# A Machine Learning Framework for Detecting Hate Speech and Fake Narratives in Hindi-English Tweets

**R.N.Yadawad**[1], **Sunil Saumya**[2], **K.N.Nivedh**[1], **Siddhaling S. Padanur**[1], **Sudev Basti**[1]
[1] Shri Dharmasthala Manjunatheshwara College of Engineering Technology, Dharwad
[2] Indian Institute of Information Technology Dharwad
rnyadawad@sdmcet.ac.in; sunil.saumya@iiitdwd.ac.in
nivedhbhat15@gmail.com; siddhaling88@gmail.com
sudevbasti0717@gmail.com

## Abstract

This paper presents a novel system developed for the Faux-Hate Shared Task at ICON 2024, addressing the detection of hate speech and fake narratives within Hindi-English code-mixed social media data. Our approach combines advanced text preprocessing, TF-IDF vectorization, and Random Forest classifiers to identify harmful content, while employing SMOTE to address class imbalance. By leveraging ensemble learning and feature engineering, our system demonstrates robust performance in detecting hateful and fake content, classifying targets, and evaluating the severity of hate speech. The results underscore the potential for real-world applications, such as moderating online platforms and identifying harmful narratives. Furthermore, we highlight ethical considerations for deploying such tools, emphasizing responsible use in sensitive domains, thereby advancing research in multilingual hate speech detection and online abuse mitigation.

## 1 Introduction

The spread of hate speech and fake narratives on social media has become a pervasive issue, significantly impacting individuals and communities. This problem is further exacerbated in code-mixed languages, where multiple languages are seamlessly interwoven within the same text. A prominent example is Hindi-English code-mixed text, commonly observed on social platforms in regions where both languages are spoken. This linguistic complexity adds layers of difficulty to traditional text analysis models, making the detection of harmful content, such as hate speech and fake narratives, much more challenging.

In recent years, numerous efforts (Ando et al., 2005; Andrew and Gao, 2007; Biradar et al., 2024b) have been made to detect hate speech and misinformation in monolingual texts. However, code-switching the act of mixing two or more languages in the same conversation—poses significant challenges for existing models, which are often optimized for a single language. Furthermore, the high level of informality and contextual variations in social media discourse adds another layer of complexity. The Faux-Hate Shared Task at ICON 2024 addresses this unique challenge by focusing on detecting fake narratives and hate speech in Hindi-English code-mixed social media data.

This paper presents a machine learning-based system designed specifically to tackle the issues associated with this task. Our approach combines TF-IDF (Term Frequency-Inverse Document Frequency) vectorization for feature extraction with Random Forest classifiers for classification. We also incorporated SMOTE (Synthetic Minority Over-sampling Technique) to handle class imbalance, a common problem in datasets where harmful content, such as hate speech and fake news, is often underrepresented.

The system is evaluated on the Faux-Hate dataset, which contains social media comments labeled for both hate speech detection and fake news detection. The results from our approach demonstrate promising performance, particularly in identifying hate speech and fake narratives across the linguistically diverse and informal nature of the dataset. Moreover, our system also identifies the target and severity of hateful speech, providing more granular insights into the content.

By leveraging state-of-the-art techniques in text preprocessing, vectorization, and classification, our work provides a robust solution for detecting harmful content in multilingual and code-mixed social media data. This approach not only contributes to the Faux-Hate Shared Task but also has potential applications in real-world scenarios, such as content moderation, social media monitoring, and automated detection of harmful speech across diverse languages.

## 2 Literature Survey

The complexity of processing code-mixed text, especially Hindi-English, is well-documented in studies like (Chakravarthi et al., 2020; Srivastava and Singh, 2022). These works highlight challenges such as transliteration, phonetic spellings, and informal expressions, which hinder effective hate speech detection. Despite advancements in multilingual hate speech detection by (Davidson et al., 2017; Malmasi and Zampieri, 2018), these studies fall short of addressing code-mixed scenarios, leaving a gap that our work aims to fill through targeted preprocessing and feature extraction techniques.

In the domain of fake news detection, (Zhou and Zafarani, 2020) emphasized the importance of linguistic feature-based approaches, such as TF-IDF, which aligns with our methodology. (Ruchansky et al., 2017) introduced hybrid models combining linguistic and network-based features, offering complementary strategies to enhance detection. Feature engineering plays a critical role in analyzing code-mixed text. (Saumya et al., 2024) extracted features for capturing the structural and lexical variations unique to code-mixed data using deep learning approach.

Addressing class imbalance is crucial for tasks like hate speech detection, where minority classes often face underrepresentation. Techniques such as SMOTE, introduced by (Chawla et al., 2002), and class weighting have been shown to enhance recall for minority classes, mitigating bias and improving overall model performance. These methods are central to our approach, enabling more equitable handling of imbalanced datasets.
This survey provides a foundation for our research by integrating insights from prior studies and addressing gaps specific to code-mixed text processing, particularly in hate speech and fake news detection.

## 3 Dataset and Task Description

The dataset utilized in this study is part of the Faux-Hate Shared Task at ICON 2024 (Biradar et al., 2024a), focusing on detecting hate speech and fake narratives in Hindi-English code-mixed text. It is structured into two main tasks, each addressing distinct classification challenges.

- **Task A: Binary Faux-Hate Detection**
  - Predict whether the content is fake (1 for fake, 0 for real).
  - Predict whether the content contains hate speech (1 for hate, 0 for non-hate).
- **Task B: Target and Severity Prediction**
  - Target: Classify the target of hate speech into one of three categories:
    - Individual (I)
    - Organization (O)
    - Religion (R)
  - Severity: Assess the severity of hate speech as:
    - Low (L)
    - Medium (M)
    - High (H)

The dataset is divided into training, validation, and test subsets. The training set contains 6,397 tweets labeled for both tasks, the validation set includes 801 labeled tweets, and the test set comprises 801 unlabeled tweets reserved for final evaluation. This structure allows for comprehensive development and testing of models designed to identify and characterize harmful content in multilingual, code-mixed social media environments.

## 4 Methodology

Our methodology involves preprocessing the dataset, extracting meaningful features, addressing class imbalance, and evaluating multiple models using robust metrics to optimize performance.

The proposed methodology begins with an extensive preprocessing phase to clean and prepare the dataset for analysis. This involved removing URLs, mentions, hashtags, and non-alphanumeric characters to eliminate noise and irrelevant information. The text was then tokenized into meaningful units, and stopwords were removed to focus on critical content, enhancing the quality of feature extraction. Features were extracted using TF-IDF vectorization, leveraging unigrams and bigrams to capture both individual words and contextual relationships within the text. To further enrich the feature set, text length was included as an additional feature, allowing the model to detect nuanced patterns often present in longer and more detailed content.
Class imbalance, a significant challenge in detecting hate speech and fake narratives, was addressed through a combination of techniques. Synthetic

Minority Over-sampling Technique (SMOTE) was applied to generate synthetic samples for the underrepresented classes, ensuring the model was exposed to a balanced dataset during training. Additionally, class weights were adjusted in several models to give higher importance to minority classes, improving recall for underrepresented categories without substantially compromising precision for the majority class.

Several models were explored to identify the optimal approach for this task. Random Forest was chosen as the primary classifier due to its robustness and ability to handle high-dimensional data, making it highly effective for text classification. Logistic Regression and Support Vector Machine (SVM) served as baseline models for performance comparison. Gradient Boosting algorithms, including XGBoost and LightGBM, were evaluated for their capability to capture complex interactions in the data and enhance overall performance. Simple feed-forward neural networks were also tested to model non-linear patterns; however, Random Forest outperformed them in this context.

The models' performance was evaluated using multiple metrics, including accuracy to assess overall correctness, and Macro F1-Score to provide a balanced measure of performance across both classes, particularly important for imbalanced datasets. Precision and recall were also examined to ensure the models could accurately identify both positive and negative samples. This comprehensive approach allowed us to address the unique challenges of code-mixed text while optimizing the detection of fake content and hate speech.

## 5 Results

The proposed approach demonstrates significant achievement in detecting fake content and hate speech in Hindi-English code-mixed text, achieving balanced performance across both tasks.

As shown in Figure 1, the model achieved an accuracy of 0.7575, demonstrating strong performance in identifying hate speech. The high Macro F1 Score indicates a balanced capability to handle both hate speech (minority class) and non-hate speech (majority class). Precision measures the proportion of tweets correctly identified as hate speech, minimizing false positives. Recall evaluates the model's ability to detect most hate speech tweets, reducing false negatives. To address class imbalance, SMOTE was employed, enhancing the model's generalization and effectiveness in detecting hate speech.

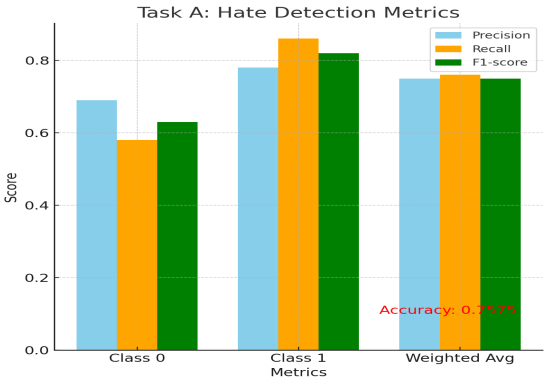


Figure 1: Hate Speech Detection (Task A)

Our performance in Task A of the ICON competition earned us a commendable 3rd place finish as shown in Table 1.

| Team | F1-Score | Rank |
|---|---|---|
| DCST Unigoa | 0.79 | 1 |
| Radicaldecoders run1 | 0.7761 | 2 |
| **Chakravyuh coders run1** | **0.7721** | **3** |

Table 1: LeaderBoard Ranking for Task A

False positives were observed in tweets containing strong but non-hateful language, while subtle or implicit hate speech was occasionally misclassified as non-hateful.

A similar experiment was conducted for fake news dection and the result is shown in Figure 2.

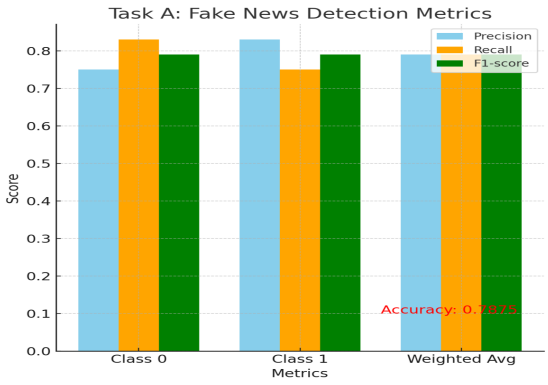


Figure 2: Fake News Detection (Task A)

As we mention in Figure 2, the model achieved an accuracy of 0.7875, reflecting strong performance in distinguishing fake narratives from real content. The results indicate the model's ability to effectively differentiate between fake

and genuine content in a multilingual, code-mixed dataset. The model also exhibited strong precision, indicating that the majority of content flagged as fake was indeed fake, thereby minimizing false positives. This is particularly important to ensure that genuine content is not incorrectly classified as fake. Additionally, the model achieved high recall, effectively identifying a substantial portion of fake content. This reduces false negatives and ensures that most instances of fake news are successfully detected, further solidifying the model's robustness and reliability in handling multilingual, code-mixed datasets.

In the next experiment, we addressed two specific tasks: Binary Faux-Hate Detection (Task A) and Target and Severity Prediction (Task B). For Task A, the goal was to classify content as Fake, Hate, or both, and our custom model, Hate-FakeNet, demonstrated strong performance. Built using a Random Forest base, enhanced with feature engineering and SMOTE to handle class imbalance, the model effectively outperformed baseline approaches such as Logistic Regression and Support Vector Machines (SVM).
For Task B, which involved predicting the Target (Individual, Organization, Religion) and Severity (Low, Medium, High) of hate speech, we introduced an improved model, Hate-FakeNet-Plus. This version utilized Random Forest and incorporated advanced techniques like Gradient Boosting Machines (XGBoost and LightGBM) to better capture complex data interactions. Optimized for multi-class classification, the model achieved high precision across multiple categories. As presented in Table 2 , these efforts culminated in securing the 13th rank, highlighting the model's capability to effectively tackle the nuanced challenges of multilingual, code-mixed hate speech detection.

| Team | F1-Score | Rank |
|---|---|---|
| DCST Unigoa | 0.6155 | 1 |
| NOVA-RMK-ADS | 0.6048 | 2 |
| Radicaldecoders run1 | 0.5947 | 3 |
| **Chakravyuh coders run1** | **0.13** | **13** |

Table 2: LeaderBoard Ranking for Task B

## 6 Conclusion

The system performed well, particularly in distinguishing explicit hate speech and fake content. However, challenges remain in identifying subtle forms of fake narratives and implicit hate speech. The detailed error analysis and confusion matrix show areas where the model misclassified subtle content.


## Acknowledgement

We would like to thank the Faux-Hate Shared Task organizers for providing the dataset and supporting materials. Their contribution has been invaluable in helping us develop and evaluate our system. We also appreciate the feedback and guidance from our mentors and colleagues, which greatly improved the quality of this work.


## References


Rie Kubota Ando, Tong Zhang, and Peter Bartlett. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of machine learning research*, 6(11).

Galen Andrew and Jianfeng Gao. 2007. Scalable training of l 1-regularized log-linear models. In *Proceedings of the 24th international conference on Machine learning*, pages 33–40.

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. 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.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020. Corpus creation for sentiment analysis in code-mixed Tamil-English text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 202–210, Marseille, France. European Language Resources association.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In

*Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Shervin Malmasi and Marcos Zampieri. 2018. Challenges in discriminating profanity from hate speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30(2):187–202.

Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 797–806.

Sunil Saumya, Abhinav Kumar, and Jyoti Prakash Singh. 2024. Filtering offensive language from multilingual social media contents: A deep learning approach. *Engineering Applications of Artificial Intelligence*, 133:108159.

Vivek Srivastava and Mayank Singh. 2022. Code-mixed nlg: Resources, metrics, and challenges. CODS-COMAD '22, page 328–332, New York, NY, USA. Association for Computing Machinery.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. 53(5).