Hoax Terminators@LT-EDI 2025: CharBERT's dominance over LLM Models in the Detection of Racial Hoaxes in Code-Mixed Hindi-English Social Media Data

Abrar Hafiz Rabbani, Diganta Das Droba, Momtazul Arefin Labib
Samia Rahman, Hasan Murad
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology (CUET), Bangladesh
{u2004038, u2004064, u1904111}@student.cuet.ac.bd
u1904022@student.cuet.ac.bd , hasanmurad@cuet.ac.bd

Abstract

This paper presents our system for the LT-EDI 2025 Shared Task on Racial Hoax Detection, addressing the critical challenge of identifying racially charged misinformation in code-mixed Hindi-English (Hinglish) social media—a low-resource, linguistically complex domain with real-world impact. We adopt a two-pronged strategy, independently fine-tuning a transformer-based model and a large language model. CharBERT was optimized using Optuna, while XLM-RoBERTa and DistilBERT were fine-tuned for the classification task. FLAN-T5-base was fine-tuned with SMOTE-based oversampling, semantic-preserving back translation, and prompt engineering, whereas LLaMA was used solely for inference. Our preprocessing included Hinglish-specific normalization, noise reduction, sentiment-aware corrections and a custom weighted loss to emphasize the minority Hoax class. Despite using FLAN-T5-base due to resource limits, our models performed well. CharBERT achieved a macro F1 of 0.70 and FLAN-T5 followed at 0.69, both outperforming baselines like DistilBERT and LLaMA-3.2-1B. Our submission ranked 4th of 11 teams, underscoring the promise of our approach for scalable misinformation detection in code-switched contexts. Future work will explore larger LLMs, adversarial training and context-aware decoding.

1 Introduction

Racial hoaxes are harmful lies that falsely tie people or groups to crimes or events, often picking on specific ethnic or social communities to spark division or fear. The rise of hateful, racially charged speech—especially on platforms like Twitter and Facebook where users often blend languages like Hinglish (a mix of Hindi and English)—poses a serious challenge. The HoaxMixPlus dataset, consisting of 5,105 YouTube comments, serves as a key benchmark for detecting such harmful content.

Detecting racial hoaxes on social networks is challenging due to the difficulty in distinguishing truth from falsehood, the sheer volume of posts, and the intentional use of humor by users (Santoso et al., 2017). Traditional systems often misclassify content due to idioms, slang and subtle contextual cues. To address this, we leverage advanced models like the transformer-based CharBERT and LLM-based FLAN-T5[1], fine-tuned with task-specific instructions, rubrics, and prompt formulations. CharBERT's character-level embeddings and FLAN-T5's contextual understanding make them well-suited for interpreting nuanced, deceptive content.

This paper presents our submission to the Racial Hoax Detection Shared Task—a robust system leveraging transformers and LLMs, structured around three contributions:

- **Augmented Training for LLM and Transformer Models:** We trained the FLAN-T5 model on an augmented dataset generated using semantic-preserving back translation and SMOTE to mitigate class imbalance and enhance generalization in the low-resource setting. For the transformer-based CharBERT model, we applied an oversampling strategy to address class imbalance.
- **Class-sensitive training:** Introduced a weighted loss function in FLAN-T5 to increase sensitivity towards minority hoax instances and improve model fairness.
- **Transformer optimization:** To attain the best classification performance, the CharBERT model's hyperparameters were tuned using Optuna, an open-source framework for hyperparameter optimization.

Our approach demonstrates promising results in detecting racial hoaxes on code-switched social

[1] https://huggingface.co/google/flan-t5-base

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 159–170

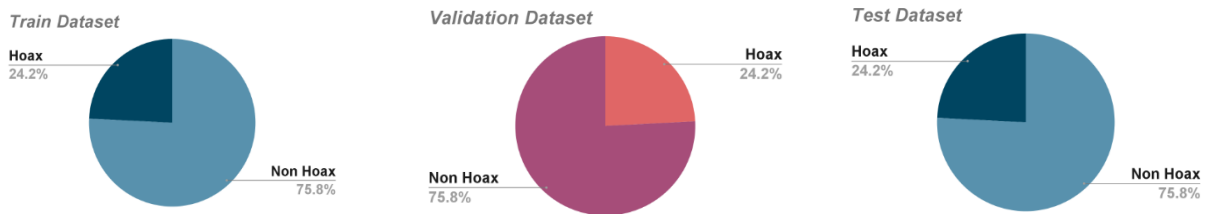

Figure 1: Dataset Distribution of Racial Hoax Dataset.

media platforms, highlighting the effectiveness of utilizing advanced transformer and LLM models alongside innovative data augmentation techniques like back translation, SMOTE and oversampling to tackle class imbalance and improve model generalization. For more details and to access the codebase, visit our project repository: GitHub Repository[2].

2 Related Work

In recent years, the detection of racial stereotypes and hoaxes in social media has become a critical focus of research. (Bosco et al., 2023) introduced a method for detecting racial stereotypes in Italian social media, focusing on the intersection of psychology and natural language processing (NLP). Building on these studies, (Schmeisser-Nieto et al., 2024) presented Stereohoax, a multilingual corpus annotated for racial hoaxes and stereotypes. Their work addresses a significant gap in understanding social media reactions to racial hoaxes and offers a valuable resource for future research. (Raza et al., 2024) explored the effectiveness of BERT-like models and large language models (LLMs) in the detection of fake news, focusing on the challenges posed by generative AI-annotated data. Their comparative evaluation provides a useful perspective on how different models perform in fake news and racial hoax detection tasks. (Banerjee et al., 2021) investigated transformer-based models for identifying hate speech and offensive content in both English and Indo-Aryan languages. Their work highlights the effectiveness of transformer models in multilingual environments, particularly in identifying harmful content in social media posts. (Guo et al., 2024) conducted a large-scale study on the use of LLMs for hate speech detection, focusing on the role of prompt engineering in improving the models' contextual understanding. Their findings suggest that LLMs can surpass traditional machine learning models in detecting hate speech when properly prompted. Recent work by (Carpenter et al., 2024) shows the effectiveness of fine-tuned FLAN-T5 models in educational settings, supporting our choice of FLAN-T5 for detecting racial hoaxes in code-mixed Hinglish. To address the issue of class imbalance, which is particularly critical in hoax detection where hoax instances are typically underrepresented, we draw inspiration from the dynamically weighted balanced (DWB) loss function proposed by (Fernando and Tsokos, 2021), which adaptively adjusts loss contributions based on class frequency and prediction confidence, enabling the model to focus on harder minority-class samples and improving generalization in imbalanced settings. Additionally, (Chakravarthi et al., 2025) presented an overview of the shared task on detecting racial hoaxes in code-mixed Hindi-English social media data, further advancing the understanding of racial hoaxes in multilingual contexts.

3 Dataset Description

The dataset (**?**) used in this shared task targets the challenge of racial hoax detection in Hinglish social media posts. It comprises real-world, user-generated content labeled with binary annotations: a label of 1 signifies the presence of a racial hoax, while 0 indicates a non-hoax instance. However, the dataset is notably imbalanced, with a significantly higher number of non-hoax examples. As illustrated in the pie charts of Figure 1, the training set contains 2,319 non-hoax cases versus only 741 hoax cases. The validation and test sets each contain 774 non-hoax and 247 hoax samples, creating an imbalance that can trip up standard classification models. These models often lean toward the majority class, making it tough to properly learn from the smaller hoax class. To tackle this, we applied methods like oversampling, SMOTE and loss function adjustments to better emphasize the minority class and enhance the model's effective-

[2] https://github.com/abrar-431/racial-hoax-detection-shared-task

ness.

4 System

In this section, we describe the methodologies employed for detecting racial hoaxes in social media content using two distinct approaches: Transformer Models and Large Language Models (LLMs).

4.1 Transformer Based Approach

Three transformer-based models were used in this study with the detailed illustrated in Figure 2 to detect racial hoaxes in Hinglish social media content.

CharBERT is used here which uses character-level embeddings to capture morphological subtleties and spelling variations in languages such as Hinglish which allow it to handle code-mixed, informal, and noisy text. This model works especially well for picking up on minute details in non-standard language usage, like that found in posts on social media [3].

DistilBERT is also used here which is a smaller and faster version of BERT. It was 60% faster and required fewer parameters while maintaining 97% of BERT's performance. This makes it perfect for real-time applications in large datasets.[4].

Finally, to address the multilingual nature of Hinglish, XLM-RoBERTa, a cross-lingual version of RoBERTa, was employed. It is proficient in understanding the contextual relationships between words in Hindi and English, having been trained on data from 100 languages. This makes it useful for cross-lingual tasks [5].

4.1.1 Data Preprocessing

The CharBERT tokenizer was used to perform tokenization because it is made to efficiently process the input data, . All characters were changed to lowercase and special characters, links, and unnecessary symbols were removed to normalize the text data. We also used oversampling techniques to address class imbalance and guarantee a balanced distribution of racial hoaxes and non-hoaxes in the training dataset. To ensure that both classes were equally represented, we resampled the dataset using RandomOverSampler [6] from the imbalanced-learn library.

[3] https://huggingface.co/imvladikon/charbert-bert-wiki
[4] https://huggingface.co/distilbert/distilbert-base-uncased
[5] https://huggingface.co/FacebookAI/xlm-roberta-base
[6] https://riverml.xyz/dev/api/imblearn/RandomOverSampler/

4.1.2 Model Architecture

The model used for this study is CharBERT, for which we have the best performance. It is a variant of the well-known BERT (Bidirectional Encoder Representations from Transformers) architecture.

- **Embedding Layer:** In CharBERT, word-level and character-level embeddings are included. The input text is first tokenized into subword units by the model. It then transforms each subword token into a dense representation that contains both the character-level embedding (which captures the morphology of words, including slang, informal language, and spelling variations) and the word-level embedding (from pre-trained word embeddings).

- **Character-Level Processing:** The emphasis on character-level information of CharBERT is the primary distinction between it and conventional BERT models. In order to identify subtle patterns in the text, such as misspellings, colloquial abbreviations, and new terms frequently found in code-mixed languages or social networks, CharBERT employs a convolutional layer.

- **Output Layer:** CharBERT model generates predictions for classification tasks by overlaying the transformer encoder with a dense output layer. Usually, a sigmoid activation function is used for binary classification tasks, or a softmax activation function for multi-class classification. To differentiate between racial hoaxes and non-hoaxes, CharBERT was optimized for binary classification in our situation.

- **Optimization:** The CharBERT model is adjusted using a cross-entropy loss function during training. To improve convergence, the Adam optimizer is used in conjunction with the learning rate scheduling to dynamically modify the learning rate.

4.1.3 Hyperparameter Optimization

To maximize the performance of transformer models, we employed an open-source hyperparameter optimization framework named Optuna [7]. A hyperparameter search space was established for learning rate, batch size, and the number of training

[7] https://optuna.org/

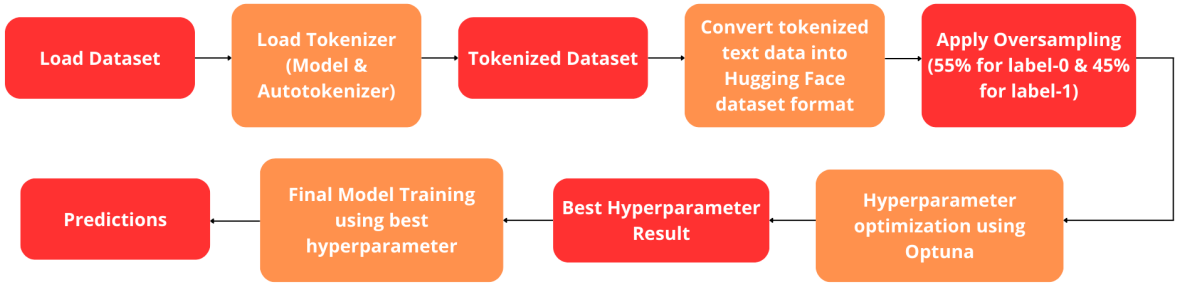

Figure 2: System flow for the Transformer-based approach.

epochs and weight decay. Maximizing the F1 score on the validation dataset served as the optimization's compass. The best model was chosen based on the evaluation outcomes of this optimization process.

4.1.4 Training

Trainer class from transformer's library was used for the training, and the recommended hyperparameters from Optuna were used. The F1 score, accuracy, precision, and recall metrics were used to monitor the model's performance, and early stopping callback was used to avoid overfitting.

4.1.5 Evaluation

The test dataset was used to assess the model following training. Performance metrics like accuracy, precision, recall, and F1 score were calculated by comparing the predictions with the true labels. The model's performance in both classes was thoroughly examined using the classification report.

4.2 Large Language Models (LLMs)

For the Large Language Models (LLMs), we employed the FLAN-T5-Base model as the primary model for racial hoax detection, with the detailed flow illustrated in Figure 3. Additionally, we utilized the Llama-3.2-1B[8] for inference to explore its capabilities in generating contextual responses. However, we primarily focused on FLAN-T5 due to its superior performance in handling code-switched text, better fine-tuning efficiency on our augmented dataset and robust generalization across diverse linguistic patterns, which were critical for detecting racial hoaxes effectively in our social media dataset.

[8] https://huggingface.co/meta-llama/Llama-3.2-1B

4.2.1 Data Augmentation

To enhance model generalization and address the significant class imbalance in the HoaxMixPlus dataset (75.8% Non-Hoax vs. 24.2% Hoax), we employed two complementary data augmentation techniques.

- **Back translation:** Back translation is a method used to generate new paraphrased samples by translating text to another language and then back to the original language. This process preserves the original meaning while altering the wording and structure.

  In our approach, we used pre-trained MarianMT models from Helsinki-NLP to translate sentences from English to Hindi[9] and then back from Hindi to English[10]. This augmentation was applied exclusively to samples labeled as Hoax (`label=1`). Using Hugging Face Transformers, we implemented a batched inference pipeline for efficient and consistent translation. The paraphrased sentences were then added to the dataset, effectively doubling the number of Hoax-labeled examples.

  This augmentation improves lexical and syntactic diversity, helping the model generalize better across different ways misinformation can be phrased. We monitored the process using the `tqdm`[11] progress bar and ensured reproducibility by shuffling the final dataset with a fixed random seed.

- **SMOTE:** While back translation introduced linguistic variety, it was not sufficient to fully address the class imbalance. Therefore, we

[9] https://huggingface.co/Helsinki-NLP/opus-mt-en-hi

[10] https://huggingface.co/Helsinki-NLP/opus-mt-hi-en

[11] https://tqdm.github.io/

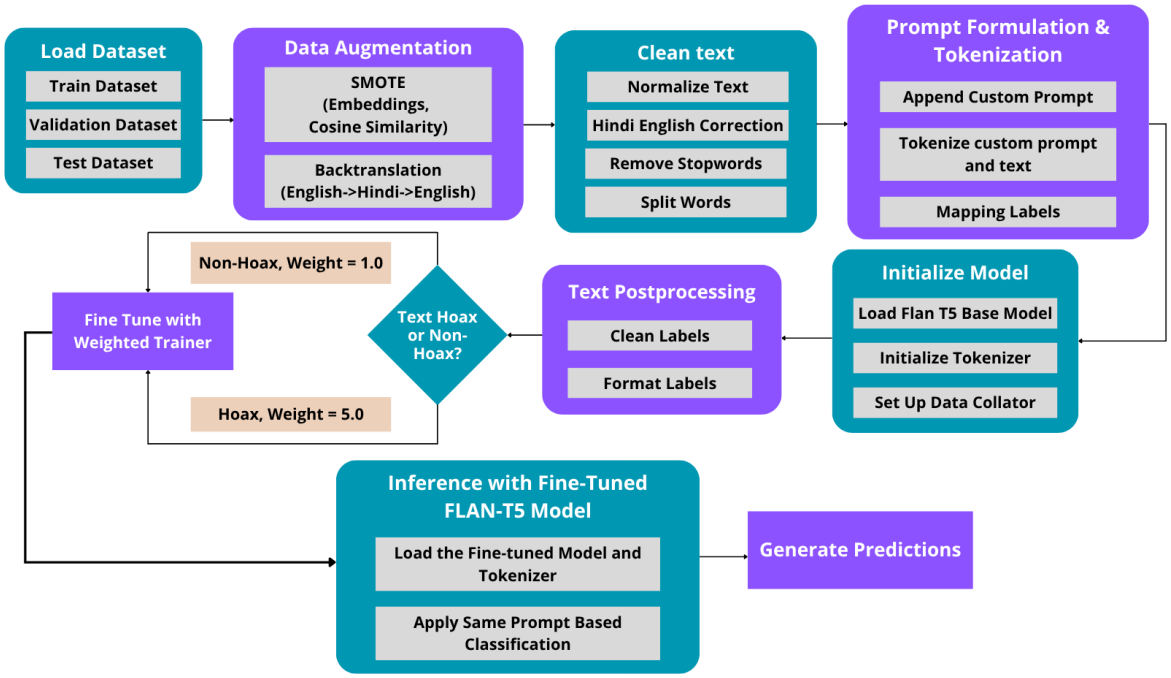

Figure 3: System flow for Fine-Tuning and Inference using the FLAN-T5 model.

also used SMOTE to synthetically generate new samples for the underrepresented Hoax class.

First, we encoded all text samples into vector representations (embeddings) using the SentenceTransformer model `all-MiniLM-L6-v2`[12]. SMOTE was then applied at the embedding level, generating new synthetic vectors by interpolating between existing Hoax samples. We set the oversampling ratio to 1.5 times the number of Non-Hoax samples to ensure balance.

To ensure that the generated vectors represented realistic content, we matched each synthetic vector to its closest original sentence using cosine similarity. This step helped maintain textual coherence in the generated samples.

In our LLM-based approach for detecting racial hoaxes on code-switched social media, we employed both back-translation and SMOTE to address the challenges of limited and imbalanced datasets. Back-translation enriched the dataset by generating diverse, semantically consistent variations of existing samples, preserving linguistic nuances critical for code-switched content. However, it alone was insufficient to handle severe class imbalances, as it primarily enhances sample diversity rather than balancing class distributions. SMOTE complemented this by synthetically generating samples for underrepresented classes, ensuring better representation of minority hoax categories. Using only one technique would have either limited diversity (with SMOTE alone) or failed to address class imbalance (with back-translation alone), compromising model performance. These augmented samples were consolidated into a Hugging Face Dataset, significantly improving class distribution and model robustness, enabling the LLM to generalize effectively across varied and imbalanced real-world scenarios.

4.2.2 Data Preprocessing

To handle the noisy, code-mixed nature of Hinglish social media text, we developed a custom preprocessing pipeline focused on normalization and token quality. We utilized textblob[13] for correcting English word fragments and estimating sentiment polarity where applicable. The pipeline involved lowercasing (while preserving sentiment-relevant punctuation), removal of numbers, URLs, emojis, and special characters. Hinglish-specific corrections were applied using a custom dictionary (e.g., "nhi" to "nahi", "pori" to "puri"), hybrid stopwords

[12] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[13] https://textblob.readthedocs.io/en/dev/

like "bhai" and "yar" were removed, and token fusion was addressed using regular expressions.

4.2.3 Model Architecture and Fine Tuning

We fine-tuned the FLAN-T5-Base model, a text-to-text transformer pretrained for instruction-following tasks, to perform binary classification of Hinglish social media posts into “Hoax” and “Non-Hoax” categories. Additionally, we leveraged the LLaMA-3.2-1B model for inference to evaluate its ability to generate contextual predictions, enhancing our exploration of LLM performance on the same dataset.

4.2.4 Prompt Engineering:

Various prompting strategies were explored—zero-shot, rubric-based, and few-shot (see Appendix A). Inputs followed the format: “Classify: <text>”, and outputs were generated as “0” (Non-Hoax) or “1” (Hoax), aligning with FLAN-T5’s instruction-tuned capabilities. To address severe class imbalance, a custom WeightedTrainer was implemented with a 5:1 loss weighting favoring the Hoax class, improving the model’s sensitivity to minority class instances. For the LLaMA-3.2-1B model, few-shot prompting was employed and proved most effective, leveraging its ability to adapt to contextual examples for improved inference performance.

4.2.5 Custom Weighted Trainer:

To deal with the strong imbalance between Hoax and Non-Hoax examples in the HoaxMixPlus dataset, we created a customized training approach that teaches the model to pay more attention to Hoax cases, which are much fewer in number. In a normal training setup, the model may become biased toward predicting the more frequent class (Non-Hoax) and ignore the less common but more important Hoax instances. To solve this, we made sure that the model gives more importance to correctly identifying Hoax examples by assigning a higher penalty when it gets them wrong. In simple terms, we told the model that misclassifying a Hoax is five times worse than misclassifying a Non-Hoax. This helped the model focus better on the minority class and significantly improved its ability to recognize misinformation, especially in real-world scenarios where such misleading content may appear less frequently but is more critical to detect.

4.2.6 Inference and Output Generation

Finally, predictions were generated using the best-performing checkpoint of the fine-tuned FLAN-T5-Base model, selected based on validation set performance. The test inputs were passed through the model in batches to ensure computational efficiency. Each output sequence generated by the model was decoded using the tokenizer to extract the predicted class labels, constrained to valid outputs—“0” representing Non-Hoax and “1” representing Hoax—to maintain label consistency. Additionally, for the LLaMA-3.2-1B model, inference was conducted using few-shot prompting, leveraging a small set of contextual examples to enhance prediction accuracy on the same dataset. The final, cleaned set of predictions was then compiled and stored in a structured CSV file format, enabling easy access for downstream evaluation, error analysis and comparison with other models. This completed a streamlined and effective end-to-end pipeline, encompassing training, evaluation, and inference.

5 Experiments and Results

In this section, we are presenting the experiments and results of our approaches-Transformer-based and Large Language Models for racial hoax detection in Hinglish social media text.

5.1 Experimental Setup

For the CharBERT model, we evaluated the performance on the HoaxMixPlus dataset, consisting of 5,105 Hinglish YouTube comments. The CharBERT model was fine-tuned for 20 epochs with a learning rate of $2e^{-5}$, weight decay of 0.01, and a batch size of 16. We used the RandomOverSampler technique used for addressing class imbalance to resample the dataset. Gradient accumulation with 4 steps was used to simulate a larger batch size, considering the memory constraints. Early stopping callback with a patience of 5 epochs was applied to prevent overfitting. The performance was monitored by using macro F1-score.

For the LLM (Flan-T5-base) model, we used a Kaggle P100 GPU to fine-tune the model on the same HoaxMixPlus dataset. The Flan-T5-base model was trained with a weighted loss function (5x for the Hoax class) to address class imbalance. Data augmentation techniques included back translation using MarianMT models[14]) and SMOTE.

[14] MarianMT Documentation

The training setup involved a learning rate of $5e^{-5}$, a batch size of 8, and early stopping based on macro F1 to select the best model checkpoint. These settings ensured that both models effectively handled the class imbalance and noisy, code-mixed nature of Hinglish data.

5.2 Parameter Setting

For the Transformer-based models, we initially used the CharBERT model and trained it for 20 epochs with a learning rate of $2e^{-5}$ and a weight decay of 0.01. The optimal learning rate schedule, warm-up ratio, and dropout values were selected automatically using Optuna framework to ensure the best hyperparameter configuration. A batch size of 16 was used, with gradient accumulation steps of 4 to simulate larger effective batch sizes, considering the memory constraints. To avoid overfitting, we applied early stopping callback with a patience of 5 epochs, monitoring the validation performance.

For the Large Language Models (LLMs), the model was trained for up to 10 epochs using the Adam optimizer, with a weight decay of 0.01 and a batch size of 8. The input sequences were truncated to 512 tokens, and early stopping was applied after 3 epochs without improvement in F1 score, ensuring that the best model checkpoint was selected. These settings were tailored to handle the complexity of detecting racial hoaxes in Hinglish social media data.

5.3 Evaluation Metrics

The macro F1 score, which balances precision and recall across both classes and is especially appropriate for our imbalanced binary classification task, was used to evaluate the performance of the Transformer-based models as well as the Large Language Models (LLMs). We provide a comprehensive classification report that includes precision, recall, and class-specific F1 scores in addition to the overall macro F1. This enables us to assess the model's ability to differentiate between hoax and non-hoax instances and identifies any remaining class-level flaws that might compromise the model's generalizability.

By concentrating on the model's advantages and disadvantages in identifying racial hoaxes, this method also aids in identifying differences in class performance in the case of LLM models. By disclosing these metrics, we hope to document the model's

5.4 Comparative Analysis

The performance of various classifiers across different model types is shown in Table 1. The results of our experiments with various transformer-based models and large language models (LLMs) reveal insightful performance trends. Among the transformer-based models, CharBERT achieved the highest macro F1 score of 0.70, alongside the highest accuracy (0.79), demonstrating its effective fine-tuning with Optuna for hyperparameter optimization. The DistilBERT models, both fine-tuned with Optuna and instructions, showed comparable performance with a macro F1 score of 0.66 and a weighted F1 of 0.77, indicating their strong performance despite being smaller variants of BERT. XLM-RoBERTa, another transformer model fine-tuned with Optuna, performed similarly to DistilBERT, with a macro F1 of 0.69 and a weighted F1 of 0.78, emphasizing its robustness for multilingual tasks.

When analyzing the performance of the FLAN-T5-Base and Llama-3.2-1B LLMs, the impact of different prompt variations becomes evident. FLAN-T5-Base, when fine-tuned with zero-shot prompting, achieved a macro F1 of 0.68. However, when fine-tuned with instructions, it showed an improved macro F1 of 0.67. The highest macro F1 score for FLAN-T5-Base was obtained when fine-tuned with Rubric, reaching 0.69. This highlights that different prompt strategies, including Rubric and few-shot prompting, can lead to varying results, with Rubric yielding the most optimal performance. On the other hand, Llama-3.2-1B, when used straight out of the box without any fine-tuning and just run in inference mode, struggled the most, hitting a low macro F1 of only 0.55. This really highlights why fine-tuning matters so much for large language models—tailoring them to specific tasks and carefully crafting prompts can make a huge difference in their performance.

To wrap it up, CharBERT and FLAN-T5-Base were the stars of the show. CharBERT delivered top-notch results with the highest macro F1 score when fine-tuned with Optuna, while FLAN-T5-Base, after being fine-tuned with the Rubric approach, achieved the best performance among the LLMs. This tells us that transformer-based models like CharBERT are strong contenders for this task, but models like FLAN-T5-Base can also excel with the right prompt tuning, especially when guided by strategies like Rubric.

Model Type	Model	Variation	F1 (Macro)	F1 (Weighted)
Transformer	**CharBERT**	**Fine-tuned with Optuna**	**0.70**	**0.78**
	DistilBERT Base	Fine-tuned with instructions	0.66	0.76
	DistilBERT	Fine-tuned with Optuna	0.66	0.77
	XLM-RoBERTa	Fine-tuned with Optuna	0.69	0.78
LLM	FLAN-T5-Base(248M)	Fine tuned with zero shot prompting	0.68	0.76
		Fine tuned with instruction	0.67	0.76
		Fine tuned with Rubric	0.69	0.79
		Fine tuned with Rubric & prompting	0.69	0.77
	Llama-3.2-1B	Inference	0.55	0.67

Table 1: Performance Evaluation of Different Models

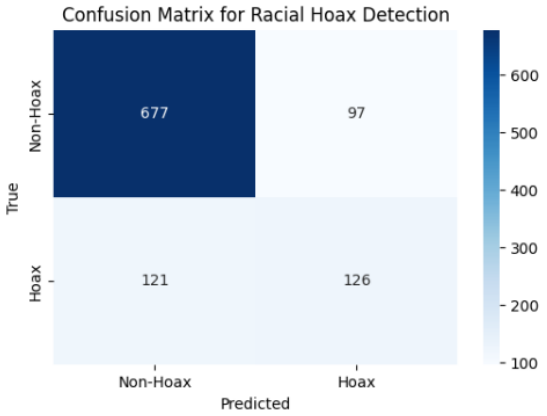

Figure 4: Confusion Matrix for CharBERT Model.

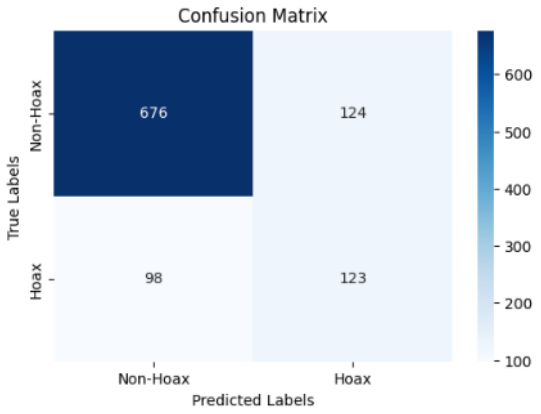

Figure 5: Confusion Matrix for LLM Model.

6 Error Analysis

In the error analysis, we evaluated the performance of the CharBERT model using the confusion matrix for racial hoax detection as shown in Figure 4. The CharBERT model correctly identified 677 non-hoax instances but misclassified 121 hoaxes as non-hoaxes and 126 non-hoaxes as hoaxes. These errors emphasize the need to reduce false negatives for better racial hoax detection. The confusion matrix in Figure 5 shows the confusion matrix with 676 true negatives, 124 false positives, 98 false negatives, and 123 true positives. These results highlight that the model performs well on Non-Hoax instances, with a solid precision of 87%, but struggles with Hoax classification, showing a lower precision of only 50%. This imbalance can be attributed to the class distribution, as Hoax examples are underrepresented in the dataset. The model tends to misclassify sarcastic posts as Non-Hoax (false positives) and hoaxes with neutral phrasing as Non-Hoax (false negatives). Augmentation strategies, such as back translation and SMOTE, helped mitigate some of these errors, but the challenge of distinguishing between subtle context variations remains. Additionally, short and noisy comments presented difficulties due to their inherent context ambiguity, further complicating accurate classification. In addition to the issues identified with sarcasm and neutral phrasing, the CharBERT model also faced difficulties when dealing with informal language and slang, common in Hinglish social media posts. This led to occasional misclassifications, especially when the context was subtle or ambiguous. Despite these challenges, the use of data augmentation techniques, like back translation and SMOTE, improved the model's performance by generating more diverse training examples. However, further improvements in handling noisy and short-text comments, along with enhancing the model's ability to detect nuanced hoaxes, will be necessary to address these shortcomings.

7 Conclusion

We introduced an innovative fine-tuned system for racial hoax detection in code-mixed Hindi-English YouTube comments. Specifically, we fine-tuned the CharBERT and Flan-T5 models on the HoaxMixPlus dataset, leveraging data augmentation techniques (including back translation and SMOTE), weighted loss optimization, and Hinglish-specific preprocessing to address challenges such as class imbalance, linguistic diversity, and contextual ambiguity. Our best-performing model, based on Flan-T5, achieved a macro F1

score of 0.69 on the test set, while the CharBERT-based model reached 0.70. These results underscore the effectiveness of integrating augmentation techniques with transformer-based architectures in low-resource, code-mixed settings. Future work will explore larger language models (LLMs), context-aware decoding tailored to Hinglish nuances, and constraint-based structured generation to further improve hoax specificity. Additionally, we plan to extend this research to other multilingual social media platforms and explore real-time detection mechanisms for dynamic hoax identification. Another promising direction involves fine-tuning models on even more diverse datasets to improve their generalization and robustness. Overall, this work demonstrates the potential of advanced NLP models in combating harmful misinformation in underrepresented languages and settings. Moreover, the findings highlight the importance of continued innovation in model architecture and training techniques to address the evolving nature of misinformation across diverse linguistic landscapes.

8 Limitations

Despite the promising results, several limitations emerged during our experiments. Data imbalance significantly impacted model performance, especially for the Hoax class. Although we employed a custom weighted loss function (with a 5:1 ratio) to prioritize hoax detection, both CharBERT and Flan-T5-base exhibited higher false negatives, indicating persistent challenges in recognizing hoax instances. CharBERT, while effective for non-hoax classification, struggled to generalize under imbalanced conditions.

Furthermore, the linguistic intricacies of code-mixed Hinglish presented challenges across both models. The Flan-T5-base model, in particular, showed sensitivity to subtle contextual shifts and the informal, slang-heavy nature of Hinglish. Our use of back translation, although beneficial for data augmentation, occasionally led to semantic drift, where paraphrased texts diverged slightly from their original meanings. Similarly, SMOTE generated synthetic samples that, due to their reliance on neighboring embeddings, often lacked linguistic diversity and richness, limiting their effectiveness in representing the true variability of hoax content.

Another key limitation stemmed from computational constraints due to restricted access to high-end GPU resources, we fine-tuned the Flan-T5-base variant rather than larger and more contextually expressive models like Flan-T5-large. This hardware limitation may have capped the model's capacity to capture deeper linguistic nuances and broader contextual signals.

Looking forward, future research should consider leveraging larger LLMs, integrating adversarial training to enhance model robustness against noisy and imbalanced data, and expanding Hinglish-specific lexicons to improve semantic understanding. Additionally, techniques like context-aware or constraint-based decoding could further enhance specificity in hoax detection by reducing ambiguity in model predictions.

References

Somnath Banerjee, Maulindu Sarkar, Nancy Agrawal, Punyajoy Saha, and Mithun Das. 2021. Exploring transformer based models to identify hate speech and offensive content in english and indo-aryan languages. In *Proceedings of FIRE 2021: Forum for Information Retrieval Evaluation*. CEUR-WS.org. Presented at FIRE 2021, Virtual Event, 13th-17th December 2021.

Cristina Bosco, Viviana Patti, Simona Frenda, Alessandra Teresa Cignarella, Marinella Paciello, and Francesca D'Errico. 2023. Detecting racial stereotypes: An italian social media corpus where psychology meets nlp. *Information Processing Management*, 60(1):103118.

Dan Carpenter, Wookhee Min, Seung Lee, Gamze Ozogul, Xiaoying Zheng, and James Lester. 2024. Assessing student explanations with large language models using fine-tuning and few-shot learning. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 403–413, Mexico City, Mexico. 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).

K. Ruwani M. Fernando and Chris P. Tsokos. 2021. Dynamically weighted balanced loss: Class imbalanced learning and confidence calibration of deep neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33(7):2940–2951. Published in IEEE Transactions on Neural Networks and Learning Systems.

Keyan Guo, Alexander Hu, Jaden Mu, Ziheng Shi, Ziming Zhao, Nishant Vishwamitra, and Hongxin Hu. 2024. An investigation of large language models for real-world hate speech detection. *arXiv preprint arXiv:2401.03346v1*. ArXiv:2401.03346v1 [cs.CY].

Shaina Raza, Drai Paulen-Patterson, and Chen Ding. 2024. Fake news detection: Comparative evaluation of bert-like models and large language models with generative ai-annotated data. *arXiv*, 2412.14276v1. License: CC BY 4.0.

Irvan Santoso, Immanuel Yohansen, Nealson, Harco Leslie Hendric Spits Warnars, and Kiyota Hashimoto. 2017. Early investigation of proposed hoax detection for decreasing hoax in social media. In *2017 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom)*, pages 123–128, Phuket, Thailand. IEEE.

Wolfgang S. Schmeisser-Nieto, Alessandra Teresa Cignarella, Tom Bourgeade, Simona Frenda, Alejandro Ariza-Casabona, Mario Laurent, Paolo Giovanni Cicirelli, Andrea Marra, Giuseppe Corbelli, Farah Benamara, Cristina Bosco, Véronique Moriceau, Marinella Paciello, Viviana Patti, Mariona Taulé, and Francesca D'Errico. 2024. Stereohoax: a multilingual corpus of racial hoaxes and social media reactions annotated for stereotypes. *Language Resources and Evaluation*, 2024:1–28.

Appendix A: Prompt Variations for Racial Hoax Detection

Prompt Type	Description	Prompt
Fine tuned with Zero Shot Prompting	A simple zero-shot prompt instructing the model to classify text as racial hoax (1) or not (0)	"Please check whether the text is racial hoax (1) or not (0):".
Fine tuned with Instructions	Instructs the model to classify Hinglish text as Hoax (1) or Non-Hoax (0) with a concise definition of both classes	"Classify the given Hinglish social media text as either 'Hoax' (1) or 'Non-Hoax' (0) for racial hoax detection. A Hoax (1) contains abusive, derogatory, or inflammatory language targeting a specific group (e.g., caste, religion, ethnicity), promotes hate or stereotypes, or includes threats or exaggerated claims to provoke fear or division. A Non-Hoax (0) is neutral, promotes unity, or discusses issues respectfully without targeting or dividing communities. Now classify this text:"
Fine tuned with Rubric	Provides a detailed rubric defining Hoax (1) and Non-Hoax (0) criteria, emphasizing strict classification for Hinglish text	"You are a binary text classifier. Classify the text strictly as Hoax (1) or Non-Hoax (0), prioritizing detection of Hoaxes. Follow this rubric tailored to Hinglish social media text: **Rubric:** - Hoax (1): Text is classified as Hoax if it: - Uses abusive, derogatory, or slang-heavy language (e.g., 'mule,' 'sale,' 'kute,' 'chamar') targeting a specific group (caste, religion, ethnicity, etc.). - Promotes hate, division, or stereotypes between communities (e.g., Hindu vs. Muslim, Dalit vs. Brahmin). - Contains threats (e.g., 'kat dalenge,' 'mita do') or exaggerated claims (e.g., conspiracies like 'gazwaehind') meant to provoke fear or anger. - Non-Hoax (0): Text is classified as Non-Hoax if it: - Encourages unity, respect, or neutral discussion across groups without hate. - Avoids abusive or inflammatory language, even if critical of issues (e.g., caste, reservation, politics). - Focuses on personal views, facts, or constructive critique without targeting or dividing communities. Now classify this text:"

Prompt Type	Description	Prompt
Fine tuned with Rubric + Few Shot	Extends the rubric-based prompt with few-shot examples from training data to guide classification of Hinglish text	"Classify the text strictly as Hoax (1) or Non-Hoax (0), prioritizing detection of Hoaxes. Follow this rubric tailored to Hinglish social media text: **Rubric:** - Hoax (1): Text is classified as Hoax if it: - Uses abusive, derogatory, or slang-heavy language (e.g., 'mule,' 'sale,' 'kute,' 'chamar') targeting a specific group (caste, religion, ethnicity, etc.). - Promotes hate, division, or stereotypes between communities (e.g., Hindu vs. Muslim, Dalit vs. Brahmin). - Contains threats (e.g., 'kat dalenge,' 'mita do') or exaggerated claims (e.g., conspiracies like 'gazwaehind') meant to provoke fear or anger. - Non-Hoax (0): Text is classified as Non-Hoax if it: - Encourages unity, respect, or neutral discussion across groups without hate. - Avoids abusive or inflammatory language, even if critical of issues (e.g., caste, reservation, politics). - Focuses on personal views, facts, or constructive critique without targeting or dividing communities. **Examples from Training Data:** 1. 'in suar ki aulad hone chali chali bache paida kar ke hindustan ka vada kar kar diya hai in a goli sa uda do' $\rightarrow$ 1 (Hoax: abusive ('suar ki aulad'), threatens violence, stereotypes population growth) 2. 'sikh hindu to ek hi hai hamara koi ghar aisa nahi hai yaha hindu sikh bacho ne shadi nahi ki ho' $\rightarrow$ 0 (Non-Hoax: promotes unity between Sikhs and Hindus, no hate) 3. 'quran or sariya kanon hi insaf kar sakta hai is duniya ka sabse mahan or prachin dharma sirf islam hai hidu to kide hai islam' $\rightarrow$ 1 (Hoax: derogatory ('kide'), promotes religious superiority and hate) 4. 'jab tak inter cast mariage ko society acept nahi karegi tab tak log kaise acept karenge ki reservation end ho' $\rightarrow$ 0 (Non-Hoax: neutral discussion on caste and reservation, no abuse or division) Now classify this text:"