MNLP@DravidianLangTech 2025: A Deep Multimodal Neural Network for Hate Speech Detection in Dravidian Languages

Shraddha Chauhan
Department of ECE
MNNIT-Allahabad
Prayagraj, Uttar Pradesh, 211004
shraddha.20224147@mnnit.ac.in

Abhinav Kumar
Department of CSE
MNNIT-Allahabad
Prayagraj, Uttar Pradesh, 211004
abhik@mnnit.ac.in

Abstract

Social media hate speech is a significant issue because it may incite violence, discrimination, and social unrest. Anonymity and reach of such platforms enable the rapid spread of harmful content, targeting individuals or communities based on race, gender, religion, or other attributes. The detection of hate speech is very important for the creation of safe online environments, protection of marginalized groups, and compliance with legal and ethical standards. This paper aims to analyze complex social media content using a combination of textual and audio features. The experimental results establish the effectiveness of the proposed approach, with F_1-scores reaching 72% for Tamil, 77% for Malayalam, and 36% for Telugu. Such results strongly indicate that multimodal methodologies have significant room for improvement in hate speech detection in resource-constrained languages and underscore the need to continue further research into this critical area.

1 Introduction

Social media has changed the way people communicate, share information, and express opinions (Saumya et al., 2024; Bhawal et al., 2021). This digital transformation has been of great benefit, but it has also enabled the spread of hate speech, misinformation, and harmful content (Saumya et al., 2024; Biradar et al., 2022). Hate speech on social media is a serious issue since it encourages discrimination, hostility, and violence, thereby negatively affecting individuals and communities (Saumya et al., 2021; Kumar et al., 2021). Such content identification and blocking are highly pertinent to ensuring safer online surroundings within multilingual regions with strong multicultural influences.

Dravidian languages, such as Tamil, Malayalam, and Telugu, represent a rich linguistic heritage spoken by millions in South India and neighboring countries. Despite their prominence, these languages remain underrepresented in computational linguistics and natural language processing (NLP) research. The unique linguistic features of Dravidian languages, such as their morphology, syntax, and phonology, make it difficult to analyze texts automatically. Moreover, hate speech in these languages is often code-mixed, using native words with English, and auditory cues in multimedia formats like speech.

This study develops hate speech detection systems for Dravidian languages by adopting a multimodal approach that incorporates both textual and auditory data. Unlike traditional text-only methods, the integration of audio features enables the detection of tonal, emotional, and contextual cues that are critical for identifying hate speech in spoken communication. By leveraging advanced deep learning techniques and multimodal data fusion, our work aims to improve the accuracy and reliability of hate speech detection in Tamil, Malayalam, and Telugu social media posts.

The rest of the structure of the paper follows the sequel: Section 2 lists the related work, Section 3 deals with the dataset and task, and Section 4 details the proposed methodology. Section 5 reports results from the proposed model, Section 6 concludes the paper, and Section 7 details the limitations of proposed model.

2 Related Work

Hate speech refers to offensive or prejudiced communication aimed at demeaning individuals or groups based on their identity or beliefs (Kumar et al., 2020; Mishra et al., 2020). (Sharma et al., 2024) provided an extensive survey relating to the study done on hate speech detection in South Asian languages and classified all studies based on their tasks, datasets, and methodologies. (Diaz-Garcia and Carvalho, 2025) provided an in-depth survey on the impact of new technologies, such as Language Models and Large Language Models, on the

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 324–329

evolution of abusive content detection.

(Sreelakshmi et al., 2024a) introduced a cost-sensitive learning approach towards hate speech and offensive language detection in code-mixed text. This improves methodologies for solving complex linguistic problems of Dravidian languages. (Premjith et al., 2024) significantly contributed to the research area of hate speech detection in Dravidian languages by performing two major tasks. They highlighted cutting-edge strategies for handling hate speech on social media platforms by presenting the results of the "HOLD-Telugu" shared task on the identification of hate and offensive language in Telugu code-mixed text. They also participated in the "MSMDA-DL" shared task, which focused on multimodal social media data analysis by combining audio and textual elements to enhance the identification of hate speech in Dravidian languages.

(Roy and Kumar, 2025) proposed a cost-sensitive learning approach for detecting code-mixed hate speech in social media posts, leveraging advanced multilingual models and machine learning classifiers to address the challenges of linguistic diversity in Dravidian languages. (Yuan and Rizoiu, 2025) proposed a multi-task learning approach to improve the generalization of hate speech detection models on different datasets. In this respect, they introduces the PubFigs dataset to analyze hate speech in political discourse.

(Singh et al., 2025) proposed a multimodal approach incorporating emotional understanding to detect offensive content, focusing on women's harassment. (Raphel et al., 2024) explored the use of multilingual transformer-based embedding models and machine learning classifiers to detect hate speech and offensive language in code-mixed Dravidian language texts.

(Sai et al., 2024) focused on breaking linguistic barriers by developing robust methodologies to detect hate speech in Telugu-English code-mixed text. Studies such as (Kakati and Dandotiya, 2024) explored ensemble methods for improved accuracy. Using multilingual BERT models, works such as (Zamir et al., 2024) and (Abitte Kanta et al., 2024) concentrated on detecting hate speech and offensive language in Telugu.

In the Malayalam code-mixed language, (Sreelakshmi et al., 2024b) presented a deep learning-based feature fusion technique designed for sentiment analysis and offensive text identification. Through the integration of several linguistic characteristics, the study emphasizes the difficulties presented by code-mixed texts and the effectiveness of sophisticated deep learning models in precisely identifying objectionable material and sentiment.

Traditional machine learning models, such as Support Vector Machines (SVM) (Cortes and Vapnik, 1995), K-Nearest Neighbors (KNN) (Cover and Hart, 1967), Random Forest (RF) (Breiman, 2001), Naive Bayes (NB) (John and Langley, 1995), and Logistic Regression (LR) (Hosmer and Lemeshow, 1989), were widely used for classification tasks, including hate speech detection (Boishakhi et al., 2021).

3 Dataset & Task

The Tamil, Malayalam, and Telugu datasets consist of 509, 883, and 551 audio and the corresponding transcript for training and testing data consists of 50 audio and the corresponding transcript for each of the languages (Lal G et al., 2025). The audio and transcripts are classified into Hate and Non-Hate categories. The hate class is further divided into subclasses: Gender (G), Political (P), Religious (R), and Personal Defamation (C), while the Non-Hate class is denoted as N.

4 Methodology

This section explores the pre-processing, feature extraction, feature fusion and training machine learning and deep learning models for classification of multimodal Hate Speech. The framework illustrating these methodologies is depicted in Figure 1. The design of the work is as follows:

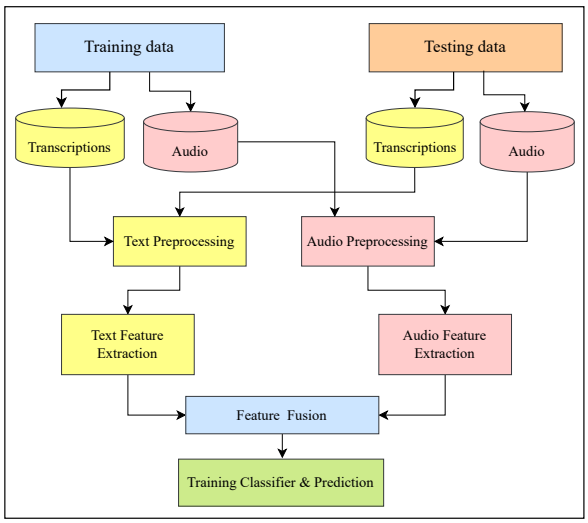

Figure 1: Overall flow diagram of the proposed framework

1. A methodology is implemented to extract audio and text features using Mel-Frequency Cepstral Coefficients (MFCCs) and XLM-RoBERTa respectively.

2. The extracted features from text and images are fused for further processing.

3. The fused features are used to train machine learning and deep learning models for classification of hate speech.

4.1 Pre-processing

Noises like stopwords, numbers, and punctuation that don't help with categorization are eliminated during pre-processing from the audio transcription. To eliminate Tamil, Malayalam, and Telugu stopwords, the github repositories'[1] Tamil, Malayalam, and Telugu are used.

4.2 Feature extraction

The pre-processed text data is then transformed into feature vectors using feature extraction techniques. This work utilizes XLM-RoBERTa to extract features from text using a transformer-based architecture designed for multilingual understanding. The process begins with tokenizing the text into subword units using Byte-Pair Encoding, ensuring effective handling of rare words and diverse languages. Each token is mapped to a high-dimensional feature that incorporates token, positional, and segment information. These features are passed through multiple transformer layers, where self-attention captures relationships between tokens, and feedforward networks refine the contextual representations. The hidden states generated in each layer provide rich contextualized features for each token. The final output is a set of dense feature vectors for Tamil, Malayalam and Telugu text.

To extract audio features, Mel-Frequency Cepstral Coefficients (MFCCs), is used for representation in audio processing. Each audio file was loaded in its original sampling rate using the Librosa library, ensuring the preservation of audio fidelity. MFCCs, capturing critical spectral characteristics of audio signals and modeled on human auditory perception, were computed with 13 coefficients per frame. These coefficients effectively summarize the power spectrum and frequency content of the audio signal. To derive a fixed-dimensional feature vector suitable for further analysis, we averaged the MFCCs over the time axis, reducing temporal variability while retaining essential features for Tamil, Malayalam, and Telugu audio. This approach ensures robust representation of audio characteristics.

[1] https://github.com/stopwords-iso/stopwords-iso

4.3 Feature Fusion & Classification

To integrate multimodal information from text and audio, we used a simple concatenation-based feature fusion strategy. The extracted text and audio features are combined along the feature dimension. This fused representation enables the model to leverage complementary insights from text and audio, improving its ability to capture multimodal context. Machine learning models like SVM, KNN, RF, NB and LR are trained on these fused features and is used for hate speech classification.

The Multimodal Classifier (MMC) integrates text and audio features for hate speech detection using a two-stage deep learning model. It consists of separate two-layer fully connected subnetworks for text and audio features, each utilizing ReLU activation, batch normalization, and dropout (0.1) for regularization. The extracted modality-specific features are concatenated and processed through a three-layer fusion network, which learns inter-modal relationships before classification using softmax activation. The model is trained for 50 epochs end-to-end with cross-entropy loss, using the Adam optimizer (learning rate = 6e-5) and batch size = 64. The hidden dimension is 256, ensuring effective feature representation and robust multimodal learning.

5 Results

Table 1 show the Accuracy (Acc), Precision (Pre), Recall (Rec), and F_1-score (F_1) achieved by various classifiers on the Tamil, Malayalam and Telugu datasets. This study evaluated both conventional machine learning and deep learning models to classify hate speech. The Multimodal Classifier (MMC) demonstrated superior performance compared to conventional machine learning models, including SVM, KNN, RF, NB and LR. The proposed Multimodal Classifier (MMC) achieved consistently higher accuracy of 72%, 78% and 37% and F_1-scores of 72%, 77% and 36% across the Tamil, Malayalam, and Telugu datasets, respec-

Table 1: Performance metrics of various models across Tamil, Malayalam and Telugu dataset.

	Tamil				Malayalam				Telugu			
Model	**Acc**	**Pre**	**Rec**	**F1**	**Acc**	**Pre**	**Rec**	**F1**	**Acc**	**Pre**	**Rec**	**F1**
SVM	0.52	0.68	0.52	0.52	0.56	0.61	0.56	0.50	0.30	0.34	0.30	0.30
KNN	0.44	0.64	0.44	0.42	0.46	0.51	0.46	0.43	0.18	0.16	0.18	0.17
RF	0.32	0.60	0.32	0.26	0.46	0.36	0.46	0.35	0.28	0.24	0.28	0.21
NB	0.42	0.50	0.42	0.40	0.50	0.49	0.50	0.48	0.12	0.10	0.12	0.10
LR	0.28	0.30	0.28	0.21	0.42	0.46	0.42	0.38	0.16	0.13	0.16	0.14
MMC	0.72	0.75	0 .72	**0.72**	0.78	0.79	0.78	**0.77**	0.37	0.36	0.37	**0.36**

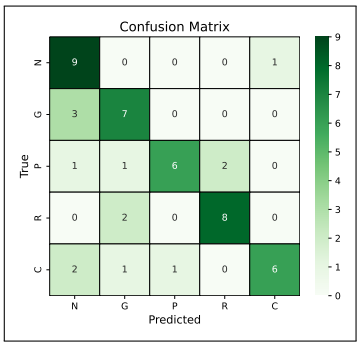

Figure 2: Confusion matrix of MMC for Tamil dataset.

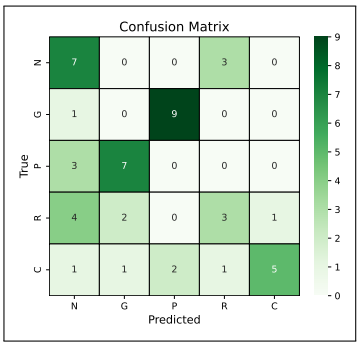

Figure 4: Confusion matrix of MMC for Telugu dataset.

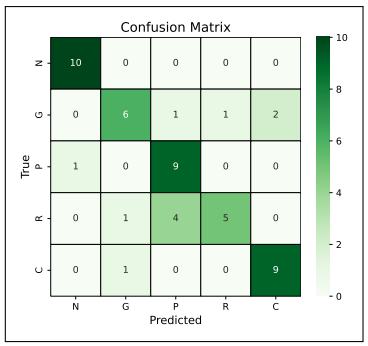

Figure 3: Confusion matrix of MMC for Malayalam dataset.

tively, effectively leveraging modality-specific features and their interactions. The confusion matrices for these models, illustrating their classification performance, are shown in Figures 2, 3, and 4, further highlighting the robustness of the MMC in handling complex multimodal data for hate speech detection tasks.

6 Conclusion

Hate speech is considered harmful as it can contribute to social divisions, violence, and harm to individuals dignity and rights. This study focused on detecting hate speech in Tamil, Malayalam, and Telugu languages using a multimodal approach that integrates text and audio features. The proposed Multimodal Classifier (MMC) outperformed traditional machine learning models such as SVM, KNN, RF, NB, and LR in terms of F_1-score across all three languages. MMC achieved an F_1-score of 72% in Tamil, 77% in Malayalam, and 36% in Telugu, significantly surpassing the performance of other models, which struggled to handle the complexities of code-mixed and linguistically diverse data. These results highlight the value of combining multiple modalities like text and audio to address the complexity of underrepresented languages. By demonstrating superior classification performance, this research emphasizes the importance of developing inclusive and robust hate speech detection systems, contributing to safer and more equitable digital spaces for diverse communities.

7 Limitations

While our multimodal approach improves hate speech detection in Tamil, Malayalam, and Telugu, certain limitations remain. Our model performs significantly better for Tamil and Malayalam but shows relatively lower performance for Telugu. This discrepancy may be due to smaller dataset

size, higher linguistic diversity, and the phonetic variations in Telugu, making it harder for the model to learn meaningful patterns. The dataset imbalance may also lead to biased predictions, affecting F_1-score for minority classes. The model struggles with dialect variations, code-mixed content, and informal social media language, which impacts its robustness in real-world scenarios. In the future, a stronger system can be developed by addressing these limitations.

The code for the proposed framework is available at:
https://github.com/Cshraddha153/A-Deep-Multimodal-Neural-Network-for-Hate-Speech-Detection-in-Dravidian-Languages.git

References

Selam Abitte Kanta, Grigori Sidorov, and Alexander Gelbukh. 2024. Selam@DravidianLangTech 2024:identifying hate speech and offensive language. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 91–95, St. Julian's, Malta. Association for Computational Linguistics.

Snehaan Bhawal, Pradeep Roy, and Abhinav Kumar. 2021. Hate speech and offensive language identification on multilingual code mixed text using bert. In *FIRE (Working Notes)*, pages 615–624.

Shankar Biradar, Sunil Saumya, Abhinav Kumar, and Ashish Singh. 2022. Pradvis vac: A socio-demographic dataset for determining the level of hatred severity in a low-resource hinglish language. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Fariha Tahosin Boishakhi, Ponkoj Chandra Shill, and Md. Golam Rabiul Alam. 2021. Multi-modal hate speech detection using machine learning. In *2021 IEEE International Conference on Big Data (Big Data)*, page 4496–4499. IEEE.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

Thomas Cover and Peter Hart. 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27.

Jose A. Diaz-Garcia and Joao Paulo Carvalho. 2025. A survey of textual cyber abuse detection using cutting-edge language models and large language models. *arXiv preprint arXiv:2501.05443*.

David W Hosmer and Stanley Lemeshow. 1989. *Applied Logistic Regression*. John Wiley & Sons.

George H John and Pat Langley. 1995. Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345. Morgan Kaufmann Publishers Inc.

Pallabi Kakati and Devendra Dandotiya. 2024. Automatic detection of hate speech in code-mixed indian languages in twitter social media interaction using dconvblstm-muril ensemble method. *Social Network Analysis and Mining*, 14(1):108.

Abhinav Kumar, Pradeep Kumar Roy, and Sunil Saumya. 2021. An ensemble approach for hate and offensive language identification in english and indo-aryan languages. In *FIRE (Working Notes)*, pages 439–445.

Abhinav Kumar, Sunil Saumya, and Jyoti Prakash Singh. 2020. NITP-AI-NLP@ HASOC-Dravidian-CodeMix-FIRE2020: a machine learning approach to identify offensive languages from dravidian code-mixed text. In *FIRE (Working notes)*, pages 384–390.

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.

Ankit Kumar Mishra, Sunil Saumya, and Abhinav Kumar. 2020. IIIT_DWD@ HASOC 2020: identifying offensive content in indo-european languages. In *FIRE (working notes)*, pages 139–144.

B Premjith, Bharathi Raja Chakravarthi, Prasanna Kumar Kumaresan, Saranya Rajiakodi, Sai Karnati, Sai Mangamuru, and Chandu Janakiram. 2024. 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.

M. Raphel, B. Premjith, K. Sreelakshmi, and B. R. Chakravarthi. 2024. Detection of hate speech and offensive language codemix text in dravidian languages using cost-sensitive learning approach. *IEEE Access*, 12:12345–12356.

Pradeep Kumar Roy and Abhinav Kumar. 2025. Ensuring safety in digital spaces: Detecting code-mixed hate speech in social media posts. *Data & Knowledge Engineering*, page 102409.

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, St. Julian's, Malta. Association for Computational Linguistics.

Sunil Saumya, Abhinav Kumar, and Jyoti Prakash Singh. 2021. Offensive language identification in dravidian code mixed social media text. In *Proceedings of the first workshop on speech and language technologies for Dravidian languages*, pages 36–45.

Sunil Saumya, Abhinav Kumar, and Jyoti Prakash Singh. 2024. Filtering offensive language from multilingual social media contents: A deep learning approach. *Engineering Applications of Artificial Intelligence*, 133:108159.

Deepawali Sharma, Tanusree Nath, Vedika Gupta, and Vivek Kumar. 2024. Hate speech detection research in south asian languages: A survey of tasks, datasets, and methods. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(1):1–25.

Gopendra Vikram Singh, Soumitra Ghosh, and Pushpak Bhattacharyya. 2025. Unmasking offensive content: A multimodal approach with emotional understanding. *Multimedia Tools and Applications*, 84(1):1–25.

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

K Sreelakshmi, B Premjith, Bharathi Raja Chakravarthi, and KP Soman. 2024b. A feature fusion and detection approach using deep learning for sentimental analysis and offensive text detection from code-mix malayalam language. *Journal of Intelligent Information Systems*.

Lanqin Yuan and Marian-Andrei Rizoiu. 2025. Generalizing hate speech detection using multi-task learning: A case study of political public figures. *Computer Speech Language*, 89:101690.

Muhammad Zamir, Moein Tash, Zahra Ahani, Alexander Gelbukh, and Grigori Sidorov. 2024. Lidoma@DravidianLangTech 2024: Identifying hate speech in Telugu code-mixed: A BERT multilingual. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 101–106, St. Julian's, Malta. Association for Computational Linguistics.