# CUET_Agile@DravidianLangTech 2025: Fine-tuning Transformers for Detecting Abusive Text Targeting Women from Tamil and Malayalam Texts

**Tareque Md Hanif**[1] and **Md Rashadur Rahman**[2]
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology
[1]tarequemd.hanif@yahoo.com
[2]rashadur@cuet.ac.bd

## Abstract

As social media has grown, so has online abuse, with women often facing harmful online behavior. This discourages their free participation and expression online. This paper outlines the approach adopted by our team for detecting abusive comments in Tamil and Malayalam. The task focuses on classifying whether a given comment contains abusive language towards women. We experimented with transformer-based models by fine-tuning Tamil-BERT for Tamil and Malayalam-BERT for Malayalam. Additionally, we fine-tuned IndicBERT v2 on both Tamil and Malayalam datasets. To evaluate the effect of pre-processing, we also conducted experiments using non-preprocessed text. Results demonstrate that IndicBERT v2 outperformed the language-specific BERT models in both languages. Pre-processing the data showed mixed results, with a slight improvement in the Tamil dataset but no significant benefit for the Malayalam dataset. Our approach secured first place in Tamil with a macro $F_1$-score of 0.7883 and second place in Malayalam with a macro $F_1$-score of 0.7234. The implementation details of the task will be found in the GitHub repository.[1]

## 1 Introduction

Over the past decade, the exponential growth of user-generated content on social media has unfortunately led to increased abusive behavior online. This includes cyberbullying, hate speech, and offensive language, often targeting various classes of people including women. These actions can lead to real-world violence and push women to the sidelines, making them feel excluded and undervalued both online and in everyday life (Kaur et al., 2021). A study in 51 countries found that 38% of women have faced online harassment. Only 25% of them reported it, and 90% reduced their online activity (Hashmi et al., 2024).

Tamil is among the oldest languages in the world, spoken by over 65 million people globally (Ramakrishnan et al., 2007). Malayalam, the official language of Kerala, has more than 37 million speakers worldwide (Rojan et al., 2020). Both Tamil and Malayalam have many dialects, making it challenging to develop NLP systems for these languages.

Developing an intelligent abuse detection model is challenging in resource-constrained languages like Tamil and Malayalam. Therefore, a shared task was organized to encourage the development of effective abuse detection models for these languages.

This shared task (Rajiakodi et al., 2025) aims to detect abusive comments targeting women in Tamil and Malayalam, sourced from YouTube comments. The dataset contains text in both languages, with each comment classified as either 'Abusive' or 'Non-Abusive'. The task focuses on identifying explicit abuse, implicit bias, stereotypes, and coded language directed at women on social media.

We fine-tuned transformer-based models for text classification. Specifically, we used Tamil-BERT (Joshi, 2022) for Tamil comments and Malayalam-BERT (Joshi, 2022) for Malayalam comments. Additionally, we fine-tuned IndicBERT v2 (Doddapaneni et al., 2023) on both Tamil and Malayalam datasets. We also experimented with training models on non-preprocessed text to analyze the impact of preprocessing.

The rest of the paper is organized into 6 sections. Section 2 reviews related work in Natural Language Processing, focusing on misogynistic text detection in Tamil, Malayalam, and other languages. Section 3 describes the dataset provided by the shared task organizers. Section 4 provides a detailed explanation of the proposed methodology and the models implemented. Section 5 presents the results and key observations. Finally, Section 6 concludes the paper.

[1]https://github.com/tmdh/DravidianLangTech-NAACL-2025-ATTW



*Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 393–397

## 2 Related Works

Recent advances in NLP have increased interest in detecting different types of hate speech, leading to many new and creative methods in this field. Offensive language detection in Tamil and Malayalam has been studied in previous research (Ponnusamy et al., 2024), but to the best of our knowledge, this is the first shared task that specifically focuses on detecting abusive texts targeting women from Dravidian texts. There have been previous shared tasks on languages other than Dravidian for misogynistic text detection, such as the Arabic Misogyny Identification (ArMI) task (Mulki and Ghanem, 2022) and the GermEval2024 shared task, GerMS-Detect (Gross et al., 2024). The ArMI task combined two subtasks: a binary classification for detecting misogynistic language and a multi-class classification for identifying seven misogynistic behaviors in 9,833 Arabic/dialectal tweets. GerMS-Detect focused on detecting sexism and misogyny in German language online news comment.

In terms of Dravidian languages, Chakravarthi et al., 2023 proposed a fusion model of MPNet (Song et al., 2020) and CNN for offensive language identification in code-mixed Tamil, Malayalam, and Kannada social media comments, achieving superior results over classical ML and transformer-based baselines.

Sreelakshmi et al., 2024 explored offensive language detection in code-mixed Tamil-English, Malayalam-English and Kannada-English using multilingual transformer embeddings with Support Vector Machine classifiers, identifying MuRIL (Khanuja et al., 2021) as the most effective model across various datasets.

The study by Vasantharajan and Thayasivam, 2021 explores offensive language detection in Tamil code-mixed YouTube comments, proposing selective translation and transliteration techniques to enhance transformer models like BERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020). Their findings highlight ULMFiT (Howard and Ruder, 2018) and mBERT-BiLSTM as the most effective models for offensive language detection.

Prior work in abuse detection has primarily focused on English, lacking substantial datasets for Indic languages. Gupta et al., 2022 proposed MACD to address this gap by introducing a large-scale multilingual abuse detection dataset and AbuseXLMR model for Indic languages.

| Class | Train | Dev | Test |
|---|---|---|---|
| Abusive | 1366 | 278 | 305 |
| Non-Abusive | 1424 | 320 | 293 |
| Total | 2790 | 598 | 598 |

Table 1: Class-wise distribution of Tamil Dataset

| Class | Train | Dev | Test |
|---|---|---|---|
| Abusive | 1531 | 303 | 323 |
| Non-Abusive | 1402 | 326 | 306 |
| Total | 2933 | 629 | 629 |

Table 2: Class-wise distribution of Malayalam Dataset

## 3 Dataset Description

The organizers of the Abusive Text Targeting Women Detection shared task provided two datasets (Priyadharshini et al., 2023, Priyadharshini et al., 2022) where one consists of Tamil texts while the other consists of Malayalam texts. Each of the texts is annotated with one of the the classes: Abusive and Non-Abusive. Table 1 displays the class-wise data distribution for the Tamil dataset, while Table 2 shows the same for the Malayalam dataset.

To provide better insights, we conducted a more in-depth analysis of the training set. Table 3 presents the detailed statistics of the training data.

## 4 Methodology

This work employed two transformer-based models on each language's dataset, both preprocessed and non-preprocessed. Firstly, we removed unwanted characters (i.e., numbers, extra spaces, and URLs) from the texts in both the Tamil and Malayalam datasets to create two preprocessed datasets.

### 4.1 Transformer Models

Recent advancements in NLP have shown that transformer-based models perform better than other approaches for text classification across different languages. In this work, we fine-tuned Tamil-

| Statistics | Abusive | Non-Abusive |
|---|---|---|
| Total words | 21166 | 19091 |
| Unique words | 9541 | 8672 |
| Max. length (words) | 48 | 48 |
| Avg. words (per text) | 15.5 | 13.4 |

Table 3: Detailed statistics of each class in the training set

| Approach | Selected Epoch | Accuracy | Precision | Recall | $F_1$-score |
|---|---|---|---|---|---|
| Tamil-BERT (Non-preprocessed) | 4 | 0.7793 | 0.7800 | 0.7786 | 0.7788 |
| Tamil-BERT (Preprocessed) | 2 | 0.7843 | 0.7861 | 0.7850 | 0.7842 |
| IndicBERT v2 (Non-preprocessed) | 3 | 0.7876 | 0.7945 | 0.7860 | 0.7857 |
| IndicBERT v2 (Preprocessed) | 2 | **0.7893** | **0.7923** | **0.7882** | **0.7883** |

Table 4: Performance comparison of various models on the test set of Tamil dataset

| Approach | Selected Epoch | Accuracy | Precision | Recall | $F_1$-score |
|---|---|---|---|---|---|
| Malayalam-BERT (Non-preprocessed) | 2 | 0.6630 | 0.7136 | 0.6692 | 0.6467 |
| Malayalam-BERT (Preprocessed) | 5 | 0.6439 | 0.6440 | 0.6426 | 0.6424 |
| IndicBERT v2 (Non-preprocessed) | 2 | **0.7234** | **0.7238** | **0.7239** | **0.7234** |
| IndicBERT v2 (Preprocessed) | 4 | 0.7122 | 0.7133 | 0.7130 | 0.7122 |

Table 5: Performance comparison of various models on the test set of Malayalam dataset

BERT[2] and Malayalam-BERT[3] on Tamil and Malayalam texts, respectively, using both preprocessed and non-preprocessed datasets. We also used IndicBERT v2[4], a multilingual model, to handle both languages.

Our classifier is based on a transformer model with a linear classification head. The architecture consists of a pre-trained BERT model followed by a fully connected layer that maps the hidden state of the [CLS] token to a two-class output. The model was trained using PyTorch Lightning (Falcon and team, 2024), which simplified the training and evaluation process. We optimized the models using the AdamW (Loshchilov and Hutter, 2019) optimizer with a learning rate of $5e-5$. The training process ran for up to five epochs, and we selected the best-performing epoch based on the highest $F_1$-score on the validation set.

For training, we used a batch size of 32 and applied cross-entropy loss. The training process logged $F_1$-score on the validation set at each epoch. Model checkpoints were saved after each epoch, and the model with the highest $F_1$ was used for evaluation.

Table 6 summarizes the hyperparameters used across all models. The selected epochs for each approach are shown in Table 4 and Table 5 for Tamil and Malayalam, respectively.

[2] https://huggingface.co/l3cube-pune/tamil-bert

[3] https://huggingface.co/l3cube-pune/malayalam-bert

[4] https://huggingface.co/ai4bharat/IndicBERTv2-MLM-only

| Hyperparameters | Values |
|---|---|
| Learning Rate | $5e-5$ |
| Batch Size | 32 |
| Max Epochs | 5 |
| Weight Decay | 0.01 |

Table 6: Hyperparameters used across models

#### 4.1.1 Tamil-BERT and Malayalam-BERT

Tamil-BERT and Malayalam-BERT are monolingual BERT models fine-tuned from the multilingual MuRIL model for the Tamil and Malayalam languages, respectively. They are trained on large monolingual corpora. These models aim to enhance performance on downstream NLP tasks for these low-resource Indian languages. (Joshi, 2022)

### 4.2 IndicBERT v2

IndicBERT v2 is a state-of-the-art multilingual language model designed specifically for Indic languages. It supports all 24 languages covered in the IndicCorp v2 dataset. The dataset includes 20.9 billion tokens from 24 languages, including Indian English. This model is a significant step forward in building robust NLU capabilities for diverse Indic languages. (Doddapaneni et al., 2023)

## 5 Results

Table 4 and 5 reports the performance comparison of the different approaches on the Tamil dataset and Malayalam dataset, respectively. The effectiveness of the models is determined based on the macro $F_1$-score.

For the Tamil dataset, IndicBERT v2 fine-tuned on the preprocessed dataset achieved the high-

est $F_1$-score of 0.7883, followed by IndicBERT v2 on the non-preprocessed dataset with an $F_1$-score of 0.7857. For the Malayalam dataset, IndicBERT v2 employed on the non-preprocessed dataset achieved the best performance with an $F_1$-score of 0.7234, while IndicBERT v2 on the pre-processed dataset also performed well with an $F_1$-score of 0.7122. It is indicating that pre-processing did not improve the performance on the Malayalam dataset. For both Tamil and Malayalam, Tamil-BERT and Malayalam-BERT did not perform well on the task, while IndicBERT v2 achieved strong performance in both languages.

## 6 Conclusion

This paper investigated two language specific transformer models and one multilingual language model to detect abuse targeted towards women from preprocessed and non-preprocessed Tamil and Malayalam texts. Among all approaches, the highest macro $F_1$-score 0.7883 for Tamil texts is obtained by IndicBERT v2 fine-tuned with preprocessed Tamil texts. For Malayalam texts, the highest macro $F_1$-score 0.7234 is gained by finetuning IndicBERT v2 with non-preprocessed Malayalam texts. Looking ahead, we plan to explore ensemble methods and other advanced transformer-based models including MuRIL and XLM-R.

## Limitations

While our model demonstrated strong performance in identifying abusive text directed at women in Tamil and Malayalam, it is important to recognize several limitations. These include the limited availability of diverse and high-quality annotated datasets for Dravidian languages, which restricts the model's ability to generalize across various dialects. Furthermore, the linguistic intricacies of Tamil and Malayalam can affect the model's effectiveness, especially in detecting implicit or subtly coded abusive language. Another challenge is the scalability of transformer-based models when handling longer texts, as they are mainly designed and optimized for shorter sequences. Moreover, fine-tuning these models demands significant GPU resources, which could restrict access for researchers with limited computational capabilities. Additionally, we did not perform an extensive hyperparameter search for critical parameters like learning rate and weight decay, which might otherwise enhance the model's performance.

## References


Bharathi Raja Chakravarthi, Manoj Balaji Jagadeeshan, Vasanth Palanikumar, and Ruba Priyadharshini. 2023. Offensive language identification in dravidian languages using mpnet and cnn. *International Journal of Information Management Data Insights*, 3(1):100151.

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.

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.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

William Falcon and The PyTorch Lightning team. 2024. Pytorch lightning.

Stephanie Gross, Johann Petrak, Louisa Venhoff, and Brigitte Krenn. 2024. GermEval2024 shared task: GerMS-detect – sexism detection in German online news fora. In *Proceedings of GermEval 2024 Task 1 GerMS-Detect Workshop on Sexism Detection in German Online News Fora (GerMS-Detect 2024)*, pages 1–9, Vienna, Austria. Association for Computational Lingustics.

Vikram Gupta, Sumegh Roychowdhury, Mithun Das, Somnath Banerjee, Punyajoy Saha, Binny Mathew, hastagiri prakash vanchinathan, and Animesh Mukherjee. 2022. Multilingual abusive comment detection at scale for indic languages. In *Advances in Neural Information Processing Systems*, volume 35, pages 26176–26191. Curran Associates, Inc.

Ehtesham Hashmi, Muhammad Mudassar Yamin, Shariq Imran, Sule Yildirim Yayilgan, and Mohib Ullah. 2024. Enhancing misogyny detection in bilingual texts using fasttext and explainable ai. In *2024 International Conference on Engineering & Computing Technologies (ICECT)*, pages 1–6.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Raviraj Joshi. 2022. L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages. *arXiv preprint arXiv:2211.11418*.

Simrat Kaur, Sarbjeet Singh, and Sakshi Kaushal. 2021. Abusive content detection in online user-generated data: A survey. *Procedia Computer Science*, 189:274–281. AI in Computational Linguistics.

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. *Preprint*, arXiv:2103.10730.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Hala Mulki and Bilal Ghanem. 2022. Working notes of the workshop arabic misogyny identification (armi-2021). In *Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '21, page 7–8, New York, NY, USA. Association for Computing Machinery.

Rahul Ponnusamy, Kathiravan Pannerselvam, Saranya R, Prasanna Kumar Kumaresan, Sajeetha Thavareesan, Bhuvaneswari S, Anshid K.a, Susminu S Kumar, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2024. From laughter to inequality: Annotated dataset for misogyny detection in Tamil and Malayalam memes. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7480–7488, Torino, Italia. ELRA and ICCL.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, Malliga Subramanian, Kogilavani Shanmugavadivel, Premjith B, Abirami Murugappan, Sai Prashanth Karnati, Rishith, Chandu Janakiram, and Prasanna Kumar Kumaresan. 2023. Findings of the shared task on Abusive Comment Detection in Tamil and Telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages DravidianLangTech 2023*. Recent Advances in Natural Language Processing.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022. Overview of abusive comment detection in Tamil-ACL 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298, Dublin, Ireland. 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.

A.G. Ramakrishnan, Lakshmish Kaushik, and Laxmi Narayana. 2007. Natural language processing for tamil tts.

Annlin Rojan, Edwin Alias, Georgy M. Rajan, Jithin Mathew, and Dhanya Sudarsan. 2020. Natural language processing based text imputation for malayalam corpora. In *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pages 161–165.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Preprint*, arXiv:2004.09297.

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.

Charangan Vasantharajan and Uthayasanker Thayasivam. 2021. Towards offensive language identification for tamil code-mixed youtube comments and posts. *SN Computer Science*, 3(1):94.