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

Billodal Roy[1], Pranav Gupta[1], Souvik Bhattacharyya[1], Niranjan Kumar M[1]
[1]Lowe's
Correspondence: {billodal.roy, pranav.gupta, souvik.bhattacharyya, niranjan.k.m}@lowes.com

Abstract

This paper describes our participation in the DravidianLangTech@NAACL 2025 shared task on hate speech detection in Dravidian languages. While the task provided both text transcripts and audio data, we demonstrate that competitive results can be achieved using text features alone. We employed fine-tuned Bidirectional Encoder Representations from Transformers (BERT) models from l3cube-pune for Malayalam, Tamil, and Telugu languages. Our system achieved notable results, securing first position for Telugu, and second position for Tamil and Malayalam tasks in the official leaderboard.

1 Introduction

The increasing volume of social media content in Dravidian languages has heightened the need for robust hate speech detection systems. The DravidianLangTech shared task at NAACL 2025 presented a multimodal challenge for hate speech detection in Malayalam, Tamil, and Telugu (Lal G et al., 2025; Premjith et al., 2024a,b; Sreelakshmi et al., 2024). These languages, with their rich morphological structures and distinct scripts, present unique challenges for automated content moderation. Our work demonstrates that while multimodal approaches are valuable, significant performance can be achieved through focused analysis of textual content alone. We utilized language-specific BERT (Devlin et al., 2019) models from l3cube-pune (Joshi, 2023), fine-tuned on the text transcripts from the task-specific dataset. This approach not only proved computationally efficient but also highly effective, suggesting that textual features capture substantial indicators of hate speech in these languages.[1]

[1]The code for this work is available at https://github.com/prannerta100/naacl2025-dravidianlangtech

2 Related Work

Research in hate speech and offensive language detection has evolved significantly, particularly for social media content. Initial approaches relied on traditional machine learning methods, utilizing handcrafted features such as n-grams and sentiment lexicons (Sreelakshmi et al., 2020). These methods, while foundational, faced limitations in capturing the contextual complexities of natural language, especially in code-mixed and low-resource scenarios.

The emergence of transformer architectures marked a significant advancement in this domain. The introduction of BERT (Devlin et al., 2019) enabled more sophisticated contextual representations, leading to substantial improvements in detection accuracy. Building on this foundation, Chakravarthi et al. (Chakravarthi et al., 2023) demonstrated enhanced performance by combining MPNet with convolutional neural networks for Dravidian languages, specifically addressing code-mixing challenges in Tamil, Malayalam, and Kannada. This work was complemented by Subramanian et al. (Subramanian et al., 2022), who focused on Tamil YouTube comments, highlighting the importance of handling class imbalance in social media content. Multilingual approaches have further advanced the field through innovative architectures. Hande et al. (Hande et al., 2022) explored multi-task learning with mBERT, simultaneously addressing sentiment analysis and offensive language detection. Roy et al. (Roy et al., 2022) proposed an ensemble framework that integrates multiple approaches, demonstrating the advantages of combining traditional and modern methodologies.

Recent work has increasingly focused on language-specific adaptations. Notable contributions include Pillai and Arun's (Pillai and Arun, 2024) investigation of feature fusion techniques for Malayalam, and IIITDWD-ShankarB's (Biradar

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 630–634

Category	Malayalam	Tamil	Telugu
Non-Hate	406	287	198
Gender	82	68	106
Political	118	33	58
Religious	91	61	72
Character	186	65	122
Total	883	514	556

Table 1: Distribution of instances across categories for each language in the dataset

and Saumya, 2022) application of mBERT to South Indian languages. Arunachalam et al. (Arunachalam and Maheswari, 2024) demonstrated the effectiveness of language-specific BERT models in achieving competitive performance using only textual features. Additional research by Sharma et al. (Sharma et al., 2023) on detecting specific forms of discriminatory content has further emphasized the importance of language-tailored approaches. This progression in the field reflects a clear shift from feature-engineered solutions to sophisticated transformer-based systems, better equipped to handle the nuances of code-mixed content and class imbalance in Dravidian language hate speech detection.

3 Dataset and Task Description

The DravidianLangTech shared task provided datasets for hate speech detection in three Dravidian languages: Malayalam, Tamil, and Telugu. Each dataset consists of text transcripts and audio recordings sourced from YouTube videos, categorized into hate and non-hate speech, with hate speech further subdivided into four categories.

3.1 Data Organization

The data follows a structured format with detailed file nomenclature containing speaker information, source identifiers, and classification labels. Each instance includes both audio recording and corresponding text transcript, though our approach utilizes only the text components.

3.2 Class Distribution

Table 1 shows the distribution of instances across different categories for each language.

3.3 Data Characteristics

A notable characteristic of the dataset is its class imbalance, with Non-Hate being the dominant category across all three languages. The distribution of hate speech subcategories varies significantly among languages, with Character Defamation being particularly prevalent in Malayalam and Telugu datasets. This imbalanced distribution presents a significant challenge for model training and necessitates careful consideration during the development of our classification approach.

4 Methodology

Our approach leverages language-specific BERT models fine-tuned for each Dravidian language, with a focus on optimizing for the inherent class imbalance in the dataset.

4.1 Model Architecture

We utilized pre-trained BERT models from l3cube-pune, specifically tailored for Dravidian languages. These models have demonstrated superior performance in capturing language-specific nuances compared to general multilingual models. The base architecture consists of the pre-trained BERT model with a classification head fine-tuned for our specific task.

Language	Base Model
Malayalam	l3cube-pune/malayalam-bert
Tamil	l3cube-pune/tamil-bert
Telugu	l3cube-pune/telugu-bert

Table 2: Language-specific BERT models

4.2 Implementation Details

The implementation utilized the Hugging Face Transformers library for model architecture and training. We maintained the original text without pre-processing, allowing the models to learn from the natural language patterns. The system was implemented using PyTorch, with training facilitated by the Transformers library's Trainer API.

Parameter	Value
Learning Rate	2e-5
Batch Size	8
Training Epochs	15-20
Label Smoothing	0.1
Weight Decay	0.005-0.01

Table 3: Training hyperparameters

4.3 Training Strategy

Our training approach evolved through systematic experimentation. Initially, we employed an 80-20 train-test split while maintaining class distribution. To address the class imbalance, we implemented label smoothing and weight decay regularization. The final models were trained on the complete dataset after parameter optimization, achieving robust performance across all categories. The training process incorporated early stopping based on evaluation loss to prevent overfitting, along with model checkpointing to retain the best-performing version. We found that a learning rate of 2e-5 with a batch size of 8 provided optimal convergence across all three languages, though Telugu required slightly higher weight decay for better generalization.

5 Results and Analysis

Our system demonstrated competitive performance across all three languages in the DravidianLangTech shared task. We achieved second rank in Malayalam and Tamil tasks, and first rank in Telugu, showcasing the effectiveness of our approach.

Language	Macro F1	Rank
Malayalam	0.7367	2/17
Tamil	0.7225	2/17
Telugu	0.3817	1/18

Table 4: Final test set performance and rankings

In the Malayalam task, our system achieved a macro F1 score of 0.7367, placing second behind SSNTrio (0.7511). The margin between the top two systems was relatively small (0.0144), indicating comparable performance levels. For Tamil, we again secured the second position with a macro F1 score of 0.7225, closely following SSNTrio (0.7332). In the Telugu task, our system outperformed all other participants with a macro F1 score of 0.3817, marginally ahead of SSNTrio (0.3758).

5.1 Error Analysis and Model Behavior

Our development set experiments showed notably different performance patterns compared to the final test set results, highlighting important insights about model generalization. During development, the Malayalam model achieved a macro F1 score of 0.80 on our test split, significantly higher than the 0.7367 obtained on the competition's test set. This performance gap suggests potential over-fitting despite our regularization efforts.

The most pronounced discrepancy appeared in the Telugu task. While our development experiments showed exceptional performance with a macro F1 score of 0.90, the final test set yielded 0.3817. This substantial difference indicates that the competition's test data likely contained more challenging or diverse examples than our training split. However, it's noteworthy that this performance level still led to a first-place ranking, suggesting that other teams faced similar generalization challenges.

The Tamil model showed the most consistent performance between development (0.52 macro F1) and final test set (0.7225) results. This consistency might be attributed to our more conservative hyperparameter choices for Tamil, particularly in terms of regularization strength.

Across all languages, we observed that the models performed most reliably on non-hate speech classification, likely due to the larger representation of this class in the training data. The detection of political hate speech proved particularly challenging, especially in Tamil where the training data was most limited for this category. These observations suggest that while our approach effectively captures general language patterns, performance on minority classes remains sensitive to data distribution shifts between training and test sets.

6 Discussion and Future Directions

The significant performance variations between our development experiments and the final test set results highlight key areas for improvement in our approach. While achieving competitive rankings, the disparity (particularly in Telugu with 0.90 in development vs 0.3817 in final test) indicates a need for more robust validation strategies.

To enhance model performance, we recommend implementing language-specific data augmentation techniques and adopting more rigorous cross-validation approaches. Our text-only implementation, while competitive, could benefit from integrating the available audio features. Recent advances in speech encoders for Indian languages, such as IndicWav2Vec (Javed et al., 2021), could provide valuable additional signals for hate speech detection.

Furthermore, focusing on Dravidian language-specific characteristics through better morphologi-

cal analysis and script handling could improve the model's understanding of regional language variations. As larger language models trained specifically on Dravidian languages become available, they may offer better feature representations for this task. These improvements, combined with effective multimodal fusion strategies, could lead to more robust and generalizable models for hate speech detection in Dravidian languages.

7 Conclusion

This paper presented our approach to hate speech detection in Dravidian languages as part of the DravidianLangTech shared task at NAACL 2025. By leveraging language-specific BERT models and implementing careful optimization strategies, we achieved competitive results across all three languages, securing first position in Telugu and second positions in both Malayalam and Tamil tasks.

Our results demonstrate that transformer-based models, even without multimodal features, can effectively detect hate speech in Dravidian languages. The performance variations between development and final test sets provided valuable insights into the challenges of model generalization in this domain. The success of our text-only approach, while encouraging, suggests potential for further improvements through multimodal integration and language-specific optimizations.

8 Limitations

While this work demonstrates effective hate speech detection in Dravidian languages using language-specific BERT models, several important limitations should be acknowledged. One key limitation is the exclusive reliance on textual features, which means that the audio components available in the dataset are not utilized. Although our approach achieved competitive results, it may miss important paralinguistic cues—such as tone, emphasis, and emotion—that could provide additional context when the text alone is ambiguous.

Another limitation is related to the computational resources required for training and inference. The language-specific BERT models, while powerful, demand significant processing power, which may restrict their use in real-time content moderation scenarios where rapid processing of large volumes of data is essential.

A further challenge lies in handling class imbalance in the training data. Despite applying regularization techniques such as label smoothing and weight decay, the models still tend to favor majority classes. This bias is particularly evident in the detection of certain types of hate speech, such as political hate speech in Tamil, where training examples are limited. This suggests that the current approach might not fully capture the nuances of minority hate speech categories.

Additionally, our models face difficulties with code-mixed content—a common characteristic of social media communication in Dravidian languages. Although language-specific BERT models capture many linguistic nuances, they may not optimally process text that switches between English and the target language or different scripts, which is increasingly prevalent online.

Finally, the observed performance disparity between the development and final test sets, especially in the Telugu task, indicates limitations in the model's ability to generalize to new, unseen data. This gap suggests that the current approach may not be robust enough to handle shifts in data distribution or novel patterns of hate speech that emerge over time.

These limitations offer clear directions for future work, including the integration of multimodal features, improved techniques for handling class imbalance and code-mixed text, and the development of more robust validation and adaptation strategies.

References

V. Arunachalam and N. Maheswari. 2024. Enhanced detection of hate speech in dravidian languages in social media using ensemble transformers. *Interdisciplinary Journal of Information, Knowledge, and Management*, 19(Article 39).

Shankar Biradar and Sunil Saumya. 2022. Iiitdwd-shankarb@dravidian-codemixi-hasoc2021: mbert based model for identification of offensive content in south indian languages. *Preprint*, arXiv:2204.10195.

Bharathi Raja Chakravarthi et al. 2023. Offensive language identification in dravidian languages using mpnet and cnn. *International Journal of Information Management Data Insights*, 3(1):100151.

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.

A. Hande, S. U. Hegde, and B. R. Chakravarthi. 2022. Multi-task learning in under-resourced dravidian languages. *Journal of Data, Information and Management*, 4:137–165.

Tahir Javed, Sumanth Doddapaneni, Abhigyan Raman, Kaushal Santosh Bhogale, Gowtham Ramesh, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2021. Towards building asr systems for the next billion users. *Preprint*, arXiv:2111.03945.

Raviraj Joshi. 2023. L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages. *Preprint*, arXiv:2211.11418.

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.

Aditya R Pillai and Biri Arun. 2024. A feature fusion and detection approach using deep learning for sentimental analysis and offensive text detection from code-mix malayalam language. *Biomedical Signal Processing and Control*, 89:105763.

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.

P. K. Roy, S. Bhawal, and C. N. Subalalitha. 2022. Hate speech and offensive language detection in dravidian languages using deep ensemble framework. *Computer Speech & Language*, 75:101386.

Deepawali Sharma, Vedika Gupta, and Vivek Kumar Singh. 2023. Detection of homophobia & transphobia in dravidian languages: Exploring deep learning methods. *Preprint*, arXiv:2304.01241.

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*.

K. Sreelakshmi, B. Premjith, and K. P. Soman. 2020. Detection of hate speech text in hindi-english code-mixed data. In *Procedia Computer Science*, volume 171, pages 737–744.

M. Subramanian, G. J. Adhithiya, S. Gowthamkrishnan, and R. Deepti. 2022. Detecting offensive tamil texts using machine learning and multilingual transformer models. In *Proceedings of the International Conference on Smart Technologies and Systems for Next Generation Computing*, pages 1–6.