SSNCSE@DravidianLangTech 2025: Multimodal Hate Speech Detection in Dravidian Languages

Sreeja K, Bharathi B
Department of Computer Science and Engineering
Sri Sivasubramania Nadar College of Engineering
sreeja2350625@ssn.edu.in
bharathib@ssn.edu.in

Abstract

Hate speech detection is a serious challenge due to the different digital media communication, particularly in low-resource languages. This research focuses on the problem of multimodal hate speech detection by incorporating both textual and audio modalities. In the context of social media platforms, hate speech is conveyed not only through text but also through audio, which may further amplify harmful content. In order to manage the issue, we provide a multiclass classification model that influences both text and audio features to detect and categorize hate speech in low-resource languages. The model uses machine learning models for text analysis and audio processing, allowing it to efficiently capture the complex relationships between the two modalities. The class weight mechanism involves avoiding overfitting. The prediction has been finalized using the majority fusion technique. Performance is measured using a macro average F1 score metric. Three languages—Tamil, Malayalam, and Telugu—have the optimal F1 scores, which are 0.59, 0.52, and 0.33.

1 Introduction

In the digital era, the analysis of multimodal social media data aligns your insights with very different types of diverse data appearing on social networks, including text, audio, and video. However, with the advent of social networks, platforms such as YouTube, Facebook, and Twitter not only aided in information sharing and networking, but also became a place where people were targeted, defamed, and marginalized based on their religion, sex, political, and personal defamation. Social networks have become increasingly integrated in this digital age; it has changed the perception of networking and socializing.

Not only humans, but chatbots can also corrupted by hate speech content. After learning foul language from user interactions, Microsoft's chatbot "Tay"(Neff and Nagy, 2016), which was designed to engage people in lighthearted and informal discussion, began using it. The hate content was too obvious for the chatbot to identify and avoid. This serves as an illustration of how important it is to identify hate speech in tweets and social networks for applications such as sentiment analysis, chatbot development, content recommendation, etc. An efficient identification guarantees a safer, more moral AI system and a blocking mechanism against the spreading of dangerous content.

Hate speech analysis models trained for such contexts must reflect features of all modalities concerned. In our case, the task is to classify multimodal (text and audio) data in Tamil, Malayalam, and Telugu into five separate hate classes: gender, political, religious, personal defamation, and non-hate.

The rest of the paper is organized as follows: Section 2 analyzes the related works done in the previous research, and Section 3 discusses the hate speech corpus in the current work. Section 4 contains a detailed discussion of the proposed models used in the current work. Section 5 explains the experimental results. Section 6 discusses the limitations. In Section 7, concludes the paper.

2 Related works

Detecting hate speech is the most effective way to make any environment safe, inclusive, and respectful, both online and offline. This will protect individuals from emotional distress, psychological suffering, and the risky transition from hostility to physical harm. The rate of division is decreased along with social integration and tolerance in communities when hate speech is recognized and suppressed. However, hate speech detection

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 218–222

in low-resource languages is challenging due to limited linguistic resources, the complexity and dynamic systems of cultures, and technological gaps. All of these challenges need strong documented work in collecting data, culturally sensitive models, and tailored approaches for fairness and effectiveness. (Lal G et al., 2025)provides an overview of the shared task on Multimodal Social Media Data Analysis in Dravidian Languages (MSMDA-DL). The paper explores multiclass hate speech detection in Dravidian languages. Detecting hate content in social media comments is not a novel concept for the English language(Kumar and Singh, 2022)(Jemima et al., 2022). Several systems have also been developed for languages other than English, such as Hindi, German (Rajalakshmi et al., 2022),(Rajalakshmi and Reddy, 2020). However, limited research focuses on identifying offensive content in low-resource Dravidian languages such as Tamil, Malayalam, and Kannada(Roy et al., 2022). The study proposes a method for identifying hate speech in low-resource languages in Tamil, Malayalam, and Telugu. The proposed model expands the task into multiclass classification, with the intent of detecting hate speech in various categories to refine the classification and enhance the detection. The switch from binary to multiclass classification identifies the potential of hate speech across different contexts and modalities. To improve the approach of (Boishakhi et al., 2021), and (Premjith et al., 2024b) which initially employed binary classification, we extend the method to handle multiclass classification. Moreover, the Binary class distinguishes only hate and non-hate. In contrast, multiclass classification categorizes the content in its target or intent, providing a deeper understanding of why it is considered hateful. This approach uses multiclass categorization for the detection and classification to prioritize and identify hate speech types. In multiclass classification the classes are imbalanced, to overcome this(Sreelakshmi et al., 2024) uses the class weight mechanism by assigning more weights to minority classes and the model pays more attention to them. Multimodal classification for abusive comment detection was discussed in (Anierudh et al., 2024).

3 Dataset Description

The task aims to develop a model for detecting Hate speech in low-resource languages namely Tamil, Malayalam, and Telugu. The dataset is sourced from the Multimodal Social Media Data Analysis in Dravidian Languages (MSMDA-DL) provided by DravidianLangTech@NAACL 2025(Premjith et al., 2024b),(Premjith et al., 2024a). The task comprises three subtasks, and each subtasks contains two modalities data like Text and Audio. Each Audio data has a corresponding Transcript in Text data. The subtasks are Multimodal Hate Speech Detection in low-resource languages namely Tamil, Malayalam, and Telugu. Each language contains 514, 863, and 556 training samples for Tamil, Malayalam, and Telugu, as well as 50 test samples for each. This is a multi-class classification task, the classes are Gender (G), Political (P), Religious (R), Personal Defamation(C), and Non-Hate(N).

Category	Tamil	Malayalam	Telugu
Non-Hate (N)	287	406	198
Personal Defamation (C)	65	186	122
Gender (G)	68	82	106
Political (P)	33	118	58
Religious (R)	61	91	72
Total	514	863	556

Table 1: Training dataset distribution for Tamil, Malayalam, and Telugu.

4 Proposed Methodology

This study proposes a systematic methodology for classifying multimodal data encompassing four low-resource languages. The entire study is divided into five stages: data preprocessing, data balancing, feature extraction, classifier modeling, and majority fusion mechanism for predictions. Each stage has been thoughtfully designed to counter various problems posed by multimodal and low-resource language data.

Data preprocessing is the most important step in preparing the input multimodal data. The data have different modalities (text and audio) that may be inconsistent or noisy. The preprocessing pipeline consists of the following steps: cleaning the data by removing noisy information, and normalizing modalities to ensure consistency. The textual data undergo techniques such as tokenization and processing by language-specific methods. This step ensures a clean, structured, and aligned dataset ready for further processing.

Class weights mechanism has been utilized to balance the classes. Class weight was guaranteed

to provide more importance for minority classes during training. The class imbalance prevents the classifiers from biasing towards majority classes by improving performance.

Feature extraction serves as the most important for training the data. Textual data is vectorized using TF-IDF vectorizer and CountVectorizer. For audio data, MFCC and Log-mel spectrograms are employed. This step guarantees that diverse modalities are successfully transformed into feature-modeling vectors that can feed into machine-learning models.

The research proposes a classifier training method. Classifier models are used to perform the multiclass classification task: Support Vector Machines(SVM), Random Forests(RF), Multi-layer Perceptron(MLP) classifier, and Logistic Regression(LR). SVM uses kernel functions to handle linear as well as non-linear relationships effectively. Random Forest is another class of techniques that exploits an ensemble-based approach, which is very robust in capturing feature interactions. The MLP classifier is a feedforward artificial neural network with input, hidden, and output layers, among other layers. It uses backpropagation for training and applies activation to capture non-linear relationships in the data. It can handle structured data effectively due to its wide versatility. Logistic Regression is a highly popular classification task, it is a simple but effective linear model for providing great interpretability and strong baseline performance. Each model is trained successfully.

Multimodal data aggregation is performed using a majority fusion mechanism to combine predictions across modalities and models. A majority voting method is used for the first time to merge the predictions of three classifiers from each modality. The final output for each modality is taken as the label predicted by the majority of the classifiers for an instance of each modality. The results of all modalities are fused again using another majority voting mechanism to produce the overall model prediction. This two-level fusion mechanism ensures that all artifacts from all modalities and classifiers are substantially fused to obtain a robust and accurate prediction system.

In summary, the proposed methodology constitutes a complete workflow for classifying multimodal data for low-resource language speakers. The issues of sparsity, imbalance, and multimodal integration are directly addressed by including preprocessing, class balancing, modality-specific feature extraction, classifier selection, and hierarchical majority fusion. In other words, the final majority fusion takes place based on modality-wise predictions, which helps the model to draw on the diverse strengths of the classifiers and modalities.

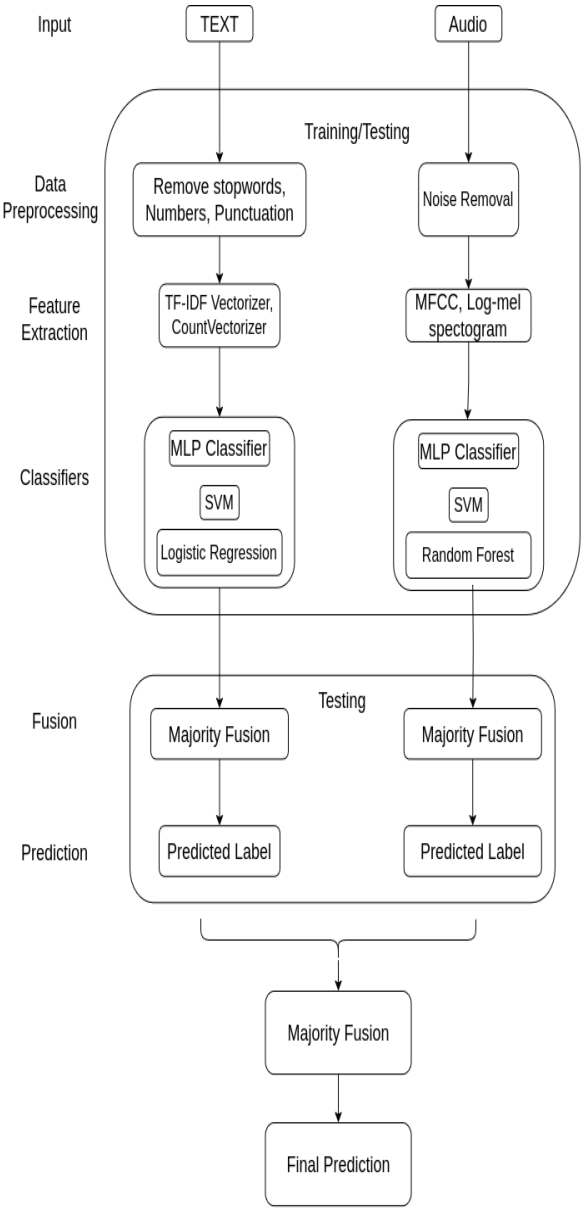

Figure 1: Architecture Diagram of the proposed work

5 Experimental Results

The performance of the Multimodal Hate Speech Detection model was evaluated with a macro F1-score. Text is trained with SVM, MLP classifier, and Logistic Regression model for feature extraction TF-IDF and Count vectorizer are used, and audio is trained with SVM, MLP classifier, and Random forest model for feature extraction MFCC and Log-mel Spectrogram are used. Using Majority Fusion technique Text and audio is fused independently. Finally, text and audio are both fused in

the Majority fusion mechanism.

Increasing the weight of a particular model in an ensemble learning system can result in huge improvements in performance, Table 2 shows the improved result, whereas Table 3 shows the results without any weight adjustment. In the weighted method described in Table 2, better-performing models will have additional influence on the final prediction, thus giving rise to better results. On the contrary, in Table 3, the models are treated uniformly; this may affect the overall accuracy downwards as well as result in less influence from more accurate models. The source code for the proposed approach and found here [1].

Metric	Tamil	Malayalam	Telugu
Accuracy	0.50	0.42	0.38
F1-Score	0.47	0.34	0.34
Precision	0.47	0.39	0.36
Recall	0.50	0.42	0.38

Table 2: Performance analysis of multimodal hate speech detection across languages without class weights

Metrics	Tamil	Malayalam	Telugu
Accuracy	0.60	0.56	0.38
F1-Score	0.59	0.52	0.33
Precision	0.63	0.57	0.33
Recall	0.60	0.56	0.38

Table 3: Performance analysis of multimodal hate speech detection across languages using class weights

6 Limitations

The current work caught several hurdles, including:

- The paper acknowledges the cruciality of obtaining sufficient and representative data for detecting hate speech in low-resource languages. The model suggests generalizability and strength may be impacted by this constraint.
- The issue of imbalance in hate speech datasets is reduced by using the application of an imbalance class weights technique, biases in the predictions will still exist to some extent, especially when it comes to the minority classes.
- The model becomes more sophisticated as text and audio modalities are added. The two-level fusion technique proposed in the research still requires additional testing before it can be used in real-world scenarios.

[1] https://github.com/SreejaKumaravel/Multimodal-Hate-Speech-Detection

7 Conclusions

The study concludes with a demonstration of the effectiveness of the majority fusion and class weighing in machine learning models for multimodal hate speech identification. In multiclass classification tasks, weighted-class models are preferred because they satisfy underrepresented classes and become sensitive enough to these class instances. With robust fusion methods capable of combining different model outputs, it is likely to obtain the optimal F1 score, which is one of the most important metrics in evaluating classification performance on imbalanced datasets. The experimental results show the promise of this method in dealing with the challenge of multimodal data and unbalanced class distribution and may lead to future advances in hate speech detection systems.

References

S Anierudh, R Abhishek, Ashwin Sundar, Amrit Krishnan, and B Bharathi. 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.

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)*, pages 4496–4499. IEEE.

P. Preethy Jemima, Bishop Raj Majumder, Bibek Kumar Ghosh, and Farazul Hoda. 2022. Hate speech detection using machine learning. In *2022 7th International Conference on Communication and Electronics Systems (ICCES)*, pages 1274–1277.

Gunjan Kumar and Jyoti Prakash Singh. 2022. Hate speech and offensive content identification in english and indo-aryan languages using machine learning models. In *FIRE (Working Notes)*, pages 542–551.

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.

Gina Neff and Peter Nagy. 2016. Talking to bots: Symbiotic agency and the case of tay. *International Journal of Communication*, 10:4915–4931.

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.

R Rajalakshmi and Yashwant Reddy. 2020. An enhanced ensemble classifier for hate and offensive content identification. *Journal of E-Technology Volume*, 11(2):71.

Ratnavel Rajalakshmi, Ankita Duraphe, and Antonette Shibani. 2022. DLRG@DravidianLangTech-ACL2022: Abusive comment detection in Tamil using multilingual transformer models. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 207–213, Dublin, Ireland. Association for Computational Linguistics.

Pradeep Kumar Roy, Snehaan Bhawal, and Chinnaudayar Navaneethakrishnan Subalalitha. 2022. Hate speech and offensive language detection in dravidian languages using deep ensemble framework. *Computer Speech Language*, 75:101386.

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*.