byteSizedLLM@DravidianLangTech 2025: Multimodal Hate Speech Detection in Malayalam Using Attention-Driven BiLSTM, Malayalam-Topic-BERT, and Fine-Tuned Wav2Vec 2.0

Durga Prasad Manukonda
ASRlytics
Hyderabad, India
mdp0999@gmail.com

Rohith Gowtham Kodali
ASRlytics
Hyderabad, India
rohitkodali@gmail.com

Daniel Iglesias
Digi Sapiens
Frankfurt, Germany
diglesias@web.de

Abstract

This research presents a robust multimodal framework for hate speech detection in Malayalam, combining fine-tuned Wav2Vec 2.0, Malayalam-Doc-Topic-BERT, and an Attention-Driven BiLSTM architecture. The proposed approach effectively integrates acoustic and textual features, achieving a macro F1-score of 0.84 on the Malayalam test set. Fine-tuning Wav2Vec 2.0 on Malayalam speech data and leveraging Malayalam-Doc-Topic-BERT significantly improved performance over prior methods using openly available models. The results highlight the potential of language-specific models and advanced multimodal fusion techniques for addressing nuanced hate speech categories, setting the stage for future work on Dravidian languages like Tamil and Telugu.

1 Introduction

Social media platforms have revolutionized digital communication, enabling the seamless exchange of multimodal content, including text, images, videos, and audio. However, the increasing prevalence of hate speech on these platforms presents significant challenges for content moderation. The detection of such content requires advanced multimodal analysis techniques that effectively integrate textual and speech features to capture intent and context.

The Shared Task on Multimodal Hate Speech Detection in Dravidian Languages (DravidianLangTech@NAACL 2025) (Lal G et al., 2025), part of the Multimodal Social Media Data Analysis (MSMDA) initiative, promotes research in analyzing complex social media data using multimodal approaches. The MSMDA-DL shared task focuses on Tamil and Malayalam, two linguistically rich Dravidian languages, emphasizing the need for innovative multimodal natural language processing (NLP) techniques. This study specifically addresses **Task 1: Multimodal Hate Speech Detection in Malayalam**, where detecting hate speech requires integrating textual embeddings and acoustic features to distinguish between non-hate and various hate categories effectively.

This paper presents a multimodal classification approach combining Attention-Driven BiLSTM, BERT-Base, and Wav2Vec models to enhance Malayalam hate speech detection. The proposed architecture captures semantic and phonetic nuances, leveraging BERT-Base for text representation and Wav2Vec for speech-based feature extraction. The study provides insights into data preprocessing, model architecture, and classification performance, contributing to the broader understanding of multimodal hate speech detection in low-resource languages.

This study details our data preprocessing, Wav2Vec fine-tuning, and multimodal classifier design, introducing optimizations that improve detection accuracy and scalability. Our results provide insights into the challenges of multimodal NLP, contributing to advancements in hate speech detection for low-resource languages.

2 Related Work

Multimodal approaches for analyzing social media data have advanced significantly, particularly for underrepresented languages like Tamil, Telugu, and Malayalam. Banerjee et al. (2020) used an autoregressive XLNet for sentiment analysis on Tamil-English and Malayalam-English datasets, highlighting the challenges of multilingual and code-mixed data.

B et al. (2022) introduced the DravidianMultiModality Dataset, incorporating textual, audio, and visual features from product and movie review videos, underscoring the benefits of multimodal sentiment analysis.

Similarly, B et al. (2023) applied multimodal deep learning to disaster response, demonstrating

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 188–193
May 3, 2025 ©2025 Association for Computational Linguistics

how text and image integration aids real-world classification tasks.

The DravidianLangTech 2024 shared tasks (B et al., 2024; Premjith et al., 2024b,a) advanced multilingual and multimodal research, focusing on sentiment analysis, hate speech detection, and language identification for Dravidian languages. These initiatives foster innovation in handling the linguistic and cultural diversity of Tamil and Malayalam.

Building on this, the MSMDA shared task in Malayalam integrates textual and audio features to enhance hate speech detection. This effort tackles challenges in code-mixed content, complex morphology, and rich phonetic structures, pushing research forward in multimodal NLP for underrepresented languages.

3 Dataset

3.1 Fine-tuning Wav2Vec for Malayalam Speech Recognition

To enhance the performance of our multimodal hate speech detection model, we fine-tuned Wav2Vec 2.0 using a Malayalam speech recognition dataset sourced from the ULCA-ASR dataset corpus[1]. This 637.88-hour unlabelled Malayalam speech dataset supports fine-tuning Wav2Vec 2.0 base, enhancing phonetic and acoustic modeling for improved speech feature extraction in downstream tasks.

3.2 Hate Speech Dataset for Multimodal Testing

The Malayalam hate speech dataset, collected from YouTube, includes 933 utterances labeled into five categories: Non-Hate (**N**) and four hate subcategories—Gender (**G**), Political (**P**), Religious (**R**), and Personal Defamation (**C**) (Sreelakshmi et al., 2024). Each sample contains both audio and text for comprehensive multimodal analysis.

The dataset is split into 883 training samples (794 train, 89 dev) and 50 test samples. This structured labeling aids in effective classification and highlights hate speech characteristics in Malayalam.

Label	Train	Test	Total
N	406	10	416
C	186	10	196
P	118	10	128
R	91	10	101
G	82	10	92
Total	883	50	933

Table 1: Label distribution for Malayalam hate speech dataset.

4 Methodology

This section presents the methodology for multimodal hate speech detection in Malayalam, integrating fine-tuned Wav2Vec 2.0 for speech, Malayalam-Doc-Topic-BERT for textual embeddings, and an attention-driven BiLSTM-BERT-Wav2Vec classifier for fusion.

4.1 Fine-tuning Wav2Vec 2.0 for Malayalam Speech Recognition

The Wav2Vec 2.0 base model was fine-tuned on the ULCA-ASR dataset corpus of unlabelled Malayalam speech. The fine-tuning process utilized Facebook's Fairseq framework, optimizing the model to capture phonetic and acoustic nuances specific to Malayalam. This adaptation allowed the model to generate robust speech embeddings for downstream tasks, addressing the rich phonetic diversity and morphological complexity of Malayalam. The fine-tuned model serves as the foundation for extracting audio features in the proposed multimodal approach.

4.2 Malayalam-Doc-Topic-BERT

For textual embeddings, we selected the IndicSBERT model, l3cube-pune malayalam-sentence-bert-nli[2] (Mirashi et al., 2024), which was further fine-tuned on the L3Cube-IndicNews Corpus. This corpus encompasses three sub-datasets: Long Document Classification (LDC), Long Paragraph Classification (LPC), and Short Headline Classification (SHC), representing different document lengths. By training on a combination of these datasets, the Malayalam-Doc-Topic-BERT model achieves consistent performance across varied text lengths. It captures contextual semantics and document-level information for Malayalam hate speech detection.

4.3 Attention-Driven BiLSTM-BERT-Wav2Vec Classifier

This study presents a hybrid Attention-Driven BiLSTM-BERT-Wav2Vec model (Liu and Guo,

[1] https://github.com/Open-Speech-EkStep/ULCA-asr-dataset-corpus

[2] https://huggingface.co/l3cube-pune/malayalam-topic-all-doc

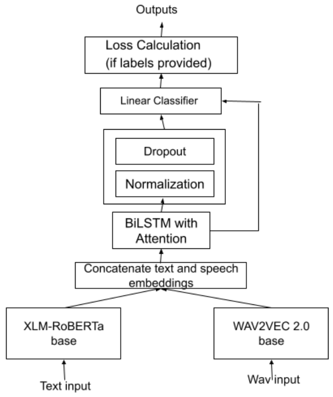

Figure 1: Architecture of the Attention-Driven BiLSTM-XLM-RoBERTa-Wav2Vec Classifier.

2019; Hochreiter and Schmidhuber, 1997; Graves and Schmidhuber, 2005; Kodali et al., 2025; Manukonda and Kodali, 2025, 2024a; Kodali and Manukonda, 2024; Manukonda and Kodali, 2024b) for multimodal classification, integrating text and speech features.

Text input is processed via a fine-tuned Malayalam-Doc-Topic-BERT, generating contextual embeddings:

$$\mathbf{X}_t = \mathbf{BERT}(input\_ids, attention\_mask) \quad (1)$$

Speech input is handled by a fine-tuned Wav2Vec model, producing acoustic embeddings:

$$\mathbf{X}_s = \mathbf{Wav2Vec}(audio\_features) \quad (2)$$

Both embeddings are concatenated:

$$\mathbf{X} = [\mathbf{X}_t; \mathbf{X}_s] \quad (3)$$

A BiLSTM extracts temporal patterns:

$$\mathbf{H}_t = [\mathbf{H}_{fwd,t}; \mathbf{H}_{bwd,t}] \quad (4)$$

An attention mechanism assigns weights α_t to focus on key features:

$$\alpha_t = \frac{\exp(\mathbf{a}_t)}{\sum_{t=1}^{T} \exp(\mathbf{a}_t)}, \quad \mathbf{H}_{attended} = \sum_{t=1}^{T} \alpha_t \cdot \mathbf{H}_t \quad (5)$$

Residual components, including layer normalization and dropout, enhance generalization, robustness, and stabilize training.

$$\mathbf{H}_{dropout} = Dropout(LayerNorm(\mathbf{H}_{attended})) \quad (6)$$

A fully connected layer maps features to classification logits:

$$\mathbf{logits} = \mathbf{W}_{cls} \cdot \mathbf{H}_{dropout} + \mathbf{b}_{cls} \quad (7)$$

The model is optimized using **cross-entropy loss**:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \quad (8)$$

This hybrid architecture (Figure 1) effectively integrates linguistic and acoustic insights, leveraging BERT, Wav2Vec, and BiLSTM with attention to enhance multimodal hate speech detection in Malayalam.

5 Experiment Setup

Our approach comprises fine-tuning Wav2Vec 2.0 base on Malayalam speech and developing a multimodal hate speech detection framework. The Wav2Vec 2.0 base model was fine-tuned on the ULCA-ASR Malayalam dataset (637.88 hours of unlabelled speech), with preprocessing steps including resampling to 16 kHz and noise reduction. Training used Fairseq, a tri-stage learning rate schedule, and ran for 50 epochs with Adam (lr = 1e-4). The best checkpoint, based on Word Error Rate (WER), was used for speech embeddings.

A hybrid model combined BERT base text embeddings with Wav2Vec base speech embeddings, processed through a BiLSTM (512 hidden units, 2 layers) with attention. Dropout (0.3) and layer normalization ensured stability. The final classifier predicted one of five labels.

Training employed PyTorch, AdamW (lr = 2e-5), and a ReduceLROnPlateau scheduler, running for 10 epochs on GPU, saving the best macro F1-score checkpoint. Evaluation measured accuracy, precision, recall, and F1-score, confirming the effectiveness of Wav2Vec-BERT fusion with BiLSTM and attention for Malayalam hate speech detection.

The model was assessed using the macro F1-score per DravidianLangTech guidelines, with scikit-learn generating precision, recall, and F1-score for all categories.

6 Results and Discussion

The fine-tuned Wav2Vec 2.0 model on the ULCA-ASR Malayalam dataset achieved a WER of 17.4%, enhancing phonetic representation for multimodal classification. Our Attention-Driven BiLSTM-BERT-Wav2Vec model attained an accuracy of

Label	**Precision**	**Recall**	**F1-score**	**Support**
C	0.77	1.00	0.87	10
G	1.00	0.70	0.82	10
N	0.83	1.00	0.91	10
P	0.88	0.70	0.78	10
R	0.80	0.80	0.80	10
Accuracy	-	-	0.84	50
Macro Avg	0.86	0.84	0.84	50
Weighted Avg	0.86	0.84	0.84	50

Table 2: Classification Report on the Test Set for Multimodal Hate Speech Detection, including precision, recall, F1-score, and support for each label.

84%, with macro and weighted F1-scores of 0.84. The best performance was observed in **C** (0.87) and **N** (0.91) categories, while **G** (0.82) showed high precision but lower recall. The **P** and **R** categories had F1-scores of 0.78 and 0.80, indicating challenges in detecting implicit hate speech.

Table 3 compares our model with top teams in DravidianLangTech@NAACL 2025. Our model outperformed all tested architectures, achieving a macro F1-score of 0.8360 due to effective integration of Malayalam-Doc-Topic-BERT and Wav2Vec 2.0 embeddings with BiLSTM and attention mechanisms. Training code and evaluation scripts are publicly available on GitHub[3], ensuring reproducibility.

Team Name	**mF1**	**Rank**
SSNTrio	0.7511	1
lowes	0.7367	2
MNLP	0.6135	3
byteSizedLLM	**0.5831**	**4**
KEC_Tech_Titans	0.5114	5
Attention-BiLSTM-BERT-Wav2Vec: 0.8360		

Table 3: Performance comparison in Multimodal Hate Speech Detection at DravidianLangTech@NAACL 2025.

Our submission ranked 4th (macro F1-score: 0.5831) as fine-tuning was incomplete at the deadline, requiring the use of an open-source Wav2Vec 2.0 base and XLM-RoBERTa base models. The `Malayalam-Doc-Topic-BERT` replacement improved performance. The top team, **SSNTrio**, achieved 0.7511.

Future work includes extending this approach to Telugu and Tamil, improving fusion techniques like hierarchical attention, and mitigating dataset imbalances.

7 Limitations and Future Work

Despite strong overall performance, our model has several limitations. First, it struggles with implicit hate speech, particularly in the **P** (Political) and **R** (Religious) categories, where nuanced language reduces recall. Second, reliance on pre-trained multilingual models limits adaptability to low-resource languages. Third, dataset imbalances affect recall, as seen in the **G** (Gender) category, which had high precision (1.00) but low recall (0.70), indicating missed instances of gender-based hate speech. Fourth, fine-tuning of Wav2Vec 2.0 was incomplete at submission, impacting final performance. Future work will focus on language-specific fine-tuning, dataset expansion, and improved multimodal fusion techniques to mitigate these limitations.

8 Conclusion

This work presents a novel multimodal framework combining Attention-Driven BiLSTM, fine-tuned Wav2Vec 2.0, and Malayalam-Doc-Topic-BERT for hate speech detection in Malayalam, achieving a macro F1-score of 0.84 and surpassing existing baselines in performance. The proposed method effectively integrates acoustic and textual features, demonstrating its ability to address the linguistic and cultural complexities of Malayalam. The use of fine-tuned Wav2Vec 2.0 and Malayalam-Doc-Topic-BERT emphasizes the importance of tailored, language-specific models for resource-scarce languages.

[3] https://github.com/mdp0999/Multimodal-Hate-Speech-in-Malayalam

References

Premjith B, Bharathi Raja Chakravarthi, Malliga Subramanian, Bharathi B, Soman Kp, Dhanalakshmi V, Sreelakshmi K, Arunaggiri Pandian, and Prasanna Kumaresan. 2022. Findings of the shared task on multimodal sentiment analysis and troll meme classification in Dravidian languages. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 254–260, Dublin, Ireland. Association for Computational Linguistics.

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.

Premjith B, Jyothish Lal G, Sowmya V, Bharathi Raja Chakravarthi, Rajeswari Natarajan, Nandhini K, Abirami Murugappan, Bharathi B, Kaushik M, Prasanth Sn, Aswin Raj R, and Vijai Simmon S. 2023. Findings of the shared task on multimodal abusive language detection and sentiment analysis in Tamil and Malayalam. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 72–79, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Shubhanker Banerjee, Arun Jayapal, and Sajeetha Thavareesan. 2020. Nuig-shubhanker@dravidian-codemix-fire2020: Sentiment analysis of code-mixed dravidian text using xlnet. *Preprint*, arXiv:2010.07773.

A. Graves and J. Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 4, pages 2047–2052 vol. 4.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Rohith Kodali and Durga Manukonda. 2024. byteSizedLLM@DravidianLangTech 2024: Fake news detection in Dravidian languages - unleashing the power of custom subword tokenization with Subword2Vec and BiLSTM. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 79–84, St. Julian's, Malta. Association for Computational Linguistics.

Rohith Gowtham Kodali, Durga Prasad Manukonda, and Daniel Iglesias. 2025. byteSizedLLM@NLU of Devanagari script languages 2025: Hate speech detection and target identification using customized attention BiLSTM and XLM-RoBERTa base embeddings. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 242–247, Abu Dhabi, UAE. International Committee on Computational Linguistics.

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.

Gang Liu and Jiabao Guo. 2019. Bidirectional lstm with attention mechanism and convolutional layer for text classification. *Neurocomputing*, 337:325–338.

Durga Manukonda and Rohith Kodali. 2024a. byteLLM@LT-EDI-2024: Homophobia/transphobia detection in social media comments - custom subword tokenization with Subword2Vec and BiLSTM. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 157–163, St. Julian's, Malta. Association for Computational Linguistics.

Durga Prasad Manukonda and Rohith Gowtham Kodali. 2024b. Enhancing multilingual natural language processing with custom subword tokenization: Subword2vec and bilstm integration for lightweight and streamlined approaches. In *2024 6th International Conference on Natural Language Processing (ICNLP)*, pages 366–371.

Durga Prasad Manukonda and Rohith Gowtham Kodali. 2025. byteSizedLLM@NLU of Devanagari script languages 2025: Language identification using customized attention BiLSTM and XLM-RoBERTa base embeddings. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 248–252, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Aishwarya Mirashi, Srushti Sonavane, Purva Lingayat, Tejas Padhiyar, and Raviraj Joshi. 2024. L3cube-indicnews: News-based short text and long document classification datasets in indic languages. *arXiv preprint arXiv:2401.02254*.

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 KP Soman. 2024. Detection of hate speech and offensive language codemix text in dravidian languages using cost-sensitive learning approach. *IEEE Access*.