LoRA adapter weight tuning with multi-task learning for Faux-Hate detection

Abhinandan Onajol, Varun Gani, Praneeta Marakatti,
Bhakti Malwankar, and Shankar Biradar
Department of Computer Science and Artificial Intelligence,
KLE Technological University's Dr. M. S. Sheshgiri College of Engineering and Technology,
Belagavi, Karnataka, India
(02fe22bci003, 02fe22bci056, 02fe22bci031)@kletech.ac.in
02fe22bci012@kletech.ac.in

Abstract

Detecting misinformation and harmful language in bilingual texts, particularly those combining Hindi and English, poses considerable difficulties. The intricacies of mixed-language content and limited available resources complicate this task even more. The proposed work focuses on unraveling deceptive stories that propagate hate. We have developed an innovative attention-weight-tuned LoRA Adopter-based model for such Faux-Hate content detection. This work is conducted as a part of the ICON 2024 shared task on Decoding Fake narratives in spreading Hateful stories. The LoRA-enhanced architecture secured 13th place among the participating teams for Task A.

1 Introduction

Social media platforms today enable individuals to express their identities and views, seek validation and feedback, promote products, and make purchases. However, this freedom of speech should be used more effectively. Many users post offensive content targeting specific groups, religions, and communities, which exposes these groups to harmful criticism and negatively affects younger audiences (Biradar et al., 2024b). Furthermore, Faux-hate narratives are frequently spread on these platforms, causing distress to the government and society alike. This faux hate, emerging from fake narratives, especially during emergencies, can gravely impact our society.

In multi-lingual societies like India, people speak 22 different languages (Gala et al., 2023). Additionally, due to the influence of English on our education system, people commonly mix the linguistic patterns of their native languages with English, leading to code-mixed text (Sheth et al., 2024) The proposed study focuses on identifying faux-hate content in low-resource Hindi-English code-mixed text. Current measures to address these issues primarily involve flagging accounts or disabling user activity, which can unintentionally suppress freedom of speech. However, this situation underscores the urgent need for a more effective strategy to combat faux hate content while allowing people to express their opinions. While hate can be conveyed through various mediums like images, animations, memes, and emoticons, our work focuses on text data.

Most of the existing work considers Fake and Hate content as separate entities and employs several single-task learning models to address these types of content. Various transformer models, such as Roberta and BERT, have been utilized. For instance, (Cooper et al., 2023), (Boucher et al., 2021), and (Ramos et al., 2024) have analyzed different techniques for hate speech detection, reviewing traditional machine learning models, deep learning models, multi-task learning methods, generative models, and transformer-based models. Their findings indicate that transformer models are well-suited for hate speech detection tasks. However, these models have primarily been trained on high-resource languages like English and tend to perform poorly with Indian regional languages.

To address the issue of limited datasets for Indian code-mixed languages and joint learning of Faux-hate content, the authors Biradar et al. (2024b) have made significant contributions by developing code-mixed language datasets. One such dataset is the Faux Hate Multi-Label Dataset (FHMLD), which includes Hindi-English code-mixed text to enhance research in this area.

To address the issue of Faux-Hate detection in low-resource Hindi-English code-mixed text, the organizers of the ICON shared task on Decoding Fake Narratives in Spreading Hateful Stories (Faux-Hate) have designed two sub-tasks. ***Task A*** focuses on developing a multi-task model that identifies fake and hate labels for a given input. At the same time, ***Task B*** aims to create a model that generates both target and severity labels. Our team participated in **Task A** and secured the 13^{th} rank among the participating teams. In our proposed work, we developed a suite for converting bilingual text into monolingual text. On top of this, we employed LoRA adapter-based attention weight tuning to enhance the model's performance.

The rest of the article is organized as follows: Section 2 delves into the methodology that is proposed for the given task, Section 3 analyzes the experimental results, and we conclude the paper alongside listing potential scope.

2 Methodology

2.1 Task and Data

The proposed method, which involved identification Faux-hate content through multi-task learning, was developed as part of the *ICON 2024 Shared Task on Decoding fake narratives in spreading Hateful stories (Faux-Hate)*[1] (Biradar et al., 2024b,a). The data set was released in three sets: train, test, and validation. The train set contained 6397 instances, the test set consisted of 801 instances, and the validation set had 801 instances separately for task A and task B. This custom dataset, created by scraping YouTube and Twitter Faux-hate comments on topics such as Religion, health, finance, entertainment, sports, and Hindi-English code mixed sentences, is highly relevant to our research on decoding fake narratives in spreading hateful stories. The detailed distribution of the corpus is presented in Table 1.

2.2 Data Preprocessing

Translating code-mixed languages directly into a single language is challenging, as it often leads to losing the original meaning. As part of our preprocessing, we have developed a suite that plays a crucial role in converting unconventional bilingual code-mixed text into monolingual English text. Initially, the dataset was translated using extit IndicTrans2 (Gala et al., 2023). The rationale behind this translation is that most pre-trained models are designed to handle high-resource English text and have achieved state-of-the-art results on such corpora.

Subsequently, we utilized the extit Natural Language Toolkit (NLTK)[2] for preprocessing techniques, such as lemmatization, to handle social media slang by converting words to their base form (e.g., "running" to "run"). Data cleaning was also applied to remove extraneous information that is not useful for final inferencing, including hyperlinks, hashtags, numerals, and emoticons. This was efficiently done using Python string operations. Additionally, we tokenized the input sentences into individual tokens, and performed padding and masking to ensure all social media text was converted to equal-length sentences. Finally, the cleaned, tokenized data with padding and masking was used as input for model building.

	Hate		Fake	
	0	1	0	1
Train	2295	4101	3110	3286
Test	287	513	383	417
Validation	287	513	377	423

Table 1: Dataset distribution for Task A.

2.3 Understanding LoRA: Efficient Model Tuning for NLP tasks

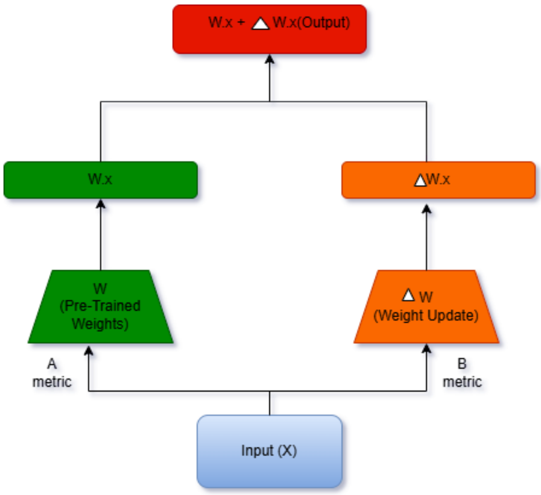

Figure 1: LoRA Adapter

2.3.1 Fine Tuning with LoRA

A parameter-efficient method for fine-tuning large language models is Low-Rank Adaptation(LoRA)(Hu et al., 2021a) it uses trainable low-rank matrices A and B to estimate weight updates ($\Delta W = A \cdot B$) shown in figure 1 while maintaining the original weights frozen. This lowers memory and computational costs by reducing the number of trainable parameters. When it comes to modifying huge models for particular purposes on constrained technology, LoRA works very well. This

[1] ICON 2024 Shared Task on Decoding fake narratives in spreading Hateful stories (Faux-Hate)

[2] https://www.nltk.org/

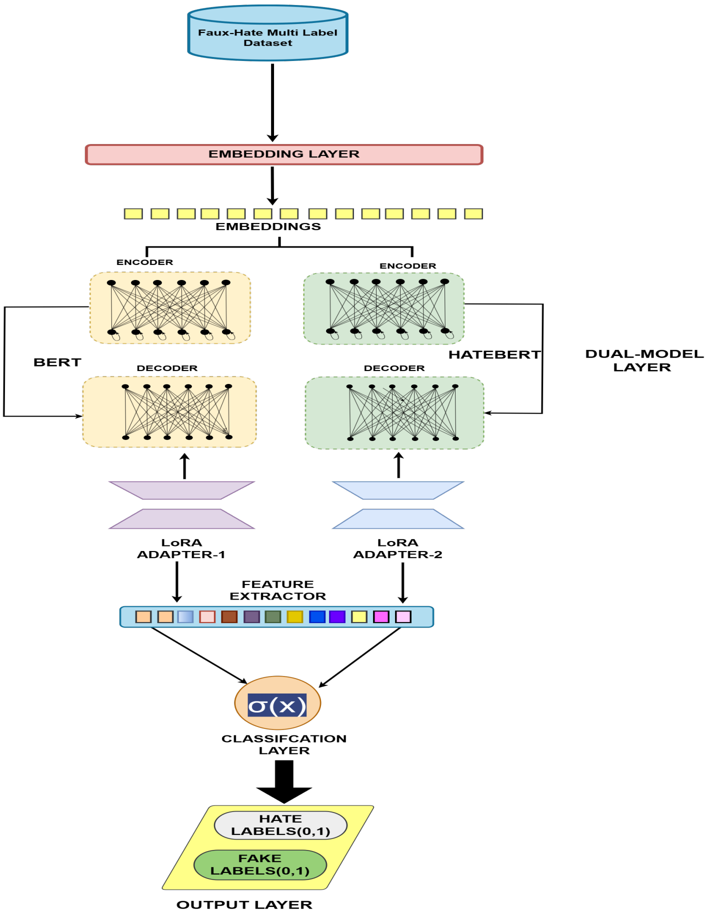

Figure 2: LoRA-Enhanced Architectur

work effectively fine-tunes pre-trained transformer models for two tasks: classifying bogus news and detecting hate speech, using LoRA. LoRA allows task-specific learning while maintaining the majority of the pre-trained weights by introducing low-rank matrices into the attention layers, which alters only a small subset of parameters.While distinct LoRA modules are set up for every activity, shared elements such as the tokenizer guarantee uniform text preprocessing and a single representation space. Compared to comprehensive fine-tuning, this parameter-efficient method uses less memory and computes less, enabling efficient adaptability to both tasks without the need for resource-intensive updates or redundancies.

Given the computational challenges of training *LLM* models, we turned to *Low-Rank Adaptation* (Hu et al., 2021b)(LoRA). This approach selectively tunes the necessary parameters of the extit BERT models while keeping the rest of the model frozen. The result is a method that delivers comparable outcomes to training the extit BERT models but at a significantly reduced computational and time cost. The hyperparameters used during the model training are selected from experimental trials, further detailed in Table tab:hyperparameters

2.4 LoRA Enhanced Architecture

The processed tokens are then passed to the *HateBert* (Caselli et al., 2021), *BERT* models (Devlin, 2018), each one of them attached to the *LoRA* adapter. Adapter 1 is applied to HateBert, allowing for a focus on hate speech detection. Adapter-1 is customized to handle the nuances of hate speech even better, whereas Adapter-2 is used to tune the BERT model to handle the fake and non-fake samples. The adapter weights tuned on custom corpus combined with original weights are then passed through the sigmoid layer for classification. The proposed work uses task-specific classification heads with mean loss functions for faux-hate classification. The detailed architecture of the model is presented in Figure 2; the reasoning behind using LoRA adopter is to avoid catastrophic forgetting of the pre-trained model and their low-resource adoption. Additionally, We have used the binary cross entropy (BCE) to find hate loss and fake loss separately. Then, we combine these individual losses to find the model's total loss. The model is then trained with a suitable number of epochs.

Hyperparameter	Value
Rank (LoRA config)	8
LoRA Alpha (LoRA config)	16
Dropout (LoRA config)	0.1
Learning Rate	1×10^{-5}
LoRA target module	`"query", "key"`

Table 2: Hyperparameters for LoRA Configuration

3 Results

The performance of the hate speech detection model and fake speech detection model is summarized in the classification report mentioned in Table 3 and Table 4, respectively. The result indicates that the model performs better for the 'Hate' class than the 'Non-Hate' class, demonstrating high effectiveness in detecting hate speech. In contrast, the fake news classification model performs well in detecting the 'Fake' class, with a balanced F1-score that indicates its reliability. The weighted and macro averages suggest that the model achieves consistent performance across both classes, with decent precision and recall.

Our team $Vector\_Visionaries$, has achieved the 13th rank ,refer table 5, these results highlight the model's capability to identify Faux-Hate content while effectively maintaining stability in performance metrics. Where execution time is concerned, the first epoch took 5 minutes to run, which was reduced to 3 minutes for the last epoch. We executed the model for 10 epochs, which took a total of 50 minutes. The source code is available in the GitHub repository.[3]

Class	Precision	Recall	F1-Score
Not Hate	0.62	0.52	0.73
Hate	0.64	0.99	0.78
Average (Macro)	0.63	0.51	0.41
Average (Weighted)	0.63	0.64	0.51

Table 3: Performance Metrics for Hate Speech Detection

4 Conclusion and Future Scope

The proposed work introduces LoRA-enhanced Architecture, a versatile approach that can be adapted for classifying Hindi-English code mixed Faux-hate comments. It is presented as part of the ICON 2024 shared task on "Decoding Fake Narratives in Spreading Hateful Stories (Faux-Hate)." Further, future work could be extended to include more efficient quantization methods to further reduce the computational cost, reinforcing the adaptability and robustness of our approach.

Class	Precision	Recall	F1-Score
Not Fake	0.59	0.45	0.51
Fake	0.58	0.70	0.63
Average (Macro)	0.58	0.58	0.57
Average (Weighted)	0.58	0.58	0.57

Table 4: Performance Metrics for Fake News Detection

Rank	TEAM	Macro F1 Score for Task A
1	DCST_unigoa	0.79
2	Radicaldecoders run1	0.7761
3	chakravyuh coders run1	0.7721
4	Tensor_Text	0.772
5	Keyboardwarriors run1	0.76
13	**Vector_Visionaries**	**0.6803**

Table 5: Macro F1 Score Rankings for Task A

References

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.

Nicholas Boucher, Ilia Shumailov, Ross J. Anderson, and Nicolas Papernot. 2021. Bad characters: Imperceptible NLP attacks. *CoRR*, abs/2106.09898.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

Portia Cooper, Mihai Surdeanu, and Eduardo Blanco. 2023. Hiding in plain sight: Tweets with hate speech

[3] GitHub Repository

masked by homoglyphs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2922–2929, Singapore. Association for Computational Linguistics.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jay Gala, Pranjal A. Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021a. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Qian Hu, Thahir Mohamed, Zheng Gao, Xibin Gao, Radhika Arava, Xiyao Ma, and Mohamed Abdel-Hady. 2021b. Collaborative data relabeling for robust and diverse voice apps recommendation in intelligent personal assistants. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 113–119, Online. Association for Computational Linguistics.

Gil Ramos, Fernando Batista, Ricardo Ribeiro, Pedro Fialho, Sérgio Moro, António Fonseca, Rita Guerra, Paula Carvalho, Catarina Marques, and Cláudia Silva. 2024. A comprehensive review on automatic hate speech detection in the age of the transformer. *Social Network Analysis and Mining*, 14(1):204.

Rajvee Sheth, Shubh Nisar, Heenaben Prajapati, Himanshu Beniwal, and Mayank Singh. 2024. Commentator: A code-mixed multilingual text annotation framework.