ANSR@DravidianLangTech 2025: Detection of Abusive Tamil and Malayalam Text Targeting Women on Social Media using RoBERTa and XGBoost

Nishanth S, Shruthi Rengarajan, S. Ananthasivan, Burugu Rahul, Sachin Kumar S
Amrita School of Artificial Intelligence, Coimbatore
Amrita Vishwa Vidyapeetham, India
{cb.en.u4aie22149, cb.en.u4aie22154,
cb.en.u4aie22148, cb.en.u4aie22161}@cb.students.amrita.edu
s_sachinkumar@cb.amrita.edu

Abstract

Abusive language directed at women on social media, often characterized by crude slang, offensive terms, and profanity, is not just harmful communication but also acts as a tool for serious and widespread cyber violence. It is imperative that this pressing issue be addressed in order to establish safer online spaces and provide efficient methods for detecting and minimising this kind of abuse. However, the intentional masking of abusive language, especially in regional languages like Tamil and Malayalam, presents significant obstacles, making detection and prevention more difficult. The system created effectively identifies abusive sentences using supervised machine learning techniques based on RoBERTa embeddings. The method aims to contribute to a safer cyber space from abusive language, which is essential for various online platforms -including but not limited to- social media, online gaming services etc. The proposed method has been ranked **8** in Malayalam and **20** in Tamil in terms of *f1* score.

Keywords: Abusive texts, Social Media, Natural Language Processing, RoBERTa, XGBoost

1 Introduction

Social media are found to be the new entertainments, information mediums, and communications alike. However, they also present online harassments by targeting women. The negative consequences such content has for victims have serious psychological, social, and professional impacts that highlight the need for effective tools to detect and mitigate abuse. Abuse appears as hate, abusive, or threatening comments directed at others as deep-rooted societal biases with the intent to promote gender inequality. Therefore, mechanisms are needed for the identification and mitigation of online abuse.

Although many work has been reported on abusive language detection for high-resource languages such as English, little work has been done on low-resource languages (Chakravarthi et al., 2023). Here, two popular South Indian languages- Tamil (Rajalakshmi et al., 2023) and Malayalam are taken(Raphel et al., 2023) -with minimal annotated datasets and tools for Natural Language Processing (NLP). Detecting abusive language is made harder due to various linguistic complexities, code mixing, and dialectal variations; therefore, this becomes an area of concern.

To address this challenge, our team participated in the Shared Task on Abusive Tamil and Malayalam Text Targeting Women on Social Media at DravidianLangTech@NAACL 2025 (Rajiakodi et al., 2025). More details about the shared task can be found at[1].

We used Machine Learning and Natural Language Processing techniques were used to build an automated detection system for abusive content directed at women (Hossain et al., 2022). Our approach is by using pre-trained word embeddings and fine-tuning an XGBoost model to achieve effective classification. We are, therefore, trying to contribute towards safer digital environments and support the efforts of moderation of content in social media with the development of Machine learning models for these low-resource languages.

2 Dataset

The data set was distributed by the shared task organisers of 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (Priyadharshini et al., 2022, 2023).

The dataset used for this study comprises sentences in Tamil and Malayalam languages, categorized into two classes: *Abusive* and *Non-Abusive* (Class distribution). The data set is split into train-

[1] https://codalab.lisn.upsaclay.fr/competitions/20701

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 746–750

Data Type	Total Sentences	Class Distribution
Training	3562	1728 : 1834
Testing	629	303 : 326

Table 1: Dataset Statistics for Malayalam

Data Type	Total Sentences	Class Distribution
Training	3388	1644 : 1744
Testing	598	293 : 305

Table 2: Dataset Statistics for Tamil

ing and test sets. Table.1 and Table.2 show the data distribution of the Malayalam and Tamil classes

It was specifically developed for evaluating the suitability of language models for identifying abusive language in low-resource Dravidian languages, ensuring near-balanced *Abusive* and *Non-Abusive* example representations to provide more efficient training and evaluation.

3 Methodology

The flowchart given in Figure.1 describes the proposed methodology used in the classification of the abusive and non-abusive comments of both Tamil and Malayalam (Kumar et al., 2017).

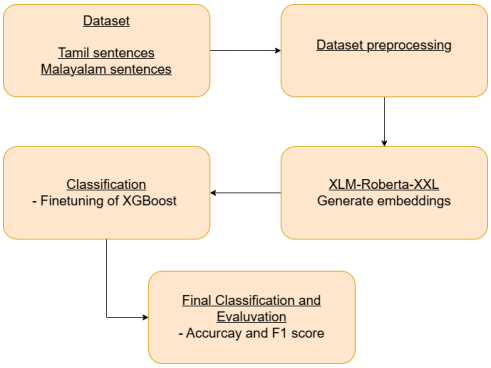

Figure 1: The proposed methodology

3.1 Data Preprocessing

The data preprocessing workflow starts with every dataset going through text cleaning, where we strip URLs, special characters, and unnecessary symbols and convert all text to lowercase to ensure uniformity and to improve model. Additionally, the target class labels (e.g., *abusive* and *non-abusive*) are standardized to lowercase for consistency.

After cleaning, the labels are mapped to numerical values: *abusive* is assigned the value `1`, and *non-abusive* is assigned the value `0`. This numerical encoding is essential for machine learning algorithms, which require numeric inputs for supervised learning tasks. Finally, the training and development datasets are concatenated into a combined dataset to ensure that the model is trained on a more diverse and comprehensive set of examples. Both Tamil and Malayalam datasets are combined as only a single model is developed for this classification task. The train-test split 80-20.

3.2 Feature Extraction and Model Preparation

XLM-RoBERTa-XXL model (Goyal et al., 2021), a state-of-the-art transformer-based model (Vaswani et al., 2017) pre-trained on multilingual corpus was used to get word embeddings for the model to be trained on.It is particularly suited for handling low-resource languages like Tamil and Malayalam. The tokenizer's job is to encode the text data into input features suitable for the transformer model, while the configuration ensures that the model's architecture and hyper-parameters align with the expected outcome, i.e., to classify abusive content. The model configuration details are shown in Table.3.

Property	Details
Parameters	10.7 billion
Number of Layers	48
Embedding Dimensions	4096

Table 3: XLM-RoBERTa-XXL Model Details

3.3 Embedding Generation

Extraction embeddings use the XLM-RoBERTa-XXL model and tokenizer to generate high dimensional vector representations of text data.

For each text sequence, the tokenizer encodes the input by truncating it to a specified maximum length of 512 tokens. The tokens are then passed through the model, and the output embeddings are obtained. A function then extracts the embedding of the first token (CLS token) from the model's output, which represents a summary of the entire sequence. These embeddings are then stored and concatenated into a single tensor, giving the final vector representation for the input data.

3.4 Machine Learning Models

Various machine learning models like logistic regression, K-Nearest Neighbors (KNN) and random forest, were explored (V P et al., 2023; Hasan et al.,

2024; K et al., 2021). The scores of these models are shown in Table.4.

Method	F1 Score
Logistic Regression	0.6618
K-Nearest Neighbors	0.5832
Random Forest	0.6735

Table 4: Scores from Different Models

After carefully testing the performance of each of the different models, XGBoost (Chen and Guestrin, 2016) stood out for the following reasons:

- **Handling Complex Relationships:** Unlike models like logistic regression and Random Forest, XGBoost captures complex patterns and interactions in the data, which is essential when working with high-dimensional word embeddings.
- **Efficiency:** XGBoost is optimised for speed and performance, making it faster to train and more efficient than models like random forest and KNN, especially while training on larger datasets like this.
- **Flexibility in Tuning:** XGBoost has a lot of hyperparameters that can be adjusted to gain better performance, including learning rate, maximum tree depth, and regularization terms.
- **Regularization:** XGBoost has built-in L1 and L2 regularization that helps to prevent overfitting, which is a major cause for concern with other models like SVM when using large embeddings.

These advantages made XGBoost an easy choice for building the pipeline.

3.5 Model Training and Submission

During the initial submission, a single XGBoost model was trained to handle both Tamil and Malayalam text simultaneously. Instead of training separate models for each language, this two birds in one shot approach shared patterns across the two languages that reduced the computational requirements (Koreddi et al., 2025).

3.6 Performance Metrics

The results from this submission was ranked as as shown in Table.5.

Language	Rank	Macro F1 Score
Malayalam	8th	0.6901
Tamil	20th	0.7201

Table 5: Submission Results in Malayalam and Tamil

4 Experimental Results

In order to determine which combination of hyperparameters produced the best outcomes, approximately 324 (36×9, different $n_{\text{estimators}}$) setups were executed. The plotted graph is shown in Figure.2. According to this analysis, the top three configurations were chosen for subsequent testing.

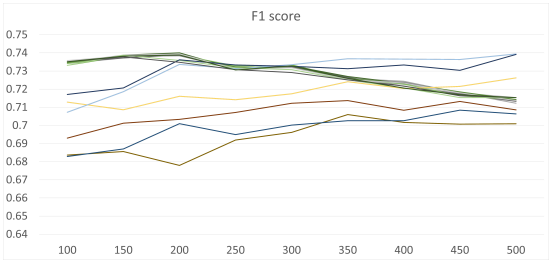

Figure 2: Accuracy Trends Across Experiments

These settings were subsequently tested using a further 30 (3×10, different $n_{\text{estimators}}$) settings to optimize the value of $n_{\text{estimators}}$ for optimal performance without sacrificing computational efficiency. The model configurations are shown in Table.6.

Hyperparameter	Value
Objective	binary:logistic
Max Depth	6
Learning Rate	0.1
Random State	42
Tree Method	hist
Device	cuda
Number of Estimators	1000
Evaluation Metric	error
Booster	dart
Subsample	0.5

Table 6: XGBoost Model Configuration after finetuning

This iterative process helped to find the highest-performing setup, significantly better than the initial baseline, while maintaining the computational speed for which XGBoost is renowned. The best F1 scores given by the model is shown in Figure.3.

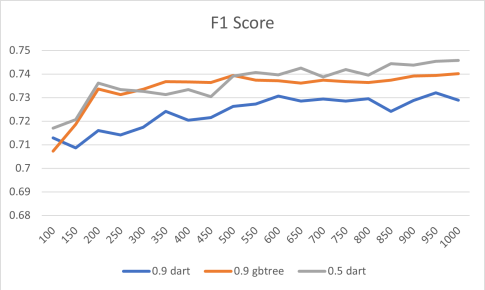

Figure 3: F1 Score of best 3 configurations

5 Evaluation

The performance of the fine-tuned XGBoost model was measured by employing the F1 score as the major indicator. The model was tested and trained with other methodologies to observe how well it would perform when processing Tamil and Malayalam text.

5.1 Comparison with Other Approaches

Prior to finalizing RoBERTa-based embeddings, extraction of embeddings from IndicBERT was attempted. When trained with XGBoost, the IndicBERT embeddings gave an F1 score in the mid-60s (~ 0.65). From this, we had a goal of reaching an F1 score in the 70s by trying more efficient embedding methods.

Shifting towards RoBERTa-based embeddings (XLM-RoBERTa), we saw significant improvement, resulting in an F1 score of **0.71** upon training and validation both on Tamil and Malayalam simultaneously. This attested that RoBERTa embeddings performed better to identify linguistic aspects pertaining to abusive language detection and was subsequently employed for initial submission.

5.2 Final Model Performance

With the optimized XGBoost model based on RoBERTa embeddings, the optimal F1 score obtained is as shown in Table.7.

Data	F1-Score
Tamil and Malayalam combined	0.745
Tamil only	0.775
Malayalam only	0.713

Table 7: Final Output Evaluation: F1 Scores Breakdown

These findings show that although a common model for both languages works quite well, training on each language separately gives a minor performance improvement. This implies that language-specific subtleties may have an effect on classification performance, and additional optimizations like language-aware pre-processing or more feature engineering may improve results (Shubhankar Barman, 2023).

The code files for this project can be accessed from[2]

6 Conclusion

This paper presented the results of the task performed as part of the Fifth Workshop on Speech and Language Technologies for Dravidian Languages on abusive text detection in Tamil and Malayalam dataset on women in social media. The conference provided the dataset for the proposed task. Out of 156 participation and 30 submissions, this proposed method was ranked 8 in Malayalam dataset and 20 in Tamil dataset.

7 Limitations

Despite getting promising results, the methodology has certain limitations that could affect performance and scalability:

- **Large Model Size:** The use of RoBERTa-XXL embeddings significantly increased the computational requirements. Due to the model's large size, standard GPUs were insufficient, and an `NVIDIA A6000` was required to handle the memory load. This makes the approach less accessible for environments with limited hardware resources.

- **High Computational Costs:** Training and fine-tuning models on such large embeddings required extensive computational time and resources, which may not be feasible for all researchers or in production environments.

References

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Shubanker Banerjee, Manoj Balaji Jagadeeshan, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Sean Benhur, and John Philip McCrae. 2023. Detecting Abusive Comments at a Fine-Grained Level in a Low-Resource Language. *Natural Language Processing Journal*, 3:100006.

[2] https://github.com/ANSR-codes/NAACL_Shared_task

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. *CoRR*, abs/1603.02754.

Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. 2021. Larger-Scale Transformers for Multilingual Masked Language Modeling. *CoRR*, abs/2105.00572.

MD. Nahid Hasan, Kazi Shadman Sakib, Taghrid Tahani Preeti, Jeza Allohibi, Abdulmajeed Atiah Alharbi, and Jia Uddin. 2024. OLF-ML: An Offensive Language Framework for Detection, Categorization, and Offense Target Identification Using Text Processing and Machine Learning Algorithms. *Mathematics*, 12(13).

Eftekhar Hossain, Omar Sharif, Mohammed Moshiul Hoque, M. Ali Akber Dewan, Nazmul Siddique, and Md. Azad Hossain. 2022. Identification of Multilingual Offense and Troll from Social Media Memes Using Weighted Ensemble of Multimodal Features. *Journal of King Saud University - Computer and Information Sciences*, 34(9):6605–6623.

Sreelakshmi K, Premjith B, and Soman Kp. 2021. Amrita_CEN_NLP@DravidianLangTech-EACL2021: Deep learning-based offensive language identification in Malayalam, Tamil and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 249–254, Kyiv. Association for Computational Linguistics.

Venkatesh Koreddi, Nalluri Manisha, Shaik Kaif, and Yeligeti Kumar. 2025. Multilingual AI System for Detecting Offensive Content Across Text, Audio, and Visual Media.

S. Sachin Kumar, M. Anand Kumar, and K. P. Soman. 2017. Sentiment Analysis of Tweets in Malayalam Using Long Short-Term Memory Units and Convolutional Neural Nets. In *Mining Intelligence and Knowledge Exploration: 5th International Conference, MIKE 2017, Hyderabad, India, December 13–15, 2017, Proceedings*, page 320–334, Berlin, Heidelberg. Springer-Verlag.

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.

Ratnavel Rajalakshmi, Srivarshan Selvaraj, Faerie Mattins R., Pavitra Vasudevan, and Anand Kumar M. 2023. Hate and Offensive Content Identification in Tamil Using Transformers and Enhanced Stemming. *Computer Speech Language*, 78:101464.

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.

Mariya Raphel, Premjith B, Sreelakshmi K, and Bharathi Raja Chakravarthi. 2023. Hate and Offensive Keyword Extraction from CodeMix Malayalam Social Media Text Using Contextual Embedding. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 10–18, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Mithun Das Shubhankar Barman. 2023. Multimodal Abusive Language Detection and Sentiment Analysis in Dravidian Languages. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*.

Abeera V P, Dr. Sachin Kumar, and Dr. Soman K P. 2023. Social media data analysis for Malayalam YouTube comments: Sentiment analysis and emotion detection using ML and DL models. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 43–51, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *CoRR*, abs/1706.03762.