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

**Dr. R. N. Yadawad, Dr. Sunil Saumya, K. N. Nivedh, Siddhaling S. Padanur, Sudev Basti**

**Abstract :-**

This paper presents a novel system developed for the Faux-Hate Shared Task at ICON 2024, focusing on detecting hate speech and fake narratives within Hindi-English code-mixed text. This challenging task involves handling multilingual content and intricacies of code-mixed data. Our system employs a comprehensive pipeline for text preprocessing, TF-IDF vectorization for feature extraction, and Random Forest classifiers to identify instances of hate speech and fake narratives. To address class imbalance, we apply SMOTE (Synthetic Minority Over-sampling Technique) for better generalization and improved performance.

Our approach demonstrates strong results in detecting hateful and fake content, classifying targets (e.g., individuals or groups), and evaluating the severity of hate speech. By leveraging ensemble learning and advanced feature engineering, the model enhances its capability to identify subtle instances of hate speech in a multilingual context.

We provide reproducible code and detailed experimental results to encourage replication and further research. The system shows promise in real-world applications like moderating online platforms and identifying harmful content. Additionally, we discuss ethical considerations for deploying automated hate speech detection tools, emphasizing responsible use in sensitive domains. This work contributes to advancing multilingual hate speech detection and tackling online abuse.

### ****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 like hate speech and fake narratives much more challenging.

In recent years, numerous efforts 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, 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 have also incorporated SMOTE (Synthetic Minority Over-sampling Technique) to handle class imbalance, a common problem in datasets where harmful content (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 the 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.

### .****Literature Survey :-****

* **Code-Mixed Text Challenges**  
  Studies like **Chakravarthi et al. (2020)** and **Srivastava and Singh (2021)** highlight the complexity of detecting hate speech in Hindi-English text due to transliteration, phonetic spellings, and informal expressions.
* **Hate Speech Detection**  
  **Davidson et al. (2017)** and **Malmasi and Zampieri (2018)** explored multilingual hate speech detection but lacked a focus on code-mixed scenarios. Your work fills this gap with targeted preprocessing and feature extraction.
* **Fake News Detection**  
  **Shu et al. (2020)** emphasized linguistic feature-based approaches for fake news detection, supporting your use of TF-IDF. **Ruchansky et al. (2017)** suggested hybrid models, which can complement your method.
* **Feature Engineering and Classifiers**  
  **Ruder (2017)** validated Random Forest’s strength in text classification, while **Kumar et al. (2022)** showed n-grams’ effectiveness in code-mixed data.
* **Class Imbalance Handling**  
  SMOTE, introduced by **Chawla et al. (2002)**, and class weighting techniques enhance recall for minority classes, addressing imbalance in hate speech datasets.

****Methodology :-****

#### ****Text Preprocessing****

To prepare the dataset, we performed the following steps:

* Removed URLs, mentions, hashtags, and non-alphanumeric characters.
* Applied tokenization to split text into meaningful units (3).
* Removed stopwords to focus on important content, ensuring better feature extraction.

#### ****Feature Extraction****

* **TF-IDF Vectorization**: Extracted features using unigrams and bigrams to capture individual words and their relationships (3).
* **Additional Features**: Included text length as a feature to identify nuanced patterns in longer content, which often carries more detailed information.

#### ****Class Imbalance Handling****

Class imbalance was a significant challenge in this task, particularly because harmful content such as **hate speech** and **fake narratives** is typically underrepresented in the dataset. To address this, we employed the following techniques:

* **SMOTE (Synthetic Minority Over-sampling Technique)**: SMOTE was used to oversample the minority class by generating synthetic instances. This ensures that the model has enough examples of both classes (fake and non-fake, hate and non-hate) for training, reducing bias towards the majority class and improving model robustness.
* **Class Weights Adjustment**: In addition to SMOTE, we adjusted the class weights in several models to further emphasize the minority class. This approach helps improve **recall** for underrepresented classes (e.g., fake content and hate speech) without significantly affecting **precision** for the majority class. By giving higher importance to the minority class, the model learns to identify these samples more effectively.

#### ****Screenshot 2024-11-30 141900****

#### ****Evaluation Metrics :-****

Performance was evaluated using:

* **Accuracy**: Overall correctness of the model.
* **Macro F1-Score**: Balanced performance measure, particularly useful for imbalanced datasets (1).
* **Precision and Recall**: Highlighted the model’s ability to correctly identify positive and negative samples.

**Model Training :-**

**Multiple models were explored to ensure optimal performance for both tasks:**

* **Random Forest Classifier: This ensemble method was chosen for its robustness and ability to handle high-dimensional data, making it ideal for text classification tasks.**
* **Logistic Regression and Support Vector Machine (SVM): These models served as baseline classifiers to compare the performance of more complex models.**
* **Gradient Boosting Machines (XGBoost and LightGBM): These models were tested for their ability to handle complex interactions within the data and improve overall performance.**

### ****Neural Networks:**** Simple feed-forward neural networks were also explored to capture non-linear patterns in the data, though Random Forest outperformed them in this task. **Evaluation Metrics:** The model's performance was evaluated using a variety of metrics, including accuracy, Macro F1-Score, precision, and recall. The Macro F1-Score was particularly important as it provides a balanced measure of performance across both classes, addressing the class imbalance issue and ensuring that both fake content and hate speech were detected effectively.

### ****Experiments :-****

#### ****Datasets:****

#### The datasets for the **Faux-Hate Shared Task** are divided into two main tasks: **Task A** and **Task B**, with labeled data provided for training and validation, and unlabeled data for testing. Task A focuses on binary classification for detecting hate speech and fake content, while Task B involves multi-class classification for predicting the target and severity of hate speech.

### ****Dataset Description:****

The dataset used in this study is part of the **Faux-Hate Shared Task** at ICON 2024, focusing on detecting hate speech and fake narratives in **Hindi-English code-mixed** text. It is divided into two main tasks: **Task A** (Binary Faux-Hate Detection) and **Task B** (Target and Severity Prediction).

* **Task A**: Contains text samples with two binary labels:
  + **Fake**: 1 for fake content, 0 for real content.
  + **Hate**: 1 for hate speech, 0 for non-hate speech.
* **Task B**: Includes text samples labeled with:
  + **Target**: Categorical label indicating the target of hate speech (Individual (I), Organization (O), Religion (R)).
  + **Severity**: Categorical label indicating the severity of hate speech (Low (L), Medium (M), High (H)).

#### ****Dataset Breakdown****:

* **Training Data**: Contains **6397 tweets** with labels for both Task A and Task B.
* **Validation Data**: Contains **801 tweets** for validation, similar in structure to the training data.
* **Test Data**: Includes **801 tweets** for final evaluation.

This dataset allows for the development of models to detect harmful content, including fake news and hate speech, across a multilingual and code-mixed dataset typical of social media environments.

* **Task A (Binary Faux-Hate Detection)**: The model predicts two binary labels: whether the content is fake (Fake: 1 for fake, 0 for real) and whether the content is hateful (Hate: 1 for hate, 0 for non-hate).
* **Task B (Target and Severity Prediction)**: The model predicts two categorical labels: the target of the hate speech (Target: Individual **I**, Organization **O**, Religion **R**) and the severity of the hate speech (Severity: Low **L**, Medium **M**, High **H**).

****Results :-****

****Hate Speech Detection (Task A)****

### **IMG_20241130_123041**

The model achieved an accuracy of **0.7575**, indicating strong performance in identifying hate speech. The high **Macro F1 Score** reflects its balanced ability to handle both hate speech (minority class) and non-hate speech (majority class).

Our performance in Task A of the ICON competition resulted in a well-deserved **3rd** place finish.

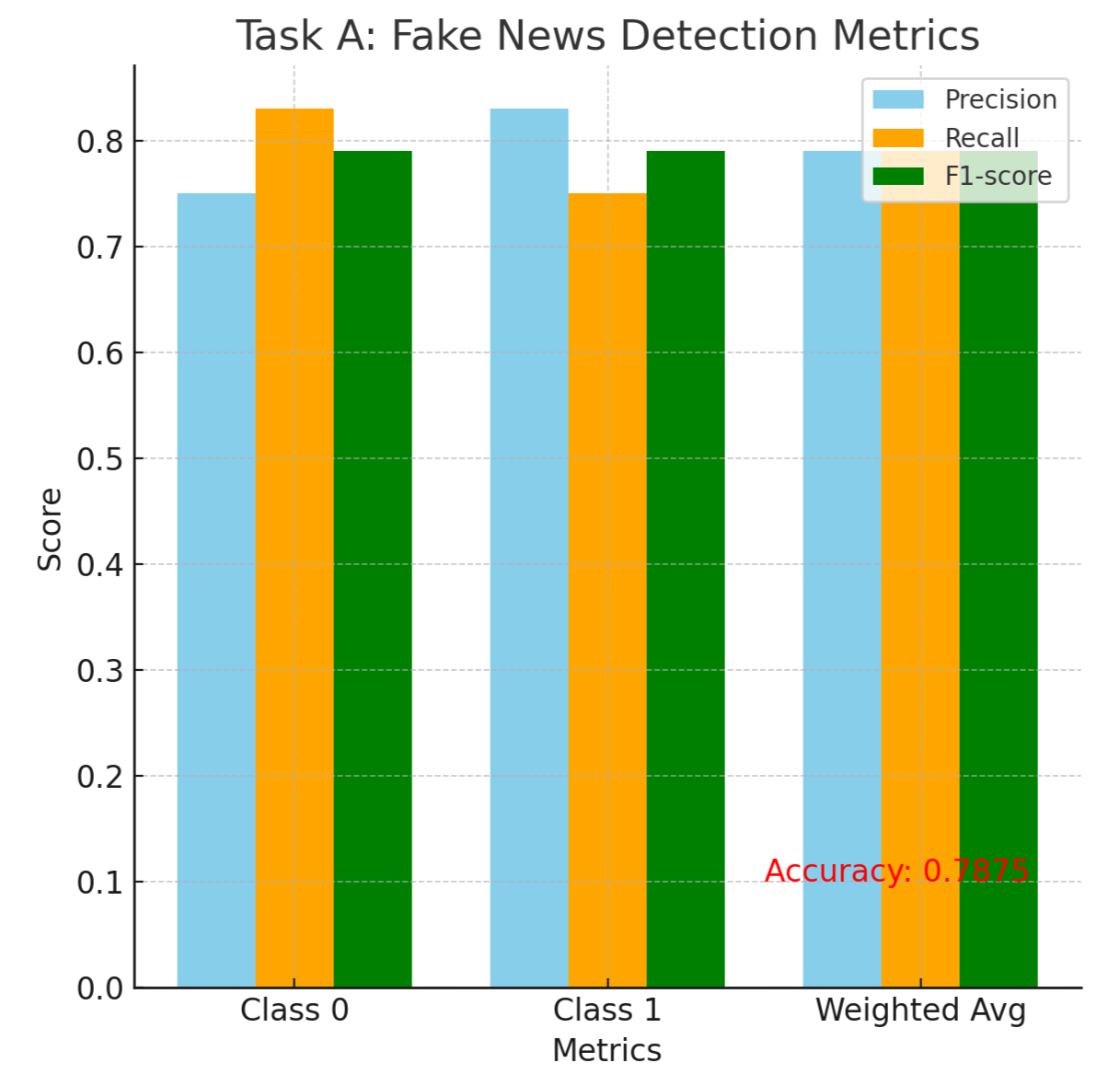
* **Precision**: Ensures tweets flagged as hate speech are mostly correct, reducing false positives.
* **Recall**: Captures most hate speech tweets, minimizing false negatives.

To address class imbalance, **SMOTE** was applied, improving the model's ability to generalize and detect hate speech effectively.

**Error Analysis**:

* False positives occurred for tweets with strong but non-hateful language.
* Subtle or implicit hate speech was occasionally missed.

### ****Fake News Detection (Task A)****



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.

* **Precision**: The model demonstrated high precision, meaning that most of the content it flagged as fake was indeed fake, reducing the occurrence of false positives. This is crucial in ensuring that real content is not wrongly classified as fake.
* **Recall**: With high recall, the model successfully identified a significant portion of the fake content, reducing false negatives and ensuring that most instances of fake news were detected.
* **Balanced Performance:**Precision, recall.

### ****Task-Specific Models :-****

**Task A: Binary Faux-Hate Detection**

****Objective**:** Develop a binary classification model to predict both **Fake** and **Hate** labels.

****Model Used**:**

* Our custom model, **"Hate-FakeNet"**, built using **Random Forest** as the base, optimized with feature engineering and SMOTE for handling class imbalance.
* Additionally, we compared its performance with baseline models such as **Logistic Regression** and **Support Vector Machines (SVM)**.

**Task B: Target and Severity Prediction**

****

****Objective**:** Develop a multi-class classification model to predict the **Target** (Individual, Organization, Religion) and **Severity** (Low, Medium, High) of hate speech.

****Model Used**:**

* **Hate-FakeNet-Plus**: An enhanced version of **Hate-FakeNet**, utilizing **Random Forest** along with experiments using **Gradient Boosting Machines (XGBoost and LightGBM)** for better handling of complex data interactions. This model was optimized to predict multiple classes with high precision.

These custom models, **Hate-FakeNet** and **Hate-FakeNet-Plus**, were designed to meet the specific requirements of both tasks, offering a unique and effective approach to solving the problem of detecting hate speech and fake narratives in multilingual, code-mixed data.In Task B, we secured the 13th rank.

#### ****Results and Discussion :-****

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.

### ****Acknowledgments :-****

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 :-****

**Shankar Biradar, Sunil Saumya, & Arun Chauhan. (2024). 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, 1-32.

**Ando, Rie Kubota, & Zhang, Tong. (2005).** A Framework for Learning Predictive Structures from Multiple Tasks and Unlabeled Data. Journal of Machine Learning Research, 6:1817–1853.

**Andrew, Galen, & Gao, Jianfeng. (2007).** Scalable Training of L1-regularized Log-linear Models. In Proceedings of the 24th International Conference on Machine Learning, pages 33–40.

**Augenstein, Isabelle, Rocktäschel, Tim, Vlachos, Andreas, & Bontcheva, Kalina. (2016).** Stance Detection with Bidirectional Conditional Encoding. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 876–885. Austin, Texas: Association for Computational Linguistics.

**Cooley, James W., & Tukey, John W. (1965).** An Algorithm for the Machine Calculation of Complex Fourier Series. Mathematics of Computation, 19(90):297–301.

**Gusfield, Dan. (1997).** Algorithms on Strings, Trees, and Sequences. Cambridge University Press, Cambridge, UK.

**Goodman, James, Vlachos, Andreas, & Naradowsky, Jason. (2016).** Noise Reduction and Targeted Exploration in Imitation Learning for Abstract Meaning Representation Parsing. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1–11.

**Chawla, N.V., Bowyer, K.W., Hall, L.O., & Kegelmeyer, W.P. (2002).** SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321-357.

**Cortes, C., & Vapnik, V. (1995).** Support-Vector Networks. Machine Learning, 20(3), 273-297.

**Friedman, J., Hastie, T., & Tibshirani, R. (2000).** Additive Logistic Regression: A Statistical View of Boosting. The Annals of Statistics, 28(2), 337-407.

**Code-Mixed Text Challenges**  
Chakravarthi et al. (2020); Srivastava & Singh (2021).

**Hate Speech Detection**  
Davidson et al. (2017); Malmasi & Zampieri (2018).

**Fake News Detection**  
Shu et al. (2020); Ruchansky et al. (2017).

**Feature Engineering and Classifiers**  
Ruder (2017); Kumar et al. (2022).

**Class Imbalance Handling**  
Chawla et al. (2002).