KEC-Elite-Analysts@LT-EDI 2025: Leveraging Deep Learning for Racial Hoax Detection in Code-Mixed Hindi-English Tweets

Malliga Subramanian[1], Aruna A [1], Amudhavan M[1], Jahaganapathi S[1], Kogilavani S V[1]

[1]*Kongu Engineering College, Erode, Tamil Nadu, India*

Abstract

Detecting misinformation in code-mixed languages, particularly Hindi-English, presents a challenge in natural language processing (NLP) (Nayak and Joshi, 2021) due to linguistic diversity on social media. This paper addresses racial hoax detection—false narratives targeting communities—in Hindi-English YouTube comments. We evaluate Logistic Regression, Random Forest, SVM, Naive Bayes, and MLP models on the HoaxMixPlus dataset from LT-EDI@LDK 2025, containing 5,105 annotated comments. Performance is measured using accuracy, precision, recall, and F1-score. Results show that neural and ensemble models outperform traditional classifiers. Future work will explore transformer models and data augmentation for improved detection in low-resource, code-mixed contexts.

1 Introduction

Racial hoax detection in NLP focuses on identifying false narratives that target specific communities. The rise of social media has intensified this issue, particularly in Hindi-English code-mixed text, where language switching and informal usage are common (Yadav et al., 2024). Traditional misinformation detection models face challenges with such multilingual, low-resource data. This study addresses the problem using the HoaxMixPlus dataset from the LT-EDI@LDK 2025 Shared Task on Racial Hoaxes. We explore machine learning approaches including Logistic Regression, SVM, Random Forest, Naive Bayes, and MLP. Experimental results highlight the effectiveness of ensemble and neural models in identifying racial hoaxes, contributing to safer online discourse in multilingual spaces.

2 Literature Survey

Racial hoax detection in Hindi-English code-mixed social media text presents unique challenges due to informal language, code-switching, and sociocultural sensitivity (Kapil and Ekbal, 2024). Earlier approaches using traditional machine learning models like Logistic Regression and SVM showed limitations in capturing the nuanced context and implicit bias present in hoax-related content. These models often struggled with language ambiguity and inconsistent grammar patterns in code-mixed data. Neural models such as MLPs and ensemble-based methods like Random Forest improved performance by learning better feature representations. However, detecting racially motivated misinformation still remains difficult due to lack of annotated datasets and subtle narrative framing. The shared task at LT-EDI@LDK 2025 introduced a benchmark dataset to address these gaps, encouraging research focused on robust detection strategies for code-mixed racial hoaxes. The overview paper presents the task setup, dataset characteristics, evaluation metrics, and a comparative analysis (Chakravarthi et al., 2025) of the approaches adopted by participating systems.

2.1 Racial Hoax Detection in Code-Mixed Text

Detecting racial hoaxes in Hindi-English code-mixed social media text is a complex task due to informal grammar, transliterations, and culturally embedded expressions (Vetagiri and Pakray, 2024). Code-mixing, where Hindi and English words are used interchangeably within a single sentence, creates additional linguistic ambiguity. Traditional methods such as rule-based filtering and basic keyword spotting fail to capture implicit narratives that spread misinformation. Machine learning models like Logistic Regression and Naive Bayes offer baseline performance but struggle with the context sensitivity required to identify hoaxes that often rely on insinuation, bias, or fabricated claims.

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 111–115

2.2 Machine Learning Approaches for Hoax Identification

Supervised learning models have been widely used for hoax and misinformation detection tasks, especially when annotated datasets are available. Models like Support Vector Machines (SVM), Random Forests, and Multi-Layer Perceptrons (MLP) are capable of learning patterns in text based on features like word frequencies, n-grams, and TF-IDF values. In the context of Hindi-English code-mixed text, these models can differentiate between hoax and non-hoax content to some extent, but they often miss deeper contextual and sociolinguistic cues (Bohra et al., 2018). Their performance is also affected by the dataset's imbalance and the informal nature of user-generated content.

2.3 Deep Learning and Contextual Modeling for Hoax Detection

Neural network-based methods, particularly MLPs, provide a significant advantage over traditional classifiers by automatically learning non-linear feature representations. However, without the use of contextual embeddings or attention mechanisms, even these models may struggle with subtle cues in hoax content. While transformer-based models are not explored in the current scope, they represent a promising direction for capturing the deeper semantic context in code-mixed racial hoax detection, particularly through transfer learning and fine-tuning (Farooqi et al., 2021) on domain-specific data.

2.4 Challenges and Future Directions in Racial Hoax Detection

The detection of racial hoaxes in code-mixed content faces several challenges, including lack of large-scale annotated datasets, underrepresentation of minority viewpoints, and subtle linguistic markers of bias (Ariza-Casabona et al., 2024). Comments that contain hoaxes may appear neutral on the surface but embed stereotypes or false attributions. Future work in this domain should focus on leveraging external knowledge sources such as hate speech lexicons and social context signals. There is also a need to explore transformer-based models that can capture deeper semantic meaning, while addressing data scarcity through transfer learning and augmentation techniques tailored for code-mixed languages.

3 Materials and Methods

This study focuses on identifying racial hoaxes in Hindi-English code-mixed social media text. The dataset used, HoaxMixPlus, comprises 3,060 annotated comments from YouTube. Racial hoaxes are challenging to detect due to implicit stereotyping, multilinguality, and informal language use (Not specified, 2021). The dataset is manually annotated, balanced across hoax and non-hoax classes, and preprocessed for modeling. Multiple machine learning models were employed, including Logistic Regression, Random Forest, Support Vector Machine (SVM), Naive Bayes, and Multi-Layer Perceptron (MLP). Performance was evaluated using accuracy, precision, recall, and F1-score.

3.1 Dataset

The dataset used in this study is derived from the HoaxMixPlus corpus, which contains 5,105 YouTube comments in code-mixed Hindi-English, addressing the complex problem of misinformation in low-resource language settings. These comments reflect user opinions on various sociopolitical and cultural contexts, particularly focusing on identity-based misinformation.

3.1.1 Dataset Size and Source

Each entry in the dataset includes two key fields: clean text and label. The clean text field represents a preprocessed code-mixed comment where emojis and punctuations have been removed to reduce noise. The label is a binary indicator, where 1 signifies a racial hoax comments that spread fabricated identity-based narratives targeting individuals or communities and 0 denotes a non-hoax, i.e., neutral or unrelated content.

The label distribution is imbalanced, with a significant number of examples labeled as non hoax, reflecting real-world data skew where harmful misinformation is relatively rare but impactful.

3.2 Preprocessing and Feature Extraction

Preprocessing steps played a vital role in managing the noisy and informal nature of social media text. The raw code-mixed Hindi-English comments were systematically cleaned by removing punctuations, emojis, URLs, and redundant whitespace, thereby standardizing the text structure while preserving the semantic integrity of the content. For feature extraction, two main techniques were employed. The first was CountVectorizer, which transformed the textual data into a bag-of-words representation

by capturing the frequency of each word without accounting for its contextual significance. The second technique was TF-IDF, which measured the importance of words based on their relative frequency across the dataset, effectively reducing the influence of commonly occurring but less informative terms. These structured vector representations were then used as inputs to various machine learning models, enabling them to identify patterns and make accurate predictions in the task of racial hoax detection.

3.3 Models and Methodology

To classify racial hoaxes in Hindi-English code-mixed social media text, we used traditional machine learning models such as Logistic Regression, Random Forest, SVM, Naive Bayes, and MLP. These models were trained using CountVectorizer and TF-IDF features to convert text into numerical form. Model performance was assessed using accuracy, precision, recall, and macro-averaged F1-score, with macro F1 being the focus due to class imbalance.

3.3.1 Hyperparameter Tuning

We tuned hyperparameters for the MLP model by varying learning rate, batch size, and hidden units. The best performance was achieved with a learning rate of 0.001, batch size of 32, and 128 hidden units, reaching 84.7% accuracy. Other configurations showed slightly lower performance, emphasizing the importance of proper tuning.

3.3.2 Preprocessing Impact

An ablation study was conducted to assess preprocessing steps. Without preprocessing, accuracy was 74.2%. Lowercasing and stopword removal gradually improved results. Applying TF-IDF significantly boosted accuracy to 81.2%, and using all steps, including stratified sampling, led to the highest accuracy of 84.7%. These results show that preprocessing plays a crucial role in effective classification.

4 Results and Discussion

This study on racial hoax detection in Hindi-English code-mixed text demonstrated that traditional machine learning models, particularly the Multi-Layer Perceptron (MLP), performed effectively when paired with proper preprocessing and feature extraction techniques. Compared to simpler models like Naive Bayes and Logistic Regression, MLP consistently achieved better accuracy due to its capacity to learn complex patterns. The best performance was observed with TF-IDF features and optimized hyperparameters, yielding an accuracy in the range of 78–80

Evaluation metrics including precision, recall, and macro-averaged F1-score showed that while simpler models could identify obvious hoaxes, they often misclassified nuanced or indirect expressions. MLP, in contrast, handled these challenges better, demonstrating the value of deep, feedforward architectures even in limited-resource, code-mixed scenarios.

4.1 Error Analysis

Understanding model limitations was key to evaluating its robustness. Through manual review of misclassified samples, several recurring issues were identified.

4.1.1 Common Misclassification Patterns

The model struggled with ambiguous or sarcastic expressions, especially when racial hoaxes were implied subtly. It often misclassified sarcastic or ironic statements as genuine due to the absence of explicit hate-related cues. Additionally, inconsistent code-switching between Hindi and English complicated contextual understanding. Comments with negations or indirect racial insinuations were also frequently misinterpreted.

4.1.2 Strategies to Address Misclassifications

To reduce misclassification, incorporating sarcasm detection and contextual sentiment cues into the model could improve accuracy. Using transformer-based architectures like mBERT or IndicBERT, trained on code-mixed data, would provide better contextual embeddings. Further, multi-label classification might help in handling complex or overlapping categories such as satirical hoaxes.

4.2 Discussion

The experimental results provided insight into the behavior and limitations of classical models for detecting racial hoaxes in code-mixed social media data. Hyperparameter tuning significantly affected model performance, as did text preprocessing steps like lowercasing, stopword removal, and TF-IDF transformation. Among the models tested, MLP with TF-IDF achieved the highest performance, with other models such as SVM and Logistic Regression trailing behind in accuracy and F1-score.

4.2.1 Computational Efficiency for Real-World Use

The MLP model showed reasonable computational efficiency for practical applications. It processed approximately 1100 tokens per second, making it viable for deployment in real-time monitoring tools on platforms like Twitter or Facebook.

Table 1: Computational Efficiency Analysis

Model	Inference Time (ms)	Memory (GB)	Tokens/sec
Naive Bayes	60	2.1	1300
Logistic Reg.	75	2.8	1200
SVM	100	3.5	1000
MLP	95	4.5	1100

4.2.2 Future Work

Future work will explore transformer-based models like DistilBERT or IndicBERT to further improve detection of complex and sarcastic racial content. Expanding the dataset to include more diverse linguistic patterns and code-switching examples will enhance generalization. Additionally, building a lightweight web-based dashboard or API would support real-time detection of racial hoaxes for social media analysts and researchers.

4.2.3 Model Performance

The best-performing MLP (Multilayer Perceptron) model achieved an impressive accuracy of approximately 80%, with a macro F1-score of 78%, precision of 76%, and recall of 77%. These results demonstrate the model's capability to handle the complexities of code-mixed data, offering a balanced and effective solution for sentiment classification tasks. The MLP's deeper architecture enabled it to better capture intricate patterns and contextual shifts present in the mixed-language data, ensuring robust performance across multiple evaluation metrics.

In comparison, Logistic Regression and SVM recorded slightly lower accuracies of 72% and 70%, respectively. While these models performed well as baselines, they were unable to match the more advanced MLP in terms of overall performance. However, they still provide useful alternatives in scenarios where model simplicity and interpretability are more important than the highest possible accuracy.

The Naive Bayes model, with an accuracy of 65%, showed significant limitations in handling the complexities of code-mixed data. Although Naive Bayes is efficient and easy to implement, it struggled to effectively capture the nuanced relationships within the mixed-language content. These findings underscore the importance of using deeper, more advanced models with appropriate preprocessing to achieve better results in code-mixed classification tasks.

Table 2: Model Performance

Model	Precision (%)	Recall (%)	F1Score (%)	Accuracy (%)
Naive Bayes	70	65	67	65
Logistic Reg.	75	72	73	72
SVM	78	70	74	70
MLP	76	77	78	80

5 Conclusion

This study addressed the challenge of detecting racial hoaxes in code-mixed Hindi-English social media content using the dataset, a collection of 5,105 annotated YouTube comments. We evaluated traditional machine learning models Logistic Regression, Random Forest, SVM, and Naive Bayes alongside a deep learning MLP model, which achieved the highest performance by effectively capturing subtle identity based misinformation patterns.

The results highlight the importance of tailored approaches for low-resource, code-mixed data where misinformation can have serious social implications. Future work will focus on expanding the dataset, incorporating transformer-based models, and optimizing hybrid architectures for improved performance.

Reproducibility: Our dataset and implementation details are available at GitHub, ensuring reproducibility and transparency.

References

A. Ariza-Casabona, W. S. Schmeisser-Nieto, M. Nofre, M. Taulé, E. Amigó, B. Chulvi, and P. Rosso. 2024. Stereohoax: A multilingual corpus of racial hoaxes and social media reactions annotated for stereotypes. *Language Resources and Evaluation*.

Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. A dataset of hindi-english code-mixed social media text for hate speech detection. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*.

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).

Zaki Mustafa Farooqi, Sreyan Ghosh, and Rajiv Ratn Shah. 2021. Leveraging transformers for hate speech detection in conversational code-mixed tweets. *arXiv preprint arXiv:2112.09986*.

Prashant Kapil and Asif Ekbal. 2024. A corpus of hindi-english code-mixed posts for hate speech detection. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*.

Ravindra Nayak and Raviraj Joshi. 2021. Contextual hate speech detection in code-mixed text using transformer-based approaches. *arXiv preprint arXiv:2110.09338*.

Not specified. 2021. Online multilingual hate speech detection: Experimenting with hindi and english social media. *Information*, 12(1):5.

Advaitha Vetagiri and Partha Pakray. 2024. Detecting hate speech and fake narratives in code-mixed hinglish social media text. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*.

Anjali Yadav, Tanya Garg, Matej Klemen, Matej Ulcar, Basant Agarwal, and Marko Robnik Sikonja. 2024. Code-mixed sentiment and hate-speech prediction. *arXiv preprint arXiv:2405.12929*.