SSNTrio @ DravidianLangTech 2025: Hybrid Approach for Hate Speech Detection in Dravidian Languages with Text and Audio Modalities

Bhuvana J
Sri Sivasubramaniya Nadar College of Engineering
bhuvanaj@ssn.edu.in

Mirnalinee T T
Sri Sivasubramaniya Nadar College of Engineering
MirnalineeTT@ssn.edu.in

Rohan R
Sri Sivasubramaniya Nadar College of Engineering
rohan2210124@ssn.edu.in

Diya Seshan
Sri Sivasubramaniya Nadar College of Engineering
diya2210208@ssn.edu.in

Avaneesh Koushik
Sri Sivasubramaniya Nadar College of Engineering
avaneesh2210179@ssn.edu.in

Abstract

This paper presents the approach and findings from the Multimodal Social Media Data Analysis in Dravidian Languages (MSMDA-DL) shared task at DravidianLangTech@NAACL 2025. The task focuses on detecting multimodal hate speech in Tamil, Malayalam, and Telugu, requiring models to analyze both text and speech components from social media content. The proposed methodology uses language-specific BERT models for the provided text transcripts, followed by multimodal feature extraction techniques, and classification using a Random Forest classifier to enhance performance across the three languages. The models achieved a macro-F1 score of 0.7332 (Rank 1) in Tamil, 0.7511 (Rank 1) in Malayalam, and 0.3758 (Rank 2) in Telugu, demonstrating the effectiveness of the approach in multilingual settings. The models performed well despite the challenges posed by limited resources, highlighting the potential of language-specific BERT models and multimodal techniques in hate speech detection for Dravidian languages.

1 Introduction

The rapid growth of social media has led to an increase in online hate speech, making automated detection a crucial task for maintaining safe digital spaces. Hate speech refers to content that promotes hate, discrimination, or offensive remarks, while non-hate speech encompasses neutral or non-offensive content. The presence of multimodal content- combining text, speech, and other media- adds complexity to this problem, particularly in underrepresented languages. The Multimodal Social Media Data Analysis in Dravidian Languages (MSMDA-DL) shared task at DravidianLangTech@NAACL 2025 focuses on multimodal hate speech detection in Tamil, Malayalam, and Telugu, presenting unique challenges due to linguistic diversity and resource limitations.

This shared task aims to develop robust models that can analyze textual as well as speech components of social media content and classify them accordingly (Premjith et al., 2024a). The task evaluates models based on their Macro Average F1 score, a common metric used in NLP to measure the performance of classification models. The datasets for Tamil, Malayalam, and Telugu pose distinct challenges, such as variations in script, phonetics, and contextual interpretation of hate speech (Sreelakshmi et al., 2024).

This paper details the methodology used to address these challenges, incorporating language-specific BERT models, followed by multimodal feature extraction, and classification using a Random Forest classifier. The results highlight the effectiveness of the approach in handling multimodal hate speech detection while also emphasizing the challenges in lower-resource languages like Telugu (Premjith et al., 2024b). The findings provide insights into improving multimodal learning for Dravidian languages and contribute to the broader field of hate speech detection in multilingual social media contexts.

2 Related Works

The detection of hate speech in social media has garnered significant attention in recent years due to the growing concerns surrounding online harassment and toxicity. Early studies primarily focused on text-based hate speech detection, employing traditional machine learning techniques such as support vector machines (SVM), random forests, and naive bayes (El-Sayed et al., 2023).

However, with the rise of deep learning, researchers began to explore neural network-based approaches for automatic feature extraction and classification. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) showed promise for handling text classification tasks, especially sentiment and hate speech analysis (Kumar, 2022). More recently, transformer-

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 532–536

based models like BERT have further advanced the field, enhancing text classification performance through contextualized word embeddings (Saleh et al., 2021).

In the context of multimodal hate speech detection, the inclusion of audio, visual, and text modalities has been explored to improve classification accuracy. These approaches are especially effective in identifying subtle forms of hate speech, where non-verbal cues and tone of speech are crucial. For instance, work on hate speech detection in video content has incorporated both speech recognition and computer vision techniques to capture the audio and visual aspects of hate speech (Das et al., 2023).

Despite advancements in multimodal hate speech detection, challenges persist, especially in low-resource languages like Tamil, Malayalam, and Telugu. While some progress has been made through language-specific models and dataset augmentation (Azam et al., 2022), hate speech detection in these languages is still under-researched. The presented work uses language-specific BERT models and multimodal feature extraction to improve performance for these languages.

3 Dataset Description

The dataset provided for this task consists of text and speech components sourced from social media platforms (Lal G et al., 2025). The dataset has been curated to reflect real-world social media discourse, ensuring a diverse representation of linguistic patterns, phonetic variations, and hate speech expressions in Tamil, Malayalam, and Telugu.

The Tamil, Malayalam, and Telugu train datasets consist of 514, 883, and 556 rows respectively. The test datasets of the 3 languages consist of 50 rows each.

Each data sample comprises:

- **Text Modality**: Transcribed text extracted from social media posts, incorporating code-mixed language, informal expressions, and slang commonly used in online communication.
- **Speech Modality**: Audio samples corresponding to spoken content, covering diverse accents, intonations, and pronunciations specific to each Dravidian language.

The dataset is annotated for hate speech classification, with labels indicating the presence or absence of hate speech. The labeling schema categorizes hate speech based on its type and severity across three languages: Malayalam, Tamil, and Telugu. Each language dataset includes two main classes - Hate and Non-Hate (N). The Hate class is further divided into four subclasses: Gender (G), Political (P), Religious (R), and Personal Defamation (C). Detailed dataset statistics and description are provided in Table 1.

Labels	Tamil	Malayalam	Telugu
G	68	82	106
R	61	91	72
P	33	118	58
C	65	186	122
N	287	406	198

Table 1: Distribution of Hate Speech Labels

4 Methodology

4.1 Dataset Preprocessing

The first step in the methodology involves preprocessing both the text and audio data. For the text data, tokenization is applied, stop words are removed, and the text is cleaned by eliminating unnecessary characters, such as punctuation and special symbols. The audio data undergoes noise reduction using spectral subtraction to improve quality and ensure consistency. In spectral subtraction, the noise spectrum is estimated from silent or low-energy portions of the audio signal, and this is subtracted from the speech signal. This results in faster training and improved accuracy, leading to more reliable predictions in tasks such as hate-speech detection.

4.2 Data Upsampling

To address the class imbalance in the dataset, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to upsample the minority class. SMOTE generates synthetic samples by interpolating between existing minority class instances and their nearest neighbors. This process balances the dataset, ensuring that the classifier is not biased towards the majority class. By incorporating these synthetic samples, the model's ability to effectively learn from both classes is enhanced, improving overall classification performance. All minority classes were upsampled to an equivalent count to match that of the majority class, ensuring a more balanced representation across all classes.

4.3 Feature Extraction

In the feature extraction phase, the preprocessed text is passed through a language specific BERT model such as Tamil BERT, Malayalam BERT or Telugu BERT, depending on the sub task, which generates embeddings that provide contextualized word representations. These embeddings are then used as the feature set for the classifier. Subsequently, features are extracted from the audio using techniques including Mel-Frequency Cepstral Coefficients (MFCC) and spectral representations, which capture key auditory information. These multimodal features, representing both textual and auditory aspects of the data, are utilized as additional input features for the classification task.

4.4 Model Building

After feature extraction, the text and audio features were combined and tested with various classification models, including Random Forest, Support Vector Machine (SVM), and other suitable algorithms, to predict the target labels based on the multimodal features.

Among these classifiers, Random Forest yielded the best results due to its ability to handle high-dimensional feature spaces by combining multiple decision trees, which reduces overfitting and improves generalization. It naturally performs feature selection, focusing on the most relevant data, further enhancing its performance.

Additionally, Random Forest is more robust compared to SVM, which requires careful parameter tuning for varying feature scales. Its ability to manage feature importance and adapt to diverse data made it the most effective model for this task.

5 Result Analysis

The performance of the model was evaluated using standard metrics, including accuracy, precision, recall, and F1-score. The results for each metric are shown in Table 2, which outlines the model's performance across different evaluation criteria.

The proposed model demonstrated strong performance across multiple low-resource languages, achieving macro-F1 scores of 0.7332 (Rank 1) in Tamil, 0.7511 (Rank 1) in Malayalam, and 0.3758 (Rank 2) in Telugu. Figures 1-3 illustrate the confusion matrix for all the three models. From these matrices, it is evident that the confusion matrices for Tamil and Malayalam are quite similar.

The model's strong macro-F1 scores for both

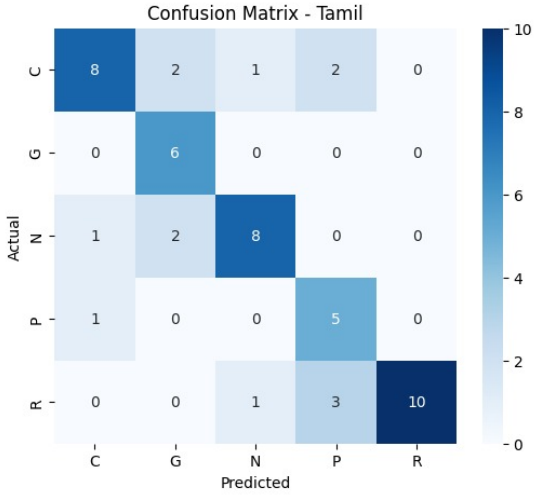

Figure 1: Confusion Matrix for Tamil Dataset

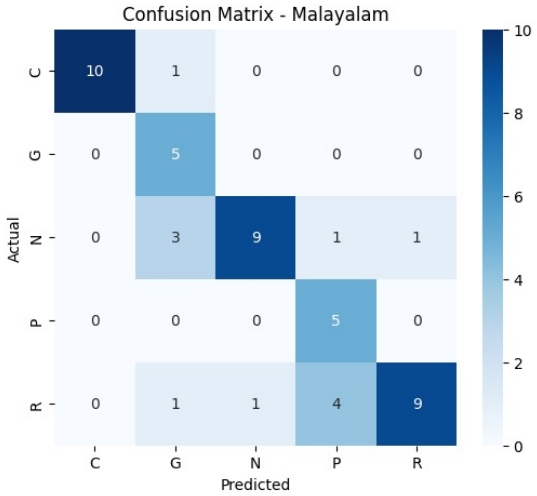

Figure 2: Confusion Matrix for Malayalam Dataset

Tamil and Malayalam suggest that the combination of BERT for text processing, MFCC-based audio features, and Random Forest for feature fusion was highly effective in capturing the linguistic and acoustic nuances of these languages. The high macro-F1 scores indicate that the model was able to balance precision and recall effectively in the presence of class imbalances.

However, the macro-F1 score for Telugu was notably lower compared to Tamil and Malayalam, which could be attributed to differences in pronunciation, phonetics, and linguistic patterns. The model can further be enhanced using advanced techniques and better fine-tuning in order to improve its performance for Telugu.

Overall, these findings demonstrate the efficacy of a multimodal approach for detecting hate speech in low-resource languages, showing that it can sur-

Language	Precision	Recall	F1 Score	Accuracy
Tamil	0.78	0.74	0.73	0.74
Malayalam	0.83	0.76	0.75	0.76
Telugu	0.43	0.36	0.38	0.36

Table 2: Performance Metrics

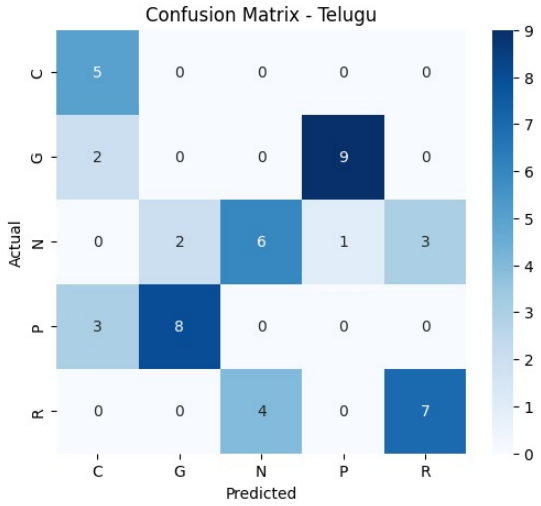

Figure 3: Confusion Matrix for Telugu Dataset

pass conventional text-only techniques by utilizing complimentary data from both modalities.

6 Conclusion

In conclusion, the proposed multimodal model for hate speech detection, which integrates both audio and text inputs, offers a promising approach for improving performance in low-resource languages. By leveraging the unique features of both modalities, this approach improves the model's ability to effectively identify hate speech, especially in underrepresented languages. The proposed method uses language-specific BERT models for text and traditional audio extraction techniques like Mel-Frequency Cepstral Coefficients (MFCC), allowing for a more robust detection process.

The results highlight a significant improvement in performance when combining audio and text, demonstrating the potential of multimodal approaches in detecting hate speech within resource-constrained environments. Overall, our approach presents a reliable and efficient solution for detecting hate speech in low-resource languages, paving the way for future advancements in the field of multimodal hate speech detection.

7 Future Enhancements

For future enhancements, one promising direction is the fine-tuning of models with domain-specific data. Although the current models have been trained on general datasets, focusing on more specific domains or social media platforms could provide more context-aware models for detecting hate speech, especially in niche topics.

Multi-modal fusion techniques also offer an exciting avenue for further enhancement. Standard techniques are already used to combine text and audio features, but exploring more sophisticated fusion strategies such as early, late, or hybrid fusion could lead to better integration of these modalities. This could improve the model's ability to effectively capture the interaction between text and speech, particularly in situations where one modality (e.g., audio) complements the other (e.g., text).

8 Limitations

While the proposed multimodal model demonstrates improved performance in hate speech detection for low-resource languages, several limitations must be considered. First, the effectiveness of the model is highly dependent on the availability and quality of labeled datasets for both text and audio modalities. Many low-resource languages lack sufficiently large and diverse datasets, which can hinder the model's generalizability.

Additionally, the reliance on language-specific BERT models may introduce biases if the pretraining data does not adequately represent different dialects, variations, or informal speech patterns. In the audio modality, challenges such as background noise, variations in pronunciation, and differences in recording quality can affect feature extraction techniques like Mel-Frequency Cepstral Coefficients (MFCC), potentially leading to misclassifications.

Moreover, multimodal models require higher computational resources compared to unimodal approaches, making real-time deployment and scalability in resource-constrained environments challenging. The fusion of audio and text features also

introduces complexities in feature alignment, particularly when dealing with asynchronous or incomplete data inputs.

Furthermore, interpretability remains a concern, as transformer-based models and deep learning approaches often function as black-box systems, making it difficult to provide clear explanations for classification decisions. Addressing these limitations through improved data augmentation, noise-robust feature extraction, and explainability techniques will be crucial for enhancing the effectiveness and practicality of multimodal hate speech detection models.

References

Ubaid Azam, Hammad Rizwan, and Asim Karim. 2022. Exploring data augmentation strategies for hate speech detection in Roman Urdu. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4523–4531, Marseille, France. European Language Resources Association.

Mithun Das, Rohit Raj, Punyajoy Saha, Binny Mathew, Manish Gupta, and Animesh Mukherjee. 2023. Hatemm: A multi-modal dataset for hate video classification. *Preprint*, arXiv:2305.03915.

Tharwat El-Sayed, Abdallah Mustafa, Ayman El-Sayed, and Mohamed Elrashidy. 2023. Hate speech detection by classic machine learning. In *2023 3rd International Conference on Electronic Engineering (ICEEM)*, pages 1–4.

Anuj Kumar. 2022. A study: Hate speech and offensive language detection in textual data by using rnn, cnn, lstm and bert model. In *2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 1–6.

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.

Hind Saleh, Areej Alhothali, and Kawthar Moria. 2021. Detection of hate speech using bert and hate speech word embedding with deep model. *Preprint*, arXiv:2111.01515.

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*.