# Hydrangea@DravidianLanTech2025: Abusive language Identification from Tamil and Malayalam Text using Transformer Models

**Shanmitha Thirumoorthy**
Vellore Institute of Technology
shanmitha.t2023@vitstudent.ac.in

**Ratnavel Rajalakshmi**
Vellore Institute of Technology
rajalakshmi.r@vit.ac.in

**Durairaj Thenmozhi**
Sri Sivasubramaniya Nadar College of Engineering
theni_d@ssn.edu.in

## Abstract

Abusive language toward women on the Internet has always been perceived as a danger to free speech and safe online spaces. In this paper, we discuss three transformer-based models - BERT, XLM-RoBERTa, and DistilBERT-in identifying gender-abusive comments in Tamil and Malayalam YouTube contents. We fine-tune and compare these models using a dataset provided by DravidianLangTech 2025 shared task for identifying the abusive content from social media. Compared to the models above, the results of XLM-RoBERTa are better and reached F1 scores of 0.7708 for Tamil and 0.6876 for Malayalam. BERT followed with scores of 0.7658 (Tamil) and 0.6671 (Malayalam). Of the DistilBERTs, performance was varyingly different for the different languages. A large difference in performance between the models, especially in the case of Malayalam, indicates that working in low-resource languages is difficult. The choice of a model is extremely critical in applying abusive language detection. The findings would be important information for effective content moderation systems in linguistically diverse contexts. In general, it would promote safe online spaces for women in South Indian language communities.

## 1 Introduction

The digital revolution has transformed social networks from a double-edged sword where it fosters democratization of communication, but also allows systemic gender-based violence through abusive content targeting women. Above 40% of the women worldwide said they had suffered online harassment; therefore, South Asian contexts are generally more vulnerable, mainly due to linguistic complexity and domination patriarchal norms. This abuse comes in the form of overt threats, veiled misogyny, and damaging stereotypes that lead to psychological trauma in the form of depression and anxiety, professional reversals, and even physical danger, which bleed into the offline world. Low-resource Dravidian languages like Tamil and Malayalam pose a heightened challenge because of morphological complexity, patterns of code-mixing, and scarce NLP resources: South Asian languages are spoken by 300 million people, but only 0.1% of AI research focuses on these languages. Advanced text classification techniques are used in automated content moderation systems, which are scalable. Especially promising for unmasking contextual abuse patterns is the transformer model. We can take the "digital silencing" effect stemming from the fact that 30% of women journalists self-censor due to online threats and develop language-specific detection frameworks, therefore cultivating safer digital ecosystems that empower rather than endanger women.

DravidianLanTech shared task (Priyadharshini et al., 2022) (Priyadharshini et al., 2023) is continuously focusing on the abusive language identification in the social media contents in Dravidian languages. DravidianLangTech 2025 (Rajiakodi et al., 2025) gives emphasis to Tamil and Malayalam languages in identifying the language targeting to women in social media. Our team Hydrangea participated in this shared task and submitted three runs for both languages. The code is found in GitHub Link

## 2 Related Works

Several research works have been carried out for detecting abusive content in social media. They used from traditional classifiers to transformer models for finding the same.

(Bansal et al., 2022) used XLM RoBERTa, a model pre-trained on more than 100 languages, and added a BiGRU layer for the classification of abusive content in 13 low-resource languages, including Hindi, Telugu, Marathi, and Tamil. Due to a robust pre-training and handling capacity of



*Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 646–650

the model, it was feasible to handle datasets with less availability of natural language processing resources that exceeded traditional methods like Naive Bayes and Logistic Regression. The study was conducted on four different transformer models namely mBERT, MurilBERT, IndicBERT, and XLM RoBERTa along with several data processing techniques like cleaning and transliteration with emoji embeddings. XLM RoBERTa performed the best where it shows superiority over language-specific models with excellent performance and efficiency. In general, the results showed that the importance of using transformer-based approaches as well as data preprocessing to improve the accuracy of models in performing multilingual tasks.

(Philipo et al., 2024) presented the evaluation of three large language models, BERT, XLM RoBERTa, and DistilBERT, as candidates for using an annotated standardized corpus of texts in recognizing cyberbullying on social media. Fine-tuned each model in such a manner that the trained models now worked as two-class classifiers - bullying or not bullying-the texts, and with optimised usage of resources along with good accuracy during classification. Key metrics was then measured on the parameters such as accuracy, precision, recall, and F1-score.

The best model was BERT in all the key metrics, performing at 95%; it seems to recognize subtle language motifs, suggestive of bullying. Both precision and recall metrics indicate excellent capacity for correctly classifying bullying as well as non-bullying instances-the model is showing good balance. XLM RoBERTa showed good performance but was less efficient compared with BERT. This accounted for slightly lower metrics overall because the dataset was monolingual. DistilBERT was designed as a lighter version of BERT and showed increased computational efficiency with faster inference times but recorded slightly lower accuracy and F1-scores. Thus BERT emerged victorious in all metrics

(Koufakou et al., 2020) evaluated the in-domain and cross-domain performance of the basic BERT model against enhanced HurtBERT based on an extensive experimental setup. In the in-domain tasks, where training and test data have similar distributions, the additional lexical features from a hate lexicon are proved to be quite useful for HurtBERT: enhancing precision, recall, and F1-score by identifying explicit abusive language patterns often missed by BERT's contextual embeddings. For cross-domain tasks, where training was done on one dataset and testing on another, HurtBERT was better in generalization because it used sentence-level lexicon encodings and word-level embeddings to adapt effectively to different linguistic styles and representations of abusive content. The problem with transformer-based models was not solved, or rather, in domains with poor amounts of labelled data or differing text styles. The above discussions reflect the opportunity of HurtBERT in practical scenarios of abusive language detection.

(Manikandan et al., 2022) used a three-stage system architecture that included pre-processing, model training, and testing using two transformer models, BERT and XLM-RoBERTa. Despite the fact that both of these models were trained on preprocessed data, in all of these metrics, XLM-RoBERTa surpassed BERT especially with homophobic content, which gave 93% accuracy against 91% BERT, the results reflected that XLM-RoBERTa works well with homophobic content and both faced the problem while handling the few samples of transphobic content.

(Gayathri et al., 2022) used support vector machine with LaBSE embedding for finding the abusive language in Tamil text. (Gayathri et al., 2024) employed a combination of statistic features and language-agnostic features and performed feature selection by using explainable AI for detecting abusive language in Tamil-English codemixed text.

## 3 Data Description

The organizer released 2 sets of dataset. Tamil data set consists of 2790 instances in the training set in which 1366 instances are Abusive category and 1424 are Non-abusive category. Malayalam training data set has a total of 2935 instances with 1531 Abusive contents and 1402 Non-abusive contents. The test data set has 598 and 629 instances for Tamil and Malayalam respectively. The data distribution shows that the data set is balanced and is not required much of balancing the data.

## 4 Methodology

The DravidianLangTech 2025 abusive language identification came equipped with a 4,543 annotated comment data set - 2,819 Tamil and 1,724 Malayalam. The data set included YouTube comments labeled with binary tags: "Abusive" and "Non-Abusive". Tamil and Malayalam are low-resource languages, and it is difficult to work

with them in the area of natural language processing. The data was gathered from YouTube comments and pre-cleaned to eliminate unwanted characters and normalize the text. The abusive and non-abusive comment distribution was not entirely balanced, but slightly more comments were non-abusive. We employed three of the most used architectures of text classification tasks for this abusive language identification task in Tamil and Malayalam: BERT, XLM-RoBERTa, and DistilBERT. Datasets compatible with each other were utilized to tune the models for classifying the comments as appropriate.

BERT is essentially a pre-trained model based on roughly 2.5 billion words taken from Wikipedia in 104 languages using a vocabulary of approximately 110,000 word pieces. BERT functions with its principal training being MLM to better handle the bidirectional words relationship. As compared to traditional RNN, it does not see one token at a time but, rather, BERT randomly masks $15\%$ of the input tokens in every layer and predicts masked tokens based on the entire input sequence. AdamW is employed for fine-tuning optimization and the learning rate is set at $2\mathrm{e}-5$.

In DistilBERT, architectural compression lowers the computational requirements of it. Knowledge distillation is employed to train a 6-layer model that mimics the output of BERT's 12-layer model.

Better performance by XLM-RoBERTa was due to pretraining on multiple languages across different linguistic structures, thus dynamically adapting the vocabulary to suit agglutinative morphological features of Tamil and Malayalam. Training was done with batch size 16, AdamW optimizer with a learning rate of $2\mathrm{e}-5$, and early stopping on validation loss.

For pre-training, BERT and XLM-RoBERTa were pre-trained for two epochs whereas DistilBERT was trained for three epochs because it has a smaller model and hence quicker training cycles.

## 5 Result and analysis

We have evaluated the three models on the datasets provided by the DravidianLangTech 2025 shared task. Tables 1 and 2 show the performance of our three models in terms of precision, recall and F1 score on the Tamil and Malayalam development data sets respectively. XLM-Roberta performed better for Tamil data set whereas BERT performed better for Malayalam data set.

Similarly for test data sets, 3 and 4 show the performance of our three models in Tamil and Malayalam respectively, where XLM-RoBERTa proved successful for both language data sets. XLM-RoBERTa showed the highest classification performance in F1-score at 0.7708 (Tamil) and 0.6876 (Malayalam) compared to BERT at 0.7658 Tamil and 0.6671 Malayalam and DistilBERT at 0.7639 Tamil and 0.4876 Malayalam. The models are very consistent with Tamil classification; all architectures showed F1-score above 0.76 while there is significant performance degradation for Malayalam, especially for DistilBERT near-random performance of 0.4876 . While BERT employed bidirectional attention to identify contextual patterns of abuse, its efficacy was lower on Malayalam with lesser pretraining exposure.

The performance difference between XLM-RoBERTa and BERT for Tamil and Malayalam may be due to various reasons. XLM-RoBERTa is particularly good at multilingual pre-training, where it is able to identify cross-lingual similarities and transfer knowledge across languages. Tamil may have been able to learn more from the cross-lingual knowledge transfer because of its linguistic proximity with other languages in the pre-training data, or being better represented in that data. On the other hand, the distinct linguistic features of Malayalam or its under-representation within the pre-training data could have impeded XLM-RoBERTa's performance in relation to BERT, which perhaps was more well-suited to pick up the particular idiosyncrasies of the Malayalam dataset.

Table 1: Performance Comparison on Tamil-English Development Set

| Model | Precision | Recall | F1 |
|---|---|---|---|
| BERT | 0.5000 | 0.5000 | 0.5000 |
| XLM-RoBERTa | 0.2900 | 0.5400 | 0.5400 |
| DistilBERT | 0.4800 | 0.4800 | 0.4700 |

Table 2: Performance Comparison on Malayalam-English Development Set

| Model | Precision | Recall | F1 |
|---|---|---|---|
| BERT | 0.5000 | 0.5000 | 0.5000 |
| XLM-RoBERTa | 0.2300 | 0.4800 | 0.4800 |
| DistilBERT | 0.4500 | 0.4700 | 0.4600 |

The F1-scores of the test case are low, and they can inform us where something went awry while

Table 3: Performance Comparison on Tamil-English Test Set

| Model | Precision | Recall | F1 |
|---|---|---|---|
| BERT | 0.7658 | 0.7658 | 0.7658 |
| XLM-RoBERTa | 0.7708 | 0.7708 | 0.7708 |
| DistilBERT | 0.7639 | 0.6307 | 0.6909 |

Table 4: Performance Comparison on Malayalam-English Test Set

| Model | Precision | Recall | F1 |
|---|---|---|---|
| BERT | 0.6671 | 0.6613 | 0.6641 |
| XLM-RoBERTa | 0.6817 | 0.6722 | 0.6769 |
| DistilBERT | 0.4798 | 0.4026 | 0.4378 |

pre-processing data and training the model. DistilBERT did not perform for Malayalam, and it may also be a matter of its smaller size not being able to keep up with the complexity of the language or of the model requiring larger training data sets due to this reason. There are further experiments to be conducted in order to ascertain the reason behind the performance of DistilBERT having a negative effect on Malayalam. There may be a reason that has something to do with too little pre-training of DistilBERT on Malayalam text data, and this implies Malayalam-specific linguistic features are under-represented in the model parameters. There may be a second reason, though, and this has something to do with the case of low-quality Malayalam data sets being used to train DistilBERT, i.e., abusive language data sets which do not allow DistilBERT to learn discriminatory patterns successfully.

## 6 Limitation

The transformer models used in our experiments lack in finding patterns from code-mixed Dravidian languages, resulting in low F1 scores. This is mainly due to the size of the data set which is not adequate to train the transformer model. This may be overcome by applying some data augmentation techniques in future.

## 7 Conclusions

Transformer models i.e. BERT, XLM-RoBERTa and DistilBERT are employed to detect the abusive content targeting women on social media. Our team Hydrangea achieved 9th and 12th ranks with macro F1 scores of 0.7708 and 0.6769 for Tamil and Malayalam, respectively in the leader board utilizing the models trained on XLM-RoBERTa. Language-agnostics embeddings can be utilized in the future to enhance the performance of our approach. In addition, Explainable AI (XAI) methods like LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) can be applied to discover how the model is making its decision by determining the words or phrases most responsible for abusive classification. Attention visualization approaches may also be utilized, especially for transformer models, to identify what aspects of the input text the model is paying attention to. These observations can then be utilized to tune feature weights and tune the model's structure, possibly leading to enhanced performance. Zero-shot or few-shot learning methods may also be employed to surmount the problem of data sparsity, by utilizing knowledge from other languages or tasks to enhance performance on Tamil and Malayalam with little training data.

## References


Vibhuti Bansal, Mrinal Tyagi, Rajesh Sharma, Vedika Gupta, and Qin Xin. 2022. A transformer based approach for abuse detection in code mixed indic languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* Just Accepted.

G L Gayathri, Krithika Swaminathan, Divyasri Krishnakumar, Thenmozhi D, and Bharathi B. 2024. Abusive comment detection in tamil code-mixed data by adjusting class weights and refining features. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* Just Accepted.

GL Gayathri, Krithika Swaminathan, K Divyasri, Thenmozhi Durairaj, and B Bharathi. 2022. Pandas@ abusive comment detection in tamil code-mixed data using custom embeddings with labse. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 112–119.

Anna Koufakou, Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. HurtBERT: Incorporating lexical features with BERT for the detection of abusive language. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 34–43, Online. Association for Computational Linguistics.

Deepalakshmi Manikandan, Malliga Subramanian, and Kogilavani Shanmugavadivel. 2022. A system for detecting abusive contents against lgbt community using deep learning based transformer models. In *FIRE (Working Notes)*, pages 106–116.

Adamu Gaston Philipo, Doreen Sebastian Sarwatt, Jianguo Ding, Mahmoud Daneshmand, and Huansheng Ning. 2024. Assessing text classification methods for

cyberbullying detection on social media platforms. *Preprint*, arXiv:2412.19928.

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.