Challenges and Insights in Identifying Hate Speech and Fake News on Social Media

Shanthi Murugan, Arthi R, Boomika E, Jeyanth S, Kaviyarasu S
R.M.K. Engineering College, Tiruvallur, Tamilnadu, India
{msi, arth22004, boom22011, jeya22018, kavi22022}.ad@rmkec.ac.in

Abstract

Social media has transformed communication, but it has also brought about a number of serious problems, most notably the proliferation of hate speech and false information. hate-related conversations are frequently fueled by misleading narratives. We address this issue by building a multiclass classification model trained on Faux Hate Multi-Label Dataset (Biradar et al. 2024) which consists of hateful remarks that are fraudulent and have a code mix of Hindi and English. Model has been built to classify Severity (Low, Medium, High) and Target (Individual, Organization, Religion) on the dataset. Performance of the model is evaluated on test dataset achieved varying scored for each. For Severity model achieves 74%, for Target model achieves 74%. The limitations and performance issues of the model has been understood and well explained.

1 Introduction

Social media platforms make it easier to communicate and share information, eventually we could see hate speech and fake news in Social Media posts, comments and blogs. Digital platforms have serious problems from hate speech, which aims to marginalize individuals or groups based on characteristics like religion, race, or political ideas, and fake news, which is defined by intentional disinformation intended to deceive. These problems worsen social and political conflicts in addition to lowering the caliber of information that is accessible online. Natural language processing (NLP) extends its application over social media narratives to address these issues by classifying them. We build a classification system with Multinomial Naive Bayes (MNB), a robust machine learning model for classification, has shown effectiveness in managing high-dimensional data, including text. We found MNB is a desirable option for addressing issues like hate speech identification and false news.

1) Task 1: We use Multinomial Naive Bayes (MNB) to classify hate and disinformation content.
2) Task 2: The same MNB algorithm with different preprocessing method is used for predicting targets, target class includes individual, groups, and region. and the intensity of hate speech.

The results highlight the significance of lightweight and interpretable models in practical applications while also enhancing the accuracy of automated systems intended to protect digital areas.

2 Related Work

Recent research highlights the intersection of hate speech and fake news, particularly in low-resource languages like Hindi-English code-mixed texts, which remain underexplored. While datasets like those from CONSTRAINT-2021 (Bhardwaj et al., 2020) and FactDRIL (Singhal et al., 2021) have addressed hate speech and fake news in Hindi separately, efforts to bridge both phenomena in multilingual contexts are limited. Contributions such as Bohra et al.'s (2018) binary classification dataset for hate speech in Hindi-English code-mixed text andw Mathur et al.'s (2018) HEOT dataset address class imbalance and real-world representation issues but remain focused on monolingual or binary tasks.

The Faux Hate Multi-Label Dataset (FHMLD) aims to address these gaps by incorporating multi-label and multi-class annotations, enabling advanced

models to tackle both hate speech and fake news simultaneously in Hindi-English code-mixed texts. Deep learning and transformer-based approaches, such as CNN-LSTM with embeddings (Mathur et al., 2018; Fharook et al., 2022), RNN-based models (Bisht et al., 2020), and refined transformers like mBERT and XML-R (Banerjee et al., 2021; Farooqi et al., 2021), have shown promise in handling such texts. However, cross-lingual and context-aware models, including BERT, ELMo, and FLAIR, have achieved macro F1 scores of around 0.71 but are yet to address the dual challenge effectively.

3 Dataset resource and data processing

To study issues such as fake news, hate speech, and harmful content, we utilized existing datasets curated from environments known for the rapid spread of misinformation and harmful rhetoric, provide a robust foundation for analyzing these critical issues. The work by Biradar et al. (2024) introduces a code-mixed dataset (Hindi - English) as part of the shared task on "Decoding Fake Narratives in Spreading Hateful Stories (Faux-Hate)." This dataset focuses on identifying and analyzing fake narratives intertwined with hateful rhetoric, making it a valuable resource for studying misinformation and hate speech in natural language.

3.1 Data Refinement

The text was tokenized, breaking down the sentences into individual words and phrases for easier analysis. Labels were assigned to each piece of content based on predefined categories such as fake, hate speech, target, severity. This step ensured that each text sample was appropriately categorized for training purposes. we represent the sentence as features by using Term Frequency and Inverse Document Frequency (TF-IDF) Vectorization method.

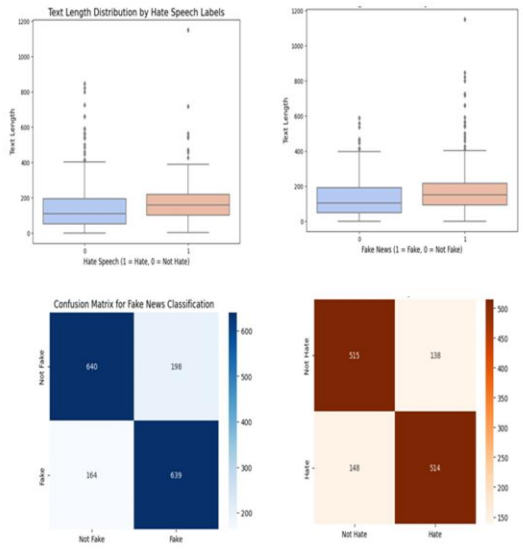
3.2 Data Analysis

The dataset includes four labels for analyzing content. Fake, which indicates whether the content is fake (1) or not fake (0); Hate, representing whether the content contains hate speech (1) or not (0). Target, which includes Individual, Group or Religious represent the target of the hateful speech. Severity labels include High, Medium, Low. To understand the distribution of these labels, we analyzed their frequency across all the samples in the dataset.

4 Methodology

4.1 Fake News and Hate Speech Classification

Visualizing data using boxplot and confusion matrix shows how text lengths (in terms of characters) are distributed between two groups: Fake News (label 1) and Not Fake News (label 0). These visual tools provide insights into potential correlations between text length and classification labels by highlighting the median, interquartile range and any outliers. From Figure 1, We can clearly observe how our dataset is distributed.



.

Figure 1: Distribution of fake news and hate speech labels

4.1.1 Training of Models

The methodology starts cleaning the text by removing special characters and mentions. Features are extracted using TF-IDF (Term Frequency-Inverse Document Frequency). Vectorization with n-gram models to represent text numerically.

SMOTE (Synthetic Minority Over-sampling Technique) is applied to address class imbalance. Multinomial Naive Bayes is chosen for its efficiency with text-based multi-tasking model, Feature selection is performed using SelectKBest and the Chi-square test to identify the most relevant features.

4.1.2 Model Performance Evaluation Metrics

Assessing the model's performance is essential to comprehending how well it works in practical applications. Additional metrics like precision, recall, F1-score, and AUC-ROC are crucial because traditional measurements like accuracy might not be enough in the situation of unbalanced data.

4.1.3 Area Under the Curve, or AUC, and the ROC Curve

The ROC Curve, which displays the balance between True Positive Rate (right detections) and The ROC Curve, which displays the balance between True Positive Rate (right detections) and False Positive Rate (false alarms), aids in assessing the model's capacity to discriminate between hate speech and non-hate speech or fake and true news. In this project, it is particularly helpful to evaluate how effectively the model handles unbalanced datasets, since a higher AUC (Area Under the Curve) in Figure 2 implies better performance. This ensures accurate categorization of both minority classes (false or hate speech) and majority classes (genuine or non-hate material).

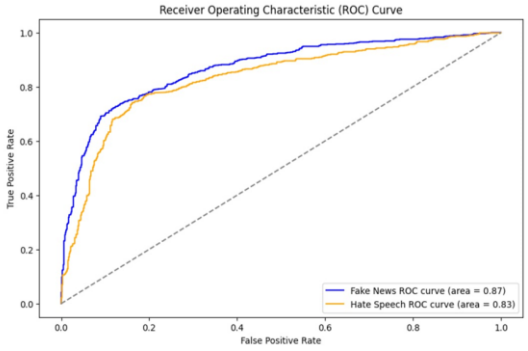


Figure 2: AUC and the ROC Curve

4.1.4 Curve of Precision-Recall

Precision-recall curves offer a clear picture of the model's performance on imbalanced datasets by highlighting its ability to handle minority classes. Precision measures the accuracy of positive predictions, while recall evaluates the ability to retrieve all relevant events. These curves aid in fine-tuning the model to strike a balance between recall and accuracy, which is essential for tasks like classifying bogus news and hate speech.

Learning curves show how the model'

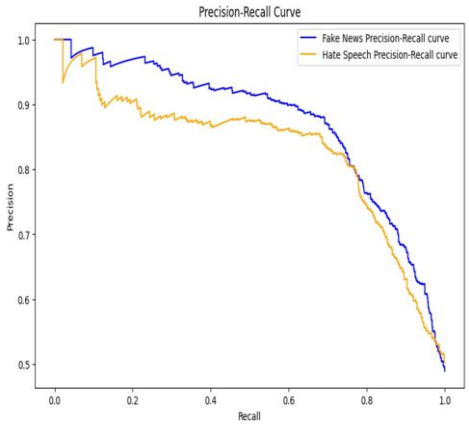


Figure 3: Precision-Recall Curve

Distribution of Classes

Significance: Class inequality is common in datasets on hate speech and fake news. Visualizing the distributions of hate vs. non-hate speech and fake vs. true news as in Figure 4 ensures both before and after oversampling (e.g., SMOTE, Random Oversampling), the model isn't biased towards the majority class. Use: These visualizations help evaluate whether oversampling methods effectively balance class distribution, enabling the model to learn from the minority class.

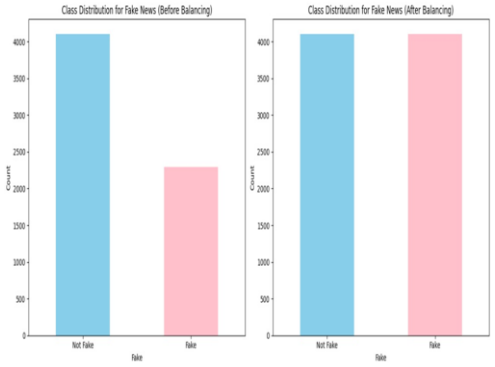


Figure 4: Class Distribution before and after balancing

4.2 Target and severity classification

Target Classification and Severity Classification are the two primary goals of the classification task. Accurately forecasting the intended outcome and its degree of severity are the main goals of these assignments. Several machine learning models were used to guarantee reliable performance, and their efficacy was examined using a range of assessment metrics and visualizations.

4.2.1. Models and Techniques Utilized

The dataset preparation involves importing, cleaning, and preprocessing text by removing irrelevant elements such as stop words and punctuation, followed by tokenizing text into words for analysis. Feature engineering is performed by transforming textual data into numerical formats using TF-IDF (Term Frequency-Inverse Document Frequency). Ensembled model (Support Vector Machine, and Random Forest Classifier) has been used for increased accuracy by averaging property. The F1 score provides a balanced metric that takes into consideration both false positives and false negatives. It is calculated as the harmonic mean of precision and recall. In situations where the data is unbalanced, this statistic is especially helpful.

Classification of the Target

The ability of the model to differentiate across target classes is demonstrated by the matrix. It detects false negatives (missed targets) and false positives (non-targets labelled as targets). This can be observed with the help of visuals as in Figure 5.

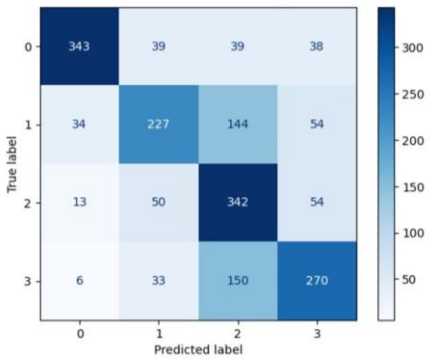


Figure 5: Confusion Matrix for Target

Classification of Severity

The model's ability to predict severity levels is demonstrated in the matrix in Figure 6. Misclassifications between severity levels may reveal areas in which the model is unable to distinguish subtle differences.

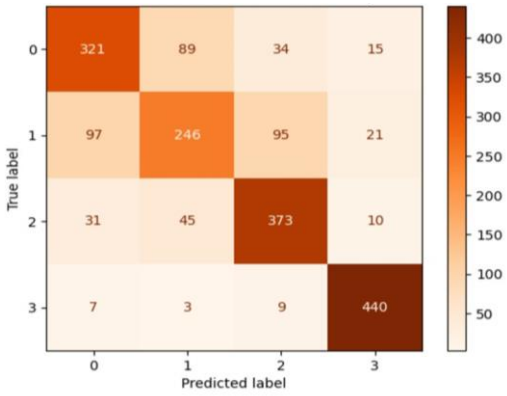


Figure 6: Confusion matrix for severity

The curve in Figure 7 demonstrates how well the model predicts different severity levels. A significant increase in the curve close to the upper-left corner indicates excellent performance.

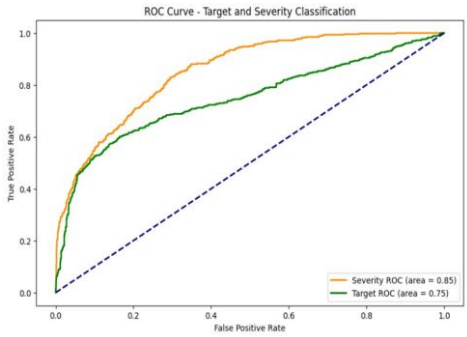


Figure 7: Roc curve for Target and Severity classification

Prediction Probability Histogram

The model's level of prediction confidence is shown graphically via the histogram in Figure 8 the predicted probability. To examine each class's projected probability distribution. It is useful to know if the model typically generates predictions that are uncertain (values near 0.5) or confident (values close to 0 or 1). For Target Classification, the model shows high confidence in its predictions if the probabilities cluster around extreme values (0 or 1). For Severity Classification: Problems distinguishing between severity levels may be highlighted by overlaps across classes in probability distributions.

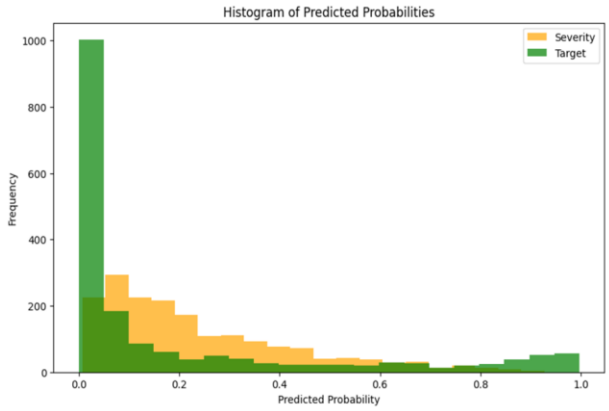


Figure 8: Histogram of Prediction Probabilities

5 Result and Findings

The performance of the model was evaluated across four key labels: Fake News, Hate Speech, Target, and Severity. Below is a summary of the results based on the metrics of Accuracy, Precision, Recall, and F1-score.

labels	Accuracy	Precision	recall	F1 score
Fake News	0.7794	0.78	0.78	0.78
Hate Speech	0.7825	0.78	0.78	0.78
Target	0.7412	0.74	0.74	0.74
Severity	0.7418	0.74	0.74	0.74

6 Conclusion

The overlap between hate speech and fake news in social media narratives necessitates integrated approaches for detection and mitigation. Our work bridges this gap by presenting a classification framework utilizing a novel Hindi-English code-mixed dataset, enabling nuanced categorization based on severity and targets. Despite achieving promising results, challenges remain in scaling to multilingual contexts, addressing computational constraints, and refining interpretability for real-world applications. These findings highlight the importance of continued research and dataset evolution to combat hate speech and misinformation effectively.

7 Limitations

This study's focus on English-Hindi code-mixed datasets limits its broader applicability to other multilingual or low-resource languages. While oversampling techniques like SMOTE help address class imbalances, they may introduce noise, leading to overfitting in minority classes. The reliance on static datasets risks model bias, impacting fairness and adaptability to evolving trends. Additionally, handling nuanced language features such as sarcasm, contextual meanings, and long or complex texts poses significant challenges. Computational resource demands also restrict the deployment in low-resource environments, and the ensemble model's lack of interpretability complicates practical applications. Ethical concerns, including privacy issues and potential censorship, further emphasize the need for cautious implementation.

References

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

Biradar, Shankar, Saumya, Sunil, and Chauhan, Arun. 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. Springer, pages 1–32.

Jafri, F. A., Siddiqui, M. A., Thapa, S., Rauniyar, K., Naseem, U., & Razzak, I. (2023). Uncovering political hate speech during Indian election campaign: A new low-resource dataset and baselines. arXiv preprint. arXiv:2306.14764.

Singh, G., & Selva, K. (2023). A comparative study of hybrid machine learning approaches for fake news detection that combine multi-stage ensemble learning and NLP-based framework. Unpublished Work.

Amutha, R., & Kumar, D. V. (2021). Ensemble-based classification of dynamic rumor detection in social networks for green communication.

Journal of Green Engineering, 11(2), 1220–1243.

Caselli, T., Basile, V., Mitrović, J., & Granitzer, M. (2021). HateBERT: Retraining BERT for abusive language detection in English. In Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021) (pp. 17–25).

Shvets, A., Fortuna, P., Soler, J., & Wanner, L. (2021). Targets and aspects in social media hate speech. In Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021) (pp. 179–190).

Singhal, S., Shah, R. R., & Kumaraguru, P. (2021). Factorization of fact-checks for low-resource Indian languages. arXiv preprint. arXiv:2102.11276.

Kar, D., Bhardwaj, M., Samanta, S., & Azad, A. P. (2021). No rumours please! A multi-indic-lingual approach for COVID fake-tweet detection. In 2021 Grace Hopper Celebration India (GHCI) (pp. 1–5). IEEE.

Chopra, S., Sawhney, R., Mathur, P., & Shah, R. R. (2020). Hindi-English hate speech detection: Author profiling, debiasing, and practical perspectives. Proceedings of the AAAI Conference on Artificial Intelligence, 34, 386–393.

Mandl, T., Modha, S., Kumar, M. A., & Chakravarthi, B. R. (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 (pp. 29–32).

Ozbay, F. A., & Alatas, B. (2020). Fake news detection within online social media using supervised artificial intelligence algorithms. Physica A: Statistical Mechanics and its Applications, 540, 123174.

Faustini, P. H. A., & Covoes, T. F. (2020). Fake news detection in multiple platforms and languages. Expert Systems with Applications, 158, 113503.

Liu, Y., & Wu, Y.-F. B. (2020). FNED: A deep network for fake news early detection on social media. ACM Transactions on Information Systems (TOIS), 38(3), 1–33.

Bohra, A., Vijay, D., Singh, V., Akhtar, S. S., & Shrivastava, M. (2018). A dataset of Hindi-English code-mixed social media text for hate speech detection. In Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (pp. 36–41).

Gollatz, K., & Jenner, L. (2018). Hate speech and fake news—how two concepts got intertwined and politicised. HIIG Digital Society Blog.

Bojanowski, P., Grave, É., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5, 135–146.

Weiss, S. M., et al. (2010). Text mining: Predictive methods for analyzing unstructured information. Springer Science & Business Media.