Team ML_Forge@DravidianLangTech 2025: Multimodal Hate Speech Detection in Dravidian Languages

Adnan Faisal, Shiti Chowdhury, Sajib Bhattacharjee,
Udoy Das[†], Samia Rahman, Momtazul Arefin Labib, Hasan Murad
Department of Computer Science and Engineering,
Chittagong University of Engineering and Technology, Bangladesh
[†]East Delta University, Bangladesh
{u2004002, u2004027, u2004003}@student.cuet.ac.bd, udoy.d@eastdelta.edu.bd,
{u1904022, u1904111}@student.cuet.ac.bd, hasanmurad@cuet.ac.bd

Abstract

Ensuring a safe and inclusive online environment requires effective hate speech detection on social media. While detection systems have significantly advanced for English, many regional languages, including Malayalam, Tamil and Telugu, remain underrepresented, creating challenges in identifying harmful content accurately. These languages present unique challenges due to their complex grammar, diverse dialects and frequent code-mixing with English. The rise of multimodal content, including text and audio, adds further complexity to detection tasks. The shared task “Multimodal Hate Speech Detection in Dravidian Languages: DravidianLangTech@NAACL 2025” has aimed to address these challenges. A Youtube-sourced dataset has been provided, labeled into five categories: Gender (G), Political (P), Religious (R), Personal Defamation (C) and Non-Hate (NH). In our approach, we have used mBERT, T5 for text and Wav2Vec2 and Whisper for audio. T5 has performed poorly compared to mBERT, which has achieved the highest F1 scores on the test dataset. For audio, Wav2Vec2 has been chosen over Whisper because it processes raw audio effectively using self-supervised learning. In the hate speech detection task, we have achieved a macro F1 score of 0.2005 for Malayalam, ranking 15th in this task, 0.1356 for Tamil and 0.1465 for Telugu, with both ranking 16th in this task.

1 Introduction

With the increasing spread of harmful content online, hate speech detection on social media has become a crucial area of research. While platforms empower users to express views, they are often exploited to propagate hate and abuse. Despite progress in English, regional languages like Malayalam, Tamil and Telugu remain under-researched, highlighting the need for more inclusive detection frameworks. These Dravidian languages, spoken in southern India, present challenges such as complex grammar, dialect variations and code-mixing (Chakravarthi et al., 2022). Existing research has faced gaps due to the lack of large, well-balanced datasets, limiting robust machine learning models (Premjith et al., 2024a). Most studies have focused on text analysis, ignoring the multimodal nature of social media content, which includes both text and audio (Chakravarthi et al., 2021).

The shared task "Multimodal Hate Speech Detection in Dravidian Languages: DravidianLangTech@NAACL 2025" has addressed key challenges in this area (Lal G et al., 2025). The dataset, sourced from YouTube, requires models to analyze both text and audio components for detecting hate speech in Malayalam, Tamil and Telugu.

In this study, we have proposed a multimodal approach using mBERT for text and Wav2Vec2 for audio. The hybrid mBERT + Wav2Vec2 model has shown improved performance, achieving F1 scores of 0.3013 for Malayalam, 0.2853 for Tamil and 0.2511 for Telugu, demonstrating the effectiveness of combining textual and acoustic features (Premjith et al., 2024b). This approach has emphasized the benefits of multimodal fusion in overcoming the limitations of single-modality models and advancing hate speech detection in underrepresented languages (B et al., 2024). The core contributions of our research work are as follows -

- We have used augmentation Technique to balance dataset for Malayalam, Tamil and Telugu languages.

- We have utilized an efficient fusion technique to improve classification and enhance model performance.

Detailed implementation information is available in the GitHub repository - https://github.com/Sojib001/MHDS

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 459–464
May 3, 2025 ©2025 Association for Computational Linguistics

2 Related Work

The rapid growth of social media has led to increased hate speech and abusive content, raising concerns about platform safety. While progress has been made in hate speech detection for high-resource languages like English, under-resourced languages such as Malayalam, Tamil and Telugu face challenges due to complex grammar, dialects, frequent English mixing and limited labeled data (Chakravarthi et al., 2021).

Unimodal approaches, relying on text-based models like BERT (Devlin et al., 2019; Liu et al., 2019) and mBERT, have analyzed linguistic features. Sreelakshmi et al. (2024) has addressed dataset imbalances in Dravidian languages and Chakravarthi et al. (2021) has demonstrated the effectiveness of transformer models for Tamil and Malayalam. However, unimodal methods have lacked the ability to incorporate signals from other modalities.

Multimodal Approaches: Combining text and audio has shown promise in hate speech detection by capturing linguistic and acoustic cues. B et al. (2024), Kiela et al. (2020) and (Anilkumar et al., 2024) have demonstrated improved detection accuracy using multimodal approaches. Chakravarthi et al. (2021) has highlighted the potential of YouTube-sourced multimodal datasets. Despite this progress, research on multimodal hate speech detection for Dravidian languages remains limited, requiring further exploration.

3 Data Description

The dataset for Multimodal Hate Speech Detection in Malayalam, Tamil and Telugu is sourced mainly from YouTube. It includes text from captions and audio from videos, covering both speech and background noise. Hate speech is categorized into Gender (G), Political (P), Religious (R), Personal Defamation (C) and Non-Hate (NH). Table 1 presents the dataset distribution across training and test sets.

Language	Training Dataset	Test Dataset
Malayalam	883	50
Tamil	514	50
Telugu	556	50
Total	**1953**	**150**

Table 1: Language-wise Distribution of Training and Test Data

Table 2 shows the distribution of five class labels—Gender (G), Political (P), Religious (R), Personal Defamation (C) and Non-Hate (NH)—across Malayalam, Tamil and Telugu, with total counts at the bottom.

Class Label	Malayalam	Tamil	Telugu
Gender (G)	82	101	63
Political (P)	118	58	33
Religious (R)	91	72	61
Personal Defamation (C)	186	122	65
Non-Hate (NH)	406	198	287
Total	**883**	**514**	**556**

Table 2: Statistical Breakdown of Class Labels Across Malayalam, Tamil and Telugu

4 Methodology

4.1 Problem Formulation

The task has been to detect hate speech in Malayalam, Tamil and Telugu across five categories: Gender, Political, Religious, Personal Defamation and Non-Hate. We have used late fusion to combine text features from mBERT and audio features from Wav2Vec2, merging them in a classification layer to improve predictions, despite challenges like language differences and code-mixing.

4.2 Data Augmentation and Preprocessing

To address class imbalance, data augmentation was applied to balance the dataset. Back translation expanded the training data by translating text to another language and back, creating variations. For audio, irrelevant features like MFCC were removed to clean the data and focus on useful information.

4.3 Uni-modal Models

4.3.1 Text-based Model

We used mBERT (multilingual BERT) and T5 for text classification due to their ability to capture contextual meanings across languages. mBERT, fine-tuned on Malayalam, Tamil and Telugu datasets, outperformed T5 in effectively classifying hate speech.

4.3.2 Audio-based Model

We have used Wav2Vec2 and Whisper for audio feature extraction, with Wav2Vec2 performing better in capturing tone, pitch and context for hate speech detection. Initially, MFCC features were included but have been removed after receiving the gold test data, improving performance with only Wav2Vec2 features.

4.4 Fusion Model

To enhance classification accuracy, we have adopted a multimodal fusion approach that integrates textual and audio features through late fusion, where mBERT (text) and Wav2Vec2 (audio) representations are combined for improved hate speech detection. mBERT has consistently outperformed T5 in text processing, while Wav2Vec2 has excelled over Whisper in capturing audio features. By merging mBERT's text embeddings and Wav2Vec2's audio features, our model effectively captures both linguistic and acoustic nuances, leading to a more robust and accurate detection system. Figure 1 illustrates the overall modeling pipeline, showcasing the integration of text and audio features through the late fusion mechanism.

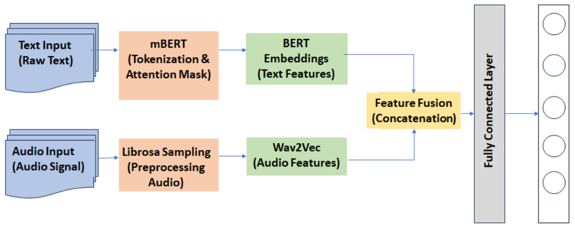


Figure 1: Abstract process of violence text detection

4.5 Evaluation Metrics

The models were evaluated using macro-F1 score, precision and recall to ensure balanced performance and accurate identification of hate speech.

5 Results and Analysis

This task has evaluated models for detecting hate speech in Malayalam, Tamil and Telugu using both text and audio data. The results have shown good performance on training data but struggles with test data, highlighting issues like overfitting, class imbalance and challenges in combining text and audio data.

5.1 Task 1: Malayalam Multimodal Hate Speech Detection

Table 3 has shown the performance of different classifiers for Malayalam. Among text-based models, mBERT has achieved the highest F1 score (0.5796), outperforming T5 (0.45). For audio models, Wav2Vec2 has performed best (F1: 0.3399), surpassing Whisper (0.30). In multimodal setups, mBERT + Wav2Vec2 have achieved the highest F1 score (0.3013), demonstrating the effectiveness of combining text and audio features. Figure 2a has represented the confusion matrix of our best-performing model.

Malayalam	**Classifier**	**P**	**R**	**F1**
Unimodal (Text)	mBERT	0.64	0.61	0.57
	T5	0.42	0.42	0.45
Unimodal (Audio)	Wav2Vec2	0.31	0.34	0.33
	Whisper	0.27	0.34	0.30
Multi-modal	**(mBERT + Wav2Vec2)**	**0.31**	**0.30**	**0.30**
	(T5 + Wav2Vec2)	0.21	0.28	0.24
	(mBERT + Whisper)	0.30	0.29	0.26

Table 3: Performance of Malayalam Classifiers (Macro Average)

5.2 Task 2: Tamil Multimodal Hate Speech Detection

Table 4 shows the classification performance for Tamil, where mBERT has achieved the highest F1 score (0.5561) for text. In the audio category, Whisper has reached an F1 score of 0.1494. The multimodal setup, combining mBERT with Wav2Vec2, has achieved an F1 score of 0.2853, demonstrating the benefits of integrating text and audio features. This fusion model has effectively combined mBERT's text embeddings with Wav2Vec2's speech representations. Figure 2b shows the confusion matrix of the best-performing model.

Tamil	**Classifier**	**P**	**R**	**F1**
Unimodal (Text)	mBERT	0.62	0.56	0.55
	T5	0.48	0.52	0.42
Unimodal (Audio)	Wav2Vec2	0.13	0.16	0.14
	Whisper	0.13	0.17	0.14
Multi-modal	**(mBERT + Wav2Vec2)**	**0.34**	**0.30**	**0.29**
	(T5 + Wav2Vec2)	0.32	0.28	0.27
	(mBERT + Whisper)	0.32	0.29	0.28

Table 4: Performance of Tamil Classifiers (Macro Average)

5.3 Task 3: Telugu Multimodal Hate Speech Detection

As shown in Table 5, different models perform differently for Telugu. mBERT achieves the best F1 score among text models at 0.3176. For audio models, Wav2Vec2 and Whisper score 0.1790 and 0.1894, respectively. Among combined models, (mBERT + Wav2Vec2) performs the best with an F1 score of 0.2511. Figure 2c represents the confusion matrix of our best performing model that combines mBERT with Wav2Vec2.

Telugu	Classifier	P	R	F1
Unimodal (Text)	mBERT	0.32	0.34	0.31
	T5	0.31	0.33	0.29
Unimodal (Audio)	Wav2Vec2	0.16	0.20	0.17
	Whisper	0.17	0.14	0.18
Multi-modal	**(mBERT + Wav2Vec2)**	**0.25**	**0.26**	**0.25**
	(T5 + Wav2Vec2)	0.21	0.25	0.19
	(mBERT + Whisper)	0.21	0.24	0.23

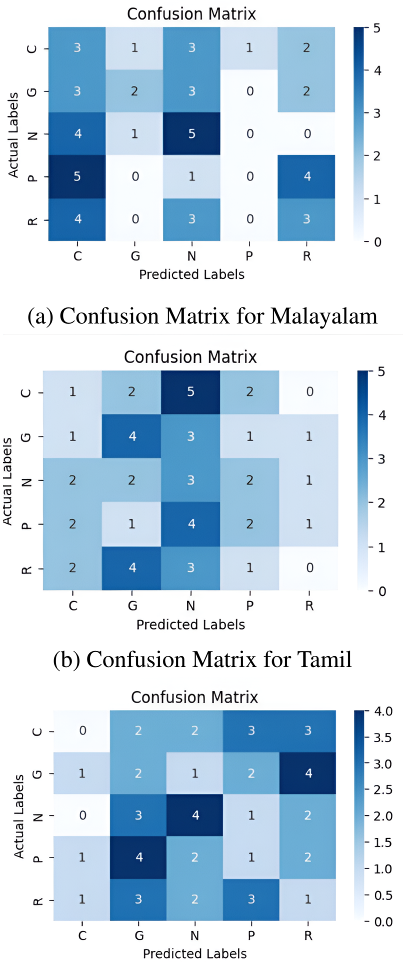
Table 5: Performance of Telugu Classifiers (Macro Average)

We have selected mBERT and Wav2Vec2 for their strong text and audio capabilities. mBERT, pretrained on Malayalam, Tamil and Telugu, has outperformed T5, while Wav2Vec2 has excelled over Whisper. Their integration has improved classification accuracy. To address overfitting, we have applied early stopping (patience = 5), L1 regularization and hyperparameter tuning. However, back translation has degraded performance.

The confusion matrices have shown classification trends, with off-diagonal elements revealing misclassifications. Figure 2 presents error patterns like C ↔ N, P ↔ R and G ↔ N. Malayalam has confused P and R, Tamil C and N and Telugu G and N, highlighting challenges in distinguishing sentiment and contextual labels.

5.4 Parameter Setting

Our best-performing model, mBERT, has used a learning rate of 0.00002, batch size 8 and the Adam optimizer, with a text input size of 512 and early stopping (patience = 5) to prevent overfitting. Wav2Vec2 has applied the same learning rate and batch size, while Whisper used a learning rate of 0.00001. T5 has followed mBERT's settings but performed worse. As shown in Table 6, the Fusion Model has combined mBERT's text features with Wav2Vec2's audio, improving multimodal classification in Malayalam, Tamil and Telugu. .



(a) Confusion Matrix for Malayalam

(b) Confusion Matrix for Tamil

(c) Confusion Matrix for Telugu

Figure 2: Confusion Matrices for Malayalam, Tamil and Telugu

Model	Learning Rate	Optimizer	Batch Size
mBERT	2e-5	AdamW	8
Wav2Vec2	2e-5	AdamW	8
Whisper	1e-5	AdamW	8
T5	2e-5	AdamW	8
Fusion Model	2e-5	Adam	8

Table 6: Key Hyperparameters for Model Training

6 Conclusion

The Shared Task on Multimodal Hate Speech Detection in Dravidian Languages: DravidianLangTech@NAACL 2025 has identified key challenges in detecting hate speech in Malayalam, Tamil and Telugu using text and audio. While transformer models have been effective, they have faced overfitting, data imbalance and multimodal integration issues. mBERT and Wav2Vec2 have shown overfitting, excelling in training but underperforming in testing. Multimodal fusion has shown promise but has struggled with noisy audio, alignment issues and class imbalance. Despite these challenges, the fusion model has been effective, leveraging mBERT's text embeddings and Wav2Vec2's speech representations to enhance classification. However, synchronization issues and noisy audio have impacted performance. Regarding back translation, it has not affected code-mixing patterns but has occasionally produced unnatural sentences, requiring manual validation. Since code-mixed texts have remained intact, their linguistic integrity has been preserved.

Limitations

The main limitation of our model is it has overfitted. It has learned noise and specific patterns in the training set that don't generalize. Not having enough training data also led our model to poor generalization. It performs worse than the unimodals in this task.

Ethical Statement

All data processing and modeling followed ethical rules for dealing with sensitive information, such as hate speech. The study seeks to improve hate speech detection while protecting rights and privacy of the people. The goal of the results is to improve moderation on online platforms and create safer spaces for users. We have recognized and handled any biases or limitations in the dataset as much as we could.

Acknowledgement

This document has been adapted from prior ACL and NAACL proceedings, including NAACL 2025 guidelines by Lal G. Jyothish, Premjith B. and Bharathi Raja Chakravarthi. We appreciate contributions from NAACL 2024 and ACL 2023 by Premjith B., Chakravarthi Bharathi Raja, Saranya Rajiakodi, Shubhankar Barman and Mithun Das. We express gratitude to the DravidianLangTech workshop organizers for promoting research in Dravidian languages. We also value insights from IEEE Access 2024 by Sreelakshmi K., Premjith B. and Bharathi Raja Chakravarthi, along with studies by Anilkumar et al. (2024), Mithun Das and Kiela et al. (2020) on multimodal abusive language detection. We acknowledge contributions from Chakravarthi et al. (2021, 2022) and Thapa et al. (2022) in multimodal sentiment analysis. Special appreciation to ACL and the NLP community for their dedication to dataset development and research on under-resourced languages.

References

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.

Premjith B, Jyothish G, Sowmya V, 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.

Bharathi Raja Chakravarthi, Jishnu Parameswaran P. K, Premjith B, K. P Soman, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Kingston Pal Thamburaj, and John P. McCrae. 2021. Dravidianmultimodality: A dataset for multi-modal sentiment analysis in tamil and malayalam. *Preprint*, arXiv:2106.04853.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John P. McCrae. 2022. Dravidiancodemix: sentiment analysis and offensive language identification dataset for dravidian languages in code-mixed text. *Language Resources and Evaluation*, 56(3):765–806.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *arXiv preprint arXiv:2005.04790*.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. In *arXiv preprint arXiv:1907.11692*.

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.

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.