LexiLogic@DravidianLangTech 2025: Detecting Misogynistic Memes and Abusive Tamil and Malayalam Text Targeting Women on Social Media

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

Abstract

Social media platforms have become a significant medium for communication and expression, but they are also plagued by misogynistic content targeting women. This study focuses on detecting misogyny in memes and abusive textual content in Tamil and Malayalam languages, which are underrepresented in natural language processing research. Leveraging advanced machine learning and deep learning techniques, we developed a system capable of identifying misogynistic memes and abusive text. By addressing cultural and linguistic nuances, our approach enhances detection accuracy and contributes to safer online spaces for women. This work also serves as a foundation for expanding misogyny detection to other low-resource languages, fostering inclusivity and combating online abuse effectively.

This paper presents our work on detecting misogynistic memes and abusive Tamil and Malayalam text targeting women on social media platforms. Leveraging the pretrained models l3cube-pune/tamil-bert and l3cube-pune/malayalam-bert, we explored various data cleaning and augmentation strategies to enhance detection performance. The models were fine-tuned on curated datasets and evaluated using accuracy, F1-score, precision, and recall. The results demonstrated significant improvements with our cleaning and augmentation techniques, yielding robust performance in detecting nuanced and culturally-specific abusive content.

Our model achieved macro F1 scores of 77.83/78.24 on L3Cube-Bert-Tamil and 78.16/77.01 on L3Cube-Bert-Malayalam, ranking 3rd and 4th on the leaderboard. For the misogyny task, we obtained 83.58/82.94 on L3Cube-Bert-Malayalam and 73.16/73.8 on L3Cube-Bert-Tamil, placing 9th in both. These results highlight our model's effectiveness in low-resource language classification.

1 Introduction

The rise of social media has enabled open communication but has also fueled an increase in toxic and abusive content, with misogyny targeting women becoming a critical concern. Harmful memes and abusive text perpetuate gender-based discrimination, impact mental health, and reinforce societal stereotypes. This issue is particularly challenging in underrepresented languages like Tamil and Malayalam, where cultural nuances and linguistic complexity hinder effective detection. Existing detection tools, designed primarily for high-resource languages, often fail to address the needs of Tamil and Malayalam speakers. Identifying such content requires specialized models that account for low-resource language challenges and cultural context.

This study focuses on detecting misogynistic memes and abusive Tamil and Malayalam text targeting women. By leveraging advanced machine learning models and incorporating cultural insights, we aim to expand NLP capabilities to underserved languages and contribute to safer, more inclusive digital platforms. We present two distinct tasks: (1) detecting misogynistic memes and (2) detecting abusive Tamil and Malayalam text specifically targeting women. Both tasks rely on a combination of linguistic nuance and contextual understanding, which are critical for effective detection.

As part of this effort, we align our work with the NAACL 2025 Dravidian language competitions (Chakravarthi et al., 2025; Rajiakodi et al., 2025), which emphasize the development of NLP solutions for low-resource Dravidian languages. These competitions serve as a platform to advance research in underrepresented languages, fostering collaboration between linguists, computer scientists, and AI researchers. By participating in these shared tasks, we aim to benchmark our models against state-of-the-art approaches and contribute to the broader initiative of improving abusive con-

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 513–517

tent detection in Tamil and Malayalam. Our work not only enhances the technological landscape for Dravidian languages but also promotes responsible AI applications in social media moderation. [1].

2 Related Work

The detection of misogynistic and abusive content has gained importance in NLP due to increasing toxicity on digital platforms. While significant progress has been made for high-resource languages like English, low-resource languages such as Tamil and Malayalam remain underexplored. This survey highlights key efforts in abuse detection, misogyny identification, and multimodal meme analysis, with a focus on Tamil and Malayalam.

Abuse and Hate Speech in Multilingual Contexts: Initial studies, such as (Waseem and Hovy, 2016) and (Wulczyn et al., 2017), advanced abuse detection in English, while (Bhattacharya et al., 2019) addressed multilingual challenges with Hindi-English datasets. Tamil and Malayalam gained attention through (Zhao and Tao, 2021), who introduced annotated datasets for Dravidian languages.

Misogyny Detection: Early research focused on textual misogyny detection (Öhman et al., 2018) and later expanded to multimodal content. For Tamil and Malayalam, leveraged pre-trained models to incorporate cultural nuances in misogyny detection.

Abusive Language Detection in Tamil and Malayalam: Recent work by (Chakravarthi et al., 2021) used transformer models like BERT for gender-specific abuse detection, tackling challenges such as dialect diversity and linguistic richness.

Multimodal Meme Classification: (Kiela et al., 2021) introduced multimodal models combining text and image embeddings, later adapted by for regional languages. Tamil and Malayalam memes, however, lack annotated datasets, limiting progress.

Low-Resource NLP Challenges: Scarcity of labeled data, dialectal complexity, and code-mixing are persistent issues(Hande et al., 2022). Approaches like transfer learning and pre-trained models, such as l3cube-pune/tamil-bert and l3cube-pune/malayalam-bert, have shown potential in overcoming these limitations (Litake et al., 2022).

Conclusion: Despite progress, Tamil and Malayalam remain underrepresented in misogyny and abuse detection research. Leveraging pre-trained models and multimodal techniques can bridge this gap, enabling safer and more inclusive online spaces.(Vaswani et al., 2023), (Devlin et al., 2019)

3 Methodology of processing

3.1 Dataset of IndicBERT

IndicXNLI(Aggarwal et al., 2022) is a benchmark dataset designed to evaluate Natural Language Inference (NLI)(Chen et al., 2018) for 11 major Indian languages, including Hindi, Tamil, Malayalam, Telugu, Kannada, Bengali, Gujarati, Punjabi, Marathi, Oriya, and Assamese. Extending the XNLI dataset(Conneau et al., 2018), it provides premise-hypothesis pairs translated into these languages to capture the linguistic and cultural diversity of the Indian subcontinent. Each instance in the dataset is labeled as entailment, contradiction, or neutral, enabling cross-lingual and multilingual evaluation. IndicXNLI supports the fine-tuning and evaluation of multilingual models such as mBERT, XLM-RoBERTa, and IndicBERT, focusing on challenges like morphological richness, dialectal variations, and code-mixed text common in Indian languages. It serves as a vital resource for assessing NLI performance in low-resource languages while identifying the limitations of existing models in handling complex linguistic structures and cultural nuances. This dataset bridges the gap in NLI research for Indian languages and informs the development of robust and inclusive NLP systems.

3.2 Tamil-BERT and Malayalam-BERT

The L3Cube-Tamil-BERT and L3Cube-Malayalam-BERT models, pre-trained on large-scale corpora specific to Tamil and Malayalam, have demonstrated strong performance on various NLP tasks. When fine-tuned on tasks such as sentiment analysis, named entity recognition, and abusive language detection, these models achieved high accuracy, surpassing many general multilingual models. For instance, L3Cube-Tamil-BERT has shown accuracies of up to 92% on sentiment classification tasks, while L3Cube-Malayalam-BERT achieved around 90% accuracy in the same domain. These models excel in understanding the unique syntactic and semantic structures of Tamil and Malayalam, improving performance on downstream tasks compared to

[1]The code for this paper is available at this GitHub repository

traditional models like mBERT or XLM-R.

In abusive language detection, these models have been shown to yield accuracy rates of around $85 - 88\%$, significantly outperforming other language-specific models that were not fine-tuned for Tamil and Malayalam. By capturing intricate language patterns and cultural context, L3Cube models provide an essential foundation for building advanced NLP systems in low-resource Indian languages, contributing to more effective applications in social media content moderation and sentiment analysis in regional languages.

3.3 Binary Classification

We conducted experiments on two significant tasks from the Dravidian Language Technology Workshop: (1) Offensive Language Identification in Dravidian Languages (Codalab Competition: [20701] and (2) Meme Classification for Tamil (Codalab Competition: [20856]. Both tasks aim to address pressing issues in regional language processing, including offensive language and meme-based toxicity detection, using Tamil and Malayalam as representative low-resource languages.

3.3.1 Offensive Language Identification (Competition 20701)

For this task, we used L3Cube-Tamil-BERT and L3Cube-Malayalam-BERT, fine-tuning them on the provided annotated datasets. The datasets consisted of social media text annotated as offensive or non-offensive. Our pre-processing involved tokenization, cleaning unwanted symbols, and normalizing code-mixed data. The models were fine-tuned with a learning rate of 2e-5 for five epochs using a cross-entropy loss function. On the test set, L3Cube-Tamil-BERT achieved macro F1 score of 78.24, and L3Cube-Malayalam-BERT achieved an F1 score of 70.01, demonstrating their ability to understand nuanced linguistic patterns in offensive content. placing it among the top-performing submissions for this task(Litake et al., 2022)..

3.3.2 Meme Classification for Tamil and Malayalam (Competition 20856)

In this task, we focused on the classification of Tamil and Malayalam memes into categories such as offensive, humorous, or neutral. We utilized L3Cube-Tamil-BERT and L3Cube-Malayalam-BERT for the textual data extracted from memes, and incorporated data augmentation techniques to improve class balance. Pre-trained Tamil-BERT and Malayalam-BERT embeddings provided contextual understanding, crucial for recognizing nuanced meanings within the textual content of memes. Our model achieved macro-F1 score of 68.707 on Tamil-BERT and 80.364 on Malayalam-BERT on test data set(Litake et al., 2022)..

Model	**Train set macro F1**	**Test set macro F1**
XLM-Bert-Tamil	72.01	73.17
Indic-Tamil	73.62	75.04
L3Cube-Bert-Tamil	**77.83**	**78.24**
XLM-Bert-Malayalam	73.52	74.01
Indic-Malayalam	73.92	74.73
L3Cube-Bert-Malayalam	**78.16**	**77.01**

Table 1: Offensive Language Identification Results on Tamil and Malayalam

Observations and Insights The results demonstrate the strength of the L3Cube models in handling low-resource Indian languages. By leveraging domain-specific embeddings, the models captured the linguistic and cultural nuances of Tamil and Malayalam, outperforming baseline multilingual models like mBERT and XLM-R. These experiments highlight the potential of language-specific pre-trained models in advancing NLP tasks for low-resource languages, contributing to safer and more inclusive digital ecosystems.

4 Results

Among the other models, IndicBERT demonstrated competitive performance, with Indic-Tamil and Indic-Malayalam achieving macro F1 scores of 75.04 and 74.73, respectively. XLM-BERT, despite being a widely used multilingual model, exhibited slightly lower performance in comparison. This suggests that models pre-trained specifically on Tamil and Malayalam data, such as L3Cube BERT and IndicBERT, have a notable advantage in handling linguistic intricacies for offensive language identification. The overall results emphasize the effectiveness of domain-specific BERT models in low-resource languages like Tamil and Malayalam. L3Cube BERT, with its targeted pretraining, outperformed general multilingual models, making it a strong candidate for applications in sentiment analysis, offensive language detection, and other NLP

Model	Train set macro F1	Test set macro F1
XLM-Bert-Malayalam	78.5	78.67
Indic-Malayalam	79.83	79.98
L3Cube-Bert-Malayalam	**83.58**	**82.94**
XLM-Bert-Tamil	67.06	69.43
Indic-Tamil	69.01	70.3
L3Cube-Bert-Tamil	**73.16**	**73.8**

Table 2: Meme Classification Results on Tamil and Malayalam

tasks in these languages. Future research could further explore hybrid approaches or fine-tuning strategies to enhance these models' performance further.

5 Conclusion

Our study highlights the significance of leveraging specialized pre-trained models such as l3cube-tamil-bert and l3cube-malayalam-bert for misogyny and abusive content detection in Tamil and Malayalam. Compared to general-purpose multilingual models like XLM-BERT and IndicBERT, the L3Cube models demonstrated superior performance, as evidenced by the results summarized in the comparison tables. These improvements underscore the importance of adopting language-specific embeddings that capture the linguistic and cultural nuances of Tamil and Malayalam.

Additionally, applying data augmentation techniques, such as synonym replacement, back-translation, and contextual data augmentation, contributed significantly to enhancing the models' performance. These methods enriched the training datasets, enabling the models to better generalize to varied and complex scenarios. By combining fine-tuned L3Cube models with robust data augmentation strategies, we achieved improved accuracy and contextual understanding, paving the way for more effective detection of misogynistic and abusive content. This work emphasizes the need for tailored approaches to address the challenges of low-resource languages, fostering safer and more inclusive digital platforms.

6 Limitations

Offensive content against specific groups such as women is a major concern on social media. While our models help in detecting inappropriate content, they rely on static datasets, which might no longer hold valid due to changing trends. Therefore, we need approaches such as active learning and continual learning for ensuring that such offensive content detectors stay up to date and have a balanced representation from new and existing social media platforms. Larger unsupervised and supervised datasets can also improve the performance metrics of such systems, especially in lower resource languages such as Tamil and Malayalam. Another important issue is that of bias- such models might inadvertently discriminate against certain social media users. Moreover, bad actors might exploit the limitations of these models to circumvent NLP-based offensive content detectors, and discover creative ways to post undesirable content.

References

Divyanshu Aggarwal, Vivek Gupta, and Anoop Kunchukuttan. 2022. Indicxnli: Evaluating multilingual inference for indian languages. *Preprint*, arXiv:2204.08776.

Paheli Bhattacharya, Kripabandhu Ghosh, Saptarshi Ghosh, Arindam Pal, Parth Mehta, Arnab Bhattacharya, and Prasenjit Majumder. 2019. Fire 2019 aila track: Artificial intelligence for legal assistance. In *Proceedings of the 11th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '19, page 4–6, New York, NY, USA. Association for Computing Machinery.

Bharathi Raja Chakravarthi, Rahul Ponnusamy, Saranya Rajiakodi, Shunmuga Priya Muthusamy Chinnan, Paul Buitelaar, Bhuvaneswari Sivagnanam, and Anshid Kizhakkeparambil. 2025. Findings of the Shared Task on Misogyny Meme Detection: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. 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.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Diana Inkpen, and Si Wei. 2018. Neural natural language inference models enhanced with external knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2406–2417, Melbourne, Australia. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

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.

Adeep Hande, Siddhanth U Hegde, Sangeetha S, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2022. The best of both worlds: Dual channel language modeling for hope speech detection in low-resourced Kannada. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 127–135, Dublin, Ireland. Association for Computational Linguistics.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2021. The hateful memes challenge: Detecting hate speech in multimodal memes. *Preprint*, arXiv:2005.04790.

Onkar Litake, Maithili Ravindra Sabane, Parth Sachin Patil, Aparna Abhijeet Ranade, and Raviraj Joshi. 2022. L3Cube-MahaNER: A Marathi named entity recognition dataset and BERT models. In *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 29–34, Marseille, France. European Language Resources Association.

Emily Öhman, Kaisla Kajava, Jörg Tiedemann, and Timo Honkela. 2018. Creating a dataset for multilingual fine-grained emotion-detection using gamification-based annotation. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 24–30, Brussels, Belgium. Association for Computational Linguistics.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Shunmuga Priya Muthusamy Chinnan, Ruba Priyadharshini, Rajameenakshi J, Kathiravan Pannerselvam, Rahul Ponnusamy, Bhuvaneswari Sivagnanam, Paul Buitelaar, Bhavanimeena K, Jananayagam V, and Kishore Kumar Ponnusamy. 2025. Findings of the Shared Task on Abusive Tamil and Malayalam Text Targeting Women on Social Media: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1391–1399, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Yingjia Zhao and Xin Tao. 2021. ZYJ@LT-EDI-EACL2021:XLM-RoBERTa-based model with attention for hope speech detection. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 118–121, Kyiv. Association for Computational Linguistics.