Transformer-driven Multi-task Learning for Fake and Hateful Content Detection

Asha Hegde[a], H L Shashirekha[b]
Department of Computer Science, Mangalore University, Mangalore, Karnataka, India
[a]hegdekasha@gmail.com, [b]hlsrekha@mangaloreuniversity.ac.in

Abstract

Social media has revolutionized communication these days in addition to facilitating the spread of fake and hate content. While fake content is the manipulation of facts by disinformation, hate content is textual violence or discrimination targeting a group or an individual. Fake narratives have the potential to spread hate content making people aggressive or hurting the sentiments of an individual or a group. Further, false narratives often dominate discussions on sensitive topics, amplifying harmful messages contributing to the rise of hate speech. Hence, understanding the relationship between hate speech driven by fake narratives is crucial in this digital age making it necessary to develop automatic tools to identify fake and hate content. In this direction, "Decoding Fake Narratives in Spreading Hateful Stories (Faux-Hate)" - a shared task organized at the International Conference on Natural Language Processing (ICON) 2024, invites researchers to tackle both fake and hate detection in social media comments, with additional emphasis on identifying the target and severity of hateful speech. The shared task consists of two subtasks - Task A (Identifying fake and hate content) and Task B (Identifying the target and severity of hateful speech). In this paper, we - team MUCS, describe the models proposed to address the challenges of this shared task. We propose two models: i) Hing_MTL - a Multi-task Learning (MTL) model implemented using pre-trained Hinglish Bidirectional Encoder Representations from Transformers (HinglishBERT), and ii) Ensemble_MTL - a MTL model implemented by ensembling two pre-trained models (HinglishBERT, and Multilingual Distiled version of BERT (MDistilBERT)), to detect fake and hate content and identify the target and severity of hateful speech. Ensemble_MTL model outperformed Hing_MTL model with macro F1 scores of 0.7589 and 0.5746 for Task A and Task B respectively, securing 6th place in both subtasks.

1 Introduction

Internet has become an integral part of modern society, serving as a primary means of communication and a platform for political and social engagement (Hegde et al., 2023a). However, the rise of digital communication has also posed significant challenges, particularly the spread of fake news, hate speech, and extremist content, misusing the anonymity of the users (Hegde and Shashirekha, 2021). Extremist groups have increasingly used the internet to disseminate their propaganda, deepening societal divides and polarizing communities. Recent studies (Gollatz and Jenner, 2018) have highlighted the connection between fake narratives and the spread of hate content on social media platforms. From this report, it is clear that a large portion of the hateful discourse circulating online is based on false information. This trend becomes especially pronounced during crises like disease outbreaks, communal violence, or public protests, where misinformation can rapidly escalate the situation.

A careful analysis of fake narratives and hate speech reveals the possibility of fake stories spreading hatred reactions which has the potential to exert a profound impact on society (Hegde et al., 2023b; Coelho et al., 2023). This warrants the pressing need for researchers to examine the relationship between fake narratives and hate speech. The research attempts in this direction can prevent the spread of hate speech driven by fake narratives keeping the social media ecosystem healthy. To address this issue, "Decoding Fake Narratives in Spreading Hateful Stories (Faux-Hate)" shared task[1] organized at ICON 2024, focuses on tackling fake and hate detection in social media comments, with additional emphasis on identifying the target and severity of hate speech (Biradar et al., 2024a).

[1] https://sai-kartheek-reddy.github.io/Shared-Task-on-Faux-Hate-Detection-at-ICON-2024/

The organizers have provided Hinglish dataset to assist in the analysis and identification of fake narratives that spread hatred on social media platforms. This shared task includes two sub-tasks: i) Task A - "Binary Faux-Hate Detection," involves identifying fake news and hate content simultaneously and ii) Task B - "Target and Severity Prediction" - involves identifying target of the content (Individual (I), Organization (O), and Religion (R)) and also severity of the content (Low (L), Medium (M), and High (H)), in the given Hinglish social media text.

The objective of this shared task is to encourage researchers to develop a single model for each task. Hence, we model these subtasks as MTL problems, where a single model is capable of performing more than one task simultaneously. With this view, we - team MUCS, proposed two models: i) Hing_MTL - a MTL system implemented using pre-trained HinglishBERT model and ii) Ensemble_MTL - a MTL model implemented by ensembling two pre-trained models (HinglishBERT and MDistlBERT), to accomplish the objectives of the shared task.

Rest of the paper is organized as follows: Related works are described in Section 2 and the proposed methodology is discussed in Section 3. Experiments and results are discussed in Section 4 and the paper concludes in Section 5 with future scopes for MTL models.

2 Related work

Until the last decade, fake news and hate speech detection were typically approached as separate binary or multi-class classification tasks (Hegde et al., 2021). Developing multiple single task models was expensive in low-resource settings. Further, training single task models with small data often lead to overfitting resulting in limited generalization. This has led the researchers to develop MTL - a single model capable of addressing multiple problems simultaneously (Waseem et al., 2018). Few of the relevant works in this field are described below:

MTL models for fake news detection - Wu et al. (2019) presented Sifted MTL for stance and fake news detection tasks using Keras embeddings and selected sharing of Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) layers. The selected sharing layer adopts gate mechanism and attention mechanism to filter and select shared feature flow between the tasks respectively. Experimenting on publicly available RumourEval[2] and PHEME[3] datasets, they obtained macro F1 scores of 0.7865 and 0.8009 for stance and fake news detection tasks respectively for RumourEval. Further, they achieved macro F1 scores of 0.8148 and 0.8127 for stance and fake news detection tasks respectively using PHEME dataset.

An integrated MTL model to perform fake news detection and news topic classification is proposed by Liao et al. (2021). To extract semantic and contextual relationships among news and news topics, the authors proposed a novel news graph that dynamically balances the tasks with a dynamic weighting strategy for the classification. They used LIAR[4] dataset to train their proposed model and obtained macro F1 scores of 0.516 and 0.645 for fake news detection and news topic classification tasks respectively. Cui and Yang (2022) introduced a multi-modal MTL model for fake news detection task considering evidence veracity classification as an auxiliary task. Their proposed approach makes use of a Convolutional Neural Network (CNN) to extract image features and a BERT encoder for textual claims and their associated evidence. Evidence veracity classification is incorporated as an auxiliary task to enhance the primary fake news detection task, sharing learned evidence representations across both the tasks. A co-attention mechanism is employed to fuse visual and textual evidence representations. Using CCMR[5] dataset for their experiments, they obtained macro F1 scores of 0.942 and 0.822 for fake news detection and evidence veracity tasks receptively.

MTL models for hate speech detection - A deep MTL model is discussed by Kapil and Ekbal (2020) to perform multiple related tasks - hate speech, racism, sexism, and offensive language classification. In order to leverage the benefits of multiple related tasks, they employed a shared private MTL framework and implemented a deep MTL model stacking CNN, CNN with attention, and GRU layers and conducted experiments on five datasets. Further, they concatenated Word2Vec[6]

[2] https://huggingface.co/datasets/strombergnlp/rumoureval_2019

[3] https://figshare.com/articles/dataset/PHEME_dataset_for_Rumour_Detection_and_Veracity_Classification/6392078

[4] https://huggingface.co/datasets/ucsbnlp/liar

[5] https://github.com/WeimingWen/CCRV

[6] https://github.com/mmihaltz/word2vec-GoogleNews-vectors/blob/master/GoogleNews-vectors-negative300.bin.gz

embeddings and character embeddings to train their proposed models. These models consistently performed well over the single task models with macro F1 scores of 0.8916, 0.9115, 0.8612, 0.9241, and 0.8535, for the detection of hate speech, offensive language, racism, communal language, and harassment content respectively. Plaza-Del-Arco et al. (2021) presented a novel study on a transformer-based MTL approach that leverages the shared affective knowledge to detect hate speech in Spanish tweets, integrating sentiment analysis and emotion analysis. The authors used transformer-based BETO model - a Spanish version of BERT model, and obtained macro F1 scores of 0.8551, 0.8603, and 0.8679, for hate speech detection, sentiment analysis, and emotion analysis tasks respectively. A Hindi hate speech detection dataset is created by Kapil et al. (2023), employing a novel hierarchical four-layer annotation scheme. The first layer addresses the binary classification of hate or not-hate and the second layer is for explicit and implicit hate. While the third layer performs multi-label classification task, the fourth layer is for named entity tagging, for the given text. Using this dataset, the authors implemented MTL models using Multilingual Representations for Indian Languages (MuRiL) and multilingual BERT pre-trained transformer models for hate speech detection. Their MTL frame work with MuRiL outperformed the single-task model with a weighted F1 score of 0.912 for hate speech detection.

In summary, models trained within the MTL framework outperform single-task models, with transformer-based architectures showing promising results. This highlights the suitability of transformer models for building effective MTL systems. Further, the limited exploration of MTL models for fake news and hate speech detection in Hinglish text underscores a significant opportunity for future research in this area.

3 Methodology

MTL is a Machine Learning (ML) technique in which a model is trained to perform multiple related tasks simultaneously, sharing certain layers and parameters in the network across different tasks. This approach improves model performance, especially when there is insufficient data for individual but related tasks (Zhang and Yang, 2021). MTL leverages shared lower-level features across tasks while allowing the model to learn higher-level features specific to each task. Using a unified framework, the model benefits from shared representations, which enhances its ability to generalize and make accurate predictions (Zhang and Yang, 2018).

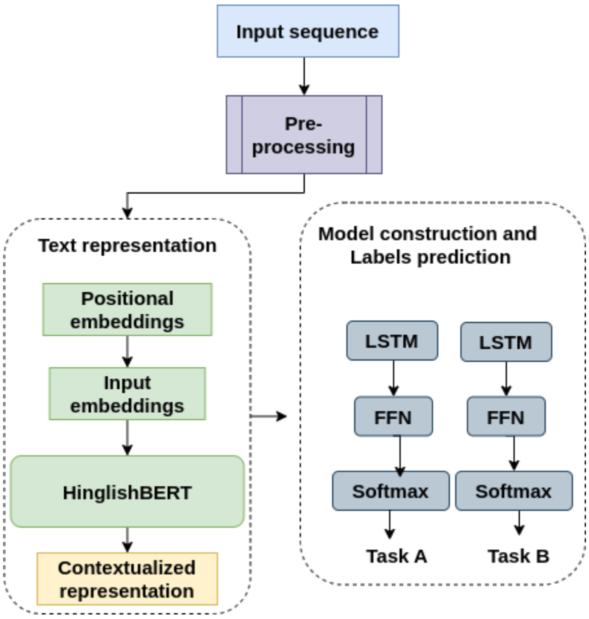


Figure 1: Framework of the proposed Hing_MTL model

The objectives of this study are to perform two tasks: i) Task A - binary classification using a single MTL model to identify fake news and hate speech content simulataneously, and ii) Task B - multi-class classification using a single MTL model for target and severity detection. To achieve these two subtasks, two distinct models - Hing_MTL and Ensemble_MTL, are proposed and descriptions of these models are given in the following subsections.

3.1 Pre-processing

Social media text often contains a large amount of irrelevant content that does not contribute to the intended task, but may negatively impact the performance of learning models (Hegde and Shashirekha, 2022). Therefore, effective pre-processing is crucial when constructing learning models for any task. In this work, user mentions (e.g., @text, #text), punctuation, and digits are removed from the given dataset to enhance the quality of the data for better model performance.

3.2 Hing_MTL Model

The proposed Hing_MTL model adopts the standard MTL framework, leveraging pre-trained transformer models. It processes input sequences to extract meaningful text representations, capturing underlying patterns, which are then shared across tasks for task-specific classification. A glimpse of the MTL framework for the proposed Hing_MTL model is shown in Figure 1.

Text representation focuses on capturing the contextual and semantic nuances of text by converting it into meaningful feature vectors (Chakravarthi et al., 2024; Hegde et al., 2023c). These representations enable the learning models to understand and generalize patterns within the data, improving their ability to make accurate predictions. By capturing deeper relationships between words and phrases, these methods enhance the model's performance in tasks like text classification. This work makes use of HinglishBERT[7] - a BERT-based model, 'BertTokenizer' and 'TFBertModel', from Hugging Face's transformers library for text representation, tokenization and loading the pre-trained HinglishBERT model respectively, for processing Hinglish tweets. HinglishBERT, pre-trained on a large corpus of Hinglish text, employs WordPiece tokenization to break words into sub-word units. 'TFBertModel' class enables the model to capture contextual information from both left and right sides of the input text, enhancing its understanding of Hinglish content.

Model Construction

As shown in Figure 1, the proposed MTL architecture utilizes contextualized representations generated by HinglishBERT to process the input sequence, which are then passed to two separate Recurrent Neural Network (RNN) modules, each equipped with Long Short-Term Memory (LSTM) cells for individual subtasks. These RNN modules use the embeddings to produce probability distributions for each task's target labels. The weights or logits produced by the LSTM layers are passed through a Feed Forward Network (FFN) to introduce non-linearity and capture more complex patterns from the input. Finally, the output logits are passed to a softmax layer to perform the final predictions by converting the logits into probability values. The overall loss L is computed as a weighted sum of individual losses for each subtask ie., $L = \sum_{i=1}^{I} w_i L_i$, allowing the model to optimize both the tasks simultaneously, where I denotes the number of labels and w_i represents the loss weights for each subtask. This design facilitates effective learning across tasks by sharing feature vectors, improving generalization and task-specific performance.

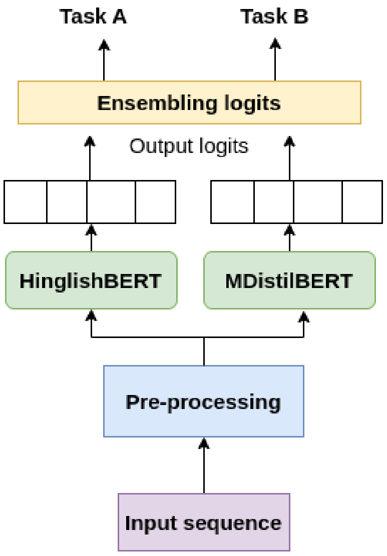


Figure 2: Framework of the proposed Ensemble_MTL model

3.3 Ensemble_MTL Model

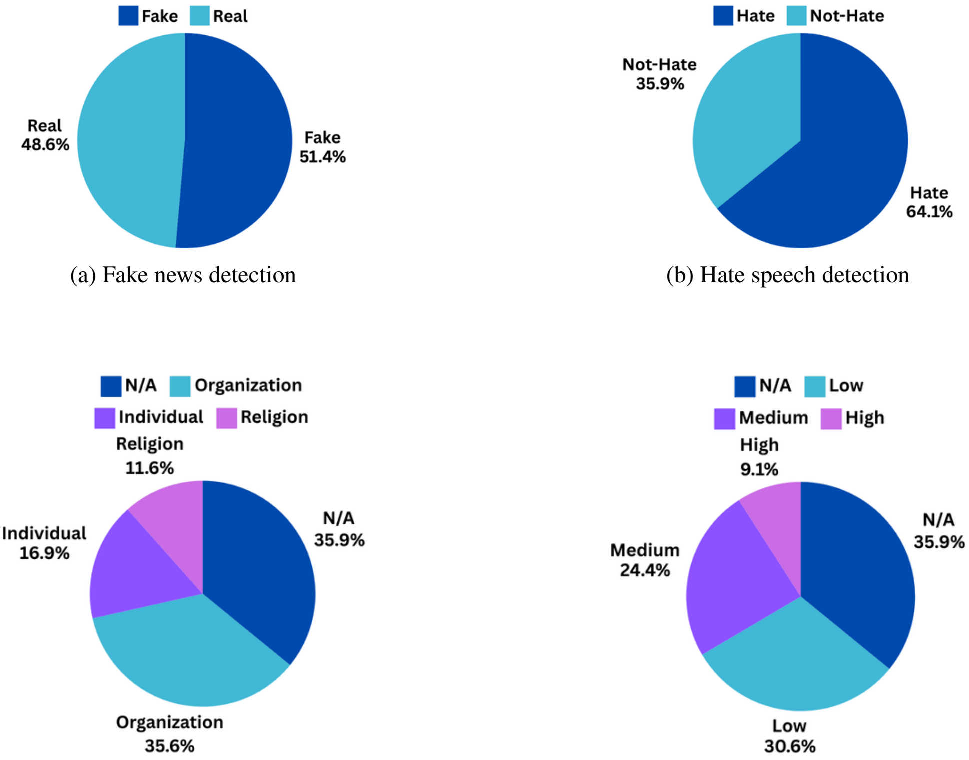
Unlike the Hing_MTL model, which relies on a single pre-trained model for MTL, the Ensemble_MTL model combines the strengths of two pre-trained models, HinglishBERT and MDistilBERT, through ensembling. Pre-processed text is independently passed through both the models to generate text representations, which are then combined via logit averaging to address the challenges of the shared task effectively. Figure 2 depicts the framework of the proposed Ensemble_MTL model.

Ensemble_MTL model makes use of HinglishBERT and MDistilBERT models to extract hidden patterns and generate contextual representations from the input text. HinglishBERT is tailored for Hinglish text, while MDistilBERT - a multilingual pre-trained model, is trained on a vast corpus of multilingual text including Hindi and English. Leveraging the transformer architecture, MDistilBERT learns contextualized word representations across multiple languages, accommodating both native scripts and romanized versions.

Model Construction

As illustrated in Figure 2, the proposed Ensemble_MTL architecture utilizes contextual representations using HinglishBERT and MDistilBERT models for the given Hinglish tweets. To achieve this, both the models are trained independently for

[7] https://huggingface.co/nirantk/hinglish-bert



(a) Fake news detection

(b) Hate speech detection

(c) Target detection

(d) Severity detection

Figure 3: Classwise distribution of Hinglish Train dataset

Hyperparameters	**Values**
Epoch	20
Batch size	16
Learning rate	1e-5
Optimizer	AdamW
Loss function	CrossEntropyLoss

Table 1: Hyperparameters and their values used to train the proposed Hing_MTL and Ensemble_MTL models

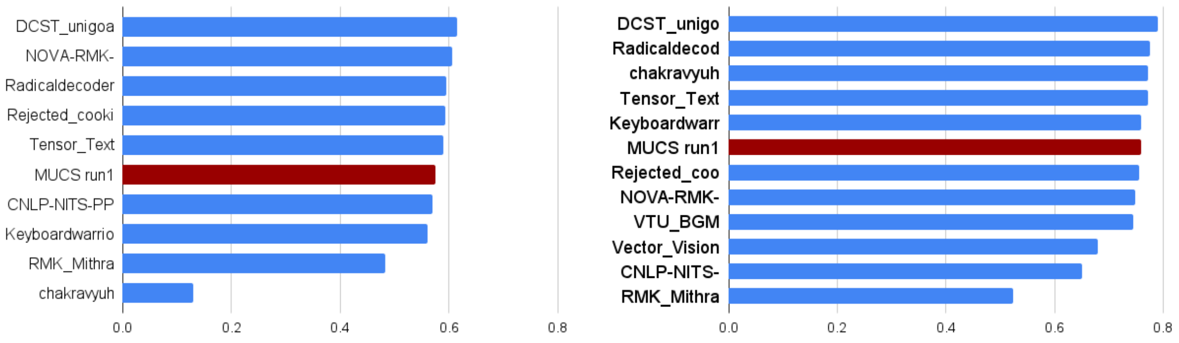
Hing_MTL				
Tasks	**Macro F1 score**	**Precision**	**Recall**	**Accuracy**
Task A	0.7026	0.7294	0.7248	0.7275
Task B	0.4359	0.4987	0.4113	0.4551
Ensemble_MTL				
Task A	**0.7589**	0.7890	0.7368	0.7714
Task B	**0.5746**	0.6460	0.6038	0.6324

Table 2: Performances of the proposed MTL models

multiple tasks and the task-specific logits are computed for each task. During inference, the ensemble method is applied by averaging the logits obtained from HinglishBERT and MDistilBERT models. This approach leverages strengths of both the pre-trained models, combining their learned representations to produce a more robust prediction. The averaged logits are then passed through an argmax operation to determine the final predicted class. This ensemble strategy helps to mitigate individual model biases and enhances overall performance by aggregating knowledge from multiple pre-trained models. Similar to Hing_MTL model, overall loss is computed for each of the models separately.

4 Experiments and Results

This paper utilizes the Hinglish tweets dataset (Biradar et al., 2024b) provided by the organizers of the shared task and Figure 3 illustrates the classwise distribution of the dataset. A series of experiments are conducted using various transformer models to perform MTL, both with a single pretrained model and an ensemble of multiple pretrained models. The models that performed well on the Validation sets are evaluated on the Test sets. To this end, Hing_MTL and Ensemble_MTL models are implemented to tackle both Task A and Task B. The hyperparameters and their values used to implement the proposed MTL models are provided in Table 1. Further, performances of the proposed models for Test sets are shown in Table 2. From the table, it is clear that the Ensemble_MTL model



(a) Task A

(b) Task B

Figure 4: Comparison of performances of the participating teams in the shared task

Task	Sl. No.	Sample comment	Actual label		Predicted label	
Task A	1.	@RanaAyyub Tablighi yahan corona kaa ilaaj dene aaye thhey?	Real	Not-Hate	Fake	Hate
	2.	Allah itna kamjor hai ki shaitan ko kuchh nahi kar sakta	Fake	Not-Hate	Fake	Hate
Task B	3.	Es dadi ko 500 or dedo over acting k	Individual	Low	N/A	N/A
	4.	Pehle wala fake news tha ya fir aaj wala... Kaunsa wala bharosa karoo	N/A	N/A	Organization	Low

Table 3: Samples of misclassification for Hinglish Test set

outperforms the Hing_MTL model. This can be attributed to the strength of the ensemble approach, where the weakness of one model is compensated by the strength of another, leading to overall improved performance. Figure 4 compares the performances of the participants with the proposed Ensemble_MTL model, measured by macro F1 scores in the shared task. This comparison highlights the effectiveness of the Ensemble_MTL model relative to the other approaches utilized by the participants in the shared task.

The misclassified comments in the given Hinglish Test set, along with the actual labels and predicted labels (obtained from Ensemble_MTL model) are shown in Table 3. From the table, it can be observed that actual labels of sample 1 are 'Real' and 'Not-Hate' for Task A. But, the model has predicted 'Fake' and 'Hate' as labels for this sample as the content words 'corona' and 'ilaaj' (Treatment) appear in the Train set in a negative tone. Further, for sample 4, the actual labels (for Target and Severity tasks) are N/A (no label specified). But, the model has predicted the labels 'Organization' and 'Low'. The reason could be the words 'news' and 'bharosa' (Trust) which appear in the context misleads the classifier, resulting in these predictions.

5 Conclusion and Future work

In this paper, we - team MUCS describe the two distinct models submitted to 'Faux-Hate' shared task organized at ICON 2024. Experiments are carried out with: i) Hing_MTL - a MTL model implemented using the pre-trained HinglishBERT model, and ii) Ensemble_MTL - a MTL model that combines HinglishBERT and MDistilBERT, to detect fake and hate content with additional emphasis on identifying the target and severity of hate speech. The proposed Ensemble_MTL model outperformed the Hing_MTL model, achieving macro F1 scores of 0.7589 and 0.5746 for Task A and Task B, respectively, securing 6th rank for both subtasks in the shared task. These results demonstrate the effectiveness of the ensemble approach in improving model performance by leveraging the strengths of multiple pre-trained models. Optimization of the ensemble method by integrating additional models and utilizing MTL framework for better generalization across diverse datasets and languages will be explored further.

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. In *Language Resources and Evaluation*, pages 1–32.

Bharathi Raja Chakravarthi, Prasanna Kumaresan, Ruba Priyadharshini, Paul Buitelaar, Asha Hegde, Hosahalli Shashirekha, Saranya Rajiakodi, Miguel Ángel García, Salud María Jiménez-Zafra, José García-Díaz, et al. 2024. Overview of Third Shared Task on Homophobia and Transphobia Detection in Social Media Comments. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 124–132.

Sharal Coelho, Asha Hegde, G Kavya, and Hosahalli Lakshmaiah Shashirekha. 2023. Mucs@ dravidianlangtech2023: Malayalam Fake News Detection using Machine Learning Approach. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 288–292.

Xinyu Cui and Li Yang. 2022. Fake News Detection in Social Media Based on Multi-modal Multi-task Learning. In *International Journal of Advanced Computer Science and Applications*. Science and Information (SAI) Organization Limited.

Kirsten Gollatz and Leontine Jenner. 2018. Hate Speech and Fake News–How Two Concepts Got Intertwined and Politicised. In *encore*, page 62.

Asha Hegde, Mudoor Devadas Anusha, and Hosahalli Lakshmaiah Shashirekha. 2021. Ensemble Based Machine Learning Models for Hate Speech and Offensive Content Identification. In *FIRE (Working Notes)*, pages 132–141.

Asha Hegde, Bharathi Raja Chakravarthi, Hosahalli Lakshmaiah Shashirekha, Rahul Ponnusamy, Subalalitha Cn, SK Lavanya, Durairaj Thenmozhi, Martha Karunakar, Shreya Shreeram, and Sarah Aymen. 2023a. Findings of the shared task on sentiment analysis in tamil and tulu code-mixed text. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 64–71.

Asha Hegde, G Kavya, Sharal Coelho, and Hosahalli Lakshmaiah Shashirekha. 2023b. MUCS@ Dravidianlangtech2023: Leveraging Learning Models to Identify Abusive Comments in Code-mixed Dravidian Languages. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 266–274.

Asha Hegde, G Kavya, Sharal Coelho, and Hosahalli Lakshmaiah Shashirekha. 2023c. MUCS@ LT-EDI2023: Homophobic/Transphobic Content Detection in Social Media Text using mBERT. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 287–294.

Asha Hegde and Hosahalli Lakshmaiah Shashirekha. 2021. Urdu Fake News Detection Using Ensemble of Machine Learning Models. In *FIRE (Working Notes)*, pages 1190–1198.

Asha Hegde and Hosahalli Lakshmaiah Shashirekha. 2022. Leveraging Dynamic Meta Embedding for Sentiment Analysis and Detection of Homophobic/Transphobic Content in Code-mixed Dravidian Languages. In *FIRE (Working Notes)*, pages 147–156.

Prashant Kapil and Asif Ekbal. 2020. A Deep Neural Network Based Multi-task Learning Approach to Hate Speech Detection. In *Knowledge-Based Systems*, page 106458. Elsevier.

Prashant Kapil, Gitanjali Kumari, Asif Ekbal, Santanu Pal, Arindam Chatterjee, and B. N. Vinutha. 2023. HHSD: Hindi Hate Speech Detection Leveraging Multi-Task Learning. In *IEEE Access*, pages 101460–101473.

Qing Liao, Heyan Chai, Hao Han, Xiang Zhang, Xuan Wang, Wen Xia, and Ye Ding. 2021. An Integrated Multi-task Model for Fake News Detection. In *IEEE Transactions on Knowledge and Data Engineering*, pages 5154–5165.

Flor Miriam Plaza-Del-Arco, M. Dolores Molina-González, L. Alfonso Ureña-López, and María Teresa Martín-Valdivia. 2021. A Multi-Task Learning Approach to Hate Speech Detection Leveraging Sentiment Analysis. In *IEEE Access*, pages 112478–112489.

Zeerak Waseem, James Thorne, and Joachim Bingel. 2018. Bridging The Gaps: Multi Task Learning for Domain Transfer of Hate Speech Detection. pages 29–55.

Lianwei Wu, Yuan Rao, Haolin Jin, Ambreen Nazir, and Ling Sun. 2019. Different Absorption from the Same Sharing: Sifted Multi-task Learning for Fake News Detection. In *arXiv preprint arXiv:1909.01720*.

Yu Zhang and Qiang Yang. 2018. An Overview of Multi-task Learning. pages 30–43.

Yu Zhang and Qiang Yang. 2021. A Survey on Multitask Learning. pages 5586–5609.