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

Ratnavel Rajalakshmi, R. Ramesh Kannan, Meetesh Saini, Bitan Mallik
School of Computer Science and Engineering
Vellore Institute of Technology, Chennai, TamilNadu
rajalakshmi.r@vit.ac.in

Abstract

Social media is a powerful communication tool and rich in diverse content requiring innovative approaches to understand nuances of the languages. Addressing challenges like hate speech necessitates multimodal analysis that integrates textual, and other cues to capture its context and intent effectively. This paper proposes a multi-modal hate speech detection system in Tamil, which uses textual and audio features for classification. Our proposed system uses a fine-tuned Indic-BERT with Whisper as a multimodal approach for hate speech detection. The fine-tuned Indic-BERT model with Whisper achieved an F1 score of 0.25 on Multimodal based approach. Our proposed approach ranked at 10th position in the shared task on Multimodal Hate Speech Detection in Dravidian languages at the NAACL 2025 Workshop DravidianLangTech.

1 Introduction

Hate speech is a serious problem that harms people and communities. It targets individuals or groups based on characteristics like race, religion, or gender, leading to violence, discrimination, and prejudice. With the rise of social media, hate speech becomes more common, creating risks for society and individuals (Sreelakshmi et al., 2024; Wickramaarachchi et al., 2023; Khanduja et al., 2024). Many studies focus on detecting hate speech in widely spoken languages like English (Khanduja et al., 2024; Conneau et al., 2020), but Dravidian languages, such as Tamil, do not receive as much attention. Tamil is a complex language with a unique structure, making it important to develop specialized hate speech detection systems for it.

This paper presents a multimodal system to detect hate speech in Tamil. Hate speech appears in different forms, including text, audio, and images. Previous research explores transformer-based models like mBERT and XLM-R for hate speech detection (Khanduja et al., 2024; Ibañez et al., 2021). Some studies also develop hate speech detection systems for low-resource languages using multimodal approaches. To handle text and audio data, we propose a system that combines a fine-tuned Indic-BERT model with Whisper as a Multimodal hate speech classification. Our proposed method obtained a F1 score of 0.25 in detection of hate speech in Tamil.

2 Related Works

Hate speech detection has traditionally focused on monolingual text, with early approaches employing machine learning algorithms such as Support Vector Machines (SVMs) and n-gram features (Warner and Hirschberg, 2012). Such methods fail to deal with the subtle semantics of hate speech. Deep learning methods, such as CNNs and LSTMs, have achieved better results by encoding contextual information (Zhang et al., 2018). More recently, transformer models like BERT have further developed the area (Devlin et al., 2019). For other low resource languages like Tamil (Rajalakshmi et al., 2023; Ganganwar and Rajalakshmi, 2022), Marathi (Rajalakshmi et al., 2021a), Telugu (Rajalakshmi et al., 2024), Hindi (Rajalakshmi and Reddy, 2019; Rajalakshmi et al., 2021b) and Multilingual languages (Reddy and Rajalakshmi, 2020) authors have worked towards Hate and offensive content identification on textual data. Recent research has investigated multimodal methods for hate speech detection, especially utilizing both the text and audio modalities. (Mahajan et al., 2024) have used models that integrates CNNs and LSTMs to handle spectrograms and text embeddings for the identification of offensive speech. Likewise, (Khanduja et al., 2024) created dataset for hate speech detection in low resource Dravidian language Telugu. Explored transformer models such as mBERT, DistilBERT, IndicBERT, NLLB, Muril, RNN+LSTM, XLM-ROBERTa, and Indic-

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 454–458

Bart. Fine-tuned mBERT model achieved a accuracy of 98.2% on the newly created hate speech Telugu dataset. For sentiment analysis of tweets on social media content, (Kannan et al., 2021) employed the IndicBERT model. (Boishakhi et al., 2021) employed a multimodal approach for hate speech detection by combining video, audio, and transcribed text. For audio, features such as MFCC, ENERGY, ZCR, and chroma were utilized, while Bag of Words and TF/IDF were applied to transcribed text. A hard voting ensemble model was used to highlight the advantages of contextual analysis in achieving more accurate hate speech classification.

3 Model Architecture

3.1 Wav2Vec2 based model

Wav2Vec2 model is pre-trained on a large corpus of multilingual speech data. Model provides robust representations of acoustic features that benefits downstream tasks. The raw audio input, preprocessed is passed into the Wav2Vec2 model, which returns a sequence of contextualized representations of hidden states. Thus, the hidden states are able to capture subtle patterns within the audio and effectively encode temporal and spectral aspects of the signal. We then perform mean pooling over the time dimension on the sequence of hidden states obtained from Wav2Vec2. This pooling operation summarizes the audio input by aggregating all the temporal information into a fixed-length representation. The pooled representation is now given as a single vector that encapsulates the core acoustic features of the audio, and then it passes through a dropout layer. This layer introduces randomness while the model learns during training, reducing overfitting and improving the generalizability of the model to unseen data. The dropout rate, set at a specific value, regulates the number of neurons randomly deactivated during training. Finally, the output of the dropout layer feeds into a fully connected linear layer. This layer maps the pooled and regularized representation to a set of logits, corresponding to the five classes of our classification task. These logits during training are utilized to calculate cross-entropy loss when compared to the ground-truth labels allowing the model fine-tune on the task particular parameters. On inference, logits are the model output used as a way to determine the predicted class label using softmax activation functions to obtain a probability distribution of the classes.

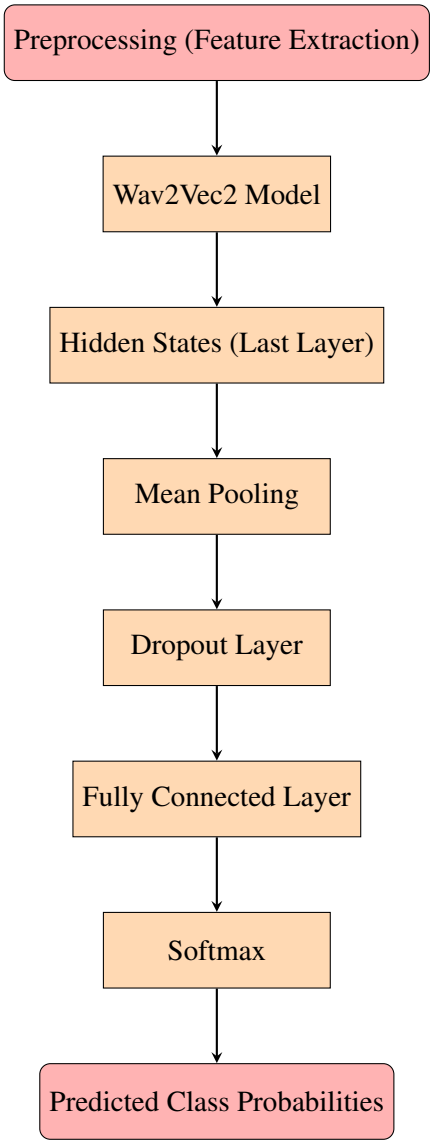

Figure 1: Flowchart of Wav2Vec2 Speech Classification Model

3.2 Indic-BERT+Whisper based model

Our proposed approach leverages Whisper for audio-to-text transcription, followed by IndicBERT for the classification of Telugu hate speech from the transcribed text. Whisper, a multilingual automatic speech recognition model, allows transcription of spoken Tamil so that subsequent textual analysis can be performed. Indic-BERT, a language model pre-trained on Indian languages exclusively, is superior to generic multilingual models like mBERT and XLM-R as it is able to capture linguistic variations specific to Tamil. This fusion overcomes the shortcomings of less specialized speech-to-text pipelines, making the transcriptions more appro-

priate for subsequent analysis. The core idea is the pooled output from the Indic-BERT model, which encapsulates the semantic understanding of the transcribed text. The pooled representation is passed through a dropout layer, which introduces stochasticity during training and prevents overfitting. The output of the dropout layer is then passed to a fully connected linear layer. This linear layer maps the contextualized text representation to the final output, which are the class logits. A cross-entropy loss is computed using the ground truth labels and predicted logits, which is used in model training. Finally, during inference, the logits are used to find class labels based on the highest probability. This architecture makes use of the capabilities of the Whisper model for speech-to-text conversion and Indic-BERT for contextual understanding in the classification task. Although Whisper is reported to have occasional transcription errors and it shows better performance in low-resource languages makes it a good choice. By integrating these two models, the proposed system efficiently handles both spoken and written Tamil, and it enhances the resilience of hate speech detection in a multimodal environment.

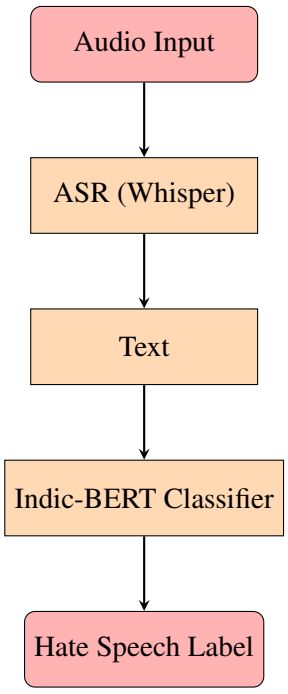

Figure 2: Flowchart of the Multimodal Hate Speech Detection System

4 Dataset Description

Multimodal Hate speech Telugu dataset contains text and audio data Tamil, Telugu and Malayalam. The audio samples are sourced from YouTube videos and encompass a variety of speakers, ensuring linguistic diversity and representation. Each utterance is labeled under one of five categories: N (Non-hate speech), G (Gender-based hate speech), P (Political hate speech), R (Religious hate speech), and C (Communal hate speech). Multimodal Hate speech Tamil dataset consists of 514 samples. It is well-structured into two categories, WAV files of speech recordings and transcriptions of text contents. The transcriptions are a textual representation of the spoken material, allowing for both linguistic analysis and machine learning use. The disbribution of the data is as shown in Figure 3. Additionally, a test dataset is provided, comprising 50 audio and text samples. However, the test labels are not included, requiring researchers to submit their model predictions for evaluation.

4.1 Data Preprocessing

The audio data is processed into raw audio signals, a format compatible with the feature extraction capabilities of the facebook/wav2vec2-large-xlsr-53 model. This process begins by loading raw audio files in .wav format and standardizing them by converting multi-channel audio to mono-channel, ensuring uniform input dimensionality. The preprocessing involves extracting the audio waveform and its corresponding sample rate. Since wav2vec2-large-xlsr-53 is trained on a 16 kHz sampling rate, input audio is resampled to match this rate. Instead of relying on manual feature engineering, the raw audio waveforms are directly passed through the model's feature extraction layers, enabling the model to learn relevant features autonomously. To meet the model's requirements, audio sequences are truncated to a maximum of 16,000 samples, equivalent to one second of audio sampled at 16 kHz, ensuring consistent input length. The audio processing pipeline in Indic-BERT employs a two-stage approach: transcription followed by classification. Initially, raw audio files are processed using the vasista22/whisper-tamil-medium model, configured specifically for Tamil transcription, to convert the audio into its textual representation with high accuracy. The resulting text undergoes preprocessing steps, including the removal of URLs and non-alphanumeric characters, as well as standardizing the input to a maximum length of 128 tokens through padding or truncation. The preprocessed text is then passed to the Indic-BERT model for classification.

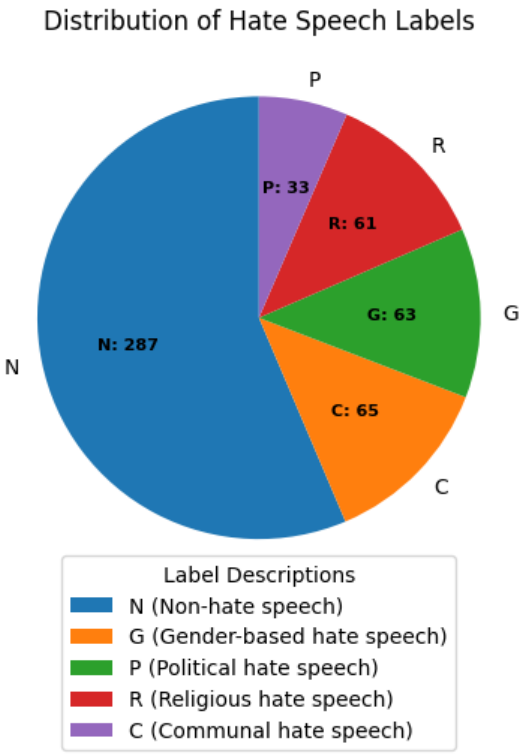

Figure 3: Distribution of Training dataset

5 Experiments and Results

The learning rate used to train both models is 3e-5 with the Adam optimizer. A linear warmup schedule is implemented during training, and both models are trained for a fixed number of epochs. At the end of each epoch, evaluation is performed to track the model progress. Gradient accumulation steps are used to optimize resource utilization for the Wav2Vec2 model. The model checkpoint with the highest held-out validation set accuracy is used in each case for the final model. For the Indic-BERT-based model, this is found at the 21st epoch, while for the Wav2Vec2-based model, training is conducted for 10 epochs. Wav2Vec2 and IndicBERT models were trained using a learning rate of 3e-5 with the Adam optimizer. The system is evaluated using a dataset from the shared task. The Wav2Vec2-based model attains an accuracy of 51% with an F1-score of 0.35 and Multimodal based Indic-BERT and Whisper obtained a F1 score of 0.25 on Tamil Hate Speech Detection. The following results were obtained after training the classification model:

Class	Prec	Recall	F1	Supp.
C	0.32	0.70	0.44	10
G	0.57	0.80	0.67	10
N	0.14	0.20	0.17	10
P	0.00	0.00	0.00	10
R	0.00	0.00	0.00	10

Metrics	Prec.	Recall	F1	Supp.
Accuracy			0.34	50
Macro Avg	0.21	0.34	0.25	50
Wei. Avg	0.21	0.34	0.25	50

Table 1: Classification report of Indic-BERT based model

6 Conclusion and Future Work

This paper explores multimodal hate speech detection in Tamil, using transcribed text and raw audio processed through separate architectures. A Whisper/Indic-BERT based multimodal approach captures textual and audio semantics and achieved a F1 score of 0.25. Wav2Vec2 focuses on only on speech features and obtained a F1 score of 0.35. Our findings laid the foundation for advanced models. The study highlights the importance of addressing challenges in audio-text integration and optimizing feature extraction. Future efforts will explore ensembling, dataset improvements, and enhanced audio pipelines, aiming to better integrate audio-text interactions for improved performance and societal impact.

References

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.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *CoRR*, abs/2006.13979.

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.

Vaishali Ganganwar and Ratnavel Rajalakshmi. 2022. Mtdot: A multilingual translation-based data augmentation technique for offensive content identification in tamil text data. *Electronics*, 11(21):3574.

Michael Ibañez, Ranz Sapinit, Lloyd Antonie Reyes, Mohammed Hussien, Joseph Marvin Imperial, and Ramón Rodriguez. 2021. Audio-based hate speech classification from online short-form videos. In *2021 International Conference on Asian Language Processing (IALP)*, pages 72–77. IEEE.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of EMNLP*.

R Ramesh Kannan, Ratnavel Rajalakshmi, and Lokesh Kumar. 2021. Indicbert based approach for sentiment analysis on code-mixed tamil tweets. In *FIRE (Working Notes)*, pages 729–736.

Namit Khanduja, Nishant Kumar, and Arun Chauhan. 2024. Telugu language hate speech detection using deep learning transformer models: Corpus generation and evaluation. *Systems and Soft Computing*, page 200112.

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.

Esshaan Mahajan, Hemaank Mahajan, and Sanjay Kumar. 2024. Ensmulhatecyb: Multilingual hate speech and cyberbully detection in online social media. *Expert Systems with Applications*, 236:121228.

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 B Yashwant Reddy. 2019. Dlrg@ hasoc 2019: An enhanced ensemble classifier for hate and offensive content identification.

Ratnavel Rajalakshmi, Faerie Mattins, S Srivarshan, Preethi Reddy, and M Anand Kumar. 2021a. Hate speech and offensive content identification in hindi and marathi language tweets using ensemble techniques. In *FIRE (Working Notes)*, pages 467–479.

Ratnavel Rajalakshmi, M Saptharishee, S Hareesh, R Gabriel, et al. 2024. Dlrg-dravidianlangtech@ eacl2024: Combating hate speech in telugu code-mixed text on social media. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 140–145.

Ratnavel Rajalakshmi, Srivarshan Selvaraj, Pavitra Vasudevan, et al. 2023. Hottest: Hate and offensive content identification in tamil using transformers and enhanced stemming. *Computer Speech & Language*, 78:101464.

Ratnavel Rajalakshmi, S Srivarshan, Faerie Mattins, E Kaarthik, and Prithvi Seshadri. 2021b. Conversational hate-offensive detection in code-mixed hindi-english tweets. In *CEUR Workshop Proceedings*, pages 1–11.

Yashwanth Reddy and Ratnavel Rajalakshmi. 2020. Dlrg@ hasoc 2020: A hybrid approach for hate and offensive content identification in multilingual tweets. In *FIRE (working notes)*, pages 304–310.

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*.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26, Montréal, Canada. Association for Computational Linguistics.

WAKM Wickramaarachchi, Sameeri Sathsara Subasinghe, KK Rashani Tharushika Wijerathna, A Sahashra Udani Athukorala, Lakmini Abeywardhana, and A Karunasena. 2023. Identifying false content and hate speech in sinhala youtube videos by analyzing the audio. In *2023 5th International Conference on Advancements in Computing (ICAC)*, pages 364–369. IEEE.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *The Semantic Web*, pages 745–760, Cham. Springer International Publishing.