to racial hoaxes: A qualitative analysis on italian sample. *Multimodal Technologies and Interaction*, 6(3):20.

Md Mizanur Rahman, Srijita Dhar, Md Mehedi Hasan, and Hasan Murad. 2025. MSM_CUET@DravidianLangTech 2025: XLM-BERT and MuRIL based transformer models for detection of abusive Tamil and Malayalam text targeting women on social media. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 243–247, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.