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

Kogilavani Shanmugavadivel[1],**Malliga Subramanian**[2],
ShahidKhan S[1], **Shri Sashmitha S**[1], **Yashica 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
{shahidkhans.22,shrisashmithas.22,yashicas.22aid}@kongu.edu

Abstract

It is difficult to detect hate speech in code-mixed Dravidian languages because the data is multilingual and unstructured. We took part in the shared task to detect hate speech in text and audio data for Tamil, Malayalam, and Telugu in this research. We tested different machine learning and deep learning models such as Logistic Regression, Ridge Classifier, Random Forest, and CNN. For Tamil, Logistic Regression gave the best macro-F1 score of 0.97 for text, whereas Ridge Classifier was the best for audio with a score of 0.75. For Malayalam, Random Forest gave the best F1-score of 0.97 for text, and CNN was the best for audio (F1-score: 0.69). For Telugu, Ridge Classifier gave the best F1-score of 0.89 for text, whereas CNN was the best for audio (F1-score: 0.87).Our findings prove that a multimodal solution efficiently tackles the intricacy of hate speech detection in Dravidian languages. In this shared task,out of 145 teams we attained the 12th rank for Tamil and 7th rank for Malayalam and Telugu.

1 Introduction

The rise of social media has facilitated global communication but also led to the spread of hate speech. Detecting and preventing hate speech is crucial for fostering a safe and inclusive online space. This challenge intensifies in multilingual and code-mixed environments, such as Tamil, Malayalam, and Telugu, where users blend local scripts with borrowed words. The complexity of these languages, along with limited annotated datasets, makes hate speech detection a vital yet challenging research area.

Multimodal approaches combining text and audio offer deeper context for understanding online speech. While text-based models analyze linguistic cues, audio models capture tonal and prosodic features to detect aggression or hostility. This study employs machine learning and deep learning techniques, including Logistic Regression, Ridge Classifier, Random Forest, and CNN, to classify hate speech data from YouTube. By integrating both modalities, the methodology addresses limitations of conventional approaches.

Findings indicate that multilingual models can accurately detect hate speech across languages. Logistic Regression and Random Forest performed well in text classification, while CNNs effectively processed audio data. The results underscore the importance of combining linguistic and acoustic features to enhance detection accuracy. By expanding the multimodal dataset for Dravidian languages, this study contributes to building robust frameworks for combating hate speech in multilingual social media.

2 Literature Survey

Rawat et al. (2024) proposed a deep NLP model combining convolutional and recurrent layers for hate speech detection on social media, achieving a macro F1 score of 0.63 on the HASOC2019 dataset. The study also explored using unlabeled data and similar corpora to improve performance and reduce overfitting. Anbukkarasi and Varadhaganapathy (2023) introduced a synonym-based Bi-LSTM model to classify hate and non-hate texts in Tamil-English code-mixed tweets using a newly designed dataset of 10,000 annotated texts, addressing challenges of limited data and code-mixed language patterns.

As part of a collaborative effort, Tash et al. (2024) investigated Tamil hate speech detection related to migration and shelter, achieving an F1 score of 0.76 with a CNN model. Premjith et al. (2023) summarized a multimodal abusive language detection and sentiment evaluation effort in Tamil and Malayalam using video, audio, and text. The findings highlighted the challenges in creating ef-

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 269–273

fective models, with results based on the macro F1-score. Poornachandran et al. (2022) emphasized evaluating regional languages like Malayalam for hate speech detection, achieving an F1 score of 0.85 with deep learning techniques on a natural Malayalam dataset.

Priyadharshini et al. (2023) presented findings on abusive remark detection in Tamil and Telugu code-mixed social media text at RANLP 2023. The project developed models evaluated using the macro F1-score. Sai et al. (2024) explored hate speech detection in Telugu-English code-mixed text for DravidianLangTech@EACL-2024, achieving a macro F1 score of 0.65, ranking 14th in the competition.

Premjith et al. (2024a) analyzed submissions for Hate and Offensive Language Detection in Telugu Codemixed Text (HOLD-Telugu) at DravidianLangTech 2024, evaluating models using the macro F1-score. Another shared project led by Premjith et al. (2024b) focused on sentiment analysis, abusive language detection, and hate speech detection in Tamil and Malayalam using multimodal data. Despite 39 participants, only two submitted results, evaluated by the macro F1-score. Sreelakshmi et al. (2024) explored multilingual transformer-based embeddings for detecting hate speech in CodeMix Dravidian languages. Their study on Kannada-English, Malayalam-English, and Tamil-English datasets found MuRIL embeddings with an SVM classifier performed best. The research also addressed class imbalance with a cost-sensitive approach and introduced a new annotated Malayalam-English CodeMix dataset extending HASOC 2021.

3 Task Description

This study investigates multimodal hate speech detection in Tamil, Malayalam, and Telugu using YouTube-sourced text and audio data. Hate speech is categorized into Gender, Political, Religious, and Personal Defamation subclasses. Text preprocessing involved Count Vectorizer and TF-IDF, while audio preprocessing extracted prosodic features. Logistic Regression, Ridge Classifier, Random Forest, and CNN were applied, with performance evaluated using the macro-F1 score Lal G et al. (2025). Among 145 teams, our system ranked 12th for Tamil and 7th for Malayalam and Telugu, demonstrating the effectiveness of integrating text and audio models for detecting hate speech in Dravidian languages.

4 Dataset Description

4.1 Text Data Description

The text dataset for Malayalam, Tamil, and Telugu categorizes records as Hate or Non-Hate. Hate includes content labeled under Gender (G), Political (P), Religious (R), and Personal Defamation (C), while Non-Hate ('N') contains content without harmful language. The training set includes 883 Malayalam, 1,397 Tamil, and 1,953 Telugu records, with smaller test sets. Table 1 details the distribution of Hate and Non-Hate classes across languages, designed for training models in hate speech detection across multilingual contexts.

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 structured similarly to the text dataset, with recordings labeled as Non-Hate or Hate. Hate includes content categorized under Gender (G), Political (P), Religious (R), and Personal Defamation (C). The training set has 883 Malayalam, 509 Tamil, and 551 Telugu recordings, with smaller test sets. Table 2 shows the distribution of Hate and Non-Hate categories across languages. This dataset helps train models for multilingual hate speech detection.

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 Data Preprocessing

Text and audio data underwent modality-specific preprocessing. Text processing included removing images, URLs, punctuation, tokenization, stopword removal, and stemming or lemmatization, followed by vectorization using Count Vectorizer and TF-IDF. Audio preprocessing involved noise reduction, normalization, and segmentation, with prosodic features like pitch and energy extracted to capture

speech tone. This approach ensured high-quality inputs for modeling.

5.2 Model Development

Logistic Regression, Ridge Classifier, Random Forest, and CNN were used for text and audio classification due to their effectiveness with high-dimensional data. These models captured linguistic and tonal features and were trained independently for Tamil, Malayalam, and Telugu to handle language-specific nuances. Class balancing, hyperparameter tuning, and cross-validation ensured robust performance.

5.3 Workflow Integration

The workflow integrates text and audio to enhance hate speech detection accuracy. Both modalities were processed separately and fed into their respective models. The outputs were analyzed to classify hate speech into categories like Gender, Political, Religious, and Personal Defamation. Figure 1 illustrates the workflow, covering preprocessing to classification. This modular design allows future experimentation with additional features or models, ensuring a comprehensive approach to multimodal hate speech detection in Dravidian languages.

6 Performance Evaluation

The performance of the fashions was evaluated based totally at the Macro-F1 score, that is a broadly used metric for category responsibilities, especially in imbalanced datasets. The fashions were educated on each text and audio records for the Tamil, Malayalam, and Telugu languages, and the respective performances are mentioned underneath.

6.1 Tamil

For text classification, Logistic Regression achieved the highest Macro-F1 score of 0.97, demonstrating strong accuracy in identifying hate speech in Tamil. Count Vectorizer and TF-IDF effectively transformed text into numerical representations, enabling the model to distinguish between Hate and Non-Hate categories. This high score indicates the model's ability to capture linguistic patterns, particularly in Gender (G), Political (P), Religious (R), and Personal Defamation (C) hate speech. Figure 2 presents the confusion matrix for the best-performing model.

For audio, Ridge Classifier performed best with a Macro-F1 score of 0.75. While CNN captured temporal speech features well, its overall performance was lower than other classifiers. Ridge Classifier's success suggests that spectral features significantly enhance hate speech detection in Tamil speech.

6.2 Malayalam

For the text modality, Random Forest achieved the highest Macro-F1 score of 0.97, demonstrating excellent performance in classifying hate speech across subclasses like Gender, Political, Religious, and Personal Defamation. As an ensemble method, Random Forest effectively leveraged features extracted through various vectorization techniques, ensuring strong predictions. Figure 3 presents the confusion matrix for the best-performing Malayalam text model.

For audio, CNN attained a Macro-F1 score of 0.69. While lower than the text score, it still showed reasonable success in detecting tonal patterns in Malayalam hate speech. The model's limitations may stem from the complexity of processing prosodic features.

6.3 Telugu

For the text modality, Ridge Classifier achieved a Macro-F1 score of 0.89, demonstrating strong performance in detecting hate speech and distinguishing between Non-Hate and Hate subclasses. The effectiveness of TF-IDF and Count-based vectorized features contributed significantly to the model's success. Figure 4 presents the confusion matrix for the best-performing Telugu text model.

For audio, CNN attained a Macro-F1 score of 0.87, effectively capturing speech dynamics in Telugu. Its strong performance highlights CNN's ability to analyze speech patterns and detect aggression or hostility.

7 Limitations

Our approach relies heavily on labeled datasets, which are limited for Dravidian languages. The complexity of prosodic features and insufficient audio samples affected audio model performance. Class imbalance in the dataset may have impacted the model's ability to generalize effectively. Expanding datasets and refining models are essential for addressing these limitations.

8 Conclusion

We applied multimodal approaches to classify hate speech in Tamil, Malayalam, and Telugu using text

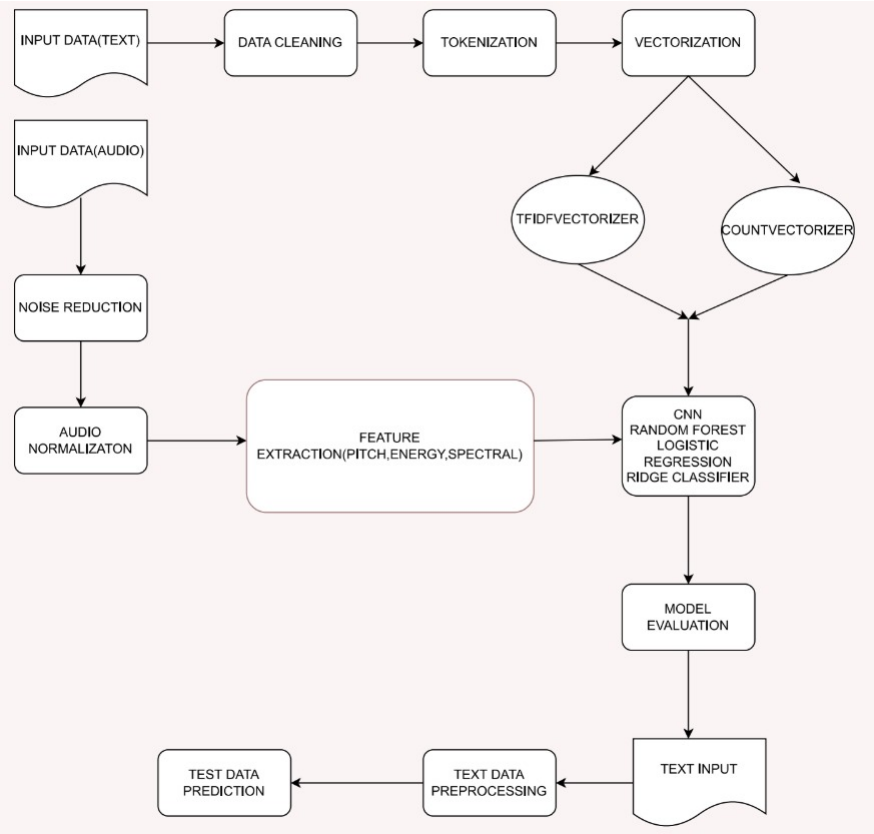

Figure 1: Proposed System Workflow

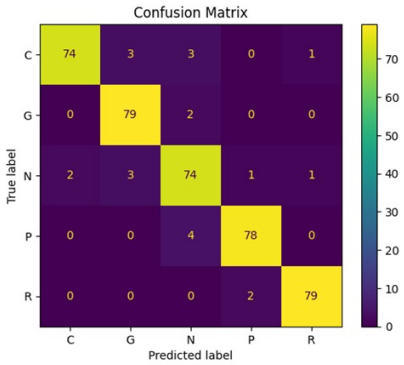

Figure 2: Confusion Matrix of Tamil-Text

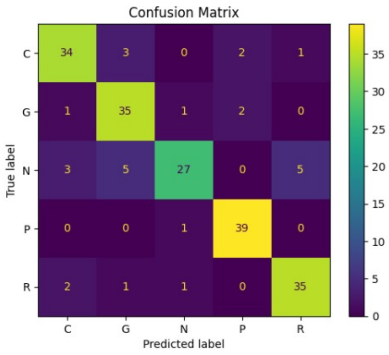

Figure 4: Confusion Matrix of Telugu-Text

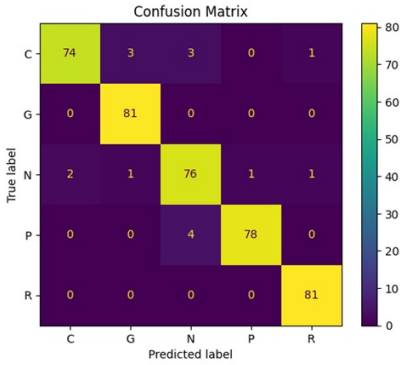

Figure 3: Confusion Matrix of Malayalam-Text

and audio data. Logistic Regression and Random Forest performed well for text, while CNN was most effective for audio, especially in Telugu. The results highlight the importance of combining linguistic and prosodic features for accurate detection. Overall, our approach shows promising results across languages. Further improvements in feature extraction and model optimization could enhance performance. The code for this shared task can be accessed at Github

References

S Anbukkarasi and S Varadhaganapathy. 2023. Deep learning-based hate speech detection in code-mixed tamil text. *IETE Journal of Research*, 69(11):7893–7898.

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.

Prabaharan Poornachandran, VG Sujadevi, Gayathri Rajendran, Vinayak Ks, Vishnu Vijayan, Arjun Ram, et al. 2022. Malhate: Hate speech detection in malayalam regional language. In *2022 IEEE 7th International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*, volume 7, pages 110–115. IEEE.

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.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Malliga S, Subalalitha Cn, Kogilavani S V, Premjith B, Abirami Murugappan, and Prasanna Kumar Kumaresan. 2023. Overview of shared-task on abusive comment detection in Tamil and Telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 80–87, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Anchal Rawat, Santosh Kumar, and Surender Singh Samant. 2024. Hate speech detection in social media: Techniques, recent trends, and future challenges. *Wiley Interdisciplinary Reviews: Computational Statistics*, 16(2):e1648.

Chava Sai, Rangoori Kumar, Sunil Saumya, and Shankar Biradar. 2024. Iiitdwd_svc@ dravidianlangtech-2024: Breaking language barriers; hate speech detection in telugu-english code-mixed text. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 119–123.

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*.

M Tash, Z Ahani, M Zamir, O Kolesnikova, and G Sidorov. 2024. Lidoma@ lt-edi 2024: Tamil hate speech detection in migration discourse. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 184–189.