JAS@DravidianLangTech 2025: Abusive Tamil Text targeting Women on Social Media

B Saathvik
saathvik2210173@ssn.edu.in

Janeshvar Sivakumar
janeshvar2210182@ssn.edu.in

Durairaj Thenmozhi
theni_d@ssn.edu.in

Sri Sivasubramaniya Nadar College of Engineering

Abstract

This paper presents our submission for Abusive Comment Detection in Tamil - DravidianLangTech@NAACL 2025. The aim is to classify whether a given comment is abusive towards women. Google's MuRIL (Khanuja et al., 2021), a transformer-based multilingual model, is fine-tuned using the provided dataset to build the classification model. The dataset is preprocessed, tokenised, and formatted for model training. The model is trained and evaluated using accuracy, F1-score, precision, and recall. Our approach achieved an evaluation accuracy of 77.76% and an F1-score of 77.65%. The lack of large, high-quality datasets for low-resource languages has also been acknowledged.

1 Introduction

Multilingualism has added a new dimension to the issue of abusive language detection despite the increasing number of efforts to prevent abusive content from being shared on social media. Social media users may find it offensive and detrimental to their mental health when other users post abusive comments on videos or in response to the comments of other users. When it comes to low-resource languages such as Tamil, the difficulty is further increased by the lack of available resources (Vegupatti et al., 2024).

Beyond being the official language of the Indian state of Tamil Nadu and the union territory of Puducherry, Tamil is also widely spoken in Malaysia, Mauritius, Fiji, and South Africa. It is also one of the official languages of Singapore and Sri Lanka. Many offensive comments can also be found in these languages on social media and there is a high demand for automated systems for categorizing the offensive and non-offensive remarks on social media comments in regional languages (Rajalakshmi et al., 2023).

In particular, social media platforms are increasingly used to target women with abusive and derogatory comments, reinforcing gender inequalities and societal biases. This form of online abuse can have serious psychological consequences.

This task is part of Abusive Comment Detection in Tamil - DravidianLangTech@NAACL 2025 (Rajiakodi et al., 2025).

The code associated with this task can be accessed through the following GitHub repository: https://github.com/saaaathvik/wise.

2 Related Work

Detecting abusive language in Tamil, particularly content targeting women, is a critical challenge due to limited resources and the complexity of the language. Several studies have explored different approaches to address this issue.

Supervised and unsupervised learning techniques have been compared for Tamil offensive language detection. "Tamil Offensive Language Detection: Supervised versus Unsupervised Learning Approaches" (Balakrishnan et al., 2023) examined traditional machine learning models such as Random Forest, SVM, and AdaBoost, while also applying K-means clustering for unsupervised learning. The results showed that clustering before classification improved detection accuracy, with ensemble models achieving 99.70% and 99.87% accuracy for balanced and imbalanced datasets. Similarly, "HOTTEST: Hate and Offensive Content Identification in Tamil" (Rajalakshmi et al., 2023) explored multiple transformer models, including MuRIL and XLM-RoBERTa, achieving an F1-score of 84% using a majority voting ensemble classifier on Tamil YouTube comments.

Transformer-based approaches have been widely used for abusive comment detection in Tamil. "Mitigating Abusive Comment Detection in Tamil Text: A Data Augmentation Approach

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 148–152

with Transformer Model" (Sheik et al., 2023) demonstrated that applying back translation and lexical replacement improved classification performance, leading to a 15-point increase in macro F1-score over existing baselines. Similarly, "Optimize_Prime@DravidianLangTech-ACL2022" (Patankar et al., 2022) investigated transformer-based models, reporting that MuRIL and XLM-RoBERTa performed best for Tamil data with macro F1-scores of 0.43 for monolingual Tamil and 0.45 for Tamil-English code-mixed data.

The challenge of detecting gendered abuse in Indic languages has been explored in "Breaking the Silence: Detecting and Mitigating Gendered Abuse in Hindi, Tamil, and Indian English Online Spaces" (Vetagiri et al., 2024), where an ensemble CNN-BiLSTM model was trained on a dataset of over 7,600 annotated social media posts. The study ranked first in the ICON 2023 shared task, highlighting the effectiveness of deep learning in handling real-world noisy text with code-switching. Another relevant work, "Brainstormers_msec at SemEval-2023 Task 10: Detection of Sexism-Related Comments in Social Media Using Deep Learning" (Mahibha et al., 2023), leveraged BERT, DistilBERT, and RoBERTa models to classify sexist comments in English social media posts, achieving macro F1-scores of 0.8073, 0.5876, and 0.3729 for sexism detection and classification tasks.

Feature extraction techniques have also been explored to improve detection in Tamil social media comments. "PANDAS@Abusive Comment Detection in Tamil Code-Mixed Data Using Custom Embeddings with LaBSE" (G L et al., 2022) introduced a hybrid approach combining TF-IDF vectorization with language-agnostic LaBSE embeddings, achieving 52% accuracy and an F1-score of 0.54 on Tamil-English code-mixed content. Similarly, "Supernova@DravidianLangTech 2023@Abusive Comment Detection in Tamil and Telugu" (Reddy et al., 2023) applied SVM classifiers with TF-IDF feature extraction to detect abusive content in Tamil, Tamil-English, and Telugu-English datasets. The study implemented preprocessing steps such as stemming, stopword removal, and special character filtering to enhance classification performance.

Additionally, "Abusive Social Media Comments Detection for Tamil and Telugu" (Vegupatti et al., 2024) employed multilingual pre-trained embeddings with BERT, demonstrating that IndicBERT and MuRIL significantly outperformed traditional classifiers for Tamil-English and Telugu-English abusive comment detection.

These studies highlight the importance of transformer models, data augmentation techniques, and customized embeddings in improving abusive language detection for Tamil, particularly in gendered abuse contexts. Our work builds upon these findings by refining abusive Tamil comment classification with a MuRIL-based transformer model, further contributing to this evolving field.

3 Dataset Analysis

The given dataset (Priyadharshini et al., 2022, 2023) comprises 2790 manually annotated YouTube comments in Tamil. In particular, it has 1424 "Non-Abusive" labeled comments and 1366 "Abusive" labeled comments. The comments contain a mix of letters, numbers, symbols, special characters, emojis, emails, and hyperlinks. The data distribution is highlighted in Figure 1 and Table 1.

Category	Count
Non-Abusive	1424
Abusive	1366
Total	**2790**

Table 1: Data description

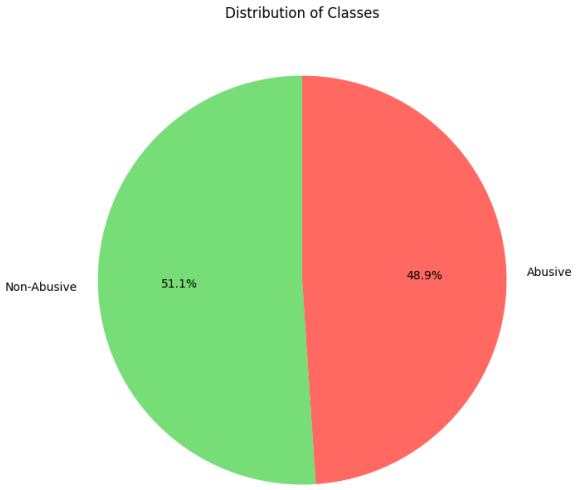

Figure 1: Data distribution

The definition and usage of stop-words is crucial for the effectiveness of such datasets. While stop-words lists for languages such as English and Spanish have been implemented in the nltk.corpus

library, the manual creation of a stop-words list for Tamil was required (Reddy et al., 2023).

This manually curated Tamil stop-words list is publicly available online [1] (see Figure 2).

'அங்கு', 'அங்கே', 'அடுத்த', 'அதற்கு', 'அதனால்', 'அதன்', 'அதிக', 'அதில்', 'அது', 'அதே', 'அதை', 'அந்த', 'அந்தக்', 'அந்தப்', 'அல்லது', 'அவரது', 'அவர்', 'அவர்கள்', 'அவள்', 'அவன்', 'அவை', 'அன்று', 'ஆகிய', 'ஆகியோர்', 'ஆகும்', 'ஆனால்', 'இங்கு', 'இங்கே', 'இடத்தில்', 'இடம்', 'இதற்கு', 'இதனால்', 'இதனை', 'இதன்', 'இதில்', 'இது', 'இதை', 'இந்த', 'இந்தக்', 'இந்தத்', 'இந்தப்', 'இப்போது', 'இரு', 'இருக்கும்', 'இருந்த', 'இருந்தது', 'இருந்து', 'இல்லை', 'இவர்', 'இவை', 'இன்னும்', 'உள்ள', 'உள்ளது', 'உள்ளன', 'உன்', 'எந்த', 'எல்லாம்', 'என', 'எனக்', 'எனக்கு', 'எனப்படும்', 'எனவும்', 'எனவே', 'எனினும்', 'எனும்', 'என்', 'என்பது', 'என்பதை', 'என்ற', 'என்று', 'என்றும்', 'என்ன', 'என்னும்', 'ஏன்', 'ஒரு', 'ஒரே', 'ஓர்', 'கொண்ட', 'கொண்டு', 'கொள்ள', 'சற்று', 'சில', 'சிறு', 'சேர்ந்த', 'தவிர', 'தனது', 'தன்', 'தான்', 'நாம்', 'நான்', 'நீ', 'பல', 'பலரும்', 'பல்வேறு', 'பற்றி', 'பற்றிய', 'பிற', 'பிறகு', 'பின்', 'பின்னர்', 'பெரும்', 'பேர்', 'போது', 'போல', 'போல்', 'போன்ற', 'மட்டுமே', 'மட்டும்', 'மற்ற', 'மற்றும்', 'மிக', 'மிகவும்', 'மீது', 'முதல்', 'முறை', 'மேலும்', 'மேல்', 'யார்', 'வந்த', 'வந்து', 'வரும்', 'வரை', 'வரையில்', 'விட', 'விட்டு', 'வேண்டும்', 'வேறு'.

Figure 2: Tamil stop-words list

4 Methodology

4.1 Preprocessing

Preprocessing of data is done to improve the efficiency of the model. The performance metrics of a model could vary drastically with efficient data preprocessing. The different steps involved in preprocessing of data are listed below (Reddy et al., 2023).

1. ***Removal of Numbers, Hyperlinks, Email Addresses, and Emojis:*** These elements do not contribute to the classification of a comment as abusive or non-abusive and are therefore removed.

2. ***Conversion of English Characters to Lowercase While Preserving Tamil Script:*** This ensures consistency in the text while maintaining the integrity of the Tamil script.

3. ***Removal of Special Characters, Punctuation, and Normalizing Spaces:*** This step enhances text consistency while preserving its meaning.

4. ***Removal of Stop Words:*** Stop words refer to frequently occurring words that lack substantial semantic meaning or contribute minimally to the holistic comprehension of a given text. By eliminating these words, the data payload is reduced, resulting in expedited processing durations and enhanced computational efficacy (Reddy et al., 2023).

4.2 Tokenization

Tokenization is a crucial preprocessing step that converts textual data into a numerical format that machine learning models can process. The "google/muril-base-cased" tokenizer (Khanuja et al., 2021) from Hugging Face's Transformers library is used. It breaks down each comment into tokens and converts them into numerical representations. Additionally, it applies truncation to ensure that input sequences do not exceed the specified length (128) and uses padding to standardize input lengths across all samples.

4.3 Transformer Model

The MuRIL (Multilingual Representations for Indian Languages) model (Khanuja et al., 2021) is a transformer-based architecture developed by Google as part of their multilingual research efforts. Its is a BERT model pre-trained on 17 Indian languages and their transliterated counterparts. By using layers of self-attention, it captures contextual relationships between words in a sentence, which is essential for tasks such as text classification. In this specific task, MuRIL's pre-trained knowledge is fine-tuned for binary classification to distinguish between abusive and non-abusive comments. Its ability to understand the linguistic and cultural context of the text makes it particularly effective for the given dataset.

The fine-tuning process for the model was carried out using the Hugging Face Trainer API. The model was trained for 5 epochs, allowing it to learn from the dataset across multiple passes. A batch size of 16 was used for both training and evaluation to ensure efficient processing while maintaining a balance between memory usage and model convergence. The training process was designed to automatically save the best-performing model based on its performance on the validation set. This was done by evaluating the model after each training epoch and selecting the version that achieved the highest F1-score, ensuring that the final model used for predictions would be the one that performed most effectively during training.

[1] Tamil Stop-Words List on GitHub

Model	Precision	Recall	F1-Score	Accuracy
Logistic Regression	0.68	0.68	0.68	0.68
SVM	0.70	0.70	0.70	0.70
LinearSVC	0.68	0.68	0.68	0.68
XGBoost	0.65	0.66	0.65	0.66
MuRIL	**0.73**	**0.83**	**0.78**	**0.78**

Table 2: Evaluation results

4.4 Alternative Classification Models

This study also explored alternative classification models, including Logistic Regression, Support Vector Machines (SVM), LinearSVC, and XGBoost. Logistic Regression serves as a foundational model for binary classification, while SVM and LinearSVC are kernel-based methods well-suited for high-dimensional feature spaces. XGBoost, an ensemble method utilizing gradient boosting, is known for its robust performance and efficiency. All models were trained on the TF-IDF (Term Frequency–Inverse Document Frequency, a statistical measure of how important a word is in a collection of text or document) transformed training data and subsequently evaluated on the validation set using classification reports.

5 Results and Analysis

The evaluation is based on Precision, Recall, F1-score, and Accuracy.

Recall measures the classifier's ability to correctly identify positives, while Precision indicates the accuracy of positive predictions.

$$Recall = \frac{TP}{TP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

The F1-score is a crucial metric in machine learning that provides a balanced measure of a model's precision and recall. The F1-score formula is derived from the harmonic mean of precision and recall.

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{3}$$

Accuracy is the proportion of all classifications that were correct, whether positive or negative.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

MuRIL consistently outperformed the other models across evaluation metrics (See Table 2). Its exceptional accuracy and ability to capture nuanced text features made it the preferred model for this study.

The evaluation results indicate that the model performed well in terms of both Precision and Recall. The F1-score was calculated at 0.7765.

For the test dataset, the model was ranked 10th in the task with an F1-score of 0.7687. This performance demonstrates the model's effectiveness in detecting abusive comments, although there is still potential for improvement.

6 Limitations

Our model's performance is influenced by dataset biases, limiting generalization across diverse scenarios. Architectural choices and loss functions may not fully capture real-world complexities, affecting robustness. Low generalizability to real-world scenarios due to the small dataset remains a challenge. Ethical concerns, including potential bias in AI decisions, require continuous monitoring. High computational demands pose scalability challenges.

7 Conclusion

This paper presents an effective approach to detecting abusive comments targeting women in Tamil using a fine-tuned MuRIL transformer model with an accuracy of 77.76% and an F1-score of 77.65%. The study highlights challenges in working with small datasets for low-resource languages and emphasizes that improving dataset quality can enhance performance. Despite these limitations, our results demonstrate the potential of transformer-based models for abusive language detection in Tamil. Future improvements, such as advanced data augmentation and fine-tuning, can further enhance performance, contributing to better automated content moderation for underrepresented languages.

References

Vimala Balakrishnan, Vithyatheri Govindan, and Kumanan N. Govaichelvan. 2023. Tamil offensive language detection: Supervised versus unsupervised learning approaches. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(4).

Gayathri G L, Krithika Swaminathan, Divyasri K, Thenmozhi Durairaj, and Bharathi B. 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, Dublin, Ireland. Association for 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.

C. Jerin Mahibha, C. M Swaathi, R. Jeevitha, R. Princy Martina, and Durairaj Thenmozhi. 2023. Brainstormers_msec at SemEval-2023 task 10: Detection of sexism related comments in social media using deep learning. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1114–1120, Toronto, Canada. Association for Computational Linguistics.

Shantanu Patankar, Omkar Gokhale, Onkar Litake, Aditya Mandke, and Dipali Kadam. 2022. Optimize_Prime@DravidianLangTech-ACL2022: Abusive comment detection in Tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 235–239, Dublin, Ireland. Association for Computational Linguistics.

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. Hottest: 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.

Ankitha Reddy, Pranav Moorthi, and Ann Maria Thomas. 2023. Supernova@DravidianLangTech 2023@abusive comment detection in Tamil and Telugu - (Tamil, Tamil-English, Telugu-English). In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 225–230, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Reshma Sheik, Raghavan Balanathan, and Jaya Nirmala S. 2023. Mitigating abusive comment detection in Tamil text: A data augmentation approach with transformer model. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 460–465, Goa University, Goa, India. NLP Association of India (NLPAI).

Mani Vegupatti, Prasanna Kumar Kumaresan, Swetha Valli, Kishore Ponnusamy, Ruba Asoka Chakravarthi, and Sajeetha Thavaresan. 2024. *Abusive Social Media Comments Detection for Tamil and Telugu*, pages 174–187.

Advaitha Vetagiri, Gyandeep Kalita, Eisha Halder, Chetna Taparia, Partha Pakray, and Riyanka Manna. 2024. Breaking the silence detecting and mitigating gendered abuse in hindi, tamil, and indian english online spaces. *arXiv preprint arXiv:2404.02013*.