Multi-Task Learning for Faux-Hate Detection in Hindi-English Code-Mixed Text

Hitesh N P[1], D Ankith[1], Poornachandra A N[1], Abhilash C B[*2]
[1]Department of Artificial Intelligence & Machine Learning
[2]Department of Computer Science and Engineering
[1,2]JSS Academy of Technical Education, Bengaluru, Karnataka, India
[1]*(hiteshnp19, ankithdadda, poornachandra308)@gmail.com*
[*2]*abhilashcb@jssateb.ac.in*

Abstract

The prevalence of harmful internet content is on the rise, especially among young people. This makes social media sites breeding grounds for hate speech and negativity even though their purpose is to create connections. The study proposes a multi-task learning model for the identification and analysis of harmful social media content. This classifies the text into fake/real and hate/non-hate categories and further identifies the target and severity of the harmful content. The proposed model showed significant improvements in performance with training on transliterated data as compared to code-mixed data. It ranked 2nd and 3rd in the ICON 2024 Faux-Hate Shared Task and the performances have made it very effective against harmful content.

1 Introduction

With rising availability of the use of smartphones and cheaper Internet, social media has become an added part of life. New technologies are gaining greater interest in the minds of the younger generation, coupled with a desire to connect with people from different walks of lives. Social media offers several benefits, such as improved communication and opportunities for network building. However, it also has its drawbacks, particularly concerning privacy. Misinformation as well as hateful narratives on social media have been prominent issues in the last decade. Such narratives have become significant and broad-ranging regarding effects on society in times of crisis.

Hate comments are fueled by misinformation that may intensify hostility in communities. It may also have the potential to polarize groups, increase tensions, and create an atmosphere that leads to conflicts. Most of the escalations driven by hate and fake narratives often tend to translate into verbal and physical violence that threatens the safety and cohesion of a community. It could cause significant psychological harm to targeted individuals, affecting the emotional toll that translates to long-term mental health issues resulting in anxiety, depression, and a sense of vulnerability. False information or fake news can mislead individuals during an emergency and distort public perception and opinion, leading to misguided beliefs and actions.

The scale of hate speech experienced during the 2020-2021 COVID-19 pandemic towards groups such as Chinese people, Asian communities, religions and other nationalities improperly blamed for the spread of the virus and social ills was unprecedented. Adding onto that is also a direct consequence, the spread of fake news, manifested in a variety of conspiracy theories and more general misinformation was also closely linked to the perpetration of violence against these targeted groups.(Pérez et al., 2023)

Most of the previous research (Biradar et al., 2024) viewed hate speech through a binary classification lens. This led to a risk of violating the use of freedom of speech by blocking all forms of possible hateful content in an unsystematic way.

Hence it is necessary that hateful and fake narratives are to be addressed in terms of both target and severity. A target comprises different individuals, groups, or any religion that might be affected by hateful or fake content. A severity measure refers to the level or intensity of damage that such content might cause (Wu et al., 2019), (Zhou and Zafarani, 2020).

As a contribution towards combating hateful content from social media and bringing about peace and harmony to society, we took part in the shared task *Decoding Fake Narratives in Spreading Hateful Stories*, called Faux-Hate, organized by *ICON 2024*. This shared task proposed to develop a hard-parameterized multi-task learning model to be able to effectively detect, classify, and analyze the fake narratives used in spreading hate while focusing on the target and severity dimensions.

2 Methodology

This section describes the approach for identifying the Faux-hate shared task.

2.1 Task and Data

The Faux-Hate shared task is divided into two sub-tasks that target the challenges of detecting and analyzing harmful online content. Task A - Binary Faux-Hate Detection is to determine if the given text samples contain fake or hateful content. Each label contains two attributes: Fake (1 for fake content, and 0 for real content) and Hate (1 for hate speech and 0 for non-hate content). The objective of this task was to develop a single multi-tasking model that outputs both the fake and hate labels for each text sample.

Task B - Target and Severity Prediction: With this task, the focus is extended to the target and the severity of the harmful content. The text samples are labeled with four categorical attributes in this dataset: Target (whether it is targeted at an individual (I), organization (O), religion (R) or N/A) and Severity (how intense the content is - Low (L), Medium (M), High (H) or N/A). The goal of this sub-task was to develop a unified model that predicts both the target and severity labels for a given text sample.

2.1.1 Dataset Analysis

The given dataset for Task-A contains the following number of instances for each class.

Fake	Fake	Non-Fake	Total
Train	4097	2291	6388
Validation	423	376	799

Table 1: Train and Validation Data for Fake and Non-Fake Labels

Hate	Hate	Non-Hate	Total
Train	3284	3104	6388
Validation	513	286	799

Table 2: Train and Validation Data for Hate and Non-Hate Labels

The given dataset for Task-B contains the following number of instances for each class.

Target	I	O	R	N/A
Train	1081	2279	741	2295
Validation	140	274	140	99

Table 3: Train and Validation Data for Target Classes

Severity	L	M	H	N/A
Train	1960	1559	582	2295
Validation	257	182	74	287

Table 4: Train and Validation Data for Severity Classes

2.1.2 Preprocessing:

To ensure the dataset was clean and ready for analysis, several preprocessing steps were applied. First we converted all the texts present in the tweet column to lowercase (to maintain uniformity), then we applied regular expressions to remove URLs, tweet mentions (such as @user), hashtags, punctuations, numeric representations, and also got rid of additional white spaces by leading and trailing methods. Lastly, we ensured encoding issues were resolved so that the text would work well with the models.

The above preprocessing steps made the dataset more refined and consistent, making it easier for models to focus on pattern recognition.

2.2 Hard Parameterized Multi-Task Learning Model

2.2.1 Overview of the Model

Multi-task learning model is a type of machine learning model where the model is trained to perform multiple tasks at the same time.

In this shared task, we present a hard parameterized Multi-Task learning model (MTL). This MTL approach benefits from shared knowledge using a common feature space, thus improving the model's ability and generalization capabilities. Separate classifiers process each task, but the shared feature representation learned by the encoder layer optimizes efficiency and accuracy across tasks. The decoder layer then classifies the input and learns over it. We have used the same architecture of MTL for both Task-A and Task-B.

2.2.2 Embedding Layer

We explored different techniques to extract linguistic patterns from the proposed dataset. The experiments were performed using syntactic and semantic features at word and sentence levels in order to develop feature sets within an embedding layer.

The methods used during the experiments involved:

XLM-Roberta (XLM-R)

XLM-R is a multilingual transformer model based on Roberta and which is trained on more than 100 languages with a masked language modeling objective, it enables to learn rich representations of text across different linguistic contexts, captures the cross-lingual relations and semantic subtleties. This makes the XLM-R suitable for our task because the dataset consists of Hindi-English code-mixed texts (Tweets).(Conneau et al., 2020)

Multilingual BERT (mBERT)

The transformer model, mBERT, is a model that is trained on 104 languages using masked language modeling. Contrasting this with XLM-R, the mBERT is trained with a relatively smaller corpus, and because of this, it makes an excellent candidate for code-mixed datasets in high-resource and low-resource languages.(Devlin et al., 2018)

HateBERT

HateBERT is a domain-specific transformer model fine-tuned on abusive and hateful language data, which makes it best suited for identifying patterns in hate speech. Since our dataset contains hate speech, we chose to use HateBERT for its ability to do well at recognizing and understanding hateful content(Caselli et al., 2021).

2.2.3 MTL Model Architecture

The preprocessed input tweets are first passed to the tokenizer, where the sentences are split into words and sub words. Each of these tokens are then passed to the embedding layer which generates the embeddings for the [CLS] tokens.

After extracting the embeddings from the embedding layer, the MTL model employs a series of common layers to refine the feature representations. These embeddings are then passed onto the Dropout layer which is used to avoid overfitting, it is then followed by two fully connected Dense Layers with ReLU activation functions, where the First Dense Layer has 512 units and the Second Dense Layer has 256 units.(Dash et al., 2024)

Subsequently, for the final output layer we used two classifiers in the model to produce task-specific outputs. These classifiers are connected to a softmax activation function under the forward pass of the model, which then classifies into the specific classes. For Task A we made use of Hate Classifier and Fake classifier that map the output of the dense layers to hate labels and fake labels respectively. Similarly, for the Task B, a Severity classifier and Target classifier were used to classify the texts into their respective labels.

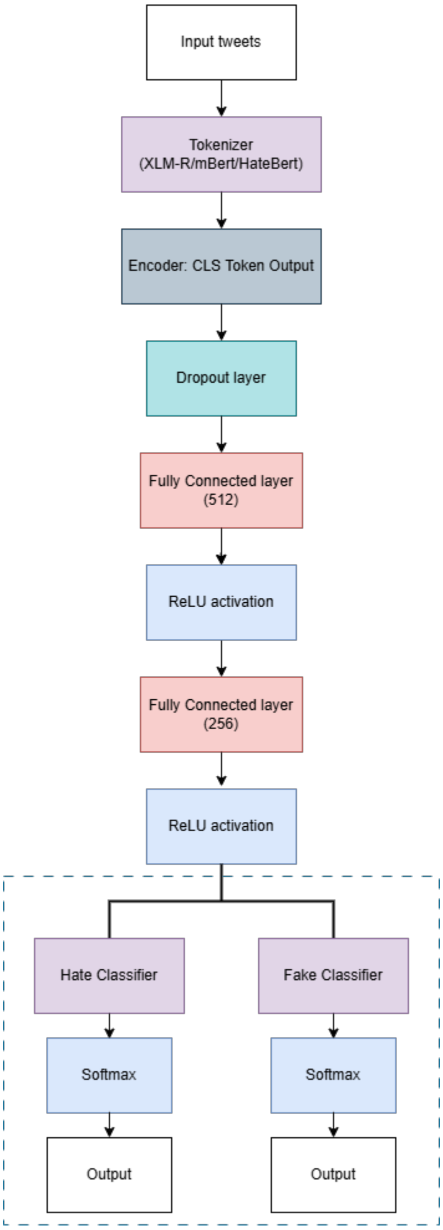

Figure 1: MTL Model architecture

The model uses CrossEntropyLoss to individually compute the loss of both tasks, which are summed together in the backpropagation step. The model uses AdamW optimizer for training and optimization with a ReduceLROnPlateau learning rate scheduler as per validation loss.

It uses a patience mechanism for early stopping, which helps to avoid overfitting, and saves the model if both tasks' accuracy thresholds are met.

2.2.4 Fine-tuning Parameters

The hyperparameters were carefully selected and fine-tuned for optimal performance. Optimization of both model complexity and training efficiency is found by governing its parameters, such as learning rate, batch size, dropout rate and others, so as to reach the best possible result considering overfitting and underfitting. The fine-tuning process enabled us to reach a version of the model that is better tuned to generalize well to the data.

Hyperparameter	**Value**
Dropout rate	0.3 to 0.4
Learning rate	1e-5 to 5e-5
Weight Decay	1e-4
Max Norm	1.0
Batch size	32
Early stopping- Patience	5

Table 5: Hyperparameters for Model Training

3 Results and Discussions

This section presents the performance evaluation of the proposed multi-task model for both tasks.

3.1 Validation data results

We trained the model for 30 epochs using the hyperparameters listed in Table 5 on code-mixed data, and the results obtained on the validation data are as follows:

Task-A	**Label**	**Accuracy**	**Macro F1**
mBert	Fake	63	61
	Hate	58	57
XLM-RoBerta	Fake	65	64
	Hate	63	58

Table 6: Performance of mBert and XLM-RoBert on Task-A (in %).

Task-B	**Label**	**Accuracy**	**Macro F1**
mBert	Target	66	59
	Severity	57	48
HateBert	Target	59	54
	Severity	53	53

Table 7: Performance of mBert and HateBert on Task-B (in %).

We failed to get reasonable results with the original data. So, we used Sanscript submodule of the *indic_transliteration* module to transliterate the data. Transliteration means changing the text from one language to another while preserving the original pronunciation and its meaning, which helps the model understand the context better.

Using the transliterated dataset, we trained the model for 30 epochs. The results obtained on the validation data are as follows:

Task-A	**Label**	**Accuracy**	**Macro F1**
mBert	Fake	69	67
mBert	Hate	78	78
XLM-R	Fake	77	**76**
XLM-R	Hate	80	**80**

Table 8: Performance of mBert and XLM-RoBert on Transliterated Task-A data(in %).

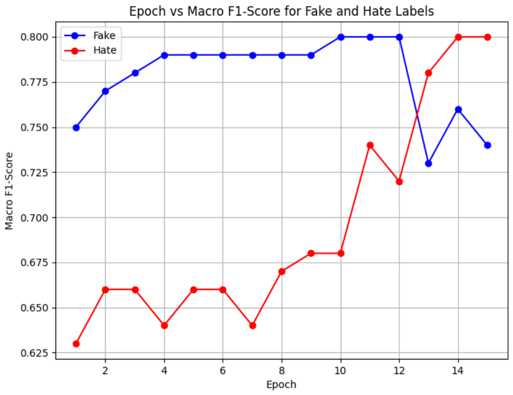

Figure 2: Macro F1 score for Fake and Hate labels

The plot in Figure 2 illustrates the Macro F1 scores for "Fake" and "Hate" labels across 15 epochs, showcasing the performance of the XLM-R model.

Task-B	**Label**	**Accuracy**	**Macro F1**
mBert	Target	70	67
	Severity	61	58
HateBert	Target	69	56
	Severity	53	51

Table 9: Performance of mBert and HateBert on Transliterated Task-B data(in %).

The plot in Figure 3 illustrates the Macro F1 scores for the "target" and "severity" labels over 8 epochs, showcasing the performance of the mBert

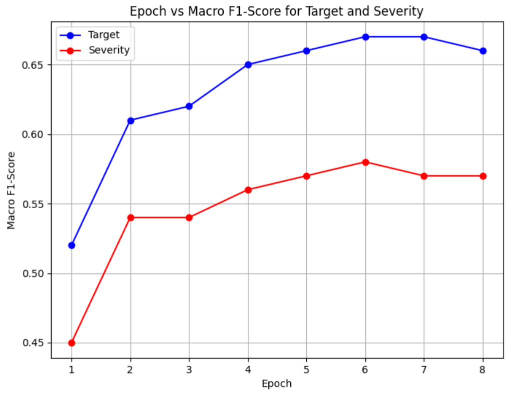

Figure 3: Macro F1 score for Target and Severity labels

model.

We were able to achieve better results on transliterated data compared to the original data. For Task-A we had a significant difference in Macro F1 score, from 57% to 80%. For Task-B there was a slight increase of 8% in Macro F1 Score.

3.2 Shared task results

We received the test data from the ICON 2024 Faux-Hate Shared Task team then cleaned and transliterated that and used our saved model to generate predictions. These predictions were submitted to the Faux-Hate team. Here are the results for the shared task:

- **Task A**:
  - Run 1: XLM-R ranked 2n**d**.
  - Run 2: mBERT ranked **8th**.
- **Task B**:
  - Run 1: mBERT ranked **3rd**.
  - Run 2: HateBERT ranked **9th**.

4 Conclusion and Future work

In conclusion, we designed our hard parameterized multi-task learning model using XLM-R, mBERT and HateBERT as encoder models for the embedding layer, then we used a dropout layer with two fully connected dense layers, each connected to a ReLU function, which was then linked to the specific label classifiers which predicted the output with the help of a softmax function. Our approach secured **2nd rank for Task A** and **3rd rank for Task B in** ICON2024 Faux-Hate Shared Task. The proposed work and model is available on Github [1]. Our future work involves further improving the architecture of the model to enhance its performance in tackling the complications of code-mixed data, exploring other embedding transformer models, and trying out ensemble learning. (Pérez et al., 2023)

Rank	TEAM	Macro F1
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
6	MUCS run1	0.7589
7	Rejected_cookies	0.7557
8	**Radicaldecoders run2**	**0.7522**
9	NOVA-RMK-ADS	0.7479
10	VTU_BGM	0.7445

Table 10: Ranking of Teams based on Macro F1 Score for Task-A.

Rank	TEAM	Macro F1
1	DCST_unigoa	0.6155
2	NOVA-RMK-ADS	0.6048
3	**Radicaldecoders run1**	**0.5947**
4	Rejected_cookies	0.5926
5	Tensor_Text	0.5887
6	MUCS run1	0.5746
7	CNLP-NITS-PP run1	0.57
8	Keyboardwarriors run1	0.56
9	**Radicaldecoders run2**	**0.5416**
10	Keyboardwarriors run2	0.54

Table 11: Ranking of Teams based on Macro F1 Score for Task-B.

Acknowledgements

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

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

References

Shankar Biradar, Sunil Saumya, and 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*, pages 1–32.

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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Sukanta Dash, Sandeep Solanki, and Soubhik Chakraborty. 2024. Deep convolutional neural networks for predominant instrument recognition in polyphonic music using discrete wavelet transform. *Circuits, Systems, and Signal Processing*, 43:1–33.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Juan Manuel Pérez, Franco M. Luque, Demian Zayat, Martín Kondratzky, Agustín Moro, Pablo Santiago Serrati, Joaquín Zajac, Paula Miguel, Natalia Debandi, Agustín Gravano, and Viviana Cotik. 2023. Assessing the impact of contextual information in hate speech detection. *IEEE Access*, 11:30575–30590.

Liang Wu, Fred Morstatter, Kathleen M Carley, and Huan Liu. 2019. Misinformation in social media: definition, manipulation, and detection. *ACM SIGKDD explorations newsletter*, 21(2):80–90.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40.