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

Kogilavani Shanmugavadivel[1], **Malliga Subramanian**[2],
Naveenram CE[1], **Vishal RS**[1], **Srinesh S**[1]
[1]Department of AI, Kongu Engineering College, Perundurai, Erode.
[2]Department of CSE, Kongu Engineering College, Perundurai, Erode.
{kogilavani.sv, mallinishanth72}@gmail.com
{naveensrn1935,rsvishaltpr,srisiva262005}@gmail.com

Abstract

Hate speech detection in code-mixed Dravidian languages presents significant challenges due to the multilingual and unstructured nature of the data. In this work, we participated in the shared task to detect hate speech in Tamil, Malayalam, and Telugu using both text and audio data. We explored various machine learning models, including Logistic Regression, Ridge Classifier, RandomForest, and Convolutional Neural Networks (CNN). For Tamil text data, Logistic Regression achieved the highest macro-F1 score of 0.97, while Ridge Classifier performed best for audio with 0.75. In Malayalam, Random Forest excelled for text with 0.97, and CNN for audio with 0.69. For Telugu, Ridge Classifier achieved 0.89 for text and CNN 0.87 for audio.These results demonstrate the efficacy of our multimodal approach in addressing the complexity of hate speech detection across the Dravidian languages.Tamil:11th rank,Malayalam :6th rank,Telugu:8th rank among 145 teams

1 Introduction

Hate speech on social media is becoming more and more troublesome, particularly in multilingual situations where users mix imported terms with native scripts, such as Telugu, Tamil, and Malayalam. Because there are few annotated datasets and linguistic variation, it is difficult to detect such speech. To improve the accuracy of detection, this study suggests a multimodal strategy that combines text and audio features. Random Forest, Ridge Classifier, and Logistic Regression are examples of text-based models that are used to assess linguistic clues. Convolutional Neural Networks (CNNs) are used to process audio inputs in order to extract prosodic.

The shortcomings of conventional text-only approaches are addressed by combining text and audio predictions. The study shows that using both modalities greatly enhances detection performance by using YouTube data. The potential of multimodal systems is demonstrated by CNNs' efficacy in audio analysis and text machine learning models. This method provides a strong framework for spotting hate speech in intricate, multilingual internet settings.

2 Literature Survey

Barman and Das (2023) developed multimodal models for abusive language detection and sentiment analysis in Tamil and Malayalam. They used MFCC for audio, ViT for images, and mBERT for text, achieving a weighted F1 score of 0.5786 in abusive language detection and securing first place.

Bala and Krishnamurthy (2023) addressed sentiment analysis in Tamil and Malayalam videos and the detection of abusive language in Tamil multimodal videos. Their models used MViT for video, OpenL3 for audio, and BERT for text, demonstrating effective multimodal fusion.

Rahman et al. (2024) applied a multimodal strategy integrating text, audio, and video for Tamil abusive language detection. They used ConvLSTM, 3D-CNN, and a hybrid 3D-CNN+BiLSTM for video, and combined textual predictions from MNB, LR, and LSTM with audio features using a late fusion model. Their best model (ConvLSTM+BiLSTM+MNB) achieved a macro F1 score of 71.43, ranking first in the task.

Premjith et al. (2024) summarized the results of a shared task on multimodal sentiment analysis, abusive language detection, and hate speech detection in Tamil and Malayalam. Despite 39 teams participating, only two submitted results, which were evaluated using macro F1-score.

Anierudh et al. (2024) focused on three tasks: (1) sentiment classification in Tamil and Malayalam (highly positive, positive, neutral, negative, highly negative); (2) abusive language detection in Tamil; (3) hate speech detection in Tamil (Caste, Offen-

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 319–323
May 3, 2025 ©2025 Association for Computational Linguistics

sive, Racist). They used machine learning models and oversampling strategies to handle dataset biases.

Rajalakshmi et al. (2024) addressed hate speech detection in code-mixed languages using transliteration. They achieved F1 scores of 0.68 (Logistic Regression) and 0.70 (Bi-GRU), contributing to research in preventing hate speech in mixed-language content.

Sreelakshmi et al. (2024) explored hate speech and offensive language (HOS) detection in Tamil-English, Malayalam-English, and Kannada-English using multilingual transformer-based embeddings. MuRIL performed best across datasets, achieving 96 accuracy in Malayalam and 72 in Tamil (DravidianLangTech 2021), and 76 in Tamil and 68 in Malayalam (HASOC 2021). A new annotated Malayalam-English test set was also introduced.

Yasaswini et al. (2021) worked on offensive language detection in Malayalam, Tamil, and Kannada at EACL 2021. They categorized social media posts into six classes using transfer learning. Their source code was released publicly.

3 Task Description

This study focuses on multimodal hate speech detection in Tamil, Malayalam, and Telugu using a dataset sourced from YouTube videos.The dataset consists of audio and text samples that have been categorized as either non-hate or hate (subclasses: political, religious, gender, and personal defamation). Vectorizers such as Count Vectorizer, TF-IDF, and Word2Vec were used to handle text data, while pre-processing was done on audio data to extract prosodic characteristics. Text was subjected to machine learning models such Random Forest, Ridge Classifier, and Logistic Regression, while audio was subjected to CNN. The advantages of a multimodal strategy were demonstrated by evaluating these models' performance using the macro-F1 score.Lal G et al. (2025) Tamil secured the 11th rank, Malayalam secured the 6th rank, and Telugu secured the 8th rank among 145 teams.

4 Dataset Description

4.1 Text Data Description

The Text dataset consists of three languages: Malayalam, Tamil, and Telugu, with each record labeled as either Non-Hate or Hate. Content that does not include offensive language is categorized as Non-Hate (abbreviated "N"), whereas content that falls within the Gender (G), Political (P), Religious (R), and Personal Defamation (C) categories is grouped together into the Hate category. Smaller test sets are available, however the training dataset consists of 883 records in Malayalam, 1397 records in Tamil, and 1953 records in Telugu. The training data for each language is broken out in depth in Table 1 below, which displays the distribution of the Non-Hate and Hate categories. With an emphasis on hate speech detection across many languages, this dataset is intended to train algorithms that categorize material as either harmful or non-harmful.

Language	Non-Hate(N)	Hate(C,G,P,R)
Malayalam	406	477
Tamil	287	491
Telugu	198	175

Table 1: Dataset Description of Text-Train

4.2 Audio Data Desciption

The Audio dataset is organized similarly to the Text dataset, with entries classified as either Hate or Non-Hate. While hateful content falls under the categories of gender (G), politics (P), religion (R), and personal defamation (C), non-hateful content is audio that does not contain damaging speech. For training, the Audio-Train dataset consists of 883 Malayalam, 509 Tamil, and 551 Telugu recordings, along with smaller test sets. The training data for each language is broken out in depth in Table 2 below, which displays the distribution of the Non-Hate and Hate categories. With an emphasis on identifying hate speech in many languages, this dataset is used to train algorithms for identifying damaging speech in audio data.

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

Table 2: Dataset Description of Audio-Train

5 Methodology

5.1 System Architecture

There are two pipelines in the system: audio and text. TF-IDF and Count Vectorizer are used for preprocessing, tokenization, and vectorization in

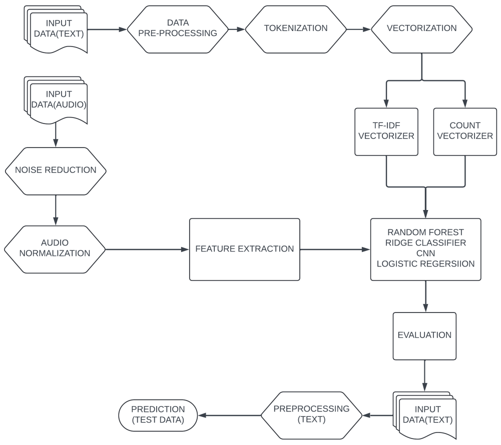

Figure 1: System Architecture

the text pipeline. Random Forest, Ridge Classifier, and Logistic Regression are then used for classification. CNN-based classification, noise reduction, and normalization are all part of the audio pipeline, and the results are combined to produce the final forecast..

5.2 Data Preprocessing

In order to create numerical representations suitable for machine learning models, the dataset underwent specific preprocessing steps tailored for each modality: for text data, we removed unnecessary punctuation, URLs, and symbols; for audio data, we segmented the audio into smaller chunks, normalized the amplitude, and reduced background noise; for audio data, we extracted prosodic features like pitch, energy, and spectral characteristics to capture tonal and temporal information relevant to hate speech detection; and for speech data, we tokenized the text into individual words, followed by stop word removal and stemming/lemmatization to reduce words to their base forms.

5.3 Model Development

We used a range of deep learning and machine learning algorithms to classify hate speech. We employed Ridge Classifier, Random Forest, and Logistic Regression for text data because of their efficacy and interpretability with high-dimensional text data. We applied Convolutional Neural Networks (CNN) to audio data in order to identify tone and temporal patterns. To take linguistic and cultural quirks into consideration, each model was taught independently for Telugu, Tamil, and Malayalam. Data imbalance was addressed by class balancing approaches, and model performance was optimized by hyperparameter tweaking.

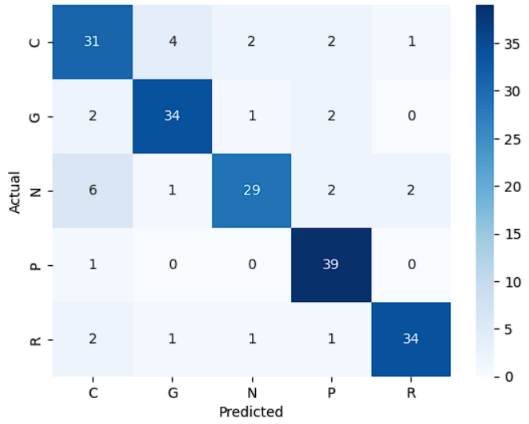

Figure 2: Confusion Matrix of Tamil-Text

6 Performance Evaluation

The performance of the models was evaluated using the Macro-F1 score. Table 3 summarizes the results for text and audio modalities across Tamil, Malayalam, and Telugu. GitHub Repository: Dravidan-LangTech

Language	Modality	Macro-F1 Score
Tamil	Text	0.97
Tamil	Audio	0.75
Malayalam	Text	0.97
Malayalam	Audio	0.69
Telugu	Text	0.89
Telugu	Audio	0.87

Table 3: Performance Metrics for Text and Audio Modalities

6.1 Tamil

Logistic Regression proved highly effective in categorizing hate speech in Tamil text, achieving a Macro-F1 score of 0.97. Using Count Vectorizer, TF-IDF, and Word2Vec techniques, the model accurately distinguished between Hate and Non-Hate categories, effectively identifying themes like gender, politics, religion, and personal defamation based on language patterns.

For audio data, the Ridge Classifier outperformed other models with a Macro-F1 score of 0.75, highlighting the importance of speech spectral features in detecting hate speech. While CNN was successful in capturing speech's temporal aspects, it performed less effectively compared to other classifiers for audio-based detection.

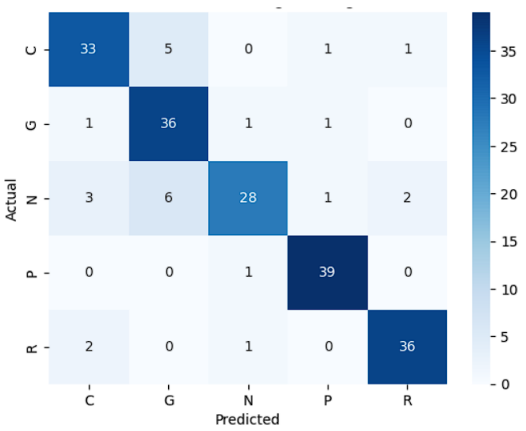

Figure 3: Confusion Matrix of Malayalam-Text

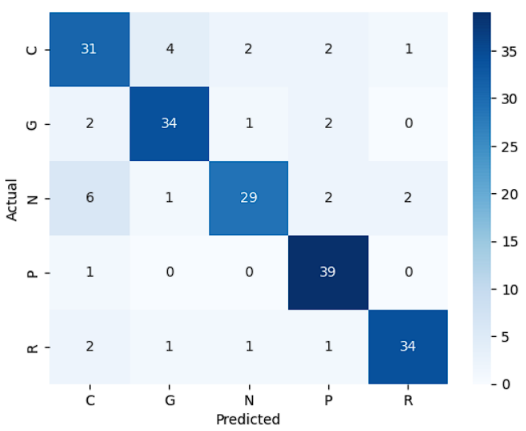

Figure 4: Confusion Matrix of Telugu-Text

6.2 Malayalam

The Random Forest Classifier had the greatest Macro-F1 score of 0.97 for the text modality in Malayalam. This outcome shows how well the model can categorize hate speech from a variety of subclasses, including gender, political, religious, and personal defamation. By utilizing the various variables that were recovered using vectorization approaches, the Random Forest model—an ensemble approach—provided excellent prediction performance.For Malayalam data, the confusion matrix for the top-performing model is shown in Figure 3.

CNN's Macro-F1 score for the audio modality was 0.69, which is less than the text performance but still shows a respectable level of success in identifying tonal characteristics linked to hate speech in Malayalam. The difficulties in utilizing CNN to analyze the rich prosodic elements of Malayalam speech may be the cause of the worse results.

6.3 Telugu

For Telugu, the text modality, the Ridge Classifier received a Macro-F1 score of 0.89. This shows that the model successfully distinguished between the Non-Hate and Hate categories, including their several subclasses, and was quite successful in detecting hate speech in Telugu. For the model to function well, the vectorized features from Word2Vec and TF-IDF were essential.The confusion matrix for the top-performing model using Telugu data is shown in Figure 4.

For the audio modality, With a Macro-F1 score of 0.87, CNN fared better than other models, demonstrating the model's capacity to accurately represent Telugu speech dynamics. CNN's excellent performance in audio categorization demonstrates its capacity to examine speech patterns and detect hostile or aggressive behavior.

7 Limitations

Although our multimodal method for identifying hate speech in Dravidian languages yields encouraging findings, there are a number of drawbacks to take into account. First, the dataset size is minimal, which could restrict how broadly the models can be applied. Second, prosodic features could be more effectively captured by further optimizing the audio data preparation pipeline. The models can also have trouble handling code-mixed content, which is prevalent on social media. Enhancing the integration of text and audio modalities and growing the dataset should be the main goals of future research.

8 Conclusion

This study successfully combined text and audio data using multimodal approaches to identify hate speech in Telugu, Tamil, and Malayalam. Text-based classification yielded strong Macro-F1 scores for machine learning models such as Random Forest and Logistic Regression. With a score of 0.87 in Telugu, CNN performed exceptionally well in audio, while Malayalam fared marginally worse. The findings emphasize how crucial prosodic characteristics are for identifying hate speech. Combining deep learning with more conventional machine learning techniques showed promise. To increase the accuracy of multilingual hate speech detection, future developments might concentrate on feature extraction optimization and model calibration.

References

S Anierudh, Abhishek R, Ashwin Sundar, Amrit Krishnan, and Bharathi B. 2024. Wit hub@DravidianLangTech-2024:multimodal social media data analysis in Dravidian languages using machine learning models. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 229–233, St. Julian's, Malta. Association for Computational Linguistics.

Abhinaba Bala and Parameswari Krishnamurthy. 2023. AbhiPaw@DravidianLangTech: Multimodal abusive language detection and sentiment analysis. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 140–146, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

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.

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, G Jyothish, V Sowmya, Bharathi Raja Chakravarthi, K Nandhini, Rajeswari Natarajan, Abirami Murugappan, Bharathi B, Saranya Rajiakodi, Rahul Ponnusamy, Jayanth Mohan, and Mekapati Reddy. 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, St. Julian's, Malta. Association for Computational Linguistics.

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.

Ratnavel Rajalakshmi, Saptharishee M, Hareesh S, Gabriel R, and Varsini Sr. 2024. DLRG-DravidianLangTech@EACL2024 : Combating hate speech in Telugu code-mixed text on social media. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 140–145, St. Julian's, Malta. Association for Computational Linguistics.

K. Sreelakshmi, B. Premjith, Bharathi Raja Chakravarthi, and K. P. Soman. 2024. Detection of hate speech and offensive language codemix text in dravidian languages using cost-sensitive learning approach. *IEEE Access*, 12:20064–20090.

Konthala Yasaswini, Karthik Puranik, Adeep Hande, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. IIITT@DravidianLangTech-EACL2021: Transfer learning for offensive language detection in Dravidian languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 187–194, Kyiv. Association for Computational Linguistics.