SSN_MMHS@DravidianLangTech 2025: A Dual Transformer Approach for Multimodal Hate Speech Detection in Dravidian Languages

Jahnavi Murali and **Rajalakshmi Sivanaiah**
Department of Computer Science and Engineering
Sri Sivasubramaniya Nadar College of Engineering
Chennai, India
jahnavi2110854@ssn.edu.in, rajalakshmis@ssn.edu.in

Abstract

The proliferation of the Internet and social media platforms has resulted in an alarming increase in online hate speech, negatively affecting individuals and communities worldwide. While most research focuses on text-based detection in English, there is an increasing demand for multilingual and multimodal approaches to address hate speech more effectively. This paper presents a methodology for multiclass hate speech classification in low-resource Indian languages namely, Malayalam, Telugu, and Tamil, as part of the shared task at DravidianLangTech 2025. Our proposed approach employs a dual transformer-based framework that integrates audio and text modalities, facilitating cross-modal learning to enhance detection capabilities[1]. Our model achieved macro-F1 scores of 0.348, 0.1631, and 0.1271 in the Malayalam, Telugu, and Tamil subtasks respectively. Although the framework's performance is modest, it provides valuable insights into the complexities of multimodal hate speech detection in low-resource settings and highlights areas for future improvement, including data augmentation and alternate fusion and feature extraction techniques.

1 Introduction

The rise of social media has amplified the spread of hate speech, making it a pervasive and pressing issue. Online hate harms both victims and observers, often leading to issues like depression, isolation, social anxiety, and loss of confidence (Walther, 2022). To address this problem and ensure a safer online space, researchers have focused extensively on developing hate speech detection methods for text-based content, while detection in audio data remains underexplored, presenting unique challenges and opportunities (Bhesra et al., 2024). Furthermore, most studies in this domain are primarily limited to English, underscoring the critical need for robust multilingual and multimodal hate speech detection systems(Chhabra and Vishwakarma, 2023; Nandi et al., 2024) in low-resource monolingual and code-mixed languages (Premjith et al., 2024a).

This work aims to advance the development of efficient multimodal fine-grained hate speech detection systems for low-resource Dravidian languages by utilizing intermediate fusion techniques to enhance cross-modal learning. The paper is organized as follows: Previous research on hate speech detection in multimodal and multilingual settings are discussed in Section 2. The datasets used and the proposed system along with the architecture diagram are outlined in Sections 3 and 4 respectively. Results obtained are presented in Section 5, and the paper concludes with intent and direction for future work.

2 Related Works

Detecting hate speech in low-resource and code-mixed languages presents unique challenges due to linguistic diversity and the lack of annotated data sets. Several studies have explored different approaches to tackle this issue. Premjith et al. (2024b) reported on the Multimodal Social Media Data Analysis in Dravidian Languages (MSMDA-DL) shared task, which addressed sentiment analysis, abusive language detection, and hate speech detection in Tamil and Malayalam using multimodal data from YouTube. The best-performing team Rahman et al. (2024) combined ConvLSTM for video, BiLSTM for audio, and Naive Bayes for text for the abusive language detection subtask, demonstrating the effectiveness of ensemble methods. Imbwaga et al. (2024) proposed an audio-based approach for English and Kiswahili, highlighting the need for improved feature modeling in low-resource speech datasets. Transformer-based models have also been widely studied, with Sivanaiah et al. (2023) and M et al. (2023) evaluating var-

[1]Code for this work is available on GitHub

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 297–301

Language	Non-Hate (N)	Hate (H)				Total
		G (Gender)	P (Political)	R (Religious)	C (Personal Defamation)	
Tamil	287	63	33	61	65	509
Telugu	198	101	58	72	122	551
Malayalam	406	82	118	91	186	883

Table 1: Dataset description: The number of instances in each class (Non-Hate and Hate subcategories) for Tamil, Telugu, and Malayalam languages.

ious transformer and machine learning techniques for code-mixed Dravidian languages. Similarly, Sreelakshmi et al. (2024) analyzed multilingual embeddings for Dravidian languages, finding that MuRIL combined with SVM performed best while employing cost-sensitive learning to address class imbalance. These studies emphasize the importance of annotated datasets and specialized embeddings for effective hate speech detection in underrepresented languages.

Beyond text-based analysis, incorporating the speech modality has also proven to be crucial for hate speech detection, as vocal tone, prosody, and speech patterns provide additional context that aids in identifying offensive content. Bhesra et al. (2024) explored a multimodal framework combining MFCC audio features with text embeddings, employing a decision-level fusion strategy to improve detection accuracy. Similarly, Rana and Jha (2022) integrated audio-based emotion features with text-based semantic representations, demonstrating the benefits of cross-modal learning. Mandal et al. (2023) introduced a multimodal Transformer-based model, leveraging log mel spectrograms for speech and tokenized embeddings for text, with a novel "Attentive Fusion layer" to effectively combine both modalities. These works highlight the potential of multimodal architectures in detecting hate speech by leveraging complementary information from audio and text.

3 Dataset Description

In this study, we used the dataset provided by the organizers (Lal G et al., 2025; Anilkumar et al., 2024), comprising hate speech utterances collected from YouTube videos in three Indian languages: Malayalam, Tamil, and Telugu. The dataset includes both audio files (.wav format) and their corresponding text transcripts (.xlsx files). It is annotated with two primary classes: Hate (H) and Non-Hate (N). The Hate class is further categorized into the following subclasses based on the type of hate speech: Gender-based Hate (G), Political Hate (P), Religious Hate (R), and Personal Defamation (C). Upon verifying the dataset, we found that out of the 556 records in the Telugu .xlsx file, five corresponding audio files were missing, bringing the total number of matched audio-text pairs to 551. Similarly, in Tamil, out of 514 records, five audio files were missing, reducing the total available Tamil dataset to 509 multimodal instances. These final counts are reflected in Table 1, which reports the dataset distribution across languages.

4 Methodology

We implemented a multimodal dual Transformer-based framework for multiclass hate speech detection, drawing significant inspiration from Mandal et al. (2023), within the novel context of low-resource Indian languages and multiclass classification. An overview of the architecture is presented in Figure 1.

4.1 Data Preprocessing

Mel-Frequency Cepstral Coefficients (MFCCs) were extracted from the audio data to capture frequency-related characteristics essential for detecting subtle variations in tone and speech patterns indicative of hate speech. Specifically, 40 MFCC features were computed on audio originally sampled at 44.1 kHz, with a Mel filterbank size (n_mels) of 128. Text data was tokenized to a maximum length of 128 tokens using the IndicBART tokenizer (Dabre et al., 2022).

4.2 Audio and Text Sampling

The audio and text data were then processed through dedicated sampling blocks. For audio data, the Speech Sampling Block involved passing the MFCC features through an LSTM layer to capture sequential dependencies, followed by the application of positional encodings to enhance temporal relationships. Specifically, the LSTM retains and updates hidden states over time via input, output,

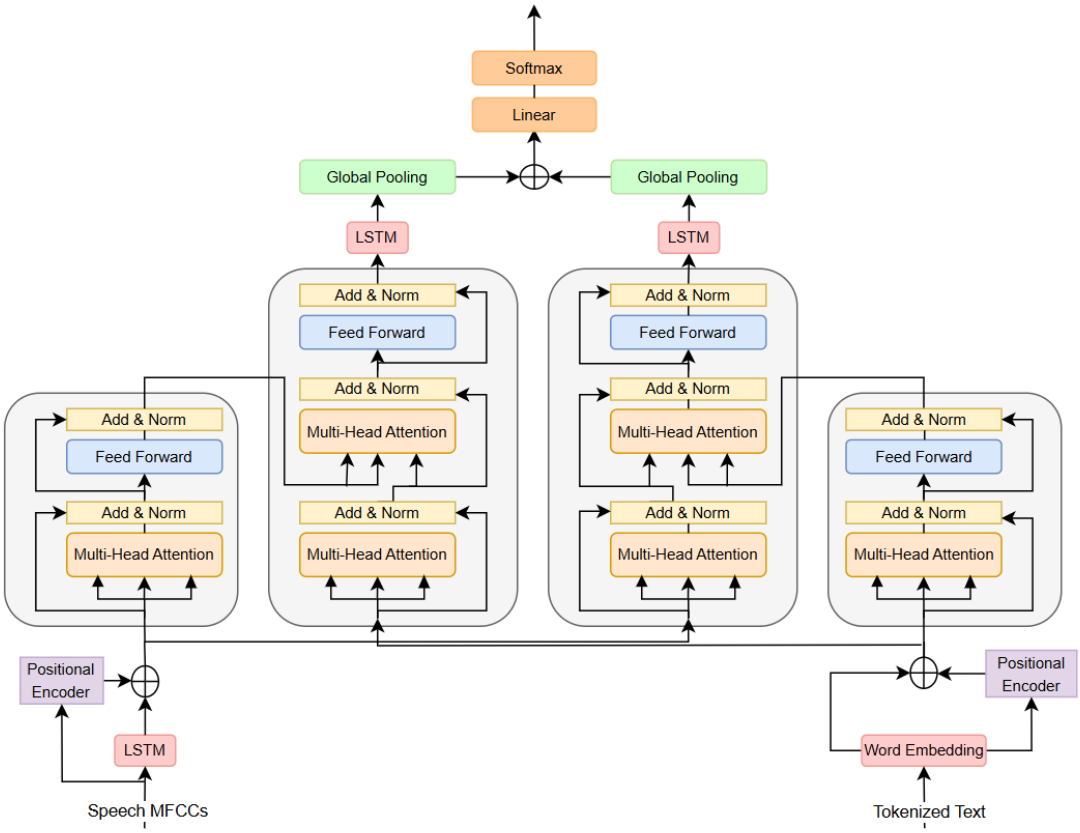

Figure 1: Model Architecture

and forget gates, allowing it to learn how consecutive MFCC frames connect. This gating mechanism leverages both short-term and long-term context, which enables effective modeling of continuous speech signals.

Similarly, the Text Sampling Block processed tokenized text sequences through an embedding layer and positional encodings. These sampling blocks ensured that both modalities were appropriately prepared for cross-modal interaction within the Transformer framework.

4.3 Dual Transformer Framework

The core architecture consists of two vanilla Transformers, introduced by Vaswani et al. (2017), each composed of two encoder and decoder layers with four attention heads (num_heads=4). The embedding size (d_model) was set to 128, and the feed-forward network (dff) had a dimension of 256. In the first transformer, the encoder processes speech features to extract contextual embeddings, which are then attended by the decoder, utilizing text features as input. This setup allows the decoder to learn from the contextual representations generated by the audio encoder. Conversely, in the second transformer, the encoder processes text features to produce contextual embeddings, while the decoder attends to these embeddings using audio features as input. This bidirectional mechanism facilitates effective cross-modal knowledge transfer, enabling each modality to enrich the other.

The outputs from both transformers are processed through separate LSTM layers, which capture long-term dependencies in the sequential data, enhancing the model's temporal understanding. These LSTM outputs then undergo global average pooling to reduce dimensionality while preserving essential information. The pooled features are subsequently concatenated, creating a unified representation of the audio and text modalities. Finally, this fused representation is passed through a dense layer with a softmax activation, classifying each input into one of five target categories.

4.4 Training Setup

The model was trained for 20 epochs using the Adam optimizer with a learning rate of 1e-4. To prevent overfitting, a dropout rate of 0.1 was applied, and an early stopping callback was employed with a patience value of 3.

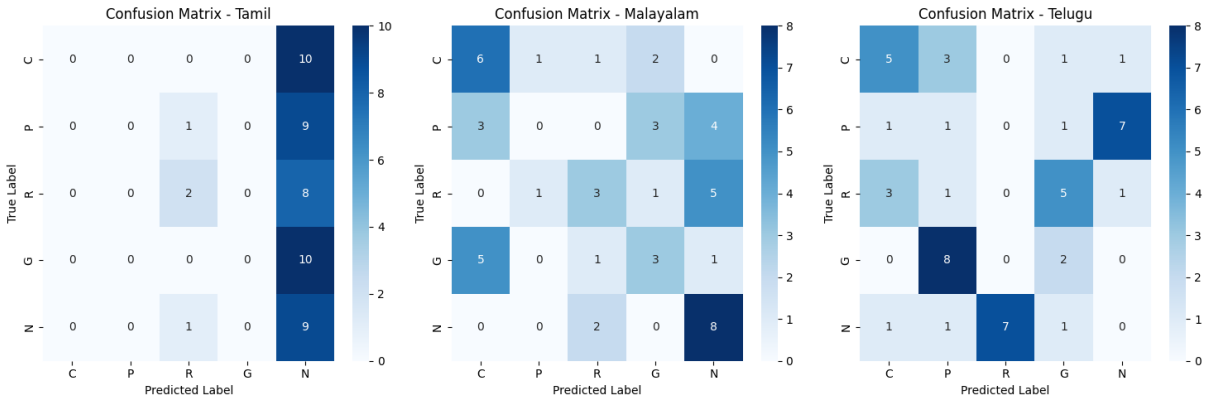

Figure 2: Confusion Matrices for Tamil, Malayalam, and Telugu Test Sets

5 Discussion of Results and Limitations

We submitted our runs for the Telugu and Malayalam subtasks, achieving macro-F1 scores of 0.1631 and 0.3480, respectively. These results placed us 14th in the Telugu subtask and 11th in the Malayalam subtask. Additionally, we conducted further experiments on the Tamil language. The macro-F1 scores for a subset of the dataset used for validation and the test set for each subtask are presented in Table 2. The variation in scores between the validation and test sets can be attributed to class imbalance. The training and validation sets contained a higher proportion of Non-Hate (N) instances, leading the model to overfit on the majority class. In contrast, the test sets had an equal distribution of 50 instances per class, exposing the model's difficulty in generalizing across underrepresented hate speech categories.

Subtask	Macro F1 Score (Val)	Macro F1 Score (Test)
Malayalam	0.6757	0.3480
Telugu	0.5146	0.1631
Tamil	0.2815	0.1271

Table 2: Macro F1 scores on validation and test sets for Malayalam, Telugu, and Tamil.

To further analyze model performance across languages, we present confusion matrices for Tamil, Malayalam, and Telugu test sets in Figure 2. These matrices provide insights into the distribution of misclassifications. The Tamil confusion matrix indicates a heavy bias toward predicting the 'N' (Non-Hate) class, resulting in a lack of diversity in predictions. Malayalam's confusion matrix, in contrast, shows a relatively balanced distribution with a few misclassifications spread across categories, indicating that the model is better at distinguishing between different classes but still struggles with borderline cases. For Telugu, the model misclassifies several instances of 'R' and 'N', suggesting that these categories are not well-separated in the learned representation.

While the model identified some patterns, its overall effectiveness was limited. The strong bias toward predicting the N (Non-Hate) class in Tamil and Telugu suggests difficulty in distinguishing hate speech categories due to insufficient minority class examples. In contrast, the slightly better performance in Malayalam highlights the impact of a larger dataset, though misclassifications persisted. These results emphasize the challenges of deep learning in low-resource languages, where data scarcity and class imbalance hinder accurate, nuanced hate speech detection.

6 Conclusion and Future Work

In this study, we explored a multimodal dual Transformer-based approach for hate speech detection in Dravidian languages. Our findings highlight the challenges of using deep learning for low-resource languages, where limited training data and class imbalance impact model performance. However, they also demonstrate the potential of multimodal approaches to leverage complementary audio and text information, offering promising directions for future advancements. We intend to explore alternative fusion and feature extraction techniques, class-weighting, and implement data augmentation strategies, such as backtranslation and Gaussian noise addition to enhance the model's ability to generalize and improve overall performance in future work.

References

Abhishek Anilkumar, Jyothish Lal G, B Premjith, and Bharathi Raja Chakravarthi. 2024. Dravlangguard: A multimodal approach for hate speech detection in dravidian social media. In *Speech and Language Technologies for Low-Resource Languages (SPELLL)*, Communications in Computer and Information Science.

Kirtilekha Bhesra, Shivam Ashok Shukla, and Akshay Agarwal. 2024. Audio vs. text: Identify a powerful modality for effective hate speech detection. In *The Second Tiny Papers Track at ICLR 2024*.

Anusha Chhabra and Dinesh Kumar Vishwakarma. 2023. A literature survey on multimodal and multilingual automatic hate speech identification. *Multimedia Systems*, 29(3):1203–1230.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh Khapra, and Pratyush Kumar. 2022. Indicbart: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics.

Joan L Imbwaga, Nagatatna B Chittaragi, and Shashidhar G Koolagudi. 2024. Automatic hate speech detection in audio using machine learning algorithms. *International Journal of Speech Technology*, 27(2):447–469.

Jyothish Lal G, B Premjith, Bharathi Raja Chakravarthi, Saranya Rajiakodi, Bharathi B, Rajeswari Natarajan, and Rajalakshmi Ratnavel. 2025. Overview of the Shared Task on Multimodal Hate Speech Detection in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Hema M, Anza Prem, Rajalakshmi Sivanaiah, and Angel Deborah S. 2023. Athena@DravidianLangTech: Abusive comment detection in code-mixed languages using machine learning techniques. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 147–151, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Atanu Mandal, Gargi Roy, Amit Barman, Indranil Dutta, and Sudip Kumar Naskar. 2023. Attentive fusion: A transformer-based approach to multimodal hate speech detection. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 720–728, Goa University, Goa, India. NLP Association of India (NLPAI).

Arpan Nandi, Kamal Sarkar, Arjun Mallick, and Arkadeep De. 2024. A survey of hate speech detection in indian languages. *Social Network Analysis and Mining*, 14(1):70.

B Premjith, Bharathi Raja Chakravarthi, Prasanna Kumar Kumaresan, Saranya Rajiakodi, Sai Karnati, Sai Mangamuru, and Chandu Janakiram. 2024a. Findings of the shared task on hate and offensive language detection in telugu codemixed text (hold-telugu)@ dravidianlangtech 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 49–55.

B Premjith, G Jyothish, V Sowmya, Bharathi Raja Chakravarthi, K Nandhini, Rajeswari Natarajan, Abirami Murugappan, B Bharathi, Saranya Rajiakodi, Rahul Ponnusamy, et al. 2024b. Findings of the shared task on multimodal social media data analysis in dravidian languages (msmda-dl)@ dravidianlangtech 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 56–61.

Md. Rahman, Abu Raihan, Tanzim Rahman, Shawly Ahsan, Jawad Hossain, Avishek Das, and Mohammed Moshiul Hoque. 2024. Binary_Beasts@DravidianLangTech-EACL 2024: Multimodal abusive language detection in Tamil based on integrated approach of machine learning and deep learning techniques. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 212–217, St. Julian's, Malta. Association for Computational Linguistics.

Aneri Rana and Sonali Jha. 2022. Emotion based hate speech detection using multimodal learning. *Preprint*, arXiv:2202.06218.

Rajalakshmi Sivanaiah, Rajasekar S, Srilakshmisai K, Angel Deborah S, and Mirnalinee ThankaNadar. 2023. Avalanche at DravidianLangTech: Abusive comment detection in code mixed data using machine learning techniques with under sampling. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 166–170, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

K Sreelakshmi, B Premjith, Bharathi Raja Chakravarthi, and KP Soman. 2024. Detection of hate speech and offensive language codemix text in dravidian languages using cost-sensitive learning approach. *IEEE Access*.

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*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Joseph B. Walther. 2022. Social media and online hate. *Current Opinion in Psychology*, 45:101298.