DravLingua@DravidianLangTech 2025: Multimodal Hate Speech Detection in Dravidian Languages using Late Fusion of Muril and Wav2Vec Models

Aishwarya S
Sri Eshwar College of Engineering
Coimbatore, Tamil Nadu, India
aishwarya.s2020cse@sece.ac.in

Abstract

Detecting hate speech on social media is increasingly difficult, particularly in low-resource Dravidian languages such as Tamil, Telugu and Malayalam. Traditional approaches primarily rely on text-based classification, often overlooking the multimodal nature of online communication, where speech plays a pivotal role in spreading hate speech. We propose a multimodal hate speech detection model using a late fusion technique that integrates Wav2Vec 2.0 for speech processing and Muril for text analysis. Our model is evaluated on the DravidianLangTech@NAACL 2025 dataset, which contains speech and text data in Telugu, Tamil, and Malayalam scripts. The dataset is categorized into six classes: Non-Hate, Gender Hate, Political Hate, Religious Hate, Religious Defamation, and Personal Defamation. To address class imbalance, we incorporate class weighting and data augmentation techniques. Experimental results demonstrate that the late fusion approach effectively captures patterns of hate speech that may be missed when analyzing a single modality. This highlights the importance of multimodal strategies in enhancing hate speech detection, particularly for low-resource languages.

1 Introduction

The rise of hate speech on social media necessitates automated detection for safer online spaces (Schmidt and Wiegand, 2017). While significant progress has been made in high-resource languages like English, research in Tamil, Malayalam, and Telugu remains limited (Zampieri et al., 2019). The linguistic complexity of Dravidian languages—rich morphology, agglutinative structures, and unique syntax—poses additional NLP challenges (Hegde et al., 2021). Hate speech is prevalent in both text and speech, especially on video-sharing and voice-based platforms (Kumar et al., 2021). Advancements in deep learning and transformer models have enabled more accurate multimodal detection (Kiela et al., 2020).

Dravidian languages suffer from insufficient labeled data, limiting supervised learning (Chakravarthi et al., 2021). Hate speech datasets are highly imbalanced, with fewer hateful instances (Saha et al., 2021), and complex linguistic features like phonetic variations and dialectal differences further challenge text and speech processing (Krishnan et al., 2022). While Muril shows promise for Indian language text processing (Khanuja et al., 2021), speech models like Wav2Vec 2.0 require adaptation for Dravidian languages.

This study introduces a multimodal hate speech detection model integrating Muril for text and Wav2Vec 2.0 for speech, employing a late fusion technique to address these challenges.

2 Related Works

Historically, hate speech detection relied on text-based models like SVMs, Naïve Bayes, and Random Forests (Davidson et al., 2017). Deep learning models, including LSTM, CNNs, and Transformers (BERT, RoBERTa, XLM-R), improved performance, especially in English (Zampieri et al., 2019), but struggle with implicit hate, sarcasm, and multimodal cues.

With the rise of speech-based platforms, self-supervised models like Wav2Vec 2.0, HuBERT, and Whisper have replaced MFCC- and HMM-based methods (Baevski et al., 2020). Multimodal approaches, such as transformers integrating text and vision (Kiela et al., 2020) and late fusion combining text and speech at the logit level (Yin and Zubiaga, 2021), have further enhanced detection.

The HOLD-Telugu shared task (Premjith et al., 2024a) highlighted transformer effectiveness in Telugu code-mixed text. Expanding on this, (Premjith et al., 2024b) demonstrated multimodal advantages in hate speech detection, while (Lal G et al., 2025)

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 50–55

introduced cost-sensitive learning for class imbalance in Dravidian text. Data augmentation techniques, including back-translation, paraphrasing, and synthetic generation, have improved text-based detection (Founta et al., 2018), while speed variation, pitch shifting, and noise injection enhance speech model robustness.

3 Dataset and Preprocessing

We use the DravidianLangTech@NAACL 2025 dataset, a benchmark for multimodal hate speech detection in Tamil, Malayalam, and Telugu. It contains text and speech samples from social media, labeled into five classes—one non-hate and four hate categories. Given the skewed class distribution (Fig. 1), specialized preprocessing and augmentation techniques are applied to improve model robustness.

3.1 Text Preprocessing

The text data is preprocessed using Unicode Normalization for consistency, Unwanted Character Removal to retain only meaningful text, Sentence Splitting and Tokenization for structured segmentation, and Stopword Removal to enhance relevance.

3.2 Speech Preprocessing

Speech preprocessing involves resampling all audio samples to 16 kHz to match Wav2Vec 2.0's default input requirements. Noise reduction is applied using spectral subtraction to remove background interference and enhance speech clarity. Finally, feature extraction is performed directly by Wav2Vec 2.0, which generates raw speech embeddings, eliminating the need for manual feature engineering techniques such as MFCCs or spectrogram analysis.

3.3 Data Augmentation

To improve model generalization, data augmentation techniques were employed separately for speech and text because to the class imbalance in the data set. We used data augmentation approaches to improve the generalization and robustness of the model for both audio and textual input.

3.3.1 Text Data Augmentation

- **Synonym Replacement:** Uses a pre-trained FastText model for contextual synonym substitution.
- **Backtranslation:** Introduces lexical and syntactic diversity via intermediate language translation.

3.3.2 Audio Data Augmentation

- **Gaussian Noise Addition:** Injects noise at varying levels (0.005, 0.01, 0.03) to enhance robustness against distortions.

4 Methodology

4.1 Text-Based Model (Muril)

Muril, a transformer-based model pre-trained on 17 Indian languages, excels in Indian language processing, particularly in Dravidian scripts and low-resource settings, outperforming mBERT and XLM-R. It is optimized for hate speech detection using the DravidianLangTech@NAACL 2025 dataset.

Fine-tuning begins with text preprocessing and tokenization using Muril's subword tokenizer. The tokenized input passes through the Muril encoder to generate contextualized embeddings, which are processed by fully connected layers and a softmax classifier to predict six hate speech classes: Non-Hate, Gender Hate, Political Hate, Religious Hate, Religious Defamation, and Personal Defamation. Training is conducted with a batch size of 32, sequence length of 128, and a 3e-5 learning rate using the AdamW optimizer for 10 epochs, with early stopping based on validation loss. Categorical cross-entropy loss is used to optimize the classification problem; the loss function is provided by:

$$\mathcal{L} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{1}$$

where y_i represents the ground-truth label, and $\hat{y}_i$ is the predicted probability for class ii.

4.2 Speech-Based Model (Wav2Vec 2.0)

Wav2Vec 2.0 (Baevski et al., 2020) is used for speech-based hate speech detection, learning speech representations directly from raw audio without phonetic transcriptions. Effective in low-resource settings, it handles dialectal variations in Tamil, Malayalam, and Telugu better than traditional MFCC-based classifiers by capturing nuanced phonetic and prosodic features.

The classification pipeline involves preprocessing (Section 3.2), followed by Wav2Vec 2.0 encoding to generate contextualized embeddings, which

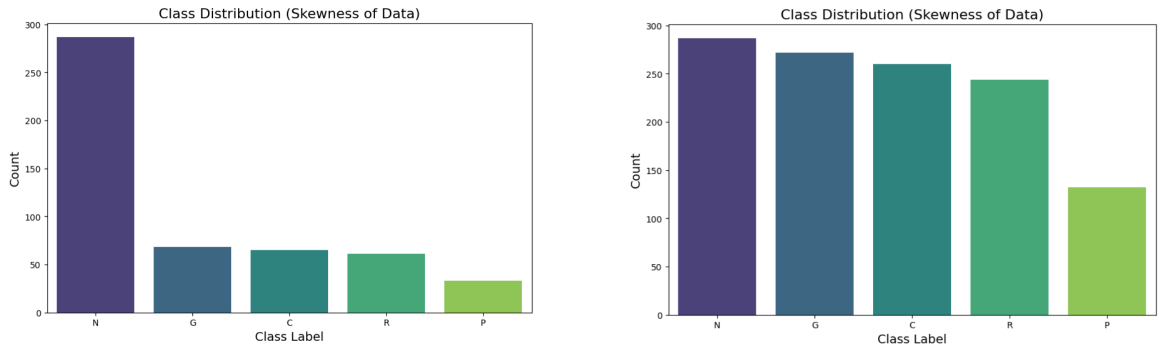

Figure 1: The left side graph depicts the unbalanced data and right side graph is the data distribution after data augmentation

are processed by fully connected layers and classified via softmax. The model is trained independently on the DravidianLangTech@NAACL 2025 dataset using the AdamW optimizer (batch size: 16, learning rate: 2e-5) for 10 epochs, with class weighting to address imbalance. Categorical cross-entropy loss is used for optimization, as in the Muril model.

4.3 Computational Cost

Training was conducted on Google Colab Free with an NVIDIA Tesla T4 GPU (16 GB VRAM), Intel Xeon CPU (2 vCPUs, 2.3 GHz), and 12 GB RAM. Fine-tuning Muril and Wav2Vec 2.0 for Tamil, Malayalam, and Telugu took approximately 1 hour per model over 10 epochs, with GPU utilization reaching 40-60% and peak memory usage of 10 GB.

5 Fusion Techniques

5.1 Early Fusion

Early fusion integrates text and speech features at the representation level by concatenating embeddings from Muril and Wav2Vec 2.0 before classification as shown in Fig. 2. This allows the model to learn cross-modal interactions early in the pipeline. The concatenated feature vector is passed through a shared neural network, which processes both modalities jointly. While early fusion enables deeper multimodal learning, it may introduce modality imbalance, where dominant features, such as text, overshadow weaker ones, such as speech. Additionally, the increased feature dimensionality can lead to overfitting and higher computational costs.

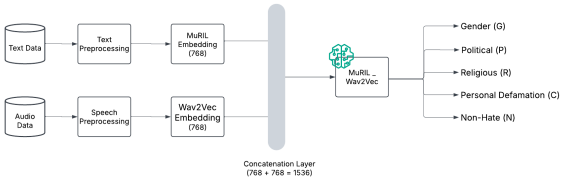

Figure 2: Early Fusion of MuRIL and Wav2Vec for Sentiment Classification

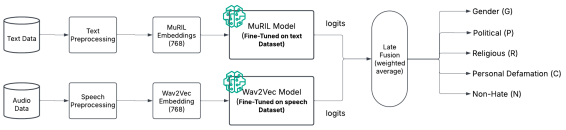

Figure 3: Late Fusion of MuRIL and Wav2Vec for Sentiment Classification

5.2 Late Fusion

Late fusion combines predictions at the decision level rather than merging raw features as shown in Fig. 3. Muril and Wav2Vec 2.0 are trained separately, generating independent class probabilities, P_t for text and P_s for speech. The final classification probability is computed as:

$$P_{\text{final}} = \alpha P_t + (1 - \alpha) P_s \tag{2}$$

where α is a tunable hyperparameter that adjusts the relative contribution of each modality. This approach allows the model to optimize each modality independently before aggregation, reducing the risk of feature redundancy and overfitting.

5.3 Comparison of Fusion Strategies

Early fusion provides stronger cross-modal feature interaction but may suffer from modality dominance and increased computational demands. In contrast, late fusion ensures independent optimization of each modality, offering greater flexibility

in weighting text and speech contributions. By exploring both techniques, we aim to determine the most effective strategy for multimodal hate speech detection.

6 Result

The results of our multimodal hate speech detection model across Tamil, Malayalam, and Telugu demonstrate variations in performance based on fusion strategies and training approaches.

Language	F1 - Train set	F1 - Test set
Tamil	0.79	0.48
Malayalam	0.83	0.51
Telugu	0.73	0.40

Table 1: Train and Test Results using Early fusion

Language	Text		Audio		F1
	Train	Test	Train	Test	Test set
Tamil	0.84	0.74	0.43	0.38	0.71
Malayalam	0.69	0.75	0.65	0.40	0.75
Telugu	0.82	0.35	0.4	0.26	0.17

Table 2: F1 Scores using class weighting (Late Fusion)

Language	Text		Audio		Late
	Train	Test	Train	Test	Fusion
Tamil	0.84	0.69	0.94	0.38	0.70

Table 3: F1 Scores - Augmented data (Late fusion)

6.1 Early Fusion Performance and Overfitting

Early fusion results (Table 1) indicate that while the model achieves relatively high F1-scores on the train set (0.79–0.83), the test set performance drops significantly (0.40–0.51), suggesting overfitting. This is likely due to the absence of explicit regularization techniques such as dropout or weight decay. The model memorizes training patterns but fails to generalize well on unseen data.

6.2 Late Fusion Generalization

Unlike early fusion, late fusion achieves better generalization without explicit regularization. This suggests that independent training of text and audio modalities before aggregation helps mitigate overfitting. Class weighting further balances the contributions of both modalities, leading to improved test performance for Tamil (0.71) and Malayalam (0.75) as shown in (Table 2). However, Telugu's performance remains weak across all modalities. Table 4 highlights that while classes like R and P perform well, G suffers from poor recall (0.30), indicating difficulty in identifying certain instances, which suggests modality-specific challenges.

Class	Precision	Recall	F1-score
C	0.58	0.70	0.64
N	0.56	0.90	0.69
R	0.82	0.90	0.86
P	0.88	0.70	0.78
G	1.00	0.30	0.46
Accuracy		0.70	
Macro Avg	0.77	0.70	0.69
Weighted Avg	0.77	0.70	0.69

Table 4: Classification report of the model trained using the late fusion-class weighting approach.

6.3 Impact of Data Augmentation

Data augmentation (Table 3) improves model robustness, particularly for Tamil, where the test F1-score for text increases to 0.69, and late fusion achieves 0.70. However, augmentation has a minimal effect on Telugu, reinforcing the hypothesis that linguistic characteristics and data sparsity play a larger role in its underperformance.

6.4 Analysis of Telugu's Underperformance

Telugu shows the weakest performance across all models, especially in late fusion (0.17), due to its high phonetic and syntactic diversity, which hinders both text and speech models. Additionally, Wav2Vec 2.0, trained on high-resource languages, struggles with Telugu's unique phonetic structure, leading to lower classification accuracy.

7 Conclusion

This study examines multimodal hate speech detection in Tamil, Malayalam, and Telugu using Muril and Wav2Vec 2.0. Comparing fusion strategies, we find that early fusion enables cross-modal interactions but suffers from overfitting, while late fusion generalizes better by optimizing text and speech models independently.

Class weighting and data augmentation enhance performance, particularly for Tamil and Malayalam, though Telugu remains challenging due to linguistic complexity and data sparsity. Future work will focus on reducing overfitting with regularization techniques, evaluating advanced transformer models, and improving interpretability for better linguistic adaptation.

The implementation of our model, including preprocessing and training scripts, is publicly available at GitHub Repository.

8 Limitations

The class disparity is still a problem in the DravidianLangTech@NAACL 2025 dataset, especially for hate categories related to politics and religion. The lack of pre-trained models for Dravidian languages, along with background noise and accent fluctuation, make speech processing difficult. Because the text-based model performs better than the speech-based model, the modalities' contributions are unbalanced, and late fusion is unable to adequately reflect their complex interconnections. Changes in hate speech patterns and a lack of discourse-level knowledge hinder generalization to real-world contexts.

9 Ethics Statement

This research focuses on improving multimodal hate speech detection while ensuring fairness, transparency, and ethical considerations. We acknowledge the potential for bias in dataset distribution, which may affect classification performance across different hate speech categories. To mitigate this, we incorporate class balancing techniques and assess misclassification trends through error analysis. All data used in this study is publicly available, and no personally identifiable information was processed. While our model aims to enhance online safety, we recognize the risks of false positives and false negatives, which highlight the need for human oversight in real-world applications. We encourage responsible AI deployment and emphasize that this work should not be used to unjustly suppress free speech but rather to foster safer online interactions, particularly in low-resource languages.

References

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Raghavendra Kumar, and E. Sherly. 2021. Dravidiancodemix: Sentiment analysis and offensive language identification dataset for dravidian languages. In *Forum for Information Retrieval Evaluation (FIRE)*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

Jayaprakash Hegde, Raghavendra Kumar, Bharathi Raja Chakravarthi, and K. P. Soman. 2021. Classification of offensive language in dravidian languages using deep learning approaches. In *Forum for Information Retrieval Evaluation (FIRE)*.

Simran Khanuja, Sudip Dandapat, Raghavendra Kumar, Sunayana Sitaram, and Monojit Choudhury. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vinayak Goswami, Davide Testuggine, and Peter West. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Prasanna Krishnan, Anand Subramanian, Bharathi Raja Chakravarthi, and K. P. Soman. 2022. Speech recognition in tamil and malayalam using self-supervised learning. In *Proceedings of the ACL Conference on Computational Linguistics*.

Anil Kumar, Ajay Kumar Ojha, Shervin Malmasi, and Marcos Zampieri. 2021. Benchmarking aggression identification in social media for low-resource languages. In *Proceedings of the Workshop on Online Abuse and Harms (ACL)*.

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.

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.

Prithviraj Saha, Binny Mathew, Narendra Ghanghor, Pawan Goyal, and Animesh Mukherjee. 2021. Hate speech detection in indic languages: A comparative study. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the International Workshop on NLP for Social Media (ACL)*.

Jun Yin and Arkaitz Zubiaga. 2021. Multimodal hate speech detection: A review and open challenges. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of NAACL-HLT*.