Faux-Hate Multitask Framework for Misinformation and Hate Speech Detection in Code-Mixed Languages

Sunil Gopal C V[1], Sudhan S[1], Shreyas Gutti Srinivas[1], Sushanth R[1], Abhilash C B[*2]
[1,2]Department of Computer Science and Engineering
[1,2]JSS Academy of Technical Education, Bengaluru, Karnataka, India
[1]*(sunilgopal63, sudhanyadav06, shreyas0337, sushanth4774)@gmail.com*
[*2]*abhilashcb@jssateb.ac.in*

Abstract

The Faux-Hate task looks at two big issues: misinformation and hate speech. It focuses on Hindi-English social media posts. This paper shares our methods for both parts of the task. For Task A, we built a special model based on XLM-RoBERTa. It has features that help us spot both fake news and hate speech at the same time. For Task B, we wanted to identify who the hate is aimed at (like individuals or groups) and how severe it is (high, medium, low). So, we added different tools to our model for this kind of sorting. To get ready for all this, we carefully cleaned the data, especially dealing with mixing languages and different spellings. In Task A, our results show that our model can clearly tell the difference between fake and real stories, as well as between hate and non-hate speech. For Task B, it does a good job with identifying targets and severity levels, giving strong predictions for multiple labels. Overall, these results show that cross-lingual models, combined with specific tweaks, can really help tackle complex text classification in languages with fewer resources.

1 Introduction

Social media has transformed the way of communication and allowed for global interaction but also become a channel for fake narratives and hate speech, which disrupts social cohesion. In linguistically diverse regions like India, where code-mixed languages such as Hindi-English are common, the complexities of transliteration and cultural nuances demand advanced approaches to handle the contextual dependencies of multilingual, low-resource text.

The two subtasks of the Faux-Hate shared task at ICON 2024 serve to bridge this challenge: Task A involves binary classification-classifying between fake/real and hate/non-hate speech. Task B calls for multi-class predictions for the target of hate (Individual, Organization, Religion, or N/A) and the severity (Low, Medium, High, or N/A). Both tasks raise challenging constraints: skewed class distributions in Task A make it hard to generalize well across underrepresented classes, such as fake but not hateful content; and Task B requires deep understanding since implicit targets often rely on sarcasm or cultural allusions Biradar et al. (2024a).

We try to overcome these challenges using multitask models based on the latest state-of-the-art cross-lingual transformer pre-trained on more than 100 languages, namely XLM-RoBERTa. To accomplish this, for Task A, we predict the fake and hate labels simultaneously with a dual-output architecture. For Task B, we extend the model with task-specific attention mechanisms in order to really boost the focus of the model on relevant features for each prediction task at hand. Preprocessing steps such as handling blank labels and removing transliteration inconsistencies improve the ability of the models to process complex code-mixed text.

This work is an extension of earlier research in the class of multilingual hate speech detection and misinformation classification. For example, frameworks such as HOLD-Z, developed especially for Telugu-English code-mixed text, show effectiveness in merging contextual embeddings with task-specific attention layers for low-resource scenarios Shaik et al. (2024). Likewise, there are studies on multilingual pre-trained models such as mBERT and XLM-RoBERTa, where their strength to capture cross-lingual patterns emerges as suitable for code-mixing scenarios (Nozza, 2021). For example, while (Biradar et al., 2021) suggest that language-specific fine-tuning is key in improving performance over lowresource Indian languages, (Saumya et al., 2022) explore ensemble methods for the detection of offensive content in YouTube comments.

Attention mechanisms, particularly self-attention in transformer architectures, have significantly improved contextual understanding

by focusing on key tokens in input sequences (Vaswani et al., 2017). In hate speech and misinformation detection, attention layers help extract relevant features while filtering out noise. For example, it was shown by (Ghosh et al., 2021) for context-dependent hate speech that hierarchical attention is effective in multi-label classification. (Zhou et al., 2021) used multi-head attention to enhance feature extraction, especially in multilingual environments with code mixing. Both of these results confirm this work's use of specific task attention layers to properly handle contextual nuances and improve classifications.

Using task-specific augmentations and preprocessing, our models reached macro F1 scores of 0.78 for Fake, 0.75 for Hate, 0.66 for Target, and 0.75 for Severity. This indicates the effectiveness of our method in dealing with misinformation and hate speech in code-mixed contexts while enabling scalable multilingual content moderation.

2 Methodology

This section describes the approach adopted towards solving the Faux-Hate shared task, focusing on identifying fake and hate speech within social media comments. Initially, a brief introduction to the problem statement and dataset is presented as it is. Consequently, approaches used to address this challenges have been considered.

2.1 Task and Dataset

The Faux-Hate shared task aims at challenging the participants to develop models that tackle the two crucial problems in social media content analysis: fake narratives detection and hate speech detection, with added emphasis on understanding the target and severity of hateful speech.

For this experiment, a dataset (Biradar et al., 2024b) with two different labels are used are used, which correspond to the two sub-tasks defined in the Faux-Hate shared task: Task A and Task B. Both datasets are specifically prepared for training purposes and contain labeled social media comments. Task A Dataset: This dataset is designed for binary classification tasks and contains two labels: Fake: Whether the content is fake (1) or real (0). Hate: Indicates whether the content contains hate speech (1) or not (0). Table 1 summarizes the distribution of samples in the Task A dataset. Task B Dataset: This dataset addresses multi-class classification tasks and contains two labels: Target: Specifies the target of hate speech, categorized into: Individual (I), Organization (O), Religion (R). Severity: Indicates the intensity of hate speech, categorized into: Low (L), Medium (M), High (H), Table 2 shows a detailed breakdown of the samples in the Task B dataset.

	Hate	**Non-Hate**	**Total**
Fake	2678	608	3286
Non-Fake	1423	1687	3110
Total	4101	2295	6396

Table 1: Binary Faux Hate dataset (Task A)

Target	**Severity**
NaN: 2295	**Nan:** 2295
O: 2279	**L:** 1960
I: 1081	**M:** 1559
R: 741	**H:** 589

Table 2: Target and Severity Prediction dataset (Task B)

2.2 Multitasking pipeline

The proposed model architecture is a multi-task attention mechanism that is tailored for binary or multi-class classification tasks. The architecture uses a shared transformer-based encoder and task-specific attention layers to enhance the learning of task-relevant features. A key aspect of this pipeline is its ability to focus on specific tokens through attention layers, improving performance across both tasks. The architecture is displayed in Figure 1.

1. Input Layer: The input consists of tokenized sequences (input-ids) and attention masks (attention-mask), which enables the model to differentiate between the padding tokens and the actual content. The inputs are preprocessed to a fixed maximum length, ensuring consistency across samples.

2. XLM-RoBERTa Encoder: The tokenized inputs are fed into a pre-trained XLM-RoBERTa model, xlm-roberta-base, to obtain contextual embeddings. The output is a tensor representing the last hidden states with a shape of (batch-size, sequence-length, hidden-dim), where the hidden dimension is 768.

3. Attention Layers: Two independent attention layers are used: one for each label-specific classification head. These attention layers

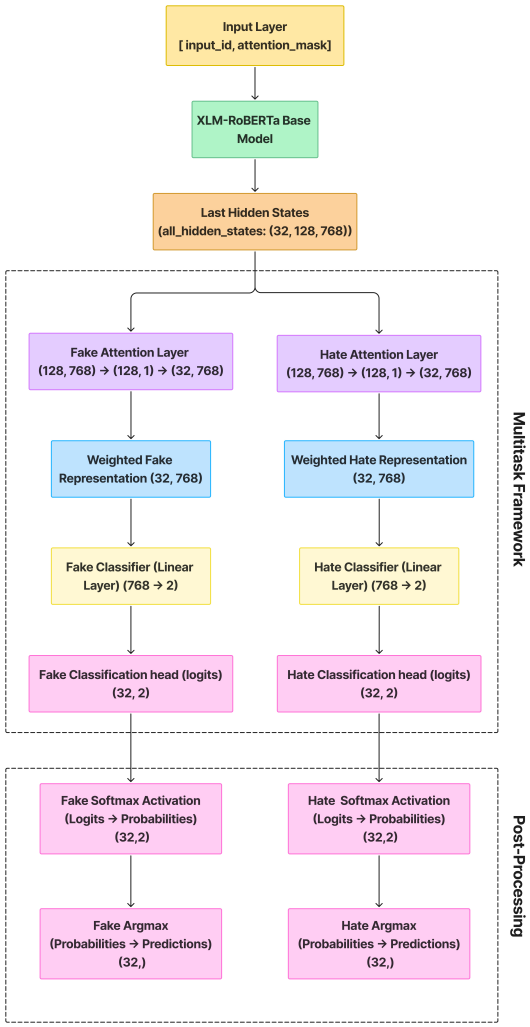

Figure 1: Multitasking Attention pipeline

compute a weighted sum of the hidden states by generating attention scores and focusing on the most relevant features in the sequence. The attention mechanism produces a weighted representation for each task, effectively summarizing the input for downstream classification.

4. Classification Heads: The weighted representations are passed through linear classifiers to generate logits for each task. Each classifier maps the 768-dimensional representation to the required output dimension (e.g., 2 for binary classification).

5. Output: The final outputs are the logits for each label (Fake and Hate or other task-specific outputs). These logits can be further processed using a loss function like cross-entropy for training.

The attention mechanism helps the model focus on crucial parts of the input text, enhancing its capacity to capture nuanced distinctions, such as context relevant to each classification label. This modular design, with a shared encoder and task-specific attention layers, allows flexibility while maintaining efficiency in the learning process.

2.3 Fine Tuning with XLM-RoBERTa

The XLM-RoBERTa-base model is selected because of its lighter architecture than XLM-RoBERTa, balancing performance with efficiency. To optimize the use of computational resources, we set the batch size to 16, so that it can process faster on a Tesla T4 GPU. Inputs were tokenized with a maximum sequence length of 180 and padded and truncated to uniformity. The model was optimized using the AdamW optimizer with a learning rate of 1 x 10^{-5}. Cross entropy loss was applied for the fake news and hate speech tasks. Training was capped at 30 epochs, with early stopping after 3 epochs of no improvement to avoid overfitting. For feature extraction enhancement, attention layers with a hidden dimension of 768 were used. The hyperparameters are described in Tabel 3.

Hyper parameter	Value
Max length	180
Training batch size	16
Validataion batch size	16
Test batch size	32
Optimizer	AdamW
Learning Rate	1 x 10^{-5}
Criterion	CrossEntropyLoss
Hidden dimension	768

Table 3: Hyper parameters for training XLM-RoBERTa-base

3 Results

This section presents the performance evaluation of the proposed multi-task shared attention model for both tasks, Task A (Faux-Hate detection) and Task B (Target and Severity prediction), in comparison with a Siamese network model for Task A. The performance is measured using the macro F1-score, a robustic metric for assessing model effectiveness in class-imbalanced datasets. The macro F1-score provides an average F1-score across all classes,

treating each class equally, regardless of its prevalence in the dataset.

3.1 Task A: Faux-Hate Detection

For Task A, which involves the binary classification of Fake and Hate speech, the shared attention model demonstrated superior performance compared to the Siamese network model. The macro F1-scores for task A is described in Table 4 as represented in Figure 2.

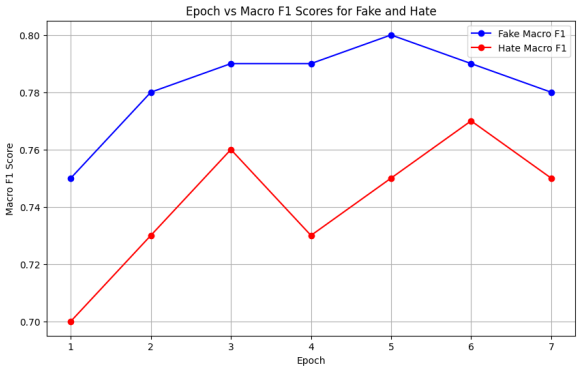

Figure 2: Macro F1 Scores by Epoch for Fake and Hate Detection

Model	Fake	Hate
Shared attention	0.80	0.76
Siamese Network	0.67	0.54

Table 4: Macro F1 score for Task A

3.2 Task B: Target and Severity Prediction

For Task B, which involves multi-class classification of Target (Individual, Organization, or Religion) and Severity (Low, Medium, or High), the shared attention model achieved the following macro F1-scores as described in Figure 3:

- Target Classification: 0.45
- Severity Classification: 0.50

The relatively lower performance in Task B compared to Task A can be associated to the increased complexity of multi-class classification and the nuanced nature of the target and severity categories. Despite this, the model's performance remains noteworthy, given the challenges associated with predicting subjective aspects like severity.

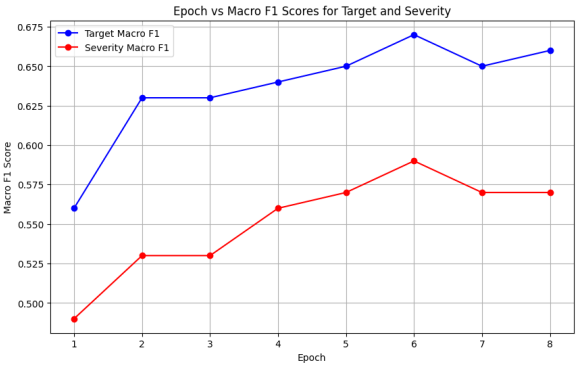

Figure 3: Macro F1 Scores for Target and Severity of Hate

3.3 Model Selection and Conclusion

Based on the evaluation results, the shared attention model was selected for both tasks due to its superior performance across all metrics. The model's ability to handle multi-task learning through task-specific attention layers proved effective, particularly in Task A, where it significantly outperformed the Siamese network model. Future work may focus on enhancing performance in Task B by incorporating additional features or leveraging advanced pre-training strategies.

This proposed work and model is uploaded to GitHub [1]. In the future, this model architectures will be fine-tuned with other types of embeddings while still finding its way through the current emerging issues of code-mixed text classification. The proposed work also has secured 4th rank in Task A and 5th rank in Task B of the Faux-Hate shared task at ICON 2024.

4 Conclusion and Future work

Attention mechanisms play a very important role in multi-task learning, especially for the complex task of text classification in areas such as detecting misinformation and hate speech. Our shared-attention-based multi-task model, which used task-specific attention layers, efficiently classified comments as fake or hate and predicted the target and severity. The model won Task A over the Siamese network with macro F1 scores of 0.80 on fake detection and 0.76 on hate detection, pushing it up to 4th rank in the rankings, though was merely moderate in Task B with F1 scores of 0.45 for target classification and 0.50 for severity prediction ranking 5th, which really illustrated challenges related to the

[1] https://github.com/SunilGopalCV/Faux-Hate

complexity and difficulty of the task.

The experiments clearly show that shared attention mechanisms play a significant role in identifying the key features across tasks, which greatly contributed to the model's success in Task A. However, moderate performance in Task B suggests further exploration is needed. Future work could focus on enhancing the granularity of attention mechanisms, incorporating pre-trained language models tailored to code-mixed text, and exploring additional contextual features to improve target and severity prediction. Additional to this, experimenting with hierarchical or task-specific encoders might further refine the shared attention framework and improve performance on all tasks.

Acknowledgment

This work was supported in part by JSS Academy of Technical Education Bangalore. We also, extend our gratitude to Innovation in Technology Lab (INiT Lab) of CSE Department, JSSATE Bangalore for the opportunity and guidance.

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. 2021. mbert based model for identification of offensive content in south indian languages. *Language Resources and Evaluation*.

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.

Shilpi Ghosh, Navonil Majumder, Alexander Gelbukh, and Erik Cambria. 2021. Analyzing hate speech in social media using hierarchical attention networks: A case study. *Cognitive Computation*, 13:1–15.

Debora Nozza. 2021. Exposing the limits of zero-shot cross-lingual hate speech detection. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 907–914.

Sunil Saumya, Vanshita Jha, and Shankar Biradar. 2022. Sentiment and homophobia detection on youtube using ensemble machine learning techniques. In *Working Notes of FIRE 2022–Forum for Information Retrieval Evaluation (Hybrid)*.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, pages 5998–6008.

Yu Zhou, Weijie Zhou, and Seung-won Kim. 2021. Multi-head attention networks for multilingual hate speech detection in low-resource settings. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 1601–1610.