MSM_CUET@DravidianLangTech 2025: XLM-BERT and MuRIL Based Transformer Models for Detection of Abusive Tamil and Malayalam Text Targeting Women on Social Media

Md Mizanur Rahman, Srijita Dhar,
Md Mehedi Hasan, Hasan Murad
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
u1904116@student.cuet.ac.bd, dsrijita2001@gmail.com,
u1904067@student.cuet.ac.bd, hasanmurad@cuet.ac.bd

Abstract

Social media has evolved into an excellent platform for presenting ideas, viewpoints, and experiences in modern society. But this large domain has also brought some alarming problems including internet misuse. Targeted specifically at certain groups like women, abusive language is pervasive on social media. The task is always difficult to detect abusive text for low-resource languages like Tamil, Malayalam, and other Dravidian languages. This paper presents a novel approach to detecting abusive Tamil and Malayalam texts targeting social media. A shared task on 'Abusive Tamil and Malayalam Text Targeting Women on Social Media Detection' has been organized by DravidianLangTech at NAACL-2025. We have implemented our model with different transformer-based models like XLM-R, MuRIL, IndicBERT, and mBERT transformers and the Ensemble method with SVM and Random Forest for training. We selected XLM-RoBERT for Tamil text and MuRIL for Malayalam text due to their superior performance compared to other models. After developing our model, we tested and evaluated it on the DravidianLangTech@NAACL 2025 shared task dataset. We found that XLM-R achieved the highest F1 score of 0.7873 on the test set, ranking 2nd among all participants for abusive Tamil text detections. In contrast, MuRIL had the highest F1 score of 0.6812 for abusive Malayalam text detections, ranking 10 th among all participants.

1 Introduction

The rise of social media has brought many benefits. Meanwhile, the prevalence of abusive text on social media has significantly increased in recent times. Abusive texting can result in cyberbullying, internet harassment, and other horrible acts. These kind words not only harm but also create a panicked and negative atmosphere for general users (Chen et al., 2017).

Abusive text detection is a critical task in social media content, especially when targeting vulnerable groups like children and women (Barker and Jurasz, 2021). There has been limited research on low-resource languages like Dravidian languages (Lee et al., 2018). Tamil and Malayalam are the two important languages in South India. In this paper, we address the abusive Tamil and Malayalam text detection. The cultural and linguistic nuances of these languages present unique challenges for natural language processing (NLP) tasks. So existing multilingual models often struggle to capture these nuances.

Our primary object of this paper is to detect abusive Tamil and Malayalam text targeting women on social media. We used different kinds of transformer models like XLM-R, MuRIL, m-BERT, and IndicBERT and ensemble methods with Random Forest, and SVM to train our model. The XLM-R and MuRIL transformer-based models were chosen because they outperformed Tamil and Malayalam texts, respectively.

The core contributions of our research work are as follows:-

1. We have developed an effective model to detect abusive Tamil and Malayalam text targeting women on social media with good generalizations.

2. We have conducted a systematic evaluation of multilingual transformer models and ensemble techniques. Comparing models for low-resource languages shows both their strengths and limitations.

The implementation details have been provided in the following GitHub repository:- https://github.com/Mizan116/DravidianLangTech-NAACL-2025.

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 330–334

2 Related Work

Abusive language detection identifies and filters damaging, insulting, or degrading information, especially on social media. Hate speech, sexism, homophobia, racism, bullying, and other verbal abuse are detected. Previously, there have been three main approaches to categorizing when it comes to studying how to identify abusive material, such as Tamil and Malayalam texts directed at women on social media.

Initially, the discipline was dominated by classical machine learning algorithms such as Support Vector Machines (SVMs) and Naïve Bayes and Random Forests for text mining techniques in Chen et al. (2017). Using manual feature engineering, abusive content detection can be achieved by extracting meaningful features such as n-grams and sentiment analysis, not handling code mixed language.

Because abusive material is complex and context-dependent. With the rise of deep learning, more powerful methods like LSTMs, especially CNNs and RNNs, are used in Founta et al. (2019), showing how Word2Vec, GloVe, and FastText which enhances the representation of abusive content, outperforming traditional feature engineering for abuse detection tasks. And findings depend on vast annotated datasets and difficulty in sarcasm and language.

Barker and Jurasz (2021) addressed the impact of text-based abuse on women on social media and the need for legislative frameworks to combat online violence, but not gender-targeted harassment in Tamil and Malayalam. Using supervised models and neural networks, The authors of Arellano et al. (2022) developed methods to recognize violent and aggressive material in Spanish, which can be applied to different languages and cultural contexts. Lexicon-based approach used in Lee et al. (2018) to enhance abusive(e.g., offensive terms, slurs) and non-abusive(e.g., polite terms or contextually mitigating words) word lists as the primary tool for detecting abusive text.

Transformer models, and neural networks, monitor relationships in sequential data like this phrase to learn context and meaning. It is a self-attention mechanism that replaces conventional RNNs and CNNs in order to capture long-range dependencies in sequences, setting the stage for models like BERT, GPT, and XLM-R, which are widely used for abusive content detection (Vaswani, 2017). Difficulties in identifying abusive comments in non-English languages, with a particular emphasis on Tamil and Malayalam.

3 Dataset and Task Overview

The DravidianLangTech 2023 (Priyadharshini et al., 2023) shared objective is to identify abusive comments in Tamil and Telugu (Priyadharshini et al., 2022) for only Tamil, including code-mixed and transliterated language, using models such as classical machine learning, deep learning, and transformers. It classified content as general abuse, which encompassed xenophobia, homophobia, and misogyny. Both of these papers do not explicitly mention the advancement of gender-specific abuse detection; rather, they concentrate on the fine-tuning of transformers for the detection of abusive comments.

We have utilized the abusive detection dataset from the DravidianLangTech@NAACL 2025 shared task, which includes two categories: Abusive and Non-abusive for both Tamil and Malayalam languages (Rajiakodi et al., 2025). The dataset is divided into three parts: training, development, and test. The test set does not have labels. Our models predicted the labels for that set. Table 1 provides an overview of the dataset, highlighting inconsistencies in label assignments. Some samples are tagged as 'Abusive', while others appear as 'abusive'. To maintain uniformity, we standardized all labels to 'Abusive' across the dataset.

2*Language	2*Split	Total Samples	2*Abusive	2*Non-Abusive
3*Tamil	Train	2790	1366	1424
	Dev	598	278	320
	Test	598	-	-
3*Malayalam	Train	2933	1531	1402
	Dev	629	303	326
	Test	629	-	-

Table 1: Category-wise distribution in the dataset

4 Methodology

This section provides an overview of the methodologies and approaches utilized to build the system using the previously described dataset and different transformer models. Methodology of our work is shown in Figure 1.

4.1 Preprocessing

The dataset is evaluated to determine its distribution and structure. The labels are encoded as Abusive: 1, Non-abusive: 0.

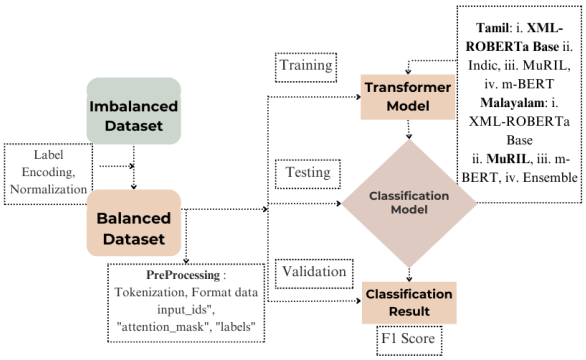

Figure 1: Methodology of our work

Model	lr	bs	ep	wd
XLM-R	2e-5	32	5	0.01
MuRIL	3e-5	32	7	0.1
Indic	2e-5	32	5	-
m-BERT	2e-5	32	5	-
Ensemble	3e-5	32	10	0.1

Table 2: Parameter setting in different model

Label normalization involves transforming variants of the abusive label to a consistent representation (e.g., 'abusive' is mapped to 'Abusive'). The normalization function applies to each sample in the dataset, ensuring label uniformity before encoding. Once standardized, labels are transformed into numerical representations for usage by machine learning models. The dataset is divided into training and validation sets using an 80%-20% ratio.

4.2 Model Selection and Training

We used XLM-RoBERTa, MuRIL, IndicBERT, and mBERT for Tamil, while for Malayalam, we employ XLM-RoBERTa, MuRIL, mBERT, and an ensemble approach. Tokenization is performed using the respective model tokenizers, ensuring uniform sequence lengths through padding and truncation. The dataset is converted into the Hugging Face Dataset format and preprocessed for PyTorch compatibility. During training, the model undergoes regular evaluations, and the best-performing model is saved. To improve efficiency, mixed precision (fp16) training is utilized. A data collator dynamically pads sequences, and evaluation metrics such as accuracy are computed to assess model performance. Early stopping is implemented to prevent overfitting by monitoring validation loss.

4.3 Evaluation and Testing

During model evaluation, we assessed performance using the new dataset for development to fine-tune hyperparameters and ensure optimal performance. Once the model had achieved satisfactory results, we proceeded with the test dataset for the final classification.

We have utilized a new test dataset, which contains unlabeled comments, is used to classify abusive and non-abusive comments. The trained model predicts the labels, distinguishing between abusive and non-abusive content. This ensured the model's ability to generalize effectively to unseen data.

5 Result and Analysis

In this section, we compare the results and analysis the different transformers performance based on some evaluation metrics. The performance of the various methods on the test set is showed in Table 2. The macro F1-score measures the supremacy of the models. However, we also consider other measures such as validation accuracy (Accuracy) and validation loss (Loss) also.

5.1 Parameter Setting

In Table 2, *lr*, *bs*, *ep*, and *wd* represents *learning_rate*, *batch_size*,*epochs*, *weight_decay* respectively. We have tuned different hyperparameters for finding the best model for the corresponding transformers.

5.2 Comparative Analysis

We have found that XLM-R has achieved the highest accuracy with 79% and an F1-score of 0.79 on the validation set for abusive Tamil text detection. For the Malayalam language, the MuRIL based model has given the highest output with 73% accuracy and an F1 score of 0.73 on the validation set. However, We have trained different transformer-based models and an ensemble method incorporating Random Forest (RF), Support Vector Machine (SVM), and MuRIL transformer altogether. However, we did not get optimal output from the ensemble methods and indic-BERT models. The model comparisons are shown in Table 3. XLM-R, MuRIL has given the close output in some cases when hyperparameters are tuned. So, after-all the MuRIL based model works finely for Malayalam whereas XML-R works better for Tamil text.

5.3 Metrics Evaluation

The performance of different model are evaluated by various metrics such as F1-score, Accuracy, Pre-

Tamil Text			
Transformer	**Loss**	**Accuracy**	**F1 Score**
XLM-R	0.3366	79%	**0.79**
MuRIL	0.6933	51%	0.51
Indic	0.9873	44%	0.42
m-BERT	0.8215	53%	0.51

Malayalam Text			
Transformer	**Loss**	**Accuracy**	**F1 Score**
XLM-R	0.4123	61%	0.60
MuRIL	0.3881	73%	**0.73**
m-BERT	0.8940	52%	0.51
Ensemble	0.6928	57%	0.56

Table 3: Comparison of different transformer models and ensemble methods

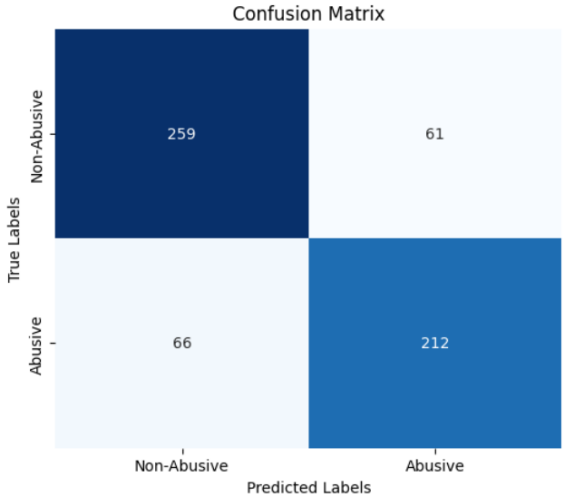

Figure 2: Confusion matrics of XLM-R transformer model for Tamil text

cision, and Recall on the test (Provided dev dataset) set. Figure 3 and Figure 3 show the confusion matrics of XLM-R and Muril based models for Tamil and Malayalam text respectively.

5.4 Error Analysis

Table 4 shows that the XLM-R based model performs well for abusive Tamil text detection and the MuRIL based model for Malayalam. In Table 4, *A*, *P*, *R*, and *F1* represents *Accuracy*, *Precision*, *Recall*, *F1_score* respectively. We have tuned different hyperparameters for finding the best model for the corresponding transformers. But in some cases, the non-abusive text is misclassified as abusive and vice versa. The confusion metrics show them well. These are due to the language morphology and lexical ambiguity, sarcasm, and irony

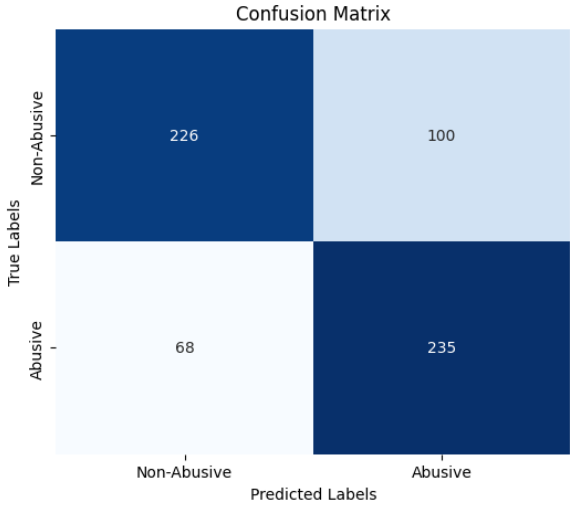

Figure 3: Confusion matrics of MuRIL transformer model for Malayalam text

Lang.	Model	A	P	R	F1
Tamil	XLM-R	79%	0.79	0.79	0.79
Malayalam	MuRIL	73%	0.74	0.73	0.73

Table 4: Evaluation metrics of our best models

when the context of the sentence is ambiguous. Rare words or Dialects may also be another reason for these misclassifications. The evaluation metrics of our best model for corresponding languages are shown in Table 4. Incorporating additional context using hierarchical models could help in better understanding the context. Fine-tuning multilingual transformers in domain-specific corpora may also improve performance.

6 Conclusion

In this research work, we have conducted a comparative study among different types of multilingual transformer and ensemble based techniques for abusive text detection in Tamil and Malayalam-two Dravidian languages. During the training and evaluation of the model, we used the Dravidian-LangTech provided annotated datasets. Although we try to use different transformers and ensemble methods with RF and SVM, MuRIL and XLM-R has given the better output comparing others. Surprisingly, ensemble techniques and Indic-BERT has performed poorly. However, we have tuned some hyperparameters for the model and got decent outputs.

Limitations

While our approach demonstrates better performance, it has certain limitations also. First of all, the provided dataset is quite small. The impact of the dataset on model development is visible in the result and error analysis section. Secondly, our model shows limitations in capturing the sarcasm, irony, or implicit abusive content. As these are low resources languages and due to their native morphology, capturing the context is challenging.

References

Luis Joaquín Arellano, Hugo Jair Escalante, Luis Villaseñor Pineda, Manuel Montes y Gómez, and Fernando Sanchez-Vega. 2022. Overview of da-vincis at iberlef 2022: Detection of aggressive and violent incidents from social media in spanish.

Kim Barker and Olga Jurasz. 2021. Text-based (sexual) abuse and online violence against women: Toward law reform? In *The Emerald International Handbook of Technology-Facilitated Violence and Abuse*, pages 247–264. Emerald Publishing Limited.

Hao Chen, Susan Mckeever, and Sarah Jane Delany. 2017. Harnessing the power of text mining for the detection of abusive content in social media. In *Advances in Computational Intelligence Systems: Contributions Presented at the 16th UK Workshop on Computational Intelligence, September 7–9, 2016, Lancaster, UK*, pages 187–205. Springer.

Antigoni Maria Founta, Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Athena Vakali, and Ilias Leontiadis. 2019. A unified deep learning architecture for abuse detection. In *Proceedings of the 10th ACM conference on web science*, pages 105–114.

Ho-Suk Lee, Hong-Rae Lee, Jun-U Park, and Yo-Sub Han. 2018. An abusive text detection system based on enhanced abusive and non-abusive word lists. *Decision Support Systems*, 113:22–31.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, Malliga Subramanian, Kogilavani Shanmugavadivel, Premjith B, Abirami Murugappan, Sai Prashanth Karnati, Rishith, Chandu Janakiram, and booktitle = Kumaresan, Prasanna Kumar". 2023. Findings of the shared task on Abusive Comment Detection in Tamil and Telugu.

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 Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.