Overview of the Shared Task on Multimodal Hate Speech Detection in Dravidian languages: DravidianLangTech@NAACL 2025

Jyothish Lal G[1], Premjith B[1], Bharathi Raja Chakravarthi[2],
Saranya Rajiakodi[3], Bharathi B[4], Rajeswari Natarajan[5], Ratnavel Rajalakshmi[6]
[1]Amrita School of Artificial Intelligence, Coimbatore, Amrita Vishwa Vidyapeetham, India,
[2]School of Computer Science, University of Galway, Ireland,
[3]Central University of Tamil Nadu, India, [4]SSN College of Engineering, Tamil Nadu, India,
[5]SASTRA University, India [6]Vellore Institute of Technology, Chennai, Tamil Nadu, India

Abstract

The detection of hate speech in social media platforms is very crucial these days. This is due to its adverse impact on mental health, social harmony, and online safety. This paper presents the overview of the shared task on Multimodal Hate Speech Detection in Dravidian Languages organized as part of DravidianLangTech@NAACL 2025. The task emphasizes detecting hate speech in social media content that combines speech and text. Here, we focus on three low-resource Dravidian languages: Malayalam, Tamil, and Telugu. Participants were required to classify hate speech in three sub-tasks, each corresponding to one of these languages. The dataset was curated by collecting speech and corresponding text from YouTube videos. Various machine learning and deep learning-based models, including transformer-based architectures and multimodal frameworks, were employed by the participants. The submissions were evaluated using the macro F1 score. Experimental results underline the potential of multimodal approaches in advancing hate speech detection for low-resource languages. Team SSNTrio achieved the highest F1 score in Malayalam and Tamil of 0.7511 and 0.7332, respectively. Team lowes scored the best F1 score of 0.3817 in the Telugu sub-task.

1 Introduction

As social networks have become an essential part of modern life, people use it to share their creations, opinions, and daily experiences (Paval et al., 2024). Although it is meant to be a platform for fun and information sharing, some use it to spread hate and profane content (Ben-David and Fernández, 2016). Hate posts usually target specific individuals, groups, or even nations. Moreover, many people use fake profiles to share such content. This harmful behavior has caused mental distress and conflicts among different groups. Addressing this issue requires better detection of hate speech, which is challenging due to the unique language styles used in online content (Schmidt and Wiegand, 2017).

In this context, manually identifying and removing hate speech is the simplest method, but it is a time consuming and tedious process (MacAvaney et al., 2019). Automated methods are becoming more popular because they are faster and more efficient. Most studies focus on detecting hate speech in written text, while progress in videos and images has been made using multimodal datasets. However, detecting hate speech in spoken language and its combination with text has not been explored much due to lack of multimodal datasets, especially for low-resource Dravidian languages (Anilkumar et al., 2024).

A lot of research is being done to recognize hate speech using images, text, and videos (Davidson et al., 2017; Safaya et al., 2020). Most studies currently focus on single modality, especially text. Neverthless, research on speech-based and multimodal approaches are also in growing phase. For example, Abhishek et al. (Anilkumar et al., 2024) used deep learning models to detect hate speech in Hindi and Marathi texts from the HASOC 2021 dataset (Velankar et al., 2021). They showed that transformer models work best, but even simple models with FastText embeddings can give strong results. Similarly, Tashvik Dhamija's (Dhamija et al., 2021) study on English tweets showed that RoBERTa embeddings combined with a decision tree algorithm performed exceptionally well.

A few other studies considered different languages and techniques for detection of hate-speech. For example, Vandan Mujadia et al. (Mujadia et al., 2019) have done hate/offensive content detection in multiple langauges such as Hindi, English, and German. The study showed best results while combining a voting system with ML classifier models. Another work by Joshi et al. (Joshi et al., 2021) showed that the BERT-based models perform better for Hindi data. Furthermore, Badjatiya et al. showed that a combination of LSTM with gradient-boosted decision trees gave good results. (Badjatiya et al., 2017). In summary, all the aforesaid studies show that advanced algorithms from the ML paradigm can improve the detection of hate speech, irrespective of the language variations.

The Shared Task on Multimodal Hate Speech Detection in Dravidian Languages, held at DravidianLangTech@NAACL 2025, is a step towards advancing the challenges in hate speech detection. This task aims to develop and test methodologies for detecting hate speech in social media using both speech and text.

2 Task Description

This shared task on hate speech detection focuses on Malayalam, Tamil, and Telugu, the three important Dravidian languages. Consequently, the task is divided into three sub-tasks as follows:

- Task 1: Multimodal hate-speech detection in Malayalam
- Task 2: Multimodal hate-speech detection in Tamil
- Task 3: Multimodal hate-speech detection in Telugu

Participants are provided with training data sets that contain multimodal content, including text and speech. The objective is to develop models capable of analyzing these components to predict the appropriate labels for hate speech detection. Model performance will be evaluated using the macro-F1 score, a widely used metric in NLP for classification tasks.

Language	Data	Label	Count	Total
Malayalam	Train	N	406	883
		C	186	
		P	118	
		R	91	
		G	82	
	Test	N	10	50
		C	10	
		P	10	
		R	10	
		G	10	
Tamil	Train	N	287	514
		C	65	
		P	33	
		R	61	
		G	68	
	Test	N	10	50
		C	10	
		P	10	
		R	10	
		G	10	
Telugu	Train	N	198	556
		C	122	
		P	58	
		R	72	
		G	106	
	Test	N	10	50
		C	10	
		P	10	
		R	10	
		G	10	

Table 1: Distribution of the hate speech data in Train and Test sets in Malayalam, Tamil, and Telugu languages. Here, the class labels N, C, P, R and G represent Non hate, Personal defamation, Political hate speech, Religious hate speech and Gender-based hate speech, respectively.

3 Dataset description

We collected our dataset from YouTube videos on channels with more than 50,000 subscribers to ensure wide reach and engagement. Instead of using a predefined list of hate speech terms, we selected videos based on the context of the spoken audio. In particular, we focused on topics that likely spark controversial or polarizing discussions. Examples include debates on the Ram Mandir inauguration, defamation of well-known figures. These topics are chosen for their high engagement and potential to attract hateful comments. By manually reviewing the audio for context and intent, we identified nuanced instances of hate speech without relying on specific keywords. Our study focuses on four types of hate speech, as defined by YouTube's Hate speech policy. They are as follows.

- **Gender-based hate speech (G):** Content targeting individuals based on their gender identity, sexual orientation, or personal relationships.
- **Political hate speech (P):** Negative remarks directed at individuals based on their nationality or political beliefs.
- **Religious hate speech (R):** Hateful content

aimed at specific individuals or communities related to their religion.

- **Personal Defamation (C):** Dehumanizing comments, such as comparisons to animals, diseases, or pests.

For sentences with multiple labels, such as personal defamation and religious content, the context of the video is used for final labeling. Further, we collected Non-hate (N) speech data from motivational videos because they are less likely to include offensive content.

4 Participants Methodology

A total of 134 teams registered for this shared task. However, only 19 teams submitted the results for atleast one sub-tasks. Precisely, 17 teams submitted the results for Malayalam and Tamil sub-task, and 18 teams submitted for Telugu sub-task. Some of the teams submitted multiple runs, of which, best run was taken as their final submission. We evaluated the submissions using the macro F1 score, and then prepared the rank list based on the results. Tables 1, 2 and 3 show the rank lists for the Malayalam, Tamil and Telugu sub-tasks, respectively. The methodologies used by each team are explained in following subsections.

4.1 SSNTrio

The team **"SSNTrio"** extracted speech features from the spectrogram and appended them to the corresponding text transcript and used a language-specific BERT model to complete the classification of hate speech. This approach achieved a macro F1 score of 0.7511, 0.7332, and 0.3758 for Task 1, Task 2, and Task 3, respectively.

4.2 lowes

The team **"lowes"** fine-tuned language open-source BERT models, which were pre-trained on the Dravidian languages by l3cube-pune. This approach achieved a macro F1 score of 0.7367, 0.7225, and 0.3817 for Task 1, Task 2, and Task 3, respectively.

4.3 MNLP

The team **"MNLP"** has used a deep learning-based model. Precisely, the model is fine-tuned for the classification task. This approach reported a macro F1 score of 0.6135, 0.4877, and 0.2184 for Task 1, 2, and 3, respectively.

Team	**Macro F1 Score**	**Rank**
SSNTrio (J et al., 2025)	0.7511	1
lowes	0.7367	2
MNLP (Chauhan and Kumar, 2025)	0.6135	3
byteSizedLLM (Manukonda et al., 2025)	0.5831	4
KEC_Tech_Titans	0.5114	5
zerowatts (Shanmugavadivel et al., 2025a)	0.4726	6
gryffindor (Shanmugavadivel et al., 2025d)	0.4725	7
VKG_VELLORE	0.4604	8
NLP_goats	0.4105	9
SSN_IT_SPEECH	0.3726	10
SSN_MMHS (Murali and Sivanaiah, 2025)	0.348	11
The Deathly Hallows (Shanmugavadivel et al., 2025b)	0.3016	12
Bright Red (Shanmugavadivel et al., 2025c)	0.2782	13
cantnlp (Wong and Li, 2025)	0.273	14
Team ML_Forge (Faisal et al., 2025)	0.2005	15
KEC-Elite-Analysts	0.0812	16
deanhthin	0.0758	17

Table 2: Rank list of Malayalam sub-task

Team	Macro F1 Score	Rank
SSNTrio (J et al., 2025)	0.7332	1
lowes	0.7225	2
The Deathly Hallows (Shanmugavadivel et al., 2025b)	0.6438	3
KEC_Tech_Titans	0.5322	4
MNLP (Chauhan and Kumar, 2025)	0.4877	5
KEC-Elite-Analysts	0.4281	6
NLP_goats	0.4049	7
VKG_VELLORE	0.3743	8
cantnlp (Wong and Li, 2025)	0.3186	9
Bright Red (Shanmugavadivel et al., 2025c)	0.3018	10
DLRG (Rajalakshmi et al., 2025)	0.2542	11
zerowatts (Shanmugavadivel et al., 2025a)	0.2432	12
gryffindor (Shanmugavadivel et al., 2025d)	0.2431	13
SSN_IT_SPEECH	0.2099	14
byteSizedLLM (Manukonda et al., 2025)	0.1596	15
Team ML_Forge (Faisal et al., 2025)	0.1346	16
deanhthin	0.0592	17

Table 3: Rank list of Tamil sub-task

Team	Macro F1 Score	Rank
lowes	0.3817	1
SSNTrio (J et al., 2025)	0.3758	2
SemanticCuet Sync_telugu (Hossain et al., 2025)	0.3514	3
VKG_VELLORE	0.3324	4
NLP_goats	0.2991	5
KEC_Tech_Titans	0.2857	6
gryffindor (Shanmugavadivel et al., 2025d)	0.264	7
zerowatts (Shanmugavadivel et al., 2025a)	0.264	8
Bright Red (Shanmugavadivel et al., 2025c)	0.251	9
byteSizedLLM (Manukonda et al., 2025)	0.2271	10
MNLP (Chauhan and Kumar, 2025)	0.2184	11
cantnlp (Wong and Li, 2025)	0.1774	12
SSN_IT_SPEECH	0.1631	13
SSN_MMHS (Murali and Sivanaiah, 2025)	0.1567	14
The Deathly Hallows (Shanmugavadivel et al., 2025b)	0.1559	15
Team ML_Forge (Faisal et al., 2025)	0.1465	16
KEC-Elite-Analysts	0.1326	17
deanhthin	0	18

Table 4: Rank list of Telugu sub-task

4.4 byteSizedLLM

The team **"byteSizedLLM"** has used a customized attention BiLSTM architecture integrated with XLM-RoBERTa base embeddings for textual features. Additionally, The XLM-RoBERTa was fine-tuned to handle the complexities associated with syntax and semantics. For audio features, they used fine-tuned wav2vec2-base multilingual speech embeddings. This approach reported a macro F1 score of 0.5831, 0.1596, and 0.2271 for Task 1, Task 2, and Task 3, respectively.

4.5 KEC_Tech_Titans

The team **"KEC_Tech_Titans"** employed pre-trained language models, including BERT, mBERT, RoBERTa, and XLNet, fine-tuned on task-specific datasets, along with CNN and BiLSTM to capture hierarchical and sequential patterns. They also utilized HAN and HGNN for attention-based feature extraction. For speech, speech-to-text models were integrated with text-based classifiers, and BiLSTM was applied for sequential feature analysis. Here, predictions from speech and text were combined using late fusion. This ensured a balanced classification. This method reported a macro F1 score of 0.5114, 0.5322, and 0.2857 for Task 1, Task 2, and Task 3, respectively.

4.6 zerowatts

The team **"zerowatts"** implemented an audio classification system by extracting acoustic features such as MFCC and spectral contrast, followed by feature normalization and label encoding to prepare data for model input. This method showed a macro F1 score of 0.4726, 0.2432, and 0.264 for Task 1, Task 2, and Task 3, respectively.

4.7 gryffindor

The team **"gryffindor"** developed an audio classification system by extracting acoustic features. The features include normalized MFCC and spectral contrast. This method reported macro F1 scores of 0.4725, 0.2431, and 0.264 for the three tasks in order.

4.8 VKG_VELLORE INSTITUTE OF TECHNOLOGY

The team **"VKG_VELLORE INSTITUTE OF TECHNOLOGY"** adopted a two-stage approach. In the first stage, language-specific models were trained using BERT embeddings. These embeddings are generated from text transcripts with pre-trained models. They have addressed the class imbalance problem through SMOTE and employed a CatBoost classifier for prediction. In the final stage, the whisper model is used for transcribing the audio. Further, the processed text was fed into the corresponding language-specific CatBoost model for classification. This method achieved a macro F1 score of 0.4604, 0.3743, and 0.3324 for the three tasks in order.

4.9 NLP_Goats

The team **"NLP_Goats"** utilized a TF-IDF-based approach for text classification, employing logistic regression to predict class labels. Text preprocessing included tokenization, stopword removal, and bigram generation, followed by TF-IDF vectorization to convert text into numerical features. To address class imbalance, oversampling techniques were applied before training the Logistic Regression model, with performance evaluated using accuracy and other classification metrics. This approach achieved a macro F1 score of 0.4105, 0.4049, and 0.2991 for Task 1, Task 2, and Task 3, respectively.

4.10 SSN_IT_SPEECH

The team **"SSN_IT_SPEECH"** employed a multimodal deep learning approach to detect hate speech by combining features from both audio and text data. Audio features are extracted using MFCC, which captures acoustic characteristics, while text features are derived using TF-IDF to analyze linguistic content. The extracted features are processed by a neural network: audio features pass through dense layers, while text features are handled by LSTM layers, which are well-suited for sequential data. This method leverages both the acoustic and textual properties of speech to achieve robust and nuanced hate speech detection. This approach achieved a macro F1 score of 0.3726, 0.2099, and 0.1631 for Task 1, Task 2, and Task 3, respectively.

4.11 SSN_MMHS

The team **"SSN_MMHS"** employed a multimodal framework for hate speech detection using two encoder-decoder transformer-based pipelines, each incorporating LSTM layers for sequential modeling. The key idea is to enable cross-modality learning by reversing the input modalities across the pipelines. In Pipeline 1, the encoder processes speech features (MFCCs), while the decoder processes text embeddings. Conversely, in Pipeline

2, the encoder processes text embeddings and the decoder processes speech features, fostering better interaction between modalities. The outputs from both pipelines are concatenated to form a unified representation, which is passed through a linear layer followed by softmax for classification. This approach achieved a macro F1 score of 0.348 and 0.1567 for Task 1 and Task 3, respectively.

4.12 The Deathly Hallows

The team **"The Deathly Hallows"** implemented a multimodal approach for classification, combining audio and text features. Audio data was augmented with techniques like noise addition, time-stretching, and pitch-shifting, and MFCCs were extracted for CNN-based classification. Text features were processed using the "*xlm-roberta-large*" model to generate embeddings, followed by an FNN for classification. Both pipelines employed robust architectures with Dropout, BatchNormalization, and Adam optimizer, ensuring accurate and generalized predictions. This approach achieved a macro F1 score of 0.3016, 0.6438, and 0.1559 for Task 1, Task 2, and Task 3, respectively.

4.13 BrightRed

The team **"BrightRed"** preprocessed text and audio data for all three languages and evaluated three models: Random Forest, LSTM, and CNN. Among these, the Random Forest model achieved the highest accuracy. This approach achieved a macro F1 score of 0.2782, 0.3018, and 0.251 for Task 1, Task 2, and Task 3, respectively.

4.14 cantnlp

The team **"cantnlp"** trained the multimodal hate speech classification model using logistic regression by transforming the audio files as melSpectrogram, which implicitly encodes linguistic information as acoustic features. They compared the performance with the text data across multiple statistical language models such as Naive Bayes Classifier for Multinomial Models, Linear Support Vector Machine, Logistic Regression, and Random Forest Classifier. This approach achieved a macro F1 score of 0.273, 0.3186, and 0.1774 for Task 1, Task 2, and Task 3, respectively.

4.15 Team ML_Forge

The team **"Team ML_Forge"** implemented a multimodal training approach, combining text and audio features. Text data was upsampled using back-translation, while audio data was processed at a 16 kHz sampling rate and augmented with variations in sound, pitch, volume, and time-stretching. Missing audio files in the Tamil and Telugu datasets were identified and addressed. Text features were extracted using the mBERT model, and audio features were processed with the wav2vec model, supplemented by MFCC features. The features from both modalities were concatenated and passed through a fully connected layer to generate final predictions. This approach achieved a macro F1 score of 0.2005, 0.1346, and 0.1465 for Task 1, Task 2, and Task 3, respectively.

4.16 KEC-Elite-Analysts

The team **"KEC-Elite-Analysts"** employed a combination of machine learning classifiers, including Random Forest, Support Vector Machine, Naive Bayes, and XGBoost. These models were used to capture diverse patterns from textual and multimodal inputs, leveraging their strengths for effective hate speech classification in underrepresented languages. This approach achieved a macro F1 score of 0.0812, 0.4281, and 0.1326 for Task 1, Task 2, and Task 3, respectively.

4.17 deanhthin

The team **"deanhthin"** utilized an LSTM model to extract text features and a CNN integrated with log-mel spectrograms to extract audio features. These features were then fused using the Tensor Fusion method. This approach achieved a macro F1 score of 0.0758 and 0.0592 for Task 1 and Task 2, respectively.

4.18 DLRG

The team **"DLRG"** used the pre-trained model "*ai4bharat/indic-bert*" for text classification and "*vasista22/whisper-tamil-medium*" for transcription of audio. Precisely, the audio was converted to text using the Whisper model, and the obtained text were classified using the Indic-BERT model. This approach achieved a macro F1 score of 0.2542 for Tamil.

4.19 SemanticCuetSync_Telugu

The team **"SemanticCuetSync_Telugu"** used "*openai/whisper-small*" for audio feature extraction and "*l3cube-pune/telugu-bert-scratch*" for textual feature extraction. The features were combined using a gated fusion approach to perform hate speech classification in Telugu. The method achieved a macro F1 score of 0.3514.

The majority of the submissions to this shared task centered around transformer-based models and multimodal frameworks. Leading teams such as SSNTrio and lowes leveraged fine-tuned BERT models augmented with speech features such as spectrograms and MFCC. Teams widely used late fusion techniques and attention mechanisms to fuse the features of text and speech data. For instance, byteSizedLLM combined XLM-RoBERTa for text with wav2vec2 for speech, while KEC_Tech_Titans integrated BERT variants with CNNs/BiLSTMs and graph networks. The efficacy of oversampling algorithms such as SMOTE was integrated into the models by some teams to address the class imbalance problem present in the data. Speech-to-text pipelines using Whisper models (DLRG, SemanticCnetSync_Telugu) and acoustic feature extraction (MFCCs, spectral contrast) paired with CNNs/LSTMs (The Deathly Hallows, zerowatts) highlighted the diversity in audio processing. The top-performing approaches showed the efficacy of fine-tuned transformers and multimodal integration, achieving superior macro F1 scores in Malayalam and Tamil, while Telugu posed greater challenges, with lower overall performance. Overall, the submissions reflected a blend of advanced deep learning architectures, traditional NLP techniques, and innovative multimodal strategies tailored to low-resource language contexts.

5 Conclusion

The shared task on Multimodal Hate Speech Detection in Dravidian languages at DravidianLangTech@NAACL 2025 is a platform to address the research in detecting hate speech in low-resource languages such as Malayalam, Tamil, and Telugu. The task highlighted the effectiveness of transformer-based models, particularly fine-tuned language-specific BERT models, and multimodal approaches that integrate both textual and acoustic features. The participation of different teams showcased various methodologies, from advanced deep learning architectures to traditional machine learning techniques, all aimed at addressing the complexities of hate speech detection in Malayalam, Tamil, and Telugu. The creation of a comprehensive, multiclass, multimodal dataset further enriches the resources available for future research. The results underscore the potential of combining textual and vocal features for robust hate speech detection, paving the way for more inclusive and accurate models in the fight against online hate speech.

Acknowledgments

This work was conducted with the financial support from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2(Insight_2).

References

Abhishek Anilkumar, Jyothish Lal G, B Premjith, and Bharathi Raja Chakravarthi. 2024. DravLangGuard: A Multimodal Approach for Hate Speech Detection in Dravidian Social Media. In *Speech and Language Technologies for Low-Resource Languages (SPELLL)*, Communications in Computer and Information Science.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep Learning for Hate Speech Detection in Tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 759–760.

Anat Ben-David and Ariadna Matamoros Fernández. 2016. Hate Speech and Covert Discrimination on Social Media: Monitoring the Facebook Pages of Extreme-Right Political Parties in Spain. *International Journal of Communication*, 10:27.

Shraddha Chauhan and Abhinav Kumar. 2025. MNLP@DravidianLangTech 2025: A Deep Multimodal Neural Network for Hate Speech Detection in Dravidian Languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Tashvik Dhamija, Anjum, and Rahul Katarya. 2021. Comparative Analysis of Machine Learning and Deep Learning Algorithms for Detection of Online Hate Speech. In *Advances in Mechanical Engineering: Select Proceedings of CAMSE 2020*, pages 509–520. Springer.

Adnan Faisal, Shiti Chowdhury, Sajib Bhattacharjee, Udoy Das, Samia Rahman, Momtazul Arefin Labib, and Hasan Murad. 2025. Team ML_Forge@DravidianLangTech 2025: Multimodal Hate Speech Detection in Dravidian Languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Md Sajjad Hossain, Symom Hossain Shohan, Ashraful Islam Paran, Jawad Hossain, and Mohammed Moshiul Hoque. 2025. SemanticCuet-Sync@DravidianLangTech 2025: Multimodal Fusion for Hate Speech Detection- A Transformer Based Approach with Cross-Modal Attention. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Bhuvana J, Mirnalinee T T, Rohan R, Diya Seshan, and Avaneesh Koushik. 2025. SSNTrio @ DravidianLangTech 2025: Hybrid Approach for Hate Speech Detection in Dravidian Languages with Text and Audio Modalities. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Ramchandra Joshi, Rushabh Karnavat, Kaustubh Jirapure, and Ravirai Joshi. 2021. Evaluation of Deep Learning Models for Hostility Detection in Hindi Text. In *2021 6th International conference for convergence in technology (I2CT)*, pages 1–5. IEEE.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate Speech Detection: Challenges and Solutions. *PloS one*, 14(8):e0221152.

Durga Prasad Manukonda, Rohith Gowtham Kodali, and Daniel Iglesias. 2025. byteSizedLLM@DravidianLangTech 2025: Multimodal Hate Speech Detection in Malayalam Using Attention-Driven BiLSTM, Malayalam-Topic-BERT, and Fine-Tuned Wav2Vec 2.0. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Vandan Mujadia, Pruthwik Mishra, and Dipti Misra Sharma. 2019. IIIT-Hyderabad at HASOC 2019: Hate Speech Detection. In *FIRE (Working Notes)*, pages 271–278.

Jahnavi Murali and Rajalakshmi Sivanaiah. 2025. SSN_MMHS@DravidianLangTech 2025: A Dual Transformer Approach for Multimodal Hate Speech Detection in Dravidian Languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Ks Paval, Vishnu Radhakrishnan, Km Krishnan, G Jyothish Lal, and B Premjith. 2024. Multimodal Fusion for Abusive Speech Detection Using Liquid Neural Networks and Convolution Neural Network. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–7. IEEE.

Ratnavel Rajalakshmi, R Ramesh Kannan, Meetesh Saini, and Bitan Mallik. 2025. DLRG@DravidianLangTech 2025: Multimodal Hate Speech Detection in Dravidian Languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059, Barcelona (online). International Committee for Computational Linguistics.

Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pages 1–10.

Kogilavani Shanmugavadivel, Malliga Subramanian, Naveenram CE, Vishal RS, and Srinesh S. 2025a. KEC_AI_ZEROWATTS@DravidianLangTech 2025: Multimodal Hate Speech Detection in Dravidian languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Kogilavani Shanmugavadivel, Malliga Subramanian, Vasantharan K, Prethish G A, and Santhosh S. 2025b. The_Deathly_Hallows@DravidianLangTech 2025: Multimodal Hate Speech Detection in Dravidian Languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Kogilavani Shanmugavadivel, Malliga Subramanian, Nishdharani P, Santhiya E, and Yaswanth Raj E. 2025c. KEC_AI_BRIGHTRED@DravidianLangTech 2025: Multimodal Hate Speech Detection in Dravidian languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Kogilavani Shanmugavadivel, Malliga Subramanian, ShahidKhan S, Shri Sashmitha S, and Yashica S. 2025d. KEC_AI_GRYFFINDOR@DravidianLangTech 2025: Multimodal Hate Speech Detection in Dravidian languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Abhishek Velankar, Hrushikesh Patil, Amol Gore, Shubham Salunke, and Raviraj Joshi. 2021. Hate and Offensive Speech Detection in Hindi and Marathi. *arXiv preprint arXiv:2110.12200*.

Sidney Wong and Andrew Li. 2025. cantnlp@DravidianLangTech2025: A Bag-of-Sounds Approach to Multimodal Hate Speech

Detection. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.