Shared Feature-Based Multitask Model for Faux-Hate Classification in Code-Mixed Text

Sanjana Kavatagi and **Rashmi Rachh** and **Prakul Hiremath**
Department of Computer Science and Engineering
Visvesvaraya Technological University, Belagavi
{kawatagi.sanjana, rashmirachh, prakulhiremath03}@gmail.com

Abstract

In recent years, the rise of harmful narratives online has highlighted the need for advanced hate speech detection models. One emerging challenge is the phenomenon of Faux Hate, a new type of hate speech that originates from the intersection of fake narratives and hate speech. Faux Hate occurs when fabricated claims fuel the generation of hateful language, often blurring the line between misinformation and malicious intent. Identifying such speech becomes especially difficult when the fake claim itself is not immediately apparent. This paper provides an overview of a shared task competition focused on detecting Faux Hate, where participants were tasked with developing methodologies to identify this nuanced form of harmful speech.

1 Introduction

The rapid increase in social media platform users has changed how information is created, shared, and consumed. This has significantly boosted both the volume and speed of information dissemination. Social media has become a crucial tool for modern communication, providing instant access to real-time updates and facilitating seamless content sharing. While this shift has made information more accessible and enhanced global connectivity, it has also led to challenges, particularly the spread of misinformation and disinformation (Biradar et al., 2023).

Among the most concerning forms of misinformation is fake news, which refers to deliberately fabricated or misleading content presented as legitimate news. The rapid spread of fake news is often exacerbated by social media algorithms, which prioritize engagement metrics over the accuracy of the information (Akash et al., 2021). The situation becomes even more alarming when fake news intersects with hate speech, forming a potent combination known as ”fake instigated hate news” (Gollatz and Jenner, 2018). This category encompasses fabricated or distorted content designed to incite hostility, discrimination, or violence against individuals or groups based on attributes such as race, religion, gender, or political beliefs. Hate speech, as defined by various legal and academic frameworks, refers to any speech, gesture, conduct, writing, or display that incites violence or prejudicial actions against others based on their identity or group membership (Shaik et al., 2024; Sai et al., 2024; Biradar et al., 2022; Kavatagi and Rachh, 2021). It perpetuates harmful stereotypes, dehumanizes marginalized groups, and exacerbates societal divisions.

The combination of fake news and hate speech can amplify social tensions, leading to real-world consequences such as violence, discrimination, and the erosion of social trust. Its widespread dissemination complicates efforts to maintain public safety and social cohesion, making it crucial to develop robust mechanisms for its detection and regulation (Biradar et al., 2024b). This dual threat of misinformation and hate speech creates unique challenges for law enforcement, policymakers, and researchers. Such content is not only deceptive but also emotionally provocative, often capitalizing on societal tensions. The deliberate creation and spread of such material threaten the social fabric, underscoring the urgent need for effective detection and mitigation mechanisms. In multilingual societies like India, code-mixed text—a blend of two or more languages within a single discourse—adds an additional layer of complexity to identifying fake news and hate speech (Bali et al., 2014). Hindi-English code-mixed communication is particularly prevalent on social media platforms, where users frequently switch between languages for expression and convenience. The informal and inconsistent nature of code-mixed text, transliteration, and frequent spelling variations pose significant challenges for traditional natural language processing

(NLP) techniques. Addressing these challenges requires sophisticated computational approaches that can effectively capture the linguistic nuances of such hybrid text. This study proposes a multitask learning model that leverages shared linguistic features to simultaneously classify code-mixed tweets as Hate vs. Non-hate and Fake vs. Non-fake. By focusing on this multi-task framework, the study aims to advance the development of effective tools to mitigate the spread of harmful content on social media.

2 Methodology

This study presents a multitasking classification model to address the challenges of detecting hate speech and fake content in Hindi-English code-mixed text. The methodology combines effective preprocessing techniques with fastText embeddings to create a shared vector space that captures code-mixed data's semantic and syntactic nuances. These shared embeddings serve as input to a multi-task framework comprising two independent Support Vector Machine (SVM) classifiers. These classifiers are designed to simultaneously categorize tweets as Hate vs. Non-hate and Fake vs. Non-fake. This approach utilizes shared linguistic features across the tasks to improve classification accuracy and computational efficiency.

2.1 Task and dataset description

The Faux-Hate shared task provides a unique platform to address the issues of fake instigated hate statements by focusing on Hindi-English code-mixed social media content. This section details the task objectives and describes the dataset used to train and evaluate models designed to effectively identify and classify hate speech and fake narratives.

2.1.1 Task description

The organizers have structured the Faux-hate shared task to challenge participants with the dual objectives of detecting both fake and hate speech in social media comments, focusing on the target and severity of hateful language (Biradar et al., 2024a). The task is divided into two subtasks. Subtask A involves binary hate speech detection, where participants receive a dataset of tweets, each labeled as either hate or non-hate and as fake or real, using categorical labels. Participants must develop a single multitasking model that can accurately identify both the fake and hate labels for each tweet.

	Hate	Non-hate	Fake	Non-fake
Train	4100	2295	3285	3110
Validation	513	287	423	377
Test	800			

Table 1: Dataset description

In subtask B, participants were given a dataset of tweets, each labeled to indicate its target: individual, organization, or religion. Additionally, each tweet was classified based on severity into three categories: low, medium, or high. Participants are tasked with developing a single model that can generate both the target and severity labels for the provided tweet samples.

2.1.2 Dataset description

The dataset for Shared Task A encompasses labeled data specifically designed for two critical binary classification tasks: identifying hate speech versus non-hate speech and discernment between fake news and legitimate news. Table 1 provides a detailed breakdown of the distribution of instances across the training, validation, and testing sets. In the training set, 4,100 records are categorized as hate speech, contrasted with 2,295 records classified as non-hate speech. This results in the dataset exhibiting a noticeable skew in favor of the hate class, which may impact model performance. Conversely, when it comes to classifying fake news, the training data presents a more balanced outlook, featuring 3,285 instances of fake news and 3,110 instances of non-fake news. The validation set further reflects this trend, consisting of 513 records designated as hate speech and 287 as non-hate speech, echoing the imbalanced distribution observed in the training data. For the validity of news classification within the validation set, 423 instances are labeled as fake news and 377 as non-fake news, ensuring consistency with the distribution patterns noted in the training dataset. Finally, the test set comprises 800 records that await evaluation. These instances will be assessed through our model to determine their appropriate labels, ultimately measuring the effectiveness of our classification approach.

2.1.3 Pre-processing

The dataset has undergone thorough preprocessing and is now ready for feature extraction. The tweets have been normalized: all words have been converted to lowercase, URLs have been removed, special characters have been eliminated, and numbers have been stripped away to reduce noise. We have tokenized the tweets into raw tokens, followed by the removal of stop words to eliminate irrelevant terms. Finally, lemmatization has been applied to reduce words to their base forms, ensuring a consistent vocabulary.

2.1.4 Feature extraction

The multitasking model employs fastText embeddings to tackle the unique challenges associated with processing Hindi-English code-mixed text. These challenges include linguistic diversity, transliteration, and variations in morphology. In this study, fastText embeddings convert the preprocessed text into fixed-length feature vectors, which serve as input for the multitasking classification framework. The detailed architecture of the proposed model is depicted in Figure 1. fastText, enhances traditional word embedding techniques by incorporating subword information, making it particularly effective for modeling morphologically rich and code-mixed languages. This approach provides robustness against spelling variations, inflections, and morphological changes. fastText generates word embeddings by breaking words down into character-level n-grams instead of treating them as individual units. For instance, the word ”pyaar” is decomposed into subwords such as ”pya,” ”yaa,” and ”aar,” allowing for meaningful representations of transliterated or out-of-vocabulary (OOV) words. This technique effectively captures both semantic and syntactic relationships across different languages. For example, the English word ”hate” and its Hindi equivalent ”nafrat” (including transliterated forms) can share similar vector representations based on their context. By creating a shared vector space, fastText harmonizes linguistic features from both Hindi and English, ensuring a unified representation of code-mixed text. The embeddings also maintain contextual relationships through the skip-gram model, where each word is predicted based on its surrounding context. These dense vector representations are computationally efficient and encode rich semantic information. Ultimately, this representation enables the effective identification of hate speech and fake content in the Hindi-English code-mixed Faux-Hate dataset. The word embeddings in the dataset are transformed into a unified vector space, ensuring that the linguistic features of codemixed text is aligned. This common representation is then used as the input for the multitasking classification framework.

2.1.5 Multi-task classifier

The shared vector space created by fastText embeddings is incorporated into a multitasking classification framework that consists of two independent Support Vector Machine (SVM) classifiers. This framework is designed to classify tweets simultaneously as either Hate or Non-hate and Fake or Non-fake, leveraging common linguistic features from both tasks. The first SVM classifier determines whether a tweet contains hate speech by analyzing linguistic cues such as offensive keywords, sentiment, and contextual relationships captured by the fastText embeddings. The second SVM classifier operates within the same shared vector space. It identifies whether a tweet contains fake information by examining features like unusual patterns, contextual mismatches, and linguistic inconsistencies encoded in the embeddings. This shared feature representation allows the model to use interrelated linguistic patterns across both tasks. For example, tweets classified as ”hate” often exhibit stylistic or lexical characteristics that may also suggest they are spreading fake content. By utilizing a shared vector space, the framework reduces redundancy, enhances computational efficiency, and improves generalization, enabling the model to manage complex inter-dependencies between the tasks effectively.

3 Results and discussion

In this section, we present the results obtained for the multitasking model developed for identifying binary hate and fake tweets simultaneously for subtask A. Organizers have used macro F1 score to evaluate the labels presented by the participants for the test dataset. The proposed multitasking model’s performance on the Faux-Hate shared task dataset demonstrates its ability to effectively classify binary hate or non-hate and fake and non-fake labels. Table 2 summarizes the F1 Scores and accuracy metrics for the validation and test datasets.

On the validation dataset, the proposed multitasking model demonstrated acceptable performance across both tasks. For the hate class, the model achieved an F1-score of 0.76, while for the non-hate class, it achieved an F1-score of 0.80 with an accuracy of 0.78, indicating its ability to identify hate content with a reasonable balance of precision and recall. For the fake class, the model obtained an F1-score of 0.79, and for the non-fake class,

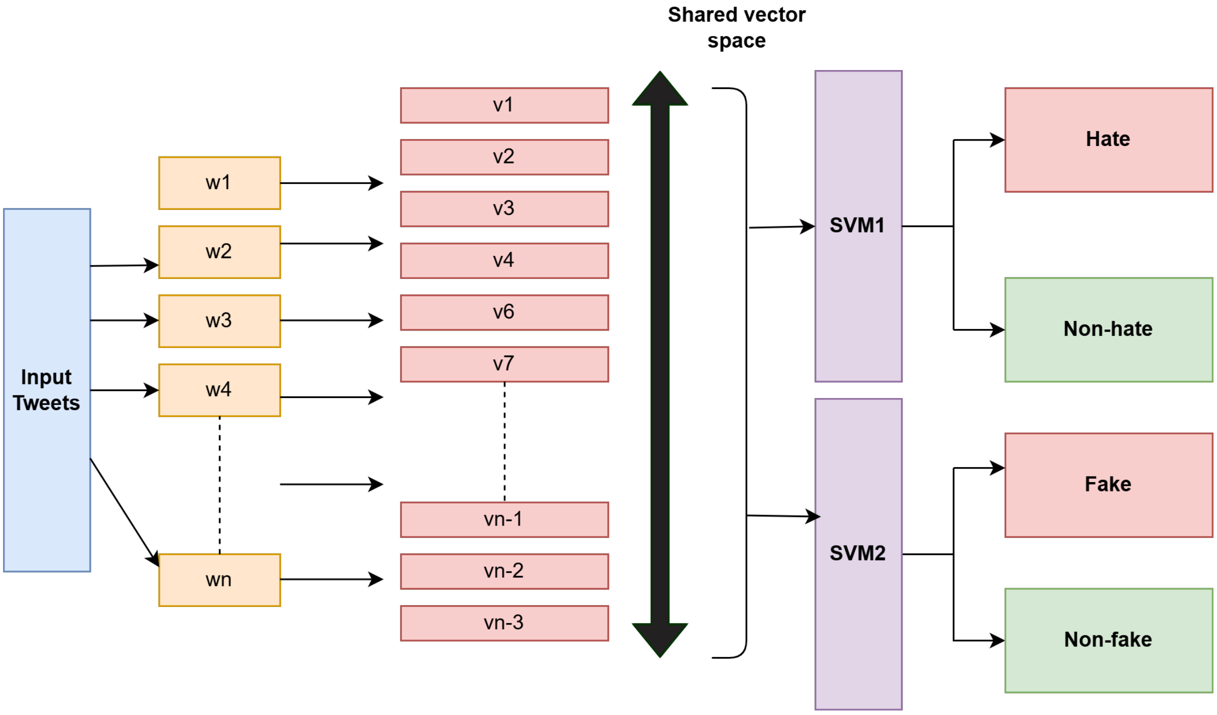

Figure 1: Architecture of the proposed multitasking model with a shared feature space.

Table 2: Results obtained for the multitask model

Class	Validation		Test	
	F1-Score	Acc	F1-Score	Acc
Hate	0.76	0.78	0.73	0.69
Non-hate	0.80		0.76	
Fake	0.79	0.81	0.72	0.71
Non-fake	0.82		0.75	

it achieved an F1-score of 0.82 with an accuracy of 0.81, showcasing its effectiveness in identifying false narratives. Notably, the non-fake class achieved the highest validation performance, with an F1-score of 0.82, reflecting the model's consistent capacity to detect real content.

On the test dataset, the model's performance indicates its ability to generalize effectively, albeit with some performance drops. For the hate class, the model recorded an F1-score of 0.73 with an accuracy of 0.69, reflecting a slight decrease compared to the validation results. This drop highlights the challenges of detecting hate content in previously unseen data, potentially due to variations in linguistic expressions or contextual nuances. For the non-hate class, the model achieved an F1-score of 0.76, which, while slightly lower than its validation performance, demonstrates its consistent ability to identify non-hate content. In the fake class, the model achieved an F1-score of 0.72, while for the non-fake class, it recorded an F1-score of 0.75 with an accuracy of 0.71. These results reflect a performance drop compared to the validation dataset, highlighting the complexities involved in identifying deceptive narratives, especially in diverse and informal text. This suggests that additional strategies, such as domain-specific feature enhancement or improved contextual representations, may be required to enhance the model's robustness further.

Table 3: Top 3 best-performing teams in the leaderboard for subtask A

Team	Macro-F1 Score	Rank
DCST_unigoa	0.79	1
Radicaldecoders	0.7761	2
Chakravyuh coders	0.7721	3
VTUBGM	**0.7445**	**10**

Table 3 showcases the performance of the top three teams in the ICON 2024 Faux Hate Subtask A based on their Macro F1scores and ranks. The winning team, DCST_unigoa, achieved the highest Macro F1-score of 0.79. Our model, VTU_BGM, achieved a Macro F1score of 0.7445, ranking 10th in the overall leaderboard. While the score reflects a competitive performance, it also highlights potential areas for improvement.

4 Conclusion

This work presents a multitasking model for identifying hate and fake content within the Faux Hate Shared Task dataset. The proposed approach leverages fastText embeddings with a shared feature space to address binary hate and fake classification tasks simultaneously. The model demonstrated competitive performance on validation and test datasets, achieving a Macro F1-score of 0.7445 on the test data and ranking 10th in the competition. These results underscore the potential of multitasking architectures in tackling complex, multi-faceted problems in natural language processing, particularly for code-mixed and informal text.

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.

Kalika Bali, Jatin Sharma, Monojit Choudhury, and Yogarshi Vyas. 2014. "i am borrowing ya mixing?" an analysis of english-hindi code mixing in facebook. In *Proceedings of the first workshop on computational approaches to code switching*, pages 116–126.

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. 2022. Fighting hate speech from bilingual hinglish speaker's perspective, a transformer-and translation-based approach. *Social Network Analysis and Mining*, 12(1):87.

Shankar Biradar, Sunil Saumya, and Arun Chauhan. 2023. Combating the infodemic: Covid-19 induced fake news recognition in social media networks. *Complex & Intelligent Systems*, 9(3):2879–2891.

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.

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

Sanjana Kavatagi and Rashmi Rachh. 2021. A context aware embedding for the detection of hate speech in social media networks. In *2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)*, pages 1–4. IEEE.

Chava Sai, Rangoori Kumar, Sunil Saumya, and Shankar Biradar. 2024. Iiitdwd_svc@ dravidianlangtech-2024: Breaking language barriers; hate speech detection in telugu-english code-mixed text. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 119–123.

Zuhair Shaik, Sai Kartheek Reddy Kasu, Sunil Saumya, and Shankar Biradar. 2024. Iiitdwd-zk@ dravidianlangtech-2024: Leveraging the power of language models for hate speech detection in telugu-english code-mixed text. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 134–139.