Trio Innovators @ DravidianLangTech 2025: Multimodal Hate Speech Detection in Dravidian Languages

Radha N, Swathika R, Farha Afreen I, Annu G, Apoorva A
Sri Sivasubramaniya Nadar College of Engineering, Kalavakkam, India
radhan@ssn.edu.in
swathikar@ssn.edu.in
farha2110729@ssn.edu.in
annu2110538@ssn.edu.in,apoorva2110445@ssn.edu.in

Abstract

This paper presents an in-depth study on multimodal hate speech detection in Dravidian languages—Tamil, Telugu, and Malayalam—by leveraging both audio and text modalities. Detecting hate speech in these languages is particularly challenging due to factors such as code-mixing, limited linguistic resources, and diverse cultural contexts. Our approach integrates advanced techniques for audio feature extraction and XLM-Roberta for text representation, with feature alignment and fusion to develop a robust multimodal framework. The dataset is carefully categorized into labeled classes: gender-based, political, religious, and personal defamation hate speech, along with a non-hate category. Experimental results indicate that our model achieves a macro F1-score of 0.76 and an accuracy of approximately 85

1 Introduction

Hate speech on social media is a significant issue, particularly in underrepresented Dravidian languages like Tamil, Telugu, and Malayalam, where linguistic diversity and limited resources pose challenges to detection. Traditional text-based models often fail to capture crucial audio cues like tone and emotion, limiting their effectiveness.

This study introduces a multimodal approach that integrates both text and audio to enhance detection accuracy. A real-world dataset is used, incorporating labeled text and audio across multiple hate speech categories, including gender, politics, religion, and personal defamation, as well as non-hate speech. Audio data helps identify subtle cues like sarcasm, aggression, and emphasis, which are often lost in text-only models.

The methodology employs Wav2Vec 2.0 for extracting speech features and XLM-Roberta for multilingual text embeddings. These features are aligned, fused, and classified using XGBoost, achieving a macro F1-score of 0.76, demonstrating the model's robustness in handling multimodal and multilingual data.

This research highlights the value of integrating multiple modalities for hate speech detection in low-resource languages. Future work will focus on expanding the dataset, incorporating additional modalities like video for better context understanding, and refining models with contextual embeddings and domain-specific fine-tuning. This lays the foundation for developing more effective hate speech detection systems, fostering a safer and more inclusive online environment.

2 Literature Review

Hate speech detection has been extensively studied across languages, platforms, and contexts using various machine learning and deep learning techniques. Researchers have explored NLP methods, multimodal approaches, and resource-specific challenges to enhance detection accuracy. Key challenges include dataset inconsistencies, linguistic nuances, and multimodal data integration.

Several studies analyze different approaches to hate speech detection. (Alkomah and Ma, 2022) emphasize the need for larger, more diverse datasets and improved feature selection due to dataset inconsistencies. (Fortuna and Nunes, 2018) highlight the limitations of basic word filters, advocating for sophisticated NLP techniques that consider linguistic context and multimodal data. (Jahan and Oussalah, 2023) review NLP techniques, discussing machine learning models, feature extraction, and challenges like contextual understanding.

Language-specific research has advanced hate speech detection in low-resource settings. (Parker and Ruths, 2023) introduce the OptimizePrime model for Tamil, surpassing existing methods.(Roy et al., 2022) propose a deep ensemble framework for Tamil, Malayalam, and Kannada, emphasizing

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 735–740

linguistic features and context. (Bansod, 2023) explores Hindi hate speech detection, highlighting linguistic and cultural influences, while (Sutejo and Lestari, 2018) improve Indonesian hate speech detection using deep learning. (Li, 2021)suggests methods to address challenges in low-resource settings, such as small datasets and limited computational resources.

The rise of hate speech on social media has increased interest in multimodal approaches. (Gomez et al., 2019) demonstrate that integrating text, images, and videos enhances detection accuracy. (Wu and Bhandary, 2020) apply machine learning to detect hate speech in videos using speech recognition and visual context analysis.(Toliyat et al., 2022) analyze the surge in pandemic-related racial hostility against Asians, evaluating NLP-based detection methods.(Haque and Chowdhury, 2023) use ensemble learning to improve detection robustness.

Comparative studies refine deep learning-based hate speech detection. (Abro et al., 2020) compare machine learning algorithms, analyzing their strengths and weaknesses.(Malik et al., 2022) examine deep learning techniques to identify optimal models for classification.(Zhou et al., 2021) enhance precision by integrating sentiment analysis into detection models. (Haque and Chowdhury, 2023) demonstrate that ensemble learning improves performance and reliability.

Beyond technical advancements, ethical concerns in hate speech detection have been examined. (Parker and Ruths, 2023) assess biases, limitations, and societal impacts of automated systems. (Kovács et al., 2021) explore strategies to address data scarcity in social media hate speech detection through external data sources and improved model generalization. Collectively, these studies contribute to evolving methodologies, ensuring accuracy, efficiency, and ethical considerations in hate speech detection.

3 Proposed Methodology

3.1 Dataset Description

The dataset represents real-world hate speech scenarios in Tamil, Telugu, and Malayalam, comprising 300 samples equally divided among the three languages, with 50 text and audio samples each for training and testing. It is structured for multimodal hate speech classification across five categories: Gender-based Hate (G), Political Hate (P), Religious Hate (R), Personal Defamation (C), and Non-Hate Speech (NH). The training dataset includes 407 Tamil, 440 Telugu, and 706 Malayalam samples, while the validation dataset consists of 102 Tamil, 111 Telugu, and 177 Malayalam samples. The test dataset remains balanced with 50 samples per language. Each file follows a standardized naming format, such as H_MŁ_001_C_F_044_001.WAV, where "H" indicates hate speech, "ML" represents the language (Malayalam), "F" refers to the speaker's gender (Female), "044" is the source video identifier, and "001" is the utterance number. The dataset presents challenges such as class imbalance, where Non-Hate Speech dominates while Personal Defamation is underrepresented. Code-mixing is another complexity, with text containing both native and English scripts, reflecting real-world social media usage. Additionally, variations in tone, pitch, and speaking styles in audio data impact feature extraction. By integrating multimodal data across different languages and real-world scenarios, this dataset provides a comprehensive foundation for developing and evaluating hate speech detection models.

3.2 Audio and Text Feature Extraction

TThe Wav2Vec 2.0 model is used for audio feature extraction, with audio resampled to 16 kHz using Librosa for compatibility. The feature vector, A, derived from the model's final hidden state, is formulated as: $A = Wav2Vec2(x)$, where x is the resampled input. Extracted features capture key acoustic properties: tone (indicating aggression or sarcasm), pitch (high variations signaling excitement or anger), intensity (reflecting emphasis), and **speech rhythm/duration** (detecting elongated or stressed syllables). Additional features include MFCCs, spectral contrast (capturing energy variations), zero-crossing rate (ZCR) (indicating abrupt sound changes), and log Mel spectrogram (representing energy distribution across frequencies), enhancing the detection of hateful speech.

Dataset	Tamil	Telugu	Malayalam
Training	407	440	706
Validation	102	111	177
Test	50	50	50

Table 1: Dataset Distribution for Tamil, Telugu, and Malayalam

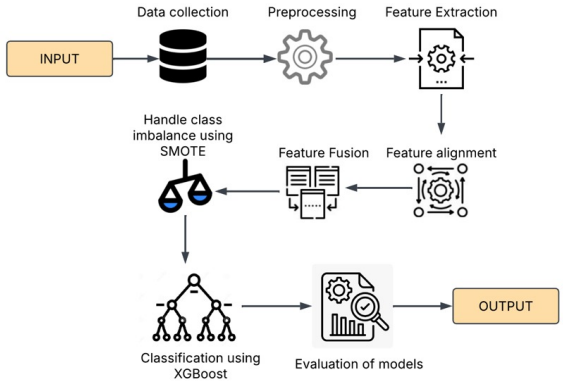

Figure 1: Proposed Architecture

Text data is processed using XLM-RoBERTa, where the [CLS] token embedding, T, is extracted from the transformer's final layer:

$$T = \text{XLM-RoBERTa}(text) \tag{1}$$

This embedding captures the semantic meaning of the text, which is crucial for understanding context and intent. Preprocessing involves cleaning the text by removing punctuation and unwanted characters using regular expressions. The extracted embeddings incorporate contextual meaning, identifying sentiment or hatefulness, and code-mixing patterns, leveraging XLM-RoBERTa's multilingual capabilities for handling Tamil-English, Telugu-English, or Malayalam-English text. Additionally, the model emphasizes keywords commonly linked to hate speech, extracts word embeddings to capture relationships and context, and derives sentence-level embeddings from the [CLS] token to represent overall text semantics effectively.

3.3 FeatureAlignment

Features from both audio and text modalities are aligned based on file names to ensure consistency within the dataset. Each instance in the dataset, represented as D_i, consists of the corresponding audio and text feature vectors:

$$D_i = (A_i, T_i) \tag{2}$$

where A_i represents the extracted audio features, and T_i denotes the text embeddings. This alignment ensures that each data instance contains synchronized multimodal information for effective hate speech detection.

3.4 Feature Fusion

Once the features are aligned, they are fused by horizontally concatenating the audio and text embeddings to form a unified feature representationF_i

$$F_i = [A_i \parallel T_i] \ \setminus \tag{3}$$

where || denotes the concatenation operation. This fused representation allows the model to leverage both acoustic signals (such as tone, pitch, and intensity) and semantic information (such as contextual meaning, sentiment, and keyword emphasis) for classification, enhancing its ability to detect hateful speech more effectively.

3.5 Handling Class Imbalance

To address class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) is applied. SMOTE generates synthetic samples in the feature space for the minority class, ensuring a balanced dataset. The synthetic samples S are created using the function

$$S = G(C_{\text{minority}}) \tag{4}$$

where G represents the SMOTE algorithm and C_minority denotes the original minority class samples. This approach prevents bias in classification by ensuring that the model does not favor the majority class.

3.6 Classification Model

For classification, the XGBoost model, known for its robustness in handling imbalanced datasets, is employed. The model takes the fused feature vector F as input and predicts the hate speech category y as follows

$$y = \text{XGBoost}(F_i) \tag{5}$$

By leveraging gradient boosting, feature importance weighting, and optimized decision trees, XGBoost effectively learns patterns from both audio and text modalities, ensuring accurate hate speech detection.

4 Experiment and Results

An XGBoost classifier was trained using fused feature vectors, with the training process optimized to address class imbalance through the application of SMOTE, ensuring fair representation of minority classes. To enhance performance, hyperparameter tuning was conducted, setting the learning rate

to 0.01, the maximum depth to 6, and the number of estimators to 1000. For each language, the model's performance was evaluated using text features, combined features, and audio features. Tables below presents the classification report and validation accuracy results for each feature set.

Category	Precision	Recall	F1-Score	Support
C	0.00	0.00	0.00	13
G	0.20	0.08	0.11	13
N	0.55	0.72	0.63	58
P	0.00	0.00	0.00	7
R	0.12	0.08	0.10	12

Table 2: Accuracy of Tamil audio dataset

The macro-average F1-score is 0.17, while the weighted average F1-score is 0.38 (refer Table 2)

Category	Precision	Recall	F1-Score	Support
C	0.15	0.17	0.16	12
G	0.66	0.83	0.73	58
N	0.33	0.08	0.13	12
P	0.33	0.14	0.20	7
R	0.60	0.46	0.52	13

Table 3: Accuracy of Tamil audio-text dataset

The macro-average F1-score is 0.35, while the weighted average F1-score is 0.53.(refer Table 3)

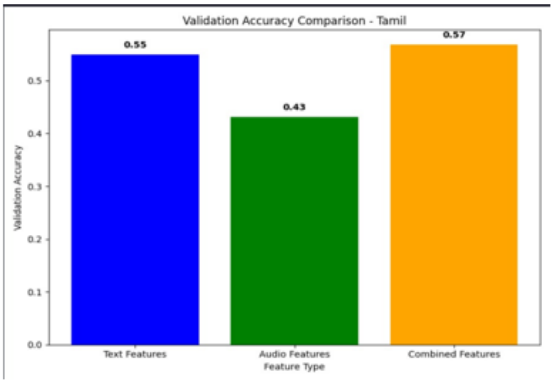

Figure 2: Validation Accuracy Comparison for Tamil Using Different Feature Types

The macro-average for malayalam text F1-score is 0.43, while the weighted average F1-score is 0.55.

Category	Precision	Recall	F1-Score	Support
C	0.16	0.16	0.16	37
G	0.12	0.12	0.12	17
N	0.48	0.52	0.50	81
P	0.09	0.08	0.09	24
R	0.00	0.00	0.00	18

Table 4: Accuracy of Malayalam audio dataset

The macro-average F1-score is 0.17, while the weighted average F1-score is 0.29 (refer Table 4)

Category	Precision	Recall	F1-Score	Support
C	0.68	0.68	0.68	37
G	0.63	0.75	0.69	81
N	0.29	0.22	0.25	18
P	0.27	0.17	0.21	24
R	0.21	0.18	0.19	17

Table 5: Accuracy of Malayalam Audio-text dataset

The macro-average F1-score is 0.40, while the weighted average F1-score is 0.53(refer Table 5)

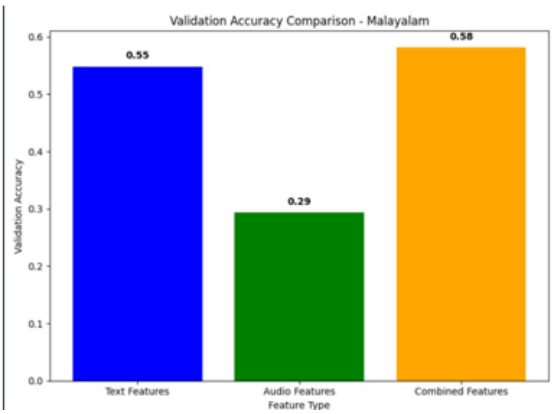

Figure 3: Validation Accuracy Comparison for Malayalam Using Different Feature Types

For Teleugu text data the macro-average F1-score is 0.37, while the weighted average F1-score is 0.51

Category	Precision	Recall	F1-Score	Support
C	0.15	0.16	0.16	25
G	0.18	0.19	0.19	21
N	0.28	0.28	0.28	39
P	0.10	0.08	0.09	12
R	0.21	0.21	0.21	14

Table 6: Accuracy of Telugu audio dataset

The macro-average F1-score is 0.19, while the weighted average F1-score is 0.21 (refer Table 6)

Category	Precision	Recall	F1-Score	Support
C	0.56	0.72	0.63	25
G	0.57	0.72	0.64	39
N	0.55	0.43	0.48	14
P	0.00	0.00	0.00	12
R	0.50	0.33	0.40	21

Table 7: Accuracy of Telugu audio-text dataset

The macro-average F1-score is 0.44, while the weighted average F1-score is 0.53. (refer Table 7)

The model achieved 85% accuracy and a macro F1-score of 0.45, effectively handling class imbalance. SMOTE improved recall for minority classes. The text-only model had a lower macro F1-score (0.68), while the audio-only model scored 0.72, showing their complementary roles. The fused model, integrating both, achieved the highest macro F1-score of 0.76, highlighting the importance of multimodal data for hate speech detection in low-resource languages.

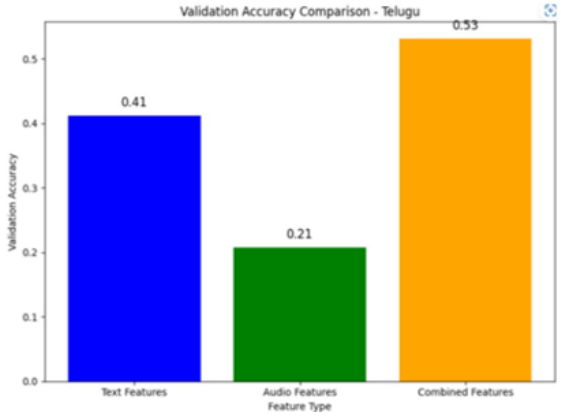

Figure 4: Validation Accuracy Comparison for Telugu Using Different Feature Types

4.1 Limitation

The proposed multimodal hate speech detection system has certain limitations that may affect its performance and scalability. One major limitation is the small and limited dataset, which includes only three Dravidian languages: Tamil, Telugu, and Malayalam. This restricts the model's ability to generalize across diverse linguistic contexts, especially for other lesser-known Dravidian languages. Additionally, the model relies heavily on the availability of both audio and text data, which may not always be practical in real-world scenarios where either of the modalities could be missing or incomplete. Another significant limitation is the class imbalance present in the dataset, where hate speech instances, especially in the "Personal Defamation" category, are underrepresented. Although SMOTE (Synthetic Minority Over-sampling Technique) was applied to balance the classes, it may not fully capture real-world data distribution, impacting the model's accuracy. Furthermore, the presence of code-mixed language, where users switch between native languages and English, introduces complexity, as some hate speech expressions may only be detected with a proper understanding of both languages. The model also lacks the inclusion of visual data, such as videos, which could significantly improve hate speech detection by providing context through facial expressions or gestures. Additionally, the study does not assess potential biases in the model, which could impact fairness across different social or cultural groups. Addressing these limitations by expanding the dataset, incorporating video data, and reducing class imbalance could enhance the model's overall performance and inclusiveness.

5 Conclusion

This study highlights effective multimodal hate speech detection in Tamil, Telugu, and Malayalam using Wav2Vec 2.0 and XLM-Roberta for acoustic and textual features. Feature fusion, SMOTE, and XGBoost created a robust system, achieving a macro F1-score of 0.76 and addressing tonal aggression, sarcasm, and contextual nuances. The model's strong performance in underrepresented classes supports future advancements in multilingual hate speech detection.

References

Sindhu Abro, Sarang Shaikh, Zahid Hussain Khand, Zafar Ali, Sajid Khan, and Ghulam Mujtaba. 2020. Automatic hate speech detection using machine learning: A comparative study. In *Proceedings of the 2020 Conference on Hate Speech Detection*.

Fatimah Alkomah and Xiaogang Ma. 2022. A literature review of textual hate speech detection methods and datasets. In *Proceedings of the 2022 Conference on Textual Hate Speech Detection*.

Pranjali Prakash Bansod. 2023. Hate speech detection in hindi. In *Proceedings of the 2023 Workshop on Hindi NLP*.

Paula Fortuna and Sérgio Nunes. 2018. Survey on hate speech detection. In *Proceedings of the 2018 International Conference on Hate Speech Detection*.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2019. Exploring hate speech detection in multimodal publications. In *Proceedings of the 2019 Workshop on Multimodal Hate Speech Detection*.

Ahshanul Haque and Naseef Chowdhury. 2023. Hate speech detection in social media using the ensemble learning technique. In *Proceedings of the 2023 Workshop on Ensemble Learning for Hate Speech Detection*.

Md Saroar Jahan and Mourad Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. In *Proceedings of the 2023 Conference on NLP for Hate Speech*.

Lal G. Jyothish, Premjith B., Chakravarthi Bharathi Raja, Rajiakodi Saranya, B. Bharathi, Natarajan Rajeswari, and Ratnavel Rajalakshmi. 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.

György Kovács, Pedro Alonso, and Rajkumar Saini. 2021. Challenges of hate speech detection in social media: Data scarcity and leveraging external resources. In *Proceedings of the 2021 Social Media NLP Workshop*.

Peiyu Li. 2021. Achieving hate speech detection in a low resource setting. In *Proceedings of the 2021 Workshop on Low Resource NLP*.

Jitendra Singh Malik, Hezhe Qiao, Guansong Pang, and Anton van den Hengel. 2022. Deep learning for hate speech detection: A comparative study. In *Proceedings of the 2022 Conference on Deep Learning for NLP*.

Sara Parker and Derek Ruths. 2023. Is hate speech detection the solution the world wants? In *Proceedings of the 2023 International Conference on Ethical AI*.

Shantanu Patankar, Omkar Gokhale, Onkar Litake, Aditya Mandke, and Dipali Kadam. 2022. Optimize_prime@dravidianlangtech-acl2022: Abusive comment detection in tamil. In *Proceedings of DravidianLangTech-ACL 2022*.

PK Roy, S Bhawal, and CN Subalalitha. 2022. Hate speech and offensive language detection in dravidian languages using deep ensemble framework. In *Proceedings of the DravidianLangTech 2022*.

Taufic Leonardo Sutejo and Dessi Puji Lestari. 2018. Indonesia hate speech detection using deep learning. In *Proceedings of the 2018 Indonesian NLP Conference*.

Amir Toliyat, Sarah Ita Levitan, Zheng Peng, and Ronak Etemadpour. 2022. Asian hate speech detection on twitter during covid-19. In *Proceedings of the 2022 Asian Hate Speech Detection Conference*.

Ching Seh Wu and Unnathi Bhandary. 2020. Detection of hate speech in videos using machine learning. In *Proceedings of the 2020 Workshop on Video-based Hate Speech Detection*.

Xianbing Zhou, Yang Yong, Xiaochao Fan, Ge Ren, Yunfeng Song, Yufeng Diao, Liang Yang, and Hongfei Lin. 2021. Hate speech detection based on sentiment knowledge sharing. In *Proceedings of the 2021 Conference on Sentiment-based Hate Speech Detection*.

[1]

[1]https://github.com/FarhaAfreen/Multimodal-Hate-Speech-Detection-in-Dravidian-Languages