cantnlp@DravidianLangTech 2025: A Bag-of-Sounds Approach to Multimodal Hate Speech Detection

Sidney Wong
Geospatial Research Institute
University of Canterbury
sidney.wong@pg.canterbury.ac.nz

Andrew Li
Lake Washington School District
landrewi@hotmail.com

Abstract

This paper presents the systems and results for the Multimodal Social Media Data Analysis in Dravidian Languages (MSMDA-DL) shared task at the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages (DravidianLangTech-2025). We took a 'bag-of-sounds' approach by training our hate speech detection system on the speech (audio) data using transformed Mel spectrogram measures. While our candidate model performed poorly on the test set, our approach offered promising results during training and development for Malayalam and Tamil. With sufficient and well-balanced training data, our results show that it is feasible to use both text and speech (audio) data in the development of multimodal hate speech detection systems.

1 Introduction

There has been increased recognition within the research field that forms of hate speech on social media are not restricted to written modalities of language, but also spoken (Chhabra and Vishwakarma, 2023) and non-linguistic (i.e., memes) modalities as well (Kiela et al., 2020). As part of Multimodal Social Media Data Analysis in Dravidian Languages shared task (Lal G et al., 2025), we propose taking a 'bag-of-sounds' approach - analogous to the bag-of-words models - to train our automatic hate speech detection system on the speech (audio) data. We do this by transforming the speech (audio) data into Mel spectrogram measures and training our classification model on the outputs.

2 Related Works

The earliest automatic hate speech detection systems relied on different linguistic features such as lexical and syntactic representations (Chen et al., 2012), template-based and parts-of-speech (POS) tagging (Warner and Hirschberg, 2012), topic-modelling (Xiang et al., 2012), or a combination of lexical, POS, character bigram and term frequency-inverse document frequency (Tf-idf) representations (Dinakar et al., 2012). With a focus on hate speech in English, the model performance of these early systems yielded moderate results with limited applications to other language conditions (Jahan and Oussalah, 2023).

The introduction of transformer-based Large Language Models (LLMs), such as BERT (Devlin et al., 2019), saw an increase of word embedding feature representations jointly with neural network models in the development of hate speech detection systems (Jahan and Oussalah, 2023). Hate speech systems are now treated as a text classification task following a standardised pipeline including data set collection and labelling, feature extraction, model learning and development, and evaluation on a multiclass or binary output (Rawat et al., 2024). Both statistical language models and LLMs are used in the development of contemporary state-of-the-art hate speech detection systems.

As with other strands of Computational Linguistics and Natural Language Processing (NLP) for social impact (Hovy and Spruit, 2016), there has been a long standing tradition of shared tasks detecting hate speech and offensive language in Indo-Aryan and Dravidian languages (Chakravarthi et al., 2021; Chakravarthi et al., 2022). The best performing models in Chakravarthi et al. (2024a) were developed using open-source multilingual transformer-based LLMs (Conneau et al., 2020; Khanuja et al., 2021). In addition to testing the usability of transformed-based LLMs in non-English conditions, these systems interrogate the efficacy of text classification in code-switching (Yasaswini et al., 2021) and script-switching (Wong and Durward, 2024) phenomena.

While previous shared tasks have focused solely on written expressions of hate speech, the first multimodal social media data analysis in Dravidian languages (MSMDA-DL) was organised by

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 621–629
May 3, 2025 ©2025 Association for Computational Linguistics

Chakravarthi et al. (2024b) with written and spoken social media language data from YouTube. The shared task provided training data with utterances in two Dravidian languages - Malayalam and Tamil - and annotated for hate speech and abusive language in YouTube videos. Only two systems were submitted as part of Chakravarthi et al. (2024b): Rahman et al. (2024) and S et al. (2024). Of interest to the current paper, Rahman et al. (2024) extracted Mel-frequency spectrogram and Mel-frequency Cepstral Coefficients (MFCCs) as acoustic features which they incorporated in their ConvLSTM.

A spectrogram is a visual representation of the acoustic frequency - or the number of vibrations in a sound wave per second - in a speech (audio) signal. The Mel-scale spectrogram, also known as Mel-frequency spectrograms or simply Mel spectrograms, is a transformation of linear machine-readable frequency measures of a spectrogram to a non-linear Mel scale which is the perceptual scale of pitch by human listeners (Stevens et al., 1937). Mel spectrograms and MFCCs (Davis and Mermelstein, 1980) have been widely used in automatic speaker recognition systems and subjective tasks such as speaker emotion recognition (Zhou et al., 2019); most recently, these acoustic measures have been included in various forms of NLP classification tasks (Arróniz and Kübler, 2023).

Rahman et al. (2024) took a Convolutional Neural Network (CNN) with Long-short term memory (LSTM), or ConvLSTM, and a hybrid 3D-CNN with LSTM approach in the development of their multimodal hate speech detection system incorporating visual, audio, and text representations. Although the shared task included training data for three Dravidian languages, only one system was designed for Tamil. During the model development phase, the system achieved a macro average F_1-score of 0.71 for Tamil. The system ranked first for Tamil in the shared task with a macro average F_1-score of 0.7143 in the test set. S et al. (2024) did not incorporate the speech (audio) components in the development of their detection system.

3 Data

The training data contained both text and speech (audio) data in three Dravidian languages: Malayalam, Tamil, and Telugu (Sreelakshmi et al., 2024). The target class labels were organised hierarchically including a binary classification with labels

Table 1: Binary Class Labels

Class	Malayalam	Tamil	Telugu
H	477	227	358
N	406	287	198

Table 2: Multiclass Class Labels

Class	Malayalam	Tamil	Telugu
C	186	65	122
N	406	287	198
P	118	33	58
R	91	61	72
G	82	68	106

'Hate' (H) and 'No Hate' (N), and a multiclass classification with five categories included Caste-related hate-speech (C), and Offensive (O), Racist (R), Sexist (S) language, and one residual non-hate speech category (N). The distribution of the target class labels in the training data for the binary classification is presented in Table 1 and for the multiclass classification in Table 2. There is significant class imbalance between target class labels between language conditions and within the training data. In addition to the class labels, the text and speech (audio) observations were identified by subject, binary gender of the speakers, source of utterance, and utterance number. We split the training data set into training and validation set, where the training set with the target labels was used to train the models and the validation set was reserved for performance evaluation.

4 Methodology

The primary purpose of the Multimodal Social Media Data Analysis in Dravidian Languages (MSMDA-DL) shared task was to develop a hate-speech detection system that can analyse the text and speech components and predict the respective labels for three Dravidian languages: Malayalam, Tamil, and Telugu. Therefore, we approached this shared task as a classification problem. We trained a suite of candidate multimodal hate speech detection system using a statistical language model approach. While transformer-based LLMs are the state-of-the-art models in hate speech detection (Chakravarthi et al., 2024a), there is limited published research testing the use of LLMs in signal processing (i.e., audio data) as existing LLMs, such as BERT (Devlin et al., 2019), are trained on word embeddings from written language data (Verma

Table 3: Macro average F_1-score on validation training data (Binary).

	NB		SVM		LR		RF	
	TEXT	SPEECH	TEXT	SPEECH	TEXT	SPEECH	TEXT	SPEECH
Malayalam	0.87	0.70	0.85	0.64	0.84	**0.90**	0.84	0.91
Tamil	0.72	0.57	**0.79**	0.64	0.76	0.67	0.72	0.75
Telugu	0.64	0.60	0.66	0.65	**0.70**	0.67	0.72	0.72

Table 4: Macro average F_1-score on validation training data (Multiclass).

	NB		SVM		LR		RF	
	TEXT	SPEECH	TEXT	SPEECH	TEXT	SPEECH	TEXT	SPEECH
Malayalam	0.32	0.34	**0.54**	0.45	**0.54**	**0.54**	0.33	0.48
Tamil	0.14	0.21	**0.46**	0.15	0.38	0.39	0.23	0.32
Telugu	0.24	0.28	**0.55**	0.33	0.53	0.38	0.41	0.34

and Pilanci, 2024). This means we cannot directly compare the model performance of text-trained or speech-trained detection systems for the purposes of this shared task. We evaluated the performance of each candidate system according to the macro average F_1-score on the training validation data before selecting and submitting the best performing candidate model. The associated code notebook and submission data can be found at the associated GitHub repository[1].

4.1 Data Preprocessing and Feature Engineering

Multimodal hate speech data is a feature of the current shared task. Prior to the model training process, we carried out the following data preprocessing and feature engineering procedures for the two modalities:

Text: The text data was supplied in the respective Indic (Brahmic) orthographies of each language condition. We applied minimal data preprocessing on the text data as we wanted to preserve the linguistic features between the text and speech (audio) data. `CountVectorizer` and `TfidfTransformer` from `sklearn.feature_extraction.text` were employed to transform text data into numerical feature vectors suitable for machine learning models.

Speech (Audio): Mel spectrograms were computed for the audio files and converted into decibel units. To ensure uniformity across inputs, all spectrograms were padded to the same shape and reshaped into flat 2D arrays for compatibility with machine learning algorithms. Additional data normalization was implemented to meet the requirements of the Multinomial Naïve Bayes algorithm, one of the machine learning algorithms analysed in this research. This process transforms the feature matrix to ensure all feature values are scaled to the range [0,1].

[1]https://github.com/sidneygjwong/cantnlp-dravidianlangtech2025

4.2 Model Training, Evaluation, and Selection Criteria

The training data was split into training and validation sets with a train:test split of 75:25. Four classification models were trained for both binary and multiclass classification tasks across all three Dravidian languages. The binary classification task served as a benchmark. The statistical methods used were as follows: Multinomial Naïve Bayes (NB), Linear Support Vector Machine (SVM), Logistic Regression Classifier (LR), and Random Forest Classifier (RF). There were 48 candidate models in total according to the following rubric: two classifications x three language conditions x two modalities x four statistical methods.

The performance of each candidate model was evaluated by macro average F_1-score. The model performance as measured by macro average F_1-score for the binary classification models are presented in Table 3 and for the multiclass classification models in Table 4. The best performing model for each language condition (row) is highlighted in **bold**. For completeness and benchmarking purposes, we have also included the model performance based on F_1-score for binary and multiclass classification models in the Appendix as shown in Tables 7 to 12.

For model evaluation, we used a macro average

Table 5: macro average F_1-score from test evaluation and rank by language.

Language	F_1-score	Rank
Malayalam	0.273	14
Tamil	0.3186	9
Telugu	0.1774	12

Table 6: Relative proportion of test ($n = 10$) to train data as a percentage (%) per class label.

Class	Malayalam	Tamil	Telugu
C	5.4	15.4	8.2
N	2.5	3.5	5.1
P	8.5	30.3	17.2
R	11.0	16.4	13.9
G	12.2	14.7	9.4

F_1 score as the primary metric. Overall, Logistic Regression (LR) achieved the highest macro average F_1-score and performed the best among all four algorithms for speech (audio) data in the binary classification candidate models as shown in Table 3. In contrast, Linear SVM (SVM) had better performance with the multiclass classification candidate models as shown in Table 4. We notice a significant drop in performance between the binary classification to the multiclass classification models with the maximum macro average F_1-score of 0.90 lowering to 0.54.

Even though the text data trained Linear SVM (SVM) models largely outperformed the speech (audio) trained candidate models, we opted for the best performing multiclass speech (audio) data trained models. We justify this decision as the purpose of the shared task was to incorporate multimodal language data and not just one modality. Where text data only encodes linguistic information, we argue speech (audio) data implicitly encodes both linguistic and paralinguistic features of hate speech. Furthermore, the performance of the text trained models Linear SVM (SVM) models only performed marginally better than our optimal model - the speech (audio) trained logistic regression (LR) models.

5 Results

The macro average F_1-scores of our candidate model on the test set are presented in Table 5. All three models performed poorly on the test set with a macro average F_1-score below chance. In contrast, the best performing model in Malayalam and Tamil was by Team `SSNTrio` who yielded a macro average F_1-score of 0.7511 for Malayalam and 0.7332 for Tamil. The best performing model in Telugu was by Team `lowes` had a macro average F_1-score of 0.3817.

6 Discussion

Based on the evaluation metrics alone, our optimal method performed poorly across all three language conditions with a macro average F_1-score below chance. With reference to Table 4 and Table 5, we see a significant drop in performance between the validation evaluation and test evaluation. This drop is particularly stark in Malayalam where the macro average F_1-score went from 0.54 to 0.273, and for Telugu from 0.38 to 0.1774. Malayalam had a median macro average F_1-score of 0.41 and average of 0.38; for Tamil a median of 0.32 and average of 0.34; and for Telugu a median of 0.24 and an average of 0.23.

While the median and mean scores suggest an improvement from Chakravarthi et al. (2024b) across the board, we argue the consistently poor performance may indicate there are underlying issues with the training data. One possible explanation for the poor model performance is the class imbalance observed in Table 2 where we see not only difference in utterances between language conditions, but also between classes especially in the minority classes. When we consider the relative proportion of utterances in the test evaluation set as shown in Table 6, some utterances in the minority classes are over represented while utterances in the majority classes are under represented.

As we refer back to the binary classification models as shown in Table 3 (which were not part of the shared task), we can see that the models performed well not only across all language conditions, but also across the different statistical methods. Even though the models performed poorly on the test set which suggests some modifications are needed to our pipeline, some of the poor model performance can be attributed to the class imbalance in the training and test data. It is possible the decline in performance from the validation to test data suggests possible over-fitting to the training set or other dataset related biases not accounted for in the current model development pipeline which will be worthy of further investigation.

Possible improvements to our existing model may include further hyper-parameter tuning such as employing optimisation techniques such as GridSearchCV or RandomisedSearchCV to fine-tune parameters for the text classifiers used in our study such as Random Forest, SVM, and Logisitic Regression. We could explore more advanced boosting algorithms like XGBoost, CatBoost, or LightGBM which may improve classification performance. Alternatively, we could look into comparing other speech feature representations such as Mel-Frequency Cepstral Coefficients (MFCCs) in addition to Mel spectrograms which have also been effective in speech-based classification tasks.

The current study provides a foundation for future work in the development of multimodal hate speech detection systems. Despite the lower than expected performance of our proposed approach when compared to other teams in the shared task, we demonstrated in this paper that speech data carries valuable extra-linguistic information for hate speech detection. We argue that further improvements in training data representation and model architecture (i.e., with state-of-the-art methodologies) may yield better performance.

7 Conclusion

While our candidate model performed poorly on the test set, our 'bag-of-sounds' approach offered promising results during training and development for Malayalam and Tamil. With sufficient and well-balanced training data, our results show that it is feasible to use both text and speech (audio) data in the development of multimodal hate speech detection systems. It is important to note that our current study intentionally avoided state-of-the-art deep learning or large language models to avoid overloading our existing approach with speculative enhancements from deep learning and transformer-based language models; however, we will look to incorporate more sophisticated models, such as ELMo (Peters et al., 2018), in future studies to determine the performance of our proposed approach alongside state-of-the-art methodologies.

Acknowledgements

We would like to thank the four anonymous peer reviewers and the organisers of the Fifth Workshop on Speech and Language Technologies for Dravidian Languages (DravidianLangTech-2025) co-located at the North American Chapter of the Association for Computational Linguistics in Albuquerque, New Mexico. We would also like to acknowledge Fulbright New Zealand | Te Tūāpapa Mātauranga o Aotearoa me Amerika and their partnership with the Ministry of Business, Innovation, and Employment | Hīkina Whakatutuki for their support through the Fulbright New Zealand Science and Innovation Graduate Award.

References

Santiago Arróniz and Sandra Kübler. 2023. Was That a Question? Automatic Classification of Discourse Meaning in Spanish. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 132–142, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Rajat Kumar Behera, Pradip Kumar Bala, Nripendra P. Rana, and Zahir Irani. 2023. Responsible natural language processing: A principlist framework for social benefits. *Technological Forecasting and Social Change*, 188:122306.

Bharathi Raja Chakravarthi, Prasanna Kumaresan, Ruba Priyadharshini, Paul Buitelaar, Asha Hegde, Hosahalli Shashirekha, Saranya Rajiakodi, Miguel Ángel García, Salud María Jiménez-Zafra, José García-Díaz, Rafael Valencia-García, Kishore Ponnusamy, Poorvi Shetty, and Daniel García-Baena. 2024a. Overview of Third Shared Task on Homophobia and Transphobia Detection in Social Media Comments. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 124–132, St. Julian's, Malta. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Thenmozhi Durairaj, John McCrae, Paul Buitelaar, Prasanna Kumaresan, and Rahul Ponnusamy. 2022. Overview of The Shared Task on Homophobia and Transphobia Detection in Social Media Comments. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 369–377, Dublin, Ireland. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Navya Jose, Anand Kumar M, Thomas Mandl, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Hariharan R L, John P. McCrae, and Elizabeth Sherly. 2021. Findings of the Shared Task on Offensive Language Identification in Tamil, Malayalam, and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 133–145, Kyiv. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Anand Kumar Madasamy, Sajeetha Thavareesan, Elizabeth Sherly, Rajeswari Nadarajan, and Manikandan Ravikiran. 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, St. Julian's, Malta. Association for Computational Linguistics.

Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting Offensive Language in Social Media to Protect Adolescent Online Safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 71–80.

Anusha Chhabra and Dinesh Kumar Vishwakarma. 2023. A literature survey on multimodal and multilingual automatic hate speech identification. *Multimedia Systems*, 29(3):1203–1230.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial Bias in Hate Speech and Abusive Language Detection Datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

S. Davis and P. Mermelstein. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366. Conference Name: IEEE Transactions on Acoustics, Speech, and Signal Processing.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. 2012. Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying. *ACM Trans. Interact. Intell. Syst.*, 2(3):18:1–18:30.

Dirk Hovy and Shannon L. Spruit. 2016. The Social Impact of Natural Language Processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Md Saroar Jahan and Mourad Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546:126232.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. MuRIL: Multilingual Representations for Indian Languages. *arXiv preprint*. ArXiv:2103.10730 [cs].

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. In *Advances in Neural Information Processing Systems*, volume 33, pages 2611–2624. Curran Associates, Inc.

Salla-Maaria Laaksonen, Jesse Haapoja, Teemu Kinnunen, Matti Nelimarkka, and Reeta Pöyhtäri. 2020. The Datafication of Hate: Expectations and Challenges in Automated Hate Speech Monitoring. *Frontiers in Big Data*, 3.

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*, Albuquerque, NM. Association for Computational Linguistics.

Nayeon Lee, Chani Jung, and Alice Oh. 2023. Hate Speech Classifiers are Culturally Insensitive. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 35–46, Dubrovnik, Croatia. Association for Computational Linguistics.

Carla Parra Escartín, Wessel Reijers, Teresa Lynn, Joss Moorkens, Andy Way, and Chao-Hong Liu. 2017. Ethical Considerations in NLP Shared Tasks. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 66–73, Valencia, Spain. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Md. Rahman, Abu Raihan, Tanzim Rahman, Shawly Ahsan, Jawad Hossain, Avishek Das, and Mohammed Moshiul Hoque. 2024. Binary_beasts@DravidianLangTech-EACL 2024: Multimodal Abusive Language Detection in Tamil based on Integrated Approach of Machine Learning and Deep Learning Techniques. In *Proceedings of the Fourth Workshop on Speech, Vision, and*

Language Technologies for Dravidian Languages, pages 212–217, St. Julian's, Malta. Association for Computational Linguistics.

Anchal Rawat, Santosh Kumar, and Surender Singh Samant. 2024. Hate speech detection in social media: Techniques, recent trends, and future challenges. *WIREs Computational Statistics*, 16(2):e1648.

Anierudh S, Abhishek R, Ashwin Sundar, Amrit Krishnan, and Bharathi B. 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, St. Julian's, Malta. Association for Computational Linguistics.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

K. Sreelakshmi, B. Premjith, Bharathi Raja Chakravarthi, and K. P. Soman. 2024. Detection of Hate Speech and Offensive Language CodeMix Text in Dravidian Languages Using Cost-Sensitive Learning Approach. *IEEE Access*, 12:20064–20090. Conference Name: IEEE Access.

S. S. Stevens, J. Volkmann, and E. B. Newman. 1937. A Scale for the Measurement of the Psychological Magnitude Pitch. *The Journal of the Acoustical Society of America*, 8(3):185–190.

Prateek Verma and Mert Pilanci. 2024. Towards Signal Processing In Large Language Models. *arXiv preprint*. ArXiv:2406.10254 [cs] version: 1.

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*, LSM '12, pages 19–26, USA. Association for Computational Linguistics.

Sidney Wong. 2024. Sociocultural Considerations in Monitoring Anti-LGBTQ+ Content on Social Media. In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 84–97, Bangkok, Thailand. Association for Computational Linguistics.

Sidney Wong and Matthew Durward. 2024. cantnlp@LT-EDI-2024: Automatic Detection of Anti-LGBTQ+ Hate Speech in Under-resourced Languages. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 177–183, St. Julian's, Malta. Association for Computational Linguistics.

Guang Xiang, Bin Fan, Ling Wang, Jason Hong, and Carolyn Rose. 2012. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, CIKM '12, pages 1980–1984, New York, NY, USA. Association for Computing Machinery.

Konthala Yasaswini, Karthik Puranik, Adeep Hande, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. IIITT@DravidianLangTech-EACL2021: Transfer Learning for Offensive Language Detection in Dravidian Languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 187–194, Kyiv. Association for Computational Linguistics.

Hengshun Zhou, Debin Meng, Yuanyuan Zhang, Xiaojiang Peng, Jun Du, Kai Wang, and Yu Qiao. 2019. Exploring Emotion Features and Fusion Strategies for Audio-Video Emotion Recognition. In *2019 International Conference on Multimodal Interaction*, ICMI '19, pages 562–566, New York, NY, USA. Association for Computing Machinery.

A Limitations

While we saw promising results in the 'bag-of-sounds' approach we proposed in the development of our hate speech detection system; there are two main limitations we wish to address in our system. The first being the use of statistical language models; and secondly, the lack of sociolinguistic input in the development of our model.

Firstly, our candidate does not use state-of-the-art modelling in the development of our system as we have not incorporated transformer-based LLMs. While we justified our reasons for excluding LLMs in Section 4 in order to maintain comparability between the two modalities, we need to determine how we might incorporate the use of LLMs in our system as existing state-of-the-art models rely on multilingual LLMs (Chakravarthi et al., 2024a). With the development of LLMs for speech (audio) signal processing (Verma and Pilanci, 2024), it may be possible for us to replicate our analysis. As we will discuss in the Ethics Statement, introducing LLMs may inadvertently introduce bias into our hate speech detection system.

This leads us to discuss the second limitation which is the lack of sociolinguistic input in the development of our system. As discussed in Section 3, each utterance was labelled for one demographic variable - the binary gender classification of the speaker. While we are aware that speech varies based on gender as a result of mechanical and social demographic differences, we have not incorporated in the development of our model. This means

Table 7: Model comparison metrics by F_1-score per class (binary) in Malayalam

	NB		SVM		LR		RF	
Class Label	TEXT	SPEECH	TEXT	SPEECH	TEXT	SPEECH	TEXT	SPEECH
H	0.89	0.71	0.87	0.78	0.86	0.92	0.87	0.92
N	0.86	0.69	0.84	0.49	0.82	0.89	0.82	0.90

Table 8: Model comparison metrics by F_1-score per class (multiclass) in Malayalam

	NB		SVM		LR		RF	
Class Label	TEXT	SPEECH	TEXT	SPEECH	TEXT	SPEECH	TEXT	SPEECH
C	0.72	0.44	0.80	0.73	0.68	0.70	0.70	0.63
N	0.00	0.20	0.21	0.08	0.35	0.31	0.00	0.15
P	0.73	0.71	0.85	0.89	0.81	0.90	0.73	0.84
R	0.14	0.19	0.38	0.55	0.42	0.58	0.13	0.45
G	0.00	0.16	0.44	0.00	0.46	0.21	0.08	0.35

Table 9: Model comparison metrics by F_1-score per class (binary) in Tamil

	NB		SVM		LR		RF	
Class Label	TEXT	SPEECH	TEXT	SPEECH	TEXT	SPEECH	TEXT	SPEECH
H	0.64	0.49	0.77	0.63	0.73	0.61	0.66	0.70
N	0.80	0.64	0.81	0.65	0.78	0.73	0.77	0.79

Table 10: Model comparison metrics by F_1-score per class (multiclass) in Tamil

	NB		SVM		LR		RF	
Class Label	TEXT	SPEECH	TEXT	SPEECH	TEXT	SPEECH	TEXT	SPEECH
C	0.00	0.00	0.16	0.00	0.07	0.08	0.11	0.10
N	0.00	0.15	0.47	0.00	0.49	0.35	0.32	0.21
P	0.70	0.69	0.81	0.74	0.76	0.72	0.72	0.79
R	0.00	0.20	0.33	0.00	0.17	0.43	0.00	0.22
G	0.00	0.00	0.53	0.00	0.39	0.36	0.00	0.26

Table 11: Model comparison metrics by F_1-score per class (binary) in Telugu

	NB		SVM		LR		RF	
Class Label	TEXT	SPEECH	TEXT	SPEECH	TEXT	SPEECH	TEXT	SPEECH
H	0.84	0.74	0.78	0.73	0.80	0.81	0.85	0.81
N	0.44	0.46	0.54	0.57	0.60	0.52	0.58	0.52

Table 12: Model comparison metrics by F_1-score per class (multiclass) in Telugu

	NB		SVM		LR		RF	
Class Label	TEXT	SPEECH	TEXT	SPEECH	TEXT	SPEECH	TEXT	SPEECH
C	0.34	0.44	0.65	0.56	0.57	0.55	0.65	0.50
N	0.33	0.14	0.59	0.28	0.52	0.34	0.47	0.27
P	0.51	0.43	0.65	0.53	0.67	0.49	0.68	0.57
R	0.00	0.32	0.36	0.13	0.38	0.24	0.12	0.37
G	0.00	0.07	0.52	0.17	0.48	0.26	0.12	0.00

it is possible that there are unexplained predictors in the data that are unaccounted for namely acoustic differences between these social categories such as gender, culture, and possibly geographic dialect bias (Wong, 2024). Therefore, future work should involve more in-depth analysis on the speech (audio) data which may justify the need to further normalise or standardise the data.

B Ethics/Broader Impact Statement

Parra Escartín et al. (2017) argued that shared tasks play an important role in Computational Linguistics and Natural Language Processing (NLP) as it helps encourage a culture within the field to develop upon the state-of-the-art. With an increased recognition of automatic hate speech detection beyond just written text to other modalities of language (Chhabra and Vishwakarma, 2023), the current shared task plays an important role in how detection can be achieved in not only multimodal but also multilingual contexts.

In spite of these benefits, there are also ethical issues and negative effects of competition in NLP shared tasks such as secretive behaviour, overlooking the relevance of ethical concerns, unconscious overlooking of ethical concerns, redundancy and replicability in the field among other concerns (Parra Escartín et al., 2017). With the 'datafication' of hate speech an increasing issue within the field of hate speech detection (Laaksonen et al., 2020), we will consider the ethics and broader impacts of our proposed system within the context of the current shared task guided by the eight principals of *Responsible NLP* (Behera et al., 2023).

Principal 1: Well-being The current system contributes to our current understanding of multimodal automatic hate speech detection in three Dravidian languages. The current system was designed alongside junior researchers which supports development in working with low- and under-resourced language condition often overlooked in NLP research.

Principal 2: Human-Centred Values The current system does not include human subjects, external annotators, or additional data from external sources. However, this is also an area of improvement where the researchers can work alongside target communities - namely Malayalam, Tamil, and Telugu speakers - in the development of a system that is fit for purpose.

Principal 3: Fairness While we have avoided to the best of our ability to not perpetuate existing prejudice towards marginalised and vulnerable communities, we are aware that hate speech training datasets are sensitive to racial (Davidson et al., 2019; Sap et al., 2019) and sociocultural (Lee et al., 2023; Wong, 2024). Therefore, we propose that further work is needed to determine the presence of underlying biases within the training data and possible downstream impacts of these biases.

Principal 4: Privacy and Security In accordance with the terms and conditions of the shared task, the authors have not re-distributed the data and have only used the data for non-commercial and academic-research purposes. We have not used the data for surveillance, analyses, or research that isolates a group of individuals for unlawful or discriminatory purposes.

Principal 5: Reliability We have provided the model performance metrics which can be found throughout the paper and in the Appendix. We acknowledge there will be variances within the metrics due to the stochastic nature of statistical language models. There is limited risk to organisers, authors, or users who wish to reproduce our systems.

Principal 6: Transparency We have described our system to the best of our ability for other researchers to reproduce our system; however, we are limited to the metadata provided of the training data provided to us by the organisers of the shared task. We have not involved additional human subjects or external annotators.

Principal 7: Interrogation We encourage readers to refer to the other system description papers associated with this shared task.

Principal 8: Accountability We encourage readers to contact the authors to discuss the contents of this paper.