CUET_Ignite@DravidianLangTech 2025: Detection of Abusive Comments in Tamil Text Using Transformer Models

MD. Mahadi Rahman , Mohammad Minhaj Uddin and **Mohammad Shamsul Arefin**
Department of Computer Science and Engineering
Chittagong University of Engineering & Technology, Chattogram-4349, Bangladesh
{u1904094, u1904118}@student.cuet.ac.bd, sarefin@cuet.ac.bd

Abstract

Abusive comment detection in low-resource languages is a challenging task particularly when addressing gender-based abuse. Identifying abusive language targeting women is crucial for effective content moderation and fostering safer online spaces. A shared task on abusive comment detection in Tamil text organized by DravidianLangTech@NAACL 2025 allowed us to address this challenge using a curated dataset. For this task, we experimented with various machine learning (ML) and deep learning (DL) models including Logistic Regression, Random Forest, SVM, CNN, LSTM, BiLSTM and transformer-based models such as mBERT, IndicBERT, XLM-RoBERTa and many more. The dataset comprised of Tamil YouTube comments annotated with binary labels, Abusive and Non-Abusive capturing explicit abuse, implicit biases and stereotypes. Our experiments demonstrated that XLM-RoBERTa achieved the highest macro F1-score(0.80), highlighting its effectiveness in handling Tamil text. This research contributes to advancing abusive language detection and natural language processing in low-resource languages particularly for addressing gender-based abuse online.

1 Introduction

Social media platforms have become integral to modern communication, offering spaces for individuals to express opinions, share experiences and engage in public discourse. However these platforms are also increasingly plagued by abusive content including hate speech, harassment and gender-based violence (Pannerselvam et al., 2023). Among the many languages used on social media, Tamil, a Dravidian language spoken by over 80 million people worldwide has seen a rise in abusive text targeting women (Chakravarthi et al., 2021). This phenomenon not only perpetuates gender-based discrimination but also poses significant challenges for natural language processing (NLP) systems tasked with detecting and reducing such content (Rajiakodi et al., 2025). The detection of abusive language in Tamil is particularly complex due to the language's rich morphology, code-mixing with English and other languages (Priyadharshini et al., 2022). Moreover cultural and contextual nuances often make it difficult for automated systems to accurately identify abusive content without misclassifying neutral text (Shanmugavadivel et al., 2022).

In our participation on Abusive Tamil and Malay alam Text Targeting Women on Social Media at DravidianLangTech@NAACL 2025 (Priyadharshini et al., 2023), we explored different models to detect abusive comments targeting women and addressed this problem with two significant contributions.

- Investigated the performance of various ML, DL and transformer-based models for detecting abusive comments.
- In particular leveraged the transformer-based XLM-RoBERTa model which demonstrated strong performance for abusive language detection in Tamil text.

This research shows that advanced models such as transformers can improve the detection of offensive comments in Tamil language. For more details, our code is available at https://github.com/MHD094/Abusive-Tamil.

2 Related Work

The detection of abusive language on social media particularly in low-resource languages like Tamil remains a critical yet underexplored area within NLP. While significant progress has been made in high-resource languages such as English, Tamil presents unique challenges including its rich morphology, informal writing styles and the prevalence

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 470–475

of code-mixing. Early research in abusive language detection focused on rule-based and machine learning methods using lexical features such as n-grams and part-of-speech tags (Rajalakshmi et al., 2022). With the advent of deep learning models and transformer-based architectures such as BERT (Devlin et al., 2019) have significantly improved abusive language detection.

In the domain of abusive language detection, (Ghanghor et al., 2021) presented a study on identifying offensive language and classifying memes in Dravidian languages particularly Tamil, Malayalam and Kannada. Gender-targeted abuse influenced by cultural nuances in Tamil remains a significant challenge. (Gong et al., 2021) tackle heterogeneous abusive language by introducing a YouTube dataset with sentence-level annotations. They propose a supervised attention model with multi-task learning, improving nuanced abuse detection. Their approach highlights the need for finer-grained annotations, relevant to Tamil abuse detection. (Mohan et al., 2025) introduced the Multimodal Tamil Hate (MATH) dataset, categorizing hate speech into offensive, sexist, racist and casteist types. This study emphasizes the need for culturally informed approaches to improve hate speech detection in Tamil.

The DravidianLangTech shared tasks (Chakravarthi et al., 2022) have further advanced research on NLP for Tamil by providing datasets for offensive language detection and sentiment analysis. These tasks have been instrumental in addressing challenges specific to Dravidian languages such as code-mixing and the compound nature of the languages. (Shanmugavadivel et al., 2022) demonstrated the effectiveness of machine learning for the sentiment analysis in Tamil code-mixed data. In spite of these advancements, challenges like the lack of large annotated datasets and the informal nature of social media writing persist. (Vetagiri et al., 2024) highlighted the need for collaborative efforts to overcome these barriers and improve abusive language detection in Tamil and other low-resource languages.

3 Task and Dataset Description

Abusive language targeting women has become a significant issue with the rise of social media, often reflecting societal biases and gender imbalances. This shared task focuses on abusive text detection in Tamil comments. It aims to identify whether a given comment contains abusive language directed at women or not. The dataset (Priyadharshini et al., 2023) and also (Priyadharshini et al., 2022) for this task comprises social media comments collected from YouTube discussions on controversial and sensitive topics where gender-based abuse is prevalent. This dataset supports accurate classification of abusive content into two binary classes as outlined below:

Abusive: Comments containing harmful or offensive language.

Non-Abusive: Comments free of offensive or abusive content.

Here, Table 1 provides the distribution of samples across training, validation and test sets. The dataset

Classes	Train	Valid	Test
Abusive	1366	278	305
Non-Abusive	1424	320	293
Total	2,790	598	598

Table 1: Dataset distribution.

is almost balanced with the Abusive class having 1,949 samples compared to 2,037 samples for the Non-Abusive class. The total dataset comprises 3,986 text which divided into training (2,790), validation (598) and test (598) sets.

4 Methodology

This section provides a concise summary of the methods and approaches adopted to address the problem outlined earlier. After thorough analysis, the transformer-based model XLM-RoBERTa demonstrates superior performance in our task. Figure 1 shows a visual representation of our methodology, highlighting the essential steps in the proposed approach.

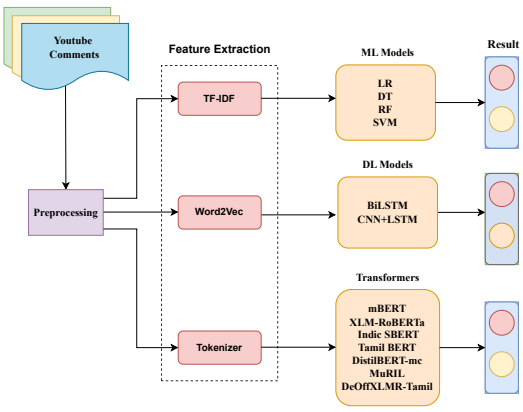

Figure 1: An abstract view of the proposed methodology

4.1 Preprocessing

Basic preprocessing steps such as removing special characters, emojis, punctuation and extra spaces were applied to clean the text. Indic-transliteration (Kunchukuttan, 2020) library used to convert code-mixed Tamil text into standardized Tamil for linguistic consistency and model compatibility.

4.2 Feature Extraction

To capture various features for different model types, three feature extraction techniques were applied. Machine learning models employ Term Frequency-Inverse Document Frequency (TF-IDF) (Salton and Buckley, 1988) to represent text features. Deep learning models utilize word embeddings generated through the Word2Vec approach (Mikolov et al., 2013) for richer semantic information. Transformer-based models employ specialized tokenizers compatible with their architectures to efficiently process input sequences.

4.3 Model Building

In our research, we examined several ML, DL and transformer-based models.

4.3.1 ML models

We trained traditional ML models such as Logistic Regression (LR), Decision Trees (DT), Random Forest (RF) and Support Vector Machines (SVM) on feature representations like TF-IDF. These models rely on statistical patterns but may face challenges in understanding complex contextual relationships in code-mixed Tamil text.

4.3.2 DL models

The deep learning models include BiLSTM and a hybrid CNN+LSTM model. These models capture the semantic structure of the text and dependencies in code-mixed Tamil using pre-trained word embeddings. Each DL model was trained for 5 epochs with a batch size of 32.

4.3.3 Transformer-based models

The transformer-based models include mBERT (Ram et al., 2024), XLM-RoBERTa (Conneau et al., 2020), Indic SBERT (Farsi et al., 2024), Tamil BERT (Raihan et al., 2024), DistilBERT-mc (Rajalakshmi et al., 2023), MuRIL (Khanuja et al., 2021), and DeOffXLMR-Tamil. Fine-tuned on our dataset with transformer-specific tokenizers, these models excel at capturing long-range dependencies and context. They improve accuracy in Tamil abusive language detection by utilizing pre-trained knowledge from multilingual datasets, handling regional dialects and code-switching effectively.

5 Results & Discussion

This section presents a comparative analysis of the performance achieved by various machine learning, deep learning and transformer-based methods for detecting abusive comments in Tamil. The evaluation highlights the effectiveness of different classifiers in predicting abusive content. Additionally, m-BERT and XLM-RoBERTa were fine-tuned by optimizing learning rates, batch sizes and epochs while maintaining the AdamW optimizer as summarized in Table 2.

Hyperparameters	**m-BERT**		**XLM-RoBERTa**	
Optimizer	AdamW	AdamW	AdamW	AdamW
Learning rate	1e-05	2e-05	3e-05	1e-05
Epochs	12	8	8	12
Batch size	32	16	16	32

Table 2: Summary of optimized hyperparameters

We fine-tuned hyperparameters including learning rates, batch sizes and epochs to improve model performance. Table 3 presents precision (P), recall (R) and macro-F1 (MF1) scores on the test dataset. Among machine learning models, Logistic Regression (LR) performed best with an MF1 of 0.71 followed by SVM (0.69). In deep learning, BiLSTM and CNN+LSTM scored 0.45 and 0.33 respectively. Transformer models outperformed all with XLM-RoBERTa achieving the highest MF1 of 0.80 followed by m-BERT, MuRIL and Indic SBERT (0.77). Tamil BERT and DistilBERT-mc showed competitive performance. XLM-RoBERTa was the most effective for Tamil abusive comment detection.

Classifier	**P**	**R**	**MF1**
LR	0.71	0.74	0.71
DT	0.61	0.61	0.61
RF	0.67	0.66	0.66
SVM	0.69	0.69	0.69
BiLSTM	0.55	0.52	0.45
CNN + LSTM	0.24	0.49	0.33
mBERT	0.77	0.77	0.77
XLM-RoBERTa	0.80	0.80	0.80
Indic SBERT	0.77	0.76	0.76
Tamil BERT	0.67	0.68	0.68
DistilBERT-mc	0.75	0.74	0.74
MuRIL	0.77	0.77	0.77
DeOffXLMR-Tamil	0.76	0.76	0.76

Table 3: Results of several models on the test dataset

5.1 Quantitative Discussion

Figure 2 represents the confusion matrix for our XLM-RoBERTa model. The results highlight the effectiveness of transformer-based architectures especially XLM-RoBERTa in identifying abusive Tamil text. The model effectively classifies 245 Non-Abusive (label-0) and 234 Abusive (label-1) instances which demonstrates strong performance. However some misclassifications occur, 48 label-0 samples are predicted as label-1 and 71 label-1 samples are incorrectly classified as label-0. These results highlight the model's overall effectiveness.

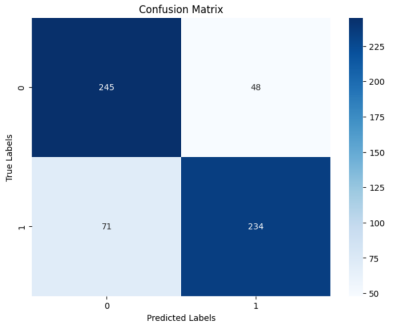

Figure 2: Confusion matrix of best performing model

5.2 Qualitative Discussion

Figure 3 displays sample predictions from our XLM-RoBERTa model. Samples 2, 3, 5, 6 and 8 are correctly classified which demonstrates the model's ability to process diverse linguistic constructs in Tamil text. However, challenges remain with samples 1, 4, and 7 which are misclassified as 0 instead of 1 due to implicit abuse, sarcasm and contextual complexity in Tamil language.

Sample Text	Actual	Predicted
சுகந்தி பொய் சொல்லுறா மோடம் பொய் சொல்லுறா (Is Sugandhi lying? Is Modem lying?)	1	0
மேடம், இப்போ பேசறதையும் நம்பாதீங்க, இதுவும் கண்டண்ட் தான் இருக்கும் (Madam, don't believe what I'm saying now, this too will be a conspiracy.)	0	0
என் வாழ்வில் இப்படி ஒரு கட்சியை பார்த்து பார்த்து சுமார் 30 தடவை பார்த்து பார்த்து சிறிது சிறிது ருசித்து சிரிப்பு வந்து விட்டது. சகோதரி லட்சுமி மிகவும் அருமையாக பேட்டி காண்பது சிரிப்பு வருது. இந்த திவ்யா ஒரு கோமாளி பெண். சுகந்தி பேச்சில் உண்மை தெரியுது. சகோதரி லட்சுமி நல்ல திரமையான ஒரு தொகுப்பாளர். கார்த்திக் எங்கே எங்கே எங்கே...பாவம் சுகந்தி (I have watched such a party in my life about 30 times, tasting it little by little, and it made me laugh. Sister Lakshmi gives an excellent interview, which is funny to watch. This Divya is a complete clown. The truth is evident in Sugandhi's speech. Sister Lakshmi is a very talented host. Karthik, where are you, where are you, where are you... Poor Sugandhi!)	0	0
பேய்க்கு பேய்க்கும் சண்ட அத ஊரே வேடிக்க பாக்குது (The city is a place where ghosts and ghosts are hunted.)	1	0
இப்படி இவங்க இல்லனா உங்களுக்கு வேலை இல்லை (Without these people, you wouldn't have a job.)	0	0
Divya எதுக்கு சுகன்யா எதிர்க்க என்ன காரணம் பாலா வுக்காக பாலா யாரு மதனுடய தம்பி...mathan யாரு அவரு mari Selva குழந்தை ய தப்ப பேசுனா...mariselva யாரு அவரு என்ன திட்டுநாரு...evlovo பிரச்சனை பிரச்சனை இருக்கு நாட்டுல Laxmi mam Enna ithu.... (Divya, why is Sukanya opposing Bala? Bala is the brother of Mathan...who is Mathan? Mari Selva, is the child talking nonsense...who is Mariselva? What is he scolding...evlovo, there is a problem in the country, Laxmi mam, Enna ithu....)	0	0
வாவ் சூப்பர் தூக்கி போட்டு மிதிக்க கூட்டங்களே அதே பெரிய விஷயம் (Wow, super, the crowds are the same big thing.)	1	0
இந்த பெண்ணுக்கு தரமான முடிவு அந்த ஊர் ஆம்பளங்க கையில இருக்கு. (The quality of this woman's life is in the hands of the town's mayor.)	1	1

Figure 3: Examples of the XLM-RoBERTa model predicted outputs

6 Conclusion

This research provides a comprehensive comparison of various machine learning, deep learning and transformer-based methods for detecting abusive comments in Tamil. The evaluation demonstrated that transformer models especially XLM-RoBERTa outperformed all other methods by achieving the highest macro-F1 score of 0.80 followed by m-BERT, MuRIL and Indic SBERT(.77). Among machine learning models, Logistic Regression showed the best performance with an MF1 of 0.71. Despite these advancements, there remain several areas for future research. Expanding datasets to include more diversity will improve model generalizability. Additionally integrating multimodal data such as images and videos could further enhance the detection of abusive content. Lastly, future research should focus on exploring the sociocultural factors driving gender-targeted abuse and developing interventions to address these issues at their core.

Limitations

One of the key limitations of this work stems from the challenges involved in preprocessing code-mixed Tamil text. The conversion of mixed language content through transliteration techniques may not fully capture all cultural context inherent in Tamil comments. This approach can lead to inaccuracies in identifying abusive language as the model may struggle with code-switching or non-standard expressions used in these comments. Furthermore the lack of sufficient high-quality annotated data for abusive language detection in low-resource languages such as Tamil restricts the model's effectiveness. The presence of implicit bias and regional dialect variations within the Tamil language adds another layer of complexity that could impact performance.

References

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Subalalitha Cn, Sangeetha S, Malliga Subramanian, Kogilavani Shanmugavadivel, Parameswari Krishnamurthy, Adeep Hande, Siddhanth U Hegde, Roshan Nayak, and Swetha Valli. 2022. Findings of the shared task on multi-task learning in Dravidian languages. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 286–291, Dublin, Ireland. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Navya Jose, Anand Kumar M, Thomas Mandl,

Prasanna Kumar Kumaresan, Rahul Ponnusamy, Hariharan R L, John P. McCrae, and Elizabeth Sherly. 2021. Findings of the shared task on offensive language identification in Tamil, Malayalam, and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 133–145, Kyiv. Association for Computational Linguistics.

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.

Salman Farsi, Asrarul Eusha, Jawad Hossain, Shawly Ahsan, Avishek Das, and Mohammed Moshiul Hoque. 2024. CUET_Binary_Hackers@DravidianLangTech EACL2024: Hate and offensive language detection in Telugu code-mixed text using sentence similarity BERT. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 193–199, St. Julian's, Malta. Association for Computational Linguistics.

Nikhil Ghanghor, Parameswari Krishnamurthy, Sajeetha Thavareesan, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2021. IIITK@DravidianLangTech-EACL2021: Offensive language identification and meme classification in Tamil, Malayalam and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 222–229, Kyiv. Association for Computational Linguistics.

Hongyu Gong, Alberto Valido, Katherine M. Ingram, Giulia Fanti, Suma Bhat, and Dorothy L. Espelage. 2021. Abusive language detection in heterogeneous contexts: Dataset collection and the role of supervised attention. *Preprint*, arXiv:2105.11119.

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, Vishnu Subramanian, and Partha Pratim Talukdar. 2021. Muril: Multilingual representations for indian languages. *ArXiv*, abs/2103.10730.

Anoop Kunchukuttan. 2020. The IndicNLP Library. https://github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Preprint*, arXiv:1301.3781.

Jayanth Mohan, Spandana Reddy Mekapati, Premjith B, Jyothish Lal G, and Bharathi Raja Chakravarthi. 2025. A multimodal approach for hate and offensive content detection in tamil: From corpus creation to model development. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* Just Accepted.

Kathiravan Pannerselvam, Saranya Rajiakodi, Rahul Ponnusamy, and Sajeetha Thavareesan. 2023. CSS-CUTN@DravidianLangTech:abusive comments detection in Tamil and Telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 306–312, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

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.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Malliga S, Subalalitha Cn, Kogilavani S V, Premjith B, Abirami Murugappan, and Prasanna Kumar Kumaresan. 2023. Overview of shared-task on abusive comment detection in Tamil and Telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 80–87, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Abu Raihan, Tanzim Rahman, Md. Rahman, Jawad Hossain, Shawly Ahsan, Avishek Das, and Mohammed Moshiul Hoque. 2024. CUET_DUO@StressIdent_LT-EDI@EACL2024: Stress identification using Tamil-Telugu BERT. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 265–270, St. Julian's, Malta. Association for Computational Linguistics.

Ratnavel Rajalakshmi, Ankita Duraphe, and Antonette Shibani. 2022. Dlrg@dravidianlangtech-acl2022: Abusive comment detection in tamil using multilingual transformer models. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 207–213. 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.

N Prabhu Ram, T Meeradevi, P Sendhuraharish, S Yogesh, and C VasanthaKumar. 2024. Multi-class emotion classification on tamil and tulu code-mixed text. In *2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS)*, volume 1, pages 231–236. IEEE.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing Management*, 24(5):513–523.

Kogilavani Shanmugavadivel, Sai Haritha Sampath, Pramod Nandhakumar, Prasath Mahalingam, Malliga Subramanian, Prasanna Kumar Kumaresan, and Ruba Priyadharshini. 2022. An analysis of machine learning models for sentiment analysis of tamil code-mixed data. *Computer Speech Language*, 76:101407.

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. *Preprint*, arXiv:2404.02013.