# Unpacking Faux-Hate: Addressing Faux-Hate Detection and Severity Prediction in Code-Mixed Hinglish Text with HingRoBERTa and Class Weighting Techniques

**Ashweta A. Fondekar**
Goa Business School,
Goa University;
Parvatibai Chowgule
College of Arts and Science
dcst.ashweta@unigoa.ac.in

**Milind M. Shivolkar**
Goa Business School,
Goa University
milind.shivolkar@unigoa.ac.in

**Dr. Jyoti D. Pawar**
Goa Business School,
Goa University
jdp@unigoa.ac.in

## Abstract

The proliferation of hate speech and fake narratives on social media poses significant societal challenges, especially in multilingual and code-mixed contexts. This paper presents our system submitted to the ICON 2024 shared task on Decoding Fake Narratives in Spreading Hateful Stories (Faux-Hate). We tackle the problem of Faux-Hate Detection, which involves detecting fake narratives and hate speech in code-mixed Hinglish text. Leveraging HingRoBERTa, a pre-trained transformer model fine-tuned on Hinglish datasets, we address two sub-tasks: Binary Faux-Hate Detection and Target and Severity Prediction. Through the introduction of class weighting techniques and the optimization of a multi-task learning approach, we demonstrate improved performance in identifying hate and fake speech, as well as in classifying their target and severity. This research contributes to a scalable and efficient framework for addressing complex real-world text processing challenges.

## 1 Introduction

Social media has revolutionized communication but has also created a breeding ground for harmful content, including hate speech. The combination of hate speech and fake narratives termed Faux-Hate poses a unique challenge as it exploits misinformation to provoke or mislead, amplifying its impact. Addressing this requires advanced models capable of understanding nuanced, code-mixed language, such as Hinglish (Hindi-English). This study focuses on developing a system to detect Faux-Hate(Biradar et al., 2024b) and classify its attributes within social media text. Using HingRoBERTa, a state-of-the-art transformer model, we target two primary objectives:

- Binary Faux-Hate Detection: Simultaneously identifying whether a post is fake or hateful.
- Target and Severity Prediction: Classifying the target (Individual, Organization, Religion) and severity (Low, Medium, High) of hateful content.

Key Contributions of This Study

- Development and benchmarking of a HingRoBERTa-based multi-tasking model capable of handling binary and multi-class labels for both tasks.
- Integration of class weighting techniques to address imbalances in the dataset, ensuring robust model performance.
- Detailed insights into the granularity of hate speech, exploring its intensity and target.

By leveraging HingRoBERTa, a state-of-the-art transformer model fine-tuned for Hinglish, and incorporating class weighting techniques, this study establishes a benchmark for addressing Faux-Hate in a low-resource language context.

The rest of the paper is organized as follows; Section 2 reviews related work, while Section 3 details the methodology, including task and dataset descriptions. Section 4 presents the experimental setup and results. Finally, Section 5 discusses the conclusion and future work, followed by a dedicated section addressing the limitations of the study.

## 2 Related work

The intersection of hate speech and fake news detection has drawn increasing attention, especially since studies in 2016 linked hate speech propagation to fabricated narratives (Gollatz and Jenner, 2018). During the U.S. presidential elections, 37 out of 49 articles on hate speech were based on fabricated content. However, significant gaps remain, particularly for low-resource languages like Hinglish.

Existing datasets like Hostile Post Detection in Hindi (Bhardwaj et al., 2020), FactDRIL (Singhal et al., 2021), IEHate (Jafri et al., 2023), HEOT (Mathur et al., 2018), and HESOC (Mandl et al., 2020) focus on either fake or hate detection, often neglecting their intersection. Most also emphasize binary hate detection, overlooking aspects like target or severity. The Faux Hate Multi-Label Dataset (FHMLD) bridges these gaps, offering multi-class annotations for hate speech target and severity in Hinglish, while linking fake narratives with hate speech.

### 2.1 Methods for Fake and Hate Content Detection

Fake content detection has benefited from deep learning models like BERT and neural networks, with ensemble approaches combining fastText, HindiBERT, and BERT achieving up to 71% accuracy (Akash et al., 2021; Mehta et al., 2021). Cross-lingual models such as mBERT and XLM-R have also been fine-tuned for Hinglish fake news detection, achieving macro F1 scores of 0.71 (Banerjee et al., 2021).

Hate speech detection has progressed from traditional methods using embeddings like Word2Vec, GloVe, and fastText with SVM and CNN-LSTM (Bisht et al., 2020; Sreelakshmi et al., 2020), to transformer-based models like BERT, IndicBERT, and XLM-R, achieving F1 scores of 0.72 (Farooqi et al., 2021). Ensemble techniques combining transformers with neural networks have further improved performance to F1 scores of 0.81 (Shekhar et al., 2021).

Despite these advances, most research focuses on monolingual or binary tasks, with limited exploration of Hinglish code-mixed text or hate speech granularity, such as target and severity (Biradar et al., 2021). The Faux Hate Multi-Label Dataset (FHMLD) and HingRoBERTa (Nayak and Joshi, 2022) address these gaps by linking fake narratives to hate speech and supporting binary and multi-class tasks. This study leverages class weighting and state-of-the-art models, setting a benchmark for Hinglish fake and hate speech detection.

## 3 Methodology

### 3.1 Task and Dataset Description

This study addresses two tasks from the ICON 2024 shared task(Biradar et al., 2024a), Decoding Fake Narratives in Spreading Hateful Stories (Faux-Hate), focusing on detecting fake content and hate speech in Hinglish, a code-mixed language. The Faux-Hate Multi-Label Dataset (FHMLD) provides a detailed resource, including binary labels for fake content (fake/real) and hate speech (hate/non-hate), as well as categorical labels for hate speech targets (Individual, Organization, Religion) and severity (Low, Medium, High).

The shared task comprises two sub-tasks:

- Task A - Binary Faux-Hate Detection: Predict fake and hate labels for each sample.
- Task B - Target and Severity Prediction: Predict target (I/O/R) and severity (L/M/H) labels.

The dataset bridges gaps in prior research by combining fake content and hate speech detection while addressing nuances like target and severity. It comprises 6,396 training, 800 validation, and 800 testing samples. Label distribution graphs for Task A provide insights into class imbalances, aiding in model development to address dataset challenges effectively.

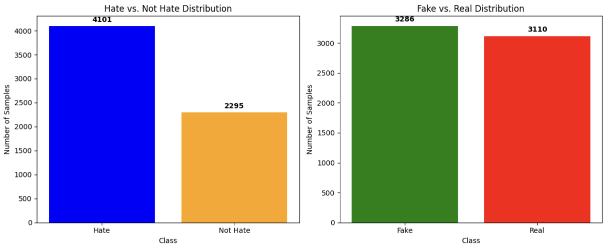

Figure 1: Label distribution in the Task A training set, highlighting the class proportions and imbalances to support model development.

### 3.2 Preprocessing of the Dataset

To ensure uniformity and improve data quality, we applied several preprocessing steps to the code-mixed Hinglish dataset, including converting text to lowercase, removing hyperlinks, wide spaces, blank lines, and alphanumeric characters.

### 3.3 Class Weighting

To handle class imbalances, class weights were computed and integrated into the loss functions. Task A: Weighted binary cross-entropy loss. Task B: Weighted categorical cross-entropy loss for Target and Severity.

### 3.4 HingRoBERTa Model

HingRoBERTa[1], a transformer model pre-trained on a large Hinglish corpus, is particularly well-suited for code-mixed text. Its architecture enables contextualized embeddings that capture linguistic nuances across languages.

### 3.5 Multi-Task Learning Framework

We designed a multi-tasking framework with two components: A shared encoder using HingRoBERTa to extract contextual embeddings. Separate classification heads for each task: Task A: Two binary classifiers for Fake and Hate labels. Task B: Two multi-class classifiers for Target and Severity.

### 3.6 Training and Fine-Tuning

The texts were tokenized using the HingRoBERTa tokenizer[2], which is specifically designed for Hinglish text. The maximum sequence length was set to 128 tokens to ensure optimal performance and avoid memory overload during training. Fine-tuning of the model was performed using the AdamW optimizer, which is well-suited for transformer-based models. The learning rate was set to 5e-5, and weight decay was applied with a value of 0.01 to help prevent overfitting and promote generalization.

Training Parameters:

- Batch Size: A batch size of 16 was used for training to strike a balance between memory usage and training efficiency.

- Epochs: The model was trained for 2 epochs to ensure convergence without overfitting on the relatively small dataset.

- Hardware: Training was conducted on an NVIDIA GPU to speed up the process and handle the computation-heavy tasks of fine-tuning transformer models.

These training settings were carefully selected to maximize performance on the specific tasks of detecting fake and hate content in Hinglish. The relatively small number of epochs and batch size were chosen to balance between training time and model performance, given the dataset size and complexity of the task.

| **Label** | **Precision** | **Recall** | **F1-Score** | **Support** |
|---|---|---|---|---|
| 0 | 0.8010 | 0.8303 | 0.8154 | 383 |
| 1 | 0.8387 | 0.8106 | 0.8244 | 417 |
| Accuracy | - | - | **0.8200** | **800** |
| Macro Avg | **0.8199** | **0.8204** | **0.8199** | **800** |

Table 1: Classification Report for Fake Detection.

| **Label** | **Precision** | **Recall** | **F1-Score** | **Support** |
|---|---|---|---|---|
| 0 | 0.7067 | 0.6969 | 0.7018 | 287 |
| 1 | 0.8317 | 0.8382 | 0.8350 | 513 |
| **Accuracy** | - | - | **0.7875** | **800** |
| **Macro Avg** | **0.7692** | **0.7675** | **0.7684** | **800** |

Table 2: Classification Report for Hate Detection.

## 4 Experimental Setup and Results

### 4.1 Evaluation Metrics

To assess the model's performance across both tasks, the following metrics were used:

- Accuracy: Measures the overall correctness of predictions.

- Precision, Recall, and F1-Score: Used to evaluate binary classification in Task A and multi-class classification in Task B.

- Macro F1-Score: Averages F1 scores across all classes, considering imbalanced datasets.

The detailed classification reports for Task A (Fake and Hate Detection) and Task B (Target and Severity Prediction), along with the combined macro F1 score of both tasks, are shown in Tables 1, 2, 3, 4 and 5, respectively. The combined Macro F1 score for both tasks is calculated as 0.7072.

### 4.2 Discussion and Analysis of Evaluation Metrics

#### 4.2.1 Fake Detection Performance

The model achieves an overall accuracy of 82.00% and a macro F1-score of 0.8199, demonstrating

| **Label** | **Precision** | **Recall** | **F1-Score** | **Support** |
|---|---|---|---|---|
| I | 0.5513 | 0.6515 | 0.5972 | 132 |
| O | 0.8397 | 0.7483 | 0.7914 | 294 |
| R | 0.6383 | 0.6897 | 0.6630 | 87 |
| nan | 0.7222 | 0.7247 | 0.7235 | 287 |
| **Accuracy** | - | - | **0.7175** | **800** |
| **Macro Avg** | **0.6879** | **0.7036** | **0.6938** | **800** |

Table 3: Classification Report for Target Detection.

[1] https://huggingface.co/l3cube-pune/hing-robert

[2] https://github.com/ashwetafondekar123/DCST_Unigoa_Faux-Hate_Shared_Task.git

balanced performance across the real (label 0) and fake (label 1) classes. Specifically:

- For real content, the model achieves a precision of 80.10%, recall of 83.03%, and an F1-score of 0.8154.
- For fake content, the model achieves a precision of 83.87%, recall of 81.06%, and an F1-score of 0.8244.

This indicates the model effectively handles both classes with minimal bias and performs slightly better at detecting fake content.

### 4.2.2 Hate Detection Performance

The hate detection task achieves an accuracy of 78.75% and a macro F1-score of 0.7684. Analysis of class-wise performance reveals:

- For non-hate speech (label 0), the model achieves an F1-score of 0.7018, with a precision of 70.67% and recall of 69.69%.
- For hate speech (label 1), the model achieves an F1-score of 0.8350, with a precision of 83.17% and recall of 83.82%.

The higher recall for hate speech suggests the model is adept at identifying hate speech instances but less effective at detecting non-hate speech, likely due to class imbalance in the dataset.

### 4.2.3 Target Prediction Performance

The target prediction task achieves an accuracy of 71.75% and a macro F1-score of 0.6938. Key insights include:

- The model performs best for the Organization (O) class with an F1-score of 0.7914, followed by Religion (R) (0.6630) and Individual (I) (0.5972).
- Recall for the Individual class (65.15%) is higher than precision (55.13%), indicating difficulties in accurately identifying this class.
- The results suggest the model is better at identifying Organization as the target, while performance on the Individual class requires further improvement.

### 4.2.4 Severity Prediction Performance

Severity prediction poses significant challenges, achieving an overall accuracy of 59.88% and a macro F1-score of 0.5643. Performance by severity level is as follows:

- Medium severity (M) achieves the highest F1-score of 0.6210, with a recall of 71.58%, reflecting strong detection of moderate severity cases.
- Low severity (L) achieves an F1-score of 0.4640, and High severity (H) achieves an F1-score of 0.4487, highlighting difficulty in distinguishing these levels due to overlapping linguistic features and limited data.

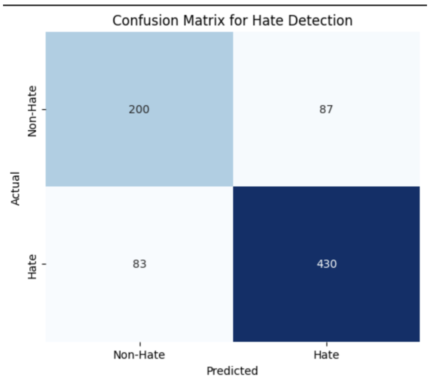

Figure 2: Confusion Matrix for Hate/Non-Hate Detection.

### 4.2.5 Combined Macro F1 Scores

A comparison of binary and multi-class tasks reveals the following:

- For binary tasks (Fake and Hate detection), the macro F1-scores are 0.8199 and 0.7684, respectively, with a combined score of 0.7988.
- For multi-class tasks (Target and Severity prediction), the macro F1-scores are 0.6938 and 0.5643, respectively, with a combined score of 0.6155.
- The disparity between binary and multi-class tasks indicates that simpler decision boundaries in binary tasks facilitate higher performance, whereas multi-class tasks require further enhancement.

| Label | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| H | 0.3977 | 0.5147 | 0.4487 | 68 |
| L | 0.5682 | 0.3922 | 0.4640 | 255 |
| M | 0.5484 | 0.7158 | 0.6210 | 190 |
| nan | 0.7222 | 0.7247 | 0.7235 | 287 |
| **Accuracy** | - | - | **0.5988** | **800** |
| **Macro Avg** | **0.5591** | **0.5868** | **0.5643** | **800** |

Table 4: Classification Report for Severity Detection.

| Task | Macro F1 Score |
|---|---|
| Hate Detection (Binary) | 0.7684 |
| Fake Detection (Binary) | 0.8199 |
| (Hate + Fake) | 0.7988 |
| Target Prediction (NA, I, O, R) | 0.6938 |
| Severity Prediction (NA, L, M, H) | 0.5643 |
| (Target + Severity) | 0.6155 |

Table 5: Macro F1 Scores for Faux-Hate Detection Tasks.

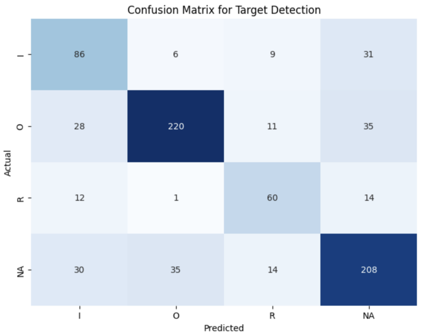

Figure 3: Confusion Matrix for Target Detection.

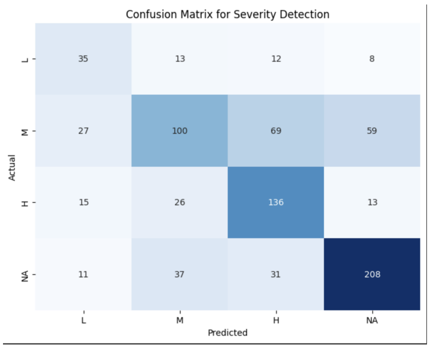

Figure 4: Confusion Matrix for Severity Detection.

### 4.3 Confusion Matrix

The confusion matrices in Figures 2–5 for Task A (Fake and Hate Detection) and Task B (Target and Severity Prediction) reveal key insights. Task A effectively differentiates between fake and hate content but shows misclassifications between fake/real and hate/non-hate categories. In Task B, Target Prediction performs well for Organization (O) but struggles with Individual (I), Religion (R), and NA cases. Severity Prediction faces difficulties distinguishing between Low (L) and Medium (M) severity. Further refinement is needed to improve precision, particularly for minority classes and NA cases.

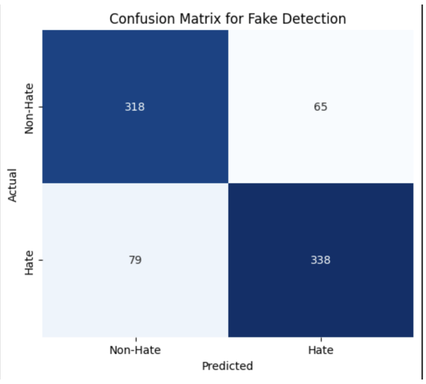

Figure 5: Confusion Matrix for Fake/Real Detection.

## 5 Conclusion

This study tackled the ICON 2024 shared task on decoding fake narratives in hateful stories (Faux-Hate), focusing on detecting fake content and hate speech in Hinglish. Using HingRoBERTa with class weighting and multi-task learning, we achieved promising results. Task A yielded macro F1 scores of 0.8199 for fake detection and 0.7684 for hate detection, while Task B achieved 0.6938 for target prediction and 0.5643 for severity classification, with challenges in distinguishing adjacent severity levels and minority classes.

Confusion matrix analysis showed strong performance in hate detection and organizational targets but highlighted areas for improvement, particularly in nuanced categories.

Future work could explore advanced architectures and techniques to address class imbalances and evolving linguistic trends in code-mixed text.

## Limitations

This study has limitations, including its focus on Hinglish, which limits generalizability, and reliance on high computational resources, restricting accessibility for resource-constrained researchers. Challenges remain in handling minority classes despite class weighting, and the model's performance on longer texts is untested, given the short social media content in the dataset. Additionally, the black-box nature of transformers hampers interpretability, and Hinglish's cultural nuances may require adaptation for broader contexts. Future work should address these to improve scalability and applicability.

## References

BS Akash, Jathin Badam, KVLN Raju, and Dipanjan Chakraborty. 2021. A poster on learnings from an attempt to build an nlp-based fake news classification system for hindi. In *Proceedings of the 4th ACM SIGCAS Conference on Computing and Sustainable Societies*, pages 397–401.

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. *arXiv preprint arXiv:2111.13974*.

Mohit Bhardwaj, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2020. Hostility detection dataset in hindi. *arXiv preprint arXiv:2011.03588*.

Shankar Biradar, Sai Kartheek Reddy Kasu, Sunil Saumya, and Md. Shad Akhtar, editors. 2024a. *Proceedings of the 21st International Conference on Natural Language Processing (ICON): Shared Task on Decoding Fake Narratives in Spreading Hateful Stories (Faux-Hate)*. Association for Computational Linguistics, AU-KBC Research Centre, MIT College, India.

Shankar Biradar, Sunil Saumya, and Arun Chauhan. 2021. Hate or non-hate: Translation based hate speech identification in code-mixed hinglish data set. In *2021 IEEE international conference on big data (Big Data)*, pages 2470–2475. IEEE.

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.

Akanksha Bisht, Annapurna Singh, HS Bhadauria, Jitendra Virmani, and Kriti. 2020. Detection of hate speech and offensive language in twitter data using lstm model. *Recent trends in image and signal processing in computer vision*, pages 243–264.

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*.

Kirsten Gollatz and Leontine Jenner. 2018. Hate speech and fake news–how two concepts got intertwined and politicised. *encore*, page 62.

Farhan Ahmad Jafri, Mohammad Aman Siddiqui, Surendrabikram Thapa, Kritesh Rauniyar, Usman Naseem, and Imran Razzak. 2023. Uncovering political hate speech during indian election campaign: A new low-resource dataset and baselines. *arXiv preprint arXiv:2306.14764*.

Thomas Mandl, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. 2020. Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german. In *Proceedings of the 12th annual meeting of the forum for information retrieval evaluation*, pages 29–32.

Puneet Mathur, Ramit Sawhney, Meghna Ayyar, and Rajiv Shah. 2018. Did you offend me? classification of offensive tweets in hinglish language. In *Proceedings of the 2nd workshop on abusive language online (ALW2)*, pages 138–148.

Manan Mehta, Utkarsh Pandey, Yash Chaudhary, Raghav Sharma, Ishaan Gill, Deepak Gupta, and Ashish Khanna. 2021. Hindi text classification: A review. In *2021 3rd international conference on advances in computing, communication control and networking (ICAC3N)*, pages 839–843. IEEE.

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.

Chander Shekhar, Bhavya Bagla, Kaushal Kumar Maurya, and Maunendra Sankar Desarkar. 2021. Walk in wild: An ensemble approach for hostility detection in hindi posts. *arXiv preprint arXiv:2101.06004*.

Shivangi Singhal, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. 2021. Factorization of fact-checks for low resource indian languages. *arXiv preprint arXiv:2102.11276*.

K Sreelakshmi, B Premjith, and KP Soman. 2020. Detection of hate speech text in hindi-english code-mixed data. *Procedia Computer Science*, 171:737–744.