The_Deathly_Hallows@DravidianLangTech 2025: Multimodal Hate Speech Detection in Dravidian Languages

Kogilavani Shanmugavadivel[1] **, Malliga Subramanian**[2]**,**
Vasantharan K[1]**,Prethish G A**[1]**,Santhosh S**[3]
[13]Department of AI, Kongu Engineering College, Perundurai, Erode.
[2]Department of CSE, Kongu Engineering College, Perundurai, Erode.
{kogilavani.sv, mallinishanth72}@gmail.com
{vasantharank.ncc, prethish0409, santhosh42169}@gmail.com

Abstract

The DravidianLangTech@NAACL 2025 shared task focused on multimodal hate speech detection in Tamil, Telugu, and Malayalam using social media text and audio. Our approach integrated advanced preprocessing, feature extraction, and deep learning models. For text, preprocessing steps included normalization, tokenization, stopword removal, and data augmentation. Feature extraction was performed using TF-IDF, Count Vectorizer, BERT-base-multilingual-cased, XLM-Roberta-Base, and XLM-Roberta-Large, with the latter achieving the best performance. The models attained training accuracies of 83% (Tamil), 88% (Telugu), and 85% (Malayalam). For audio, Mel Frequency Cepstral Coefficients (MFCCs) were extracted and enhanced with augmentation techniques such as noise addition, time-stretching, and pitch-shifting. A CNN-based model achieved training accuracies of 88% (Tamil), 88% (Telugu), and 93% (Malayalam). Macro F1 scores ranked Tamil 3rd (0.6438), Telugu 15th (0.1559), and Malayalam 12th (0.3016). Our study highlights the effectiveness of text-audio fusion in hate speech detection and underscores the importance of preprocessing, multimodal techniques, and feature augmentation in addressing hate speech on social media.

1 Introduction

Social media has revolutionized communication, allowing global information sharing. However, hate speech targeting groups based on race, religion, gender, or political views has also surged, presenting serious societal issues. Detecting hate speech, especially in low resource languages, is crucial for machine learning and NLP. Dravidian languages (Tamil, Telugu, and Malayalam) pose significant challenges due to their linguistic complexity and limited computational resources. Furthermore, these languages are often code-mixed with English, complicating NLP tasks. Effective hate speech detection requires multimodal models, as social media content is often a combination of text, audio, and images.

The DravidianLangTech@NAACL 2025 Multimodal Hate Speech Detection Shared Task addressed these challenges by providing datasets in Malayalam, Tamil, and Telugu, categorizing hate speech into Gender, Political, Religious, Personal Defamation, and Non-Hate Lal G et al. 2025.

We used extensive preprocessing, such as IndicNormalizerFactory for normalization, 'indic_tokenize.trivial_tokenize' for tokenization, stopword removal, and nlpaug for data augmentation. For audio, we extracted Mel-Frequency Cepstral Coefficients (MFCCs) using librosa and applied techniques like noise addition, time-stretching, and pitch-shifting. Feature extraction was done using TF-IDF, Count Vectorization, and XLM-Roberta-Large embeddings, while a CNN-based model analyzed MFCCs for audio classification.

Our models ranked third in Tamil, twelfth in Malayalam, and fifteenth in Telugu in the shared task. Despite challenges like class imbalances and noisy annotations, our results demonstrate that advanced preprocessing, feature extraction, and deep learning can effectively tackle these multimodal issues. This paper discusses our methods, challenges, and lessons learned in improving hate speech detection for low-resource languages.

2 Literature Review

Sreelakshmi et al. 2024 used machine learning classifiers and multilingual embeddings for hate speech detection in CodeMix Dravidian languages. MuRIL performed best for Malayalam, and a cost-sensitive strategy addressed class imbalance. A new annotated Malayalam-English dataset was introduced.

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 713–717

Premjith et al. 2024a highlighted challenges in Telugu CodeMix text and the need for hate speech detection. MPNet and CNNs were used in the HOLD-Telugu shared task, utilizing macro F1-scores for binary classification. Premjith et al. 2024b showed monolingual models' limitations for Tamil and Malayalam and emphasized multilingual BERT and multimodal analysis (text, audio, video) for better performance. Chakravarthi et al. 2023 proposed a fusion model combining MPNet and CNN for detecting offensive language in CodeMix Dravidian languages, achieving high F1-scores (Tamil: 0.85, Malayalam: 0.98, Kannada: 0.76). Imbwaga et al. 2024 explored hate speech detection in English and Kiswahili audio. Spectral, temporal, and prosodic features were extracted, with XGBoost excelling in Kiswahili and Random Forest in English, highlighting language-specific classifier importance. Premjith et al. 2023 discussed a Dravidian Languages Workshop task on multimodal abusive language detection and sentiment analysis. Sixty teams participated, using macro F1-scores for evaluation. Barman and Das 2023 addressed abusive language detection in Tamil and Malayalam, achieving an F1-score of 0.5786 using mBERT, ViT, and MFCC.

An et al. 2024 introduced two explainable audio hate speech detection methods: End-to-End (E2E) and a cascade approach. E2E performed better, with frame-level justifications improving accuracy. Koreddi et al. 2024 presented an AI system for detecting objectionable content across text, audio, and visual media, integrating speech recognition, OCR, NLP, and Google Translator. BERT was used for text analysis, enhancing online content safety. Narula and Chaudhary 2024 studied hate speech detection challenges in Hindi, focusing on dialects, code-switching, and Romanized Hindi. Advanced machine learning was suggested to combat misinformation and societal unrest.

3 Task Description

The Multimodal Hate Speech Detection Shared Task at DravidianLangTech@NAACL 2025 challenges researchers to detect hate speech in Tamil, Malayalam, and Telugu using multimodal social media data. Participants receive training data with text and speech components to classify hate speech into gender, political, religious, and personal defamation categories. Models are evaluated using the macro-F1 score.

4 Dataset Description

Anilkumar et al. 2024 present a dataset comprising YouTube videos in Tamil, Malayalam, and Telugu, incorporating text and audio features. Hate speech is classified into gender (G), political (P), religious (R), and personal defamation (C) categories. Text preprocessing includes tokenization, stopword removal, and augmentation, while audio features are extracted using MFCC with enhancements like noising, time-stretching, and pitch-shifting.

Language	Non-Hate	Hate (C,G,P,R)
Malayalam	406	477
Tamil	287	227
Telugu	198	358

Table 1: Text Data Distribution

Language	Non-Hate	Hate (C,G,P,R)
Malayalam	406	477
Tamil	287	222
Telugu	198	353

Table 2: Audio Data Distribution

5 Methodology

The multimodal hate speech detection approach categorizes content as hate or non-hate speech, with subclasses for gender, political, religious, and personal defamation. Text and audio data from YouTube videos undergo preprocessing, including tokenization, stopword removal, augmentation, and MFCC-based audio feature extraction with noise addition and pitch/time-stretching. Text features are derived using TF-IDF, Count Vectorizer, and transformer embeddings (e.g., XLM-Roberta), while MFCCs are used for audio. A fully connected model classifies text, while CNN handles audio. Optimization techniques such as dropout, and early stopping enhance performance. Models are evaluated using the macro F1 score, addressing linguistic and multimodal challenges while ensuring scalability.

5.1 Data Preprocessing

Preprocessing enhances model performance by cleaning, normalizing, and augmenting text and audio data, ensuring consistency and meaningful feature extraction.

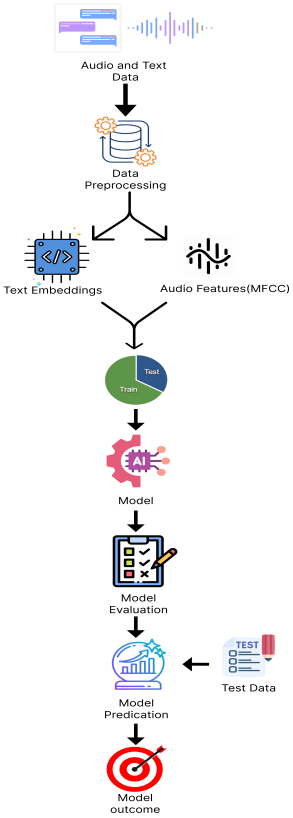

Figure 1: Proposed Model Workflow

5.1.1 Text Preprocessing and Feature Extraction

Tamil, Telugu, and Malayalam text is normalized using IndicNormalizerFactory to handle script variations. Tokenization is performed with 'indic_tokenize.trivial_tokenize', and stopwords are removed using custom lists (Malayalam) and advertools (Tamil, Telugu). nlpaug is used for data augmentation (synonym replacement, word insertion). For feature extraction, TF-IDF, Count Vectorizer, and transformer embeddings (XLM-Roberta-Base, XLM-Roberta-Large, BERT Multilingual Cased) are used. XLM-Roberta-Large provides the best accuracy due to its strong multilingual capabilities.

5.1.2 Audio Preprocessing and Feature Extraction

Audio features are extracted using MFCCs, which mimic human auditory perception. Augmentation techniques (noise addition, pitch/time-stretching) improve model robustness. These processed features ensure stability under varying conditions. These processed features were essential in training effective multimodal models.

5.2 Model Architecture

The multimodal system classifies hate speech using separate architectures for text and audio.

5.2.1 Text Classification Model

XLM-Roberta-Large and BERT-Multilingual Cased embeddings capture contextual features. The model includes dense layers with ReLU, batch normalization, dropout, and a softmax output layer for classification (Gender, Political, Religious, Personal Defamation Hate, Non-Hate). Optimized with Adam, categorical cross-entropy, and evaluated using macro-F1 score.

5.2.2 Audio Classification Model

A CNN-based model processes 2D MFCC features from speech data. Convolutional layers with increasing filters extract hierarchical patterns, while dropout after pooling prevents overfitting. A dense layer refines features before the softmax output. Training is optimized with ReduceLROnPlateau, EarlyStopping, ModelCheckpoint.

Together, these models form a robust multimodal hate speech detection system, excelling in both text and audio classification.

6 Limitations

Although this study of demonstrates strong performance, certain limitations remain. The primary challenge is the limited dataset size, which affects the model's ability to generalize across diverse hate speech patterns. Additionally, class imbalance in the dataset impacts the fair representation of all categories, leading to biased classification. While data augmentation helps improve model robustness, it cannot fully compensate for the lack of real-world diversity in the dataset. Future work can focus on expanding the dataset and implementing more effective balancing techniques to enhance performance.

7 Performance Evaluation

The models were evaluated using classification metrics such as precision, recall, F1-score, and accuracy to assess their effectiveness in detecting hate and non-hate speech across languages. Precision measured the model's ability to minimize false positives, while recall indicated its capacity to detect relevant samples. The F1-score, a harmonic mean of precision and recall, was crucial for handling class imbalances. Accuracy represented the proportion of correctly predicted instances. The best per-

Models Used	Tamil	Telugu	Malayalam
BERT-base-multilingual-cased	67%	71%	73%
TF-IDF Vectorizer	55%	58%	58%
CountVectorizer	59%	60%	58%
XLM-Roberta-base	73%	71%	72%
XLM-Roberta-large	83%	88%	85%

Table 3: Performance of Different Models in Tamil, Telugu, and Malayalam (text)

Models Used	Tamil	Telugu	Malayalam
Without Preprocessing	64%	54%	85%
TTSAudio + Dynamic Normalizer	58%	54%	80%
Data Augmentation	85%	82%	90%
Data Augmentation with BatchNormalization	88%	88%	93%

Table 4: Performance of Different Models in Tamil, Telugu, and Malayalam (audio)

formance in text classification was achieved using XLM-Roberta-Large embeddings, while MFCCs and augmentation improved audio classification. These metrics provided a comprehensive evaluation of the models across multiple modalities and languages.

8 Conclusion

This study aimed to develop a robust multimodal hate speech detection system for Malayalam, Tamil, and Telugu. Using textual data with "XLM-Roberta-Large" embeddings and audio data with "MFCCs" and data augmentation, the system categorized hate speech into gender, political, religious, and personal defamation. The models achieved third place in the shared challenge, with Tamil scoring 0.6438, and Telugu and Malayalam scoring 0.1559 and 0.3016, respectively. Despite challenges like unbalanced datasets and complex cultural contexts, this work demonstrates how multimodal techniques can address hate speech detection in under-represented languages. Future work could explore larger datasets, additional modalities, specialized architectures for multimodal tasks, and better domain adaptation strategies to further enhance performance. This study provides a foundation for ensuring linguistic inclusion, improving online safety, and enabling fully automated hate speech detection. The source code for our approach is available at https://github.com/vasantharan/Multimodal_Hate_Speech_Detection_in_Dravidian_languages.

References

Jinmyeong An, Wonjun Lee, Yejin Jeon, Jungseul Ok, Yunsu Kim, and Gary Geunbae Lee. 2024. An investigation into explainable audio hate speech detection. *arXiv preprint arXiv:2408.06065*.

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.

Shubhankar Barman and Mithun Das. 2023. hate-alert@ dravidianlangtech: Multimodal abusive language detection and sentiment analysis in dravidian languages. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 217–224.

Bharathi Raja Chakravarthi, Manoj Balaji Jagadeeshan, Vasanth Palanikumar, and Ruba Priyadharshini. 2023. Offensive language identification in dravidian languages using mpnet and cnn. *International Journal of Information Management Data Insights*, 3(1):100151.

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.

Venkatesh Koreddi, Nalluri Manisha, Shaik Mohammad Kaif Mohammad Kaif, and Yeligeti Tejaswa Sai Kumar. 2024. Multilingual ai system for detecting offensive content across text, audio, and visual media. *Engineering Research Express*.

Jyothish Lal G, B Premjith, Bharathi Raja Chakravarthi, Saranya Rajiakodi, Bharathi B, Rajeswari Nataraj an, 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.

Rachna Narula and Poonam Chaudhary. 2024. A comprehensive review on detection of hate speech for multi-lingual data. *Social Network Analysis and Mining*, 14(1):1–35.

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.

B Premjith, V Sowmya, Bharathi Raja Chakravarthi, Rajeswari Natarajan, K Nandhini, Abirami Murugappan, B Bharathi, M Kaushik, Prasanth Sn, et al. 2023. Findings of the shared task on multimodal abusive language detection and sentiment analysis in tamil and malayalam. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 72–79.

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*.