KEC_AI_BRIGHTRED@DravidianLangTech 2025: Multimodal Hate Speech Detection in Dravidian languages

Kogilavani Shanmugavadivel[1] **, Malliga Subramanian**[2]**,**
Nishdharani P[1]**,Santhiya E**[1]**,Yaswanth Raj E**[1]
[1]Department of AI, Kongu Engineering College, Perundurai, Erode.
[2]Department of CSE, Kongu Engineering College, Perundurai, Erode.
{kogilavani.sv, mallinishanth72}@gmail.com
{nishdharanip05, santhiyae587, yaswanthraje2004}@gmail.com

Abstract

Hate speech detection in multilingual settings presents significant challenges due to linguistic variations and speech patterns across different languages. This study proposes a fusion-based approach that integrates audio and text features to enhance classification accuracy in Tamil, Telugu, and Malayalam. We extract Mel-Frequency Cepstral Coefficients and their delta variations for speech representation, while text-based features contribute additional linguistic insights. Several models were evaluated, including BiLSTM, Capsule Networks with Attention, Capsule-GRU, ConvLSTM-BiLSTM, and Multinomial Naïve Bayes, to determine the most effective architecture. Experimental results demonstrate that Random Forest performs best for text classification, while CNN achieves the highest accuracy for audio classification. The model was evaluated using the Macro F1 score and ranked ninth in Tamil with a score of 0.3018, ninth in Telugu with a score of 0.251, and thirteenth in Malayalam with a score of 0.2782 in the Multimodal Social Media Data Analysis in Dravidian Languages shared task at DravidianLangTech@NAACL 2025. By leveraging feature fusion and optimized model selection, this approach provides a scalable and effective framework for multilingual hate speech detection, contributing to improved content moderation on social media platforms.

1 Introduction

With the rise of social media and digital communication, hate speech has become a major concern, particularly in multilingual communities. Traditional hate speech detection methods primarily rely on text analysis, but spoken content, such as audio messages and voice notes, also plays a crucial role in spreading harmful discourse. Detecting hate speech in languages like Tamil, Telugu, and Malayalam presents unique challenges due to code-mixing, informal language structures, and phonetic variations.

This study addresses these challenges by incorporating both text and audio-based features to improve classification accuracy. We extract MFCC and delta features from speech data and apply various deep learning and machine learning models to analyze textual content. By evaluating models such as BiLSTM, Capsule Networks, GRU, ConvLSTM, and Naïve Bayes, we identify the most effective approach for each modality. Our results demonstrate that Random Forest performs best for text-based hate speech detection, while CNN excels in audio-based classification. This research contributes to enhancing multilingual content moderation by leveraging both acoustic and linguistic features for more robust hate speech detection.

2 Literature Survey

(Lal G et al., 2025), presented an overview of the shared task on multimodal hate speech detection in Tamil, Telugu, and Malayalam at DravidianLangTech@NAACL 2025. It discusses dataset creation, preprocessing techniques, and model performance in identifying hate speech across text and audio modalities.(Premjith et al., 2024) analyzed multimodal social media data, including text, audio, and video from platforms like Twitter and YouTube. Their study focused on sentiment analysis, abusive language detection, and hate speech detection. The results were presented for Dravidian languages.(Sreelakshmi et al., 2024) showed that MuRIL embeddings with an SVM (RBF kernel) performed well across six datasets. The highest accuracies were 66% (Kannada), 72% (Tamil), and 96% (Malayalam) for DravidianLangTech 2021. HASOC 2021 achieved 68% (Malayalam) and 76% (Tamil), while HASOC 2020 reached 92% (Malayalam). (Mohan et al., 2023) proposed a multimodal approach for hate speech detection in Tamil us-

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 763–767

ing TimeSformer for video, Wav2vec2 for audio, and BERT-based models for text. They achieved 81.82% accuracy (F1: 68.65%) for text, 63.63% accuracy (F1: 50.60%) for audio, and 45.45% accuracy (F1: 33.64%) for video. By combining features from all modalities, they achieved 81.82% accuracy and a 66.67% F1 score. (Arunachalam and Maheswari, 2024) proposed a method to detect hateful remarks in Dravidian languages on social media. Using mBERT with CATBOOST and GSCV, they achieved F1 scores of 0.94 (Tamil), 0.98 (Malayalam), and 0.82 (Kannada) on the Dravidian Code-Mix FIRE 2021 dataset. Their approach effectively analyzed YouTube comments using various preprocessing techniques and binary classifiers. (Roy et al., 2022) developed a deep ensemble framework using deep learning and transformer models to detect offensive posts in Tamil-Malayalam code-mixed text. Their approach achieved weighted F1-scores of 0.802 (Malayalam) and 0.933 (Tamil). The model outperformed state-of-the-art methods on these datasets. (Dhanya and Balakrishnan, 2021) promoted the creation of an automated hate speech detection system for Malayalam by presenting a survey.

3 Task Description

The task aims to develop an effective multimodal hate speech detection system that can process and classify hate speech in both textual and audio formats across Tamil, Telugu, and Malayalam. The challenge is divided into three key components: the first focuses on detecting hate speech from textual data by classifying transcripts into "hate speech" or "non-hate speech." The second component deals with audio-based classification, where audio features (e.g., MFCCs, spectral features) are extracted and used to identify hate speech. The core of the task involves multimodal fusion, where the outputs of text-based classification using Random Forest and audio-based classification using CNN are combined to enhance overall detection accuracy. The models will be evaluated using metrics such as accuracy, precision, recall, and F1-score on both individual and multimodal data.

4 Dataset description

The dataset used for this task consists of multimodal data from social media platforms, specifically targeting hate speech detection in Tamil, Telugu, and Malayalam. It contains both text and audio data, with each instance representing a piece of social media content that could potentially contain hate speech.

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

The dataset is categorized into two main classes: Hate and Non-Hate. The hate speech instances are further classified into four subclasses based on their nature: Gender (G), Political (P), Religious (R), and Personal Defamation (C).

5 Methodology

In this section, we outline the approach taken for multimodal hate speech detection using both text and audio data.

5.1 Preprocessing Data

5.1.1 Text

Preprocessing Tamil, Malayalam, and Telugu texts involves data cleaning (removal of special characters, numbers, and extra spaces), tokenization using language-specific tools, normalization for spelling variations, and case conversion. TF-IDF represents text numerically, while stemming or lemmatization reduces words to root forms. SMOTE addresses class imbalance by generating synthetic samples. These steps ensure a clean and balanced dataset for effective hate speech detection.

5.1.2 Audio

Preprocessing Tamil, Telugu, and Malayalam audio involves normalization for consistent volume, noise reduction using spectral gating, and resampling (e.g., 16 kHz). Silence trimming removes pauses, and phoneme segmentation improves accuracy. Key acoustic features like MFCCs and Mel Spectrograms are extracted, while speaker normalization minimizes variability. Data augmentation

(e.g., pitch shifting, time-stretching) enhances robustness, ensuring effective hate speech detection in Dravidian languages.

5.2 Models Developed and Evaluated

We explored and compared several models to address the task of multimodal hate speech detection.

5.2.1 Random Forest (Text Classification)

We implemented a Random Forest classifier, an ensemble method known for handling noisy data and capturing complex feature relationships. It achieved the highest accuracy on text data. The model was trained on extracted text features, demonstrating strong performance.

5.2.2 Convolutional Neural Network (CNN) (Audio Classification)

For audio classification, a CNN was used to process MFCC-extracted features, capturing spatial hierarchies and detecting hate speech patterns. The model showed high accuracy. This section outlines the approach for multimodal hate speech detection using text and audio data.

5.2.3 Bi-directional LSTM (BiLSTM)

We experimented with BiLSTM networks for text but found they underperformed compared to the Random Forest model. While BiLSTM captures long-range dependencies, it did not generalize well for hate speech detection in this dataset.

5.2.4 Capsule Networks with Attention-based BiLSTM

We evaluated a Capsule Network with Attention-based BiLSTM to capture spatial hierarchies and key features. Despite its theoretical benefits, it did not outperform the simpler Random Forest or CNN models.

5.2.5 Capsule Networks with GRU

We tested a Capsule Network with GRU, leveraging GRUs for sequential data processing. However, the integration did not improve accuracy, and the model performed worse than others.

5.2.6 ConvLSTM + BiLSTM

We tested the ConvLSTM + BiLSTM model, combining Convolutional LSTM with BiLSTM to capture spatial and temporal dependencies. However, its complexity led to overfitting on the training data. As a result, it performed worse than simpler models with lower accuracy.

5.2.7 Multinomial Naive Bayes (Text Classification)

The Multinomial Naive Bayes model for text classification, it performed poorly due to its inability to handle data noise and its assumption of feature independence, making it unsuitable for multimodal hate speech detection.

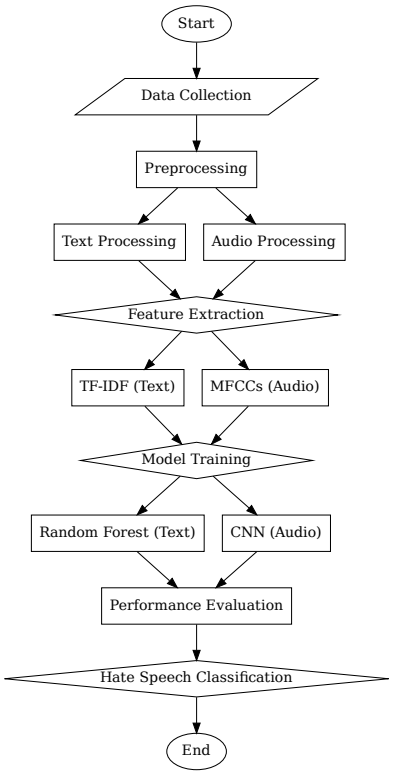

Figure 1: Proposed Model Workflow

5.3 Model Selection

The Random Forest model for text and the CNN model for audio were chosen because of their higher classification task accuracy. Convolutional Neural Networks (CNNs) greatly improved the detection of hate speech in spoken content by exhibiting exceptional efficiency in extracting crucial audio information, such as temporal fluctuations and frequency patterns. CNNs were able to capture complicated audio representations that were frequently missed by standard models by utilizing many layers of feature extraction.

On the other hand, because of its capacity to handle high-dimensional data and capture linguistic subtleties, the Random Forest model demonstrated remarkable efficacy in text classification. By combining several decision trees, Random For-

est's ensemble learning feature enabled strong generalization and enhanced resistance to overfitting. This made it particularly adept at distinguishing subtle variations in textual content, such as sarcasm, implicit biases, and contextual dependencies—factors that are crucial for accurately identifying hate speech in written form.

5.4 Performance Comparison

5.4.1 Text Classification:

The Random Forest classifier performed the best in terms of training accuracy, surpassing BiLSTM and Naive Bayes.

Class/Metric	Precision	Recall	F1-Score
Tamil (Text)			
C	0.93	0.93	0.93
G	0.88	0.89	0.89
N	1.00	0.98	0.99
P	1.00	0.98	0.99
R	0.98	0.95	0.96
Accuracy	-	-	0.94
Macro Avg	0.94	0.94	0.94
Weighted Avg	0.94	0.94	0.94
Telugu (Text)			
C	0.89	0.95	0.92
G	0.74	0.97	0.84
N	0.85	0.79	0.82
P	1.00	0.88	0.93
R	0.97	0.84	0.90
Accuracy	-	-	0.89
Macro Avg	0.89	0.89	0.89
Weighted Avg	0.90	0.88	0.89
Malayalam (Text)			
C	0.87	0.87	0.87
G	0.94	0.88	0.91
N	0.55	0.79	0.65
P	1.00	0.73	0.85
R	1.00	0.73	0.85
Accuracy	-	-	0.82
Macro Avg	0.87	0.82	0.83
Weighted Avg	0.87	0.82	0.83

Table 3: Detailed Classification Report for Tamil, Telugu, and Malayalam (Text)

5.4.2 Audio Classification:

The CNN model outperformed all other deep learning models, including BiLSTM and ConvLSTM-based models, which struggled with audio data.

Class/Metric	Precision	Recall	F1-Score
Tamil (Audio)			
C	1.00	0.20	0.33
G	0.00	0.22	0.36
P	1.00	0.58	0.73
R	0.77	0.77	0.77
N	0.67	0.96	0.79
Accuracy	-	-	0.71
Macro Avg	0.85	0.53	0.58
Weighted Avg	0.76	0.71	0.66
Telugu (Audio)			
C	0.70	0.64	0.67
G	0.71	0.50	0.59
N	0.68	0.90	0.77
P	0.91	0.67	0.77
R	0.80	0.57	0.67
Accuracy	-	-	0.72
Macro Avg	0.76	0.69	0.72
Weighted Avg	0.73	0.72	0.71
Malayalam (Audio)			
C	0.00	0.00	0.00
G	0.33	0.11	0.17
P	0.33	0.17	0.22
R	0.78	0.54	0.64
N	0.63	0.96	0.76
Accuracy	-	-	0.62
Macro Avg	0.42	0.36	0.36
Weighted Avg	0.52	0.62	0.53

Table 4: Detailed Classification Report for Tamil, Telugu, and Malayalam (Audio)

5.5 Performance Evaluation

The performance of the various models used for multimodal hate speech detection, comparing the accuracy of both text and audio classification models. The models explored include BiLSTM, Capsule-based models, ConvLSTM, Multinomial Naive Bayes, Random Forest and CNN. This section presents accuracy results for text models in Tamil, Telugu, and Malayalam. Random Forest outperformed all models, achieving the highest accuracy in Tamil (0.9373), Telugu (0.8838), and Malayalam (0.8153). Multinomial Naive Bayes performed reasonably but was outperformed by Random Forest. BiLSTM and other models like Capsule + Attention-based BiLSTM and ConvLSTM + BiLSTM showed poor performance, with BiLSTM scoring just 0.087 across all languages. CNN for Audio achieved good accuracy, with the highest in Malayalam (0.7577), followed by Telugu (0.7207) and Tamil (0.7059).

Model	Tamil	Telugu	Malayalam
Random Forest (Text)	0.9373	0.8838	0.8153
CNN (Audio)	0.7059	0.7207	0.7577
Capsule+Attention based BiLSTM (Text)	0.5922	0.5434	0.5263
Capsule+GRU (Text)	0.5437	0.5260	0.5132
ConvLSTM+ BiLSTM(Text)	0.5340	0.5421	0.6051
Multinomial Naive Bayes (Text)	0.6699	0.7021	0.6901

Table 5: Accuracy Comparison

6 Limitations

There are limitations to this study that need more research. Performance could be improved by enhancing the fusion strategy with transformer-based solutions. Optimization is required, according to the F1-scores (Tamil: 0.3018, Telugu: 0.251, Malayalam: 0.2782). Complex designs such as ConvLSTM and BiLSTM performed worse than simpler models. Future research should improve models, hone fusion techniques, compare to the most advanced methods, and guarantee reproducibility through open-source implementation.

7 Conclusion

In conclusion, the hate speech recognition system for Tamil, Telugu, and Malayalam showed promising results by integrating preprocessing techniques for both audio and text. The Random Forest model effectively captured semantic features for text, achieving high accuracy, while the CNN model extracted key features from audio signals, also yielding high accuracy. The fusion of these models enhanced performance, enabling more accurate and context-aware predictions by utilizing both modalities. The source code for our approach is available at https://github.com/NishdharaniP/Multimodal_hatespeech_detection.git.

References

V Arunachalam and N Maheswari. 2024. Enhanced detection of hate speech in dravidian languages in social media using ensemble transformers. *Interdisciplinary Journal of Information, Knowledge, and Management*, 19:036.

LK Dhanya and Kannan Balakrishnan. 2021. Hate speech detection in asian languages: a survey. In *2021 international conference on communication, control and information sciences (ICCISc)*, volume 1, pages 1–5. IEEE.

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.

Jayanth Mohan, Spandana Reddy Mekapati, and Bharathi Raja Chakravarthi. 2023. A multimodal approach for hate and offensive content detection in tamil: From corpus creation to model development. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

B Premjith, G Jyothish, V Sowmya, Bharathi Raja Chakravarthi, K Nandhini, Rajeswari Natarajan, Abirami Murugappan, B Bharathi, Saranya Rajiakodi, Rahul Ponnusamy, et al. 2024. 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.

Pradeep Kumar Roy, Snehaan Bhawal, and Chinnaudayar Navaneethakrishnan Subalalitha. 2022. Hate speech and offensive language detection in dravidian languages using deep ensemble framework. *Computer Speech & Language*, 75:101386.

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*.