TriVector@DravidianLangTech 2026: Abusive Tamil Text Detection on Social Media Using Lexicon-Augmented Transformers

Oarisa Rebayet, Tahmima Hoque Eid, Fawzia Tabassum, Hasan Murad
Department of Computer Science and Engineering,
Chittagong University of Engineering and Technology, Bangladesh
{u2104129, u2104114, u2104074}@student.cuet.ac.bd, hasanmurad@cuet.ac.bd

Abstract

Abusive comment detection in low-resource languages poses significant challenges, particularly when targeting gender-based abuse on social media platforms. This work presents our system for 'Abusive Tamil text targeting women on social media' at DravidianLangTech@ACL 2026. We introduce nine handcrafted lexicon features integrated with pretrained multilingual transformer embeddings and evaluate their effectiveness in classifying Tamil online comments as abusive or non-abusive. To better understand their impact, we compare model performance with and without these lexical attributes across multiple transformer architectures. Our best-performing model, XLM-RoBERTa-Large, achieved a macro F1-score of 81.71%, securing 15th rank in the competition. The findings indicate that larger multilingual models generalize more effectively to unseen data compared to smaller domain-specific models, while the addition of lexical features yields only mild gains.

1 Introduction

Social media has become a widespread platform for global communication. At the same time, it has emerged as a primary point where people are subject to harassment. Tamil, a widely spoken Dravidian language, has seen a rise in abusive comments toward women. This may lead to severe psychological and social consequences, which emphasizes the urgent need for effective detection mechanisms. However, social media text is highly informal, characterized by frequent multilingual code-switching. This involvement of code-mixed data has introduced a complex subdomain within natural language processing (NLP). Building automated systems to detect abusive Tamil text is difficult due to the language's intricate morphological features and English code-mixing. This often causes models to misclassify neutral text as abusive.

To solve this challenge, we have taken part in Abusive Tamil Text Targeting Women on Social Media at DravidianLangTech@ACL 2026 (Rajiakodi et al., 2026), which aims to classify abusive text. This work conducts a systematic contrastive evaluation of whether statistically derived lexical features continue to add value on top of large multilingual transformers in a low-resource, code-mixed setting. Our contributions are as follows:

- We propose a lexicon-augmented transformer architecture that fuses contextual representations from pretrained language models with nine lexical features.
- We compare four pretrained models of varying scale and domain backgrounds across multiple system configurations.

2 Literature Review

The detection of abusive content in the Tamil language remains an underexplored area in NLP. Notable progress has been achieved in high-resource languages, but Tamil continues to pose significant challenges due to its complex morphological structure.

Early research on abusive language detection in Tamil relied on traditional machine learning approaches, which included models like SVM and Logistic regression (S et al., 2025). Further work compared supervised and unsupervised methods for Tamil offensive language detection (Balakrishnan et al., 2023). The DravidianLangTech shared tasks (Chakravarthi et al., 2021) played a crucial role in advancing NLP research for Tamil by providing benchmarked datasets. Transformer-based models like BERT, m-BERT, and XLNET achieved a weighted F1-score of 96% on Tamil–English code-mixed data, compared to 59% on pure Tamil text (B and Varsha, 2022).

Multiple multilingual transformer models were evaluated on the DravidianLangTech benchmark,

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 420–428

where mBERT achieved the best performance on the monolingual Tamil dataset, obtaining a weighted F1-score of 70% (Rajalakshmi et al., 2022). It was further found that combining language-specific stemming with transformer embeddings such as XLM-RoBERTa and MuRIL meaningfully improves offensive content identification in Tamil (Rajalakshmi et al., 2023).

More recently, a variety of architectures like CNN, BiLSTM, and XLM-RoBERTa have been explored for Tamil detection (Rahman et al., 2025). A hybrid model structure has been introduced that combines transformer embeddings with attention mechanisms (Kodali et al., 2025). In the broader Dravidian language context, Ghanghor et al. (2021) presented a study on offensive language identification and meme classification in Tamil, Malayalam, and Kannada. Gong et al. (2021) handled heterogeneous abusive language by introducing a YouTube dataset with sentence-level annotations and a supervised attention model with multi-task learning. Despite these advancements, Vetagiri et al. (2024) pointed out challenges such as the lack of large annotated datasets and the informal nature of social media text, and then emphasized the need for more effective approaches to accurately detect offensive comments in Tamil.

Maheshwari et al. (2025) highlight the role of domain-aware multilingual lexicon generation in capturing context-sensitive word semantics in low-resource languages.

While most existing studies have focused on improving contextual representations using larger pretrained transformers, comparatively little attention has been paid to whether statistically derived lexical features can still offer additional benefits. By conducting such a contrastive evaluation, we explore not only the overall differences in model performance but also the extent to which lexical features continue to make meaningful contributions, especially in cases where even large contextual embeddings may miss subtle but important cues.

3 Data Description

Aiming to classify abusive comments directed toward women, the task organizers created an approximately balanced dataset comprising Tamil online comments. This dataset enables precise classification of content into the categories: Abusive and Non-Abusive. Beyond the class distribution, we also examine the linguistic characteristics of the dataset. Basic corpus statistics such as total and unique word counts, the proportion of Tamil and English tokens and the average word count provide a better understanding of the dataset.

Table 1 presents the distribution and key corpus statistics across the training, validation, and test set.

Statistic	Train	Validation	Test
Total Samples	2,921	731	913
Total Words	42,564	10,322	12,745
Unique Words	15,401	5,342	6,266
Tamil Word Count	40,192	9,729	11,939
English Word Count	1,471	394	536
Avg Word Count	14.57	14.12	13.96

Table 1: Corpus Statistics.

4 Methodology

This section presents a concise overview of the methods and approaches employed to address the problem described earlier. Following extensive analysis, the transformer-based model XLM-RoBERTa-Large exhibited the best performance for our task. Figure 1 provides a visual representation of the proposed methodology, highlighting the key steps in the approach. The implementation and source code are publicly available on GitHub.[1]

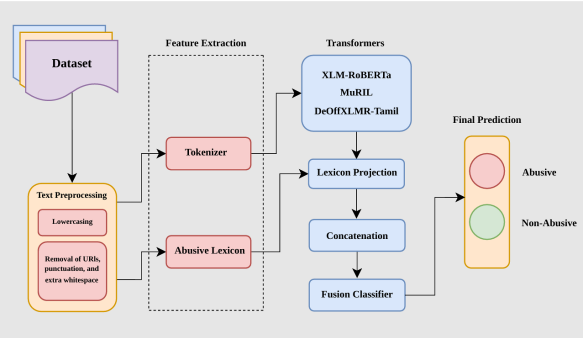

Figure 1: An abstract view of the proposed lexicon-augmented transformer system.

4.1 Text Preprocessing

Basic text preprocessing was performed to clean the dataset. This included lowercasing all text, removal of URLs, punctuation, and special characters. The same normalization procedure was applied consistently across all data splits.

[1] https://github.com/OarisaR/TriVector-abusive-tamil-detection

4.2 Lexicon Construction

We use Pointwise Mutual Information (PMI) to quantify how strongly each word is associated with the abusive class relative to the non-abusive class. For every word w occurring in at least 3 abusive documents:

$$\text{PMI}(w) = \log \frac{P(w \mid \text{Abusive})}{P(w \mid \text{Non-Abusive})} \quad (1)$$

If any word w frequently occurs in abusive comments but rarely found in non-abusive ones, the ratio exceeds 1 and PMI is positive, which indicates w is a discriminative abusive term. The top 150 words by PMI score are chosen for the abusive lexicon.

Severity Scoring

Each lexicon word is assigned a severity score that reflects *how strongly* abusive it is, by rescaling its PMI from $[-3, 3]$ to $[0, 5]$:

$$\text{sev}(w) = \text{clip}\left(\frac{\text{PMI}(w) + 3}{6} \cdot 5,\ 0,\ 5\right) \quad (2)$$

This lets the model distinguish mildly offensive words from strongly abusive ones, rather than treating all lexicon matches equally.

Lexical Feature Vector

Nine scalar features are extracted per text using the abusive lexicon and sentiment word list, as described in Table 2.

Index	Feature	Description
1	Abusive match count	Number of PMI-lexicon tokens present
2	Severity sum	Sum of severity scores of matched tokens
3	Positive word count	Count of positive-sentiment words
4	Negative word count	Count of negative-sentiment words
5	Negation count	Count of negation words
6	Intensifier count	Count of intensifier words
7	Sarcasm heuristic	1 if positive, negative, *and* negation all co-occur
8	Tamil ratio	Fraction of characters in Tamil Unicode block
9	English ratio	Fraction of characters that are ASCII letters

Table 2: Nine lexical features extracted per input

4.3 Model Architecture

We experiment with multiple transformer-based pretrained language models: XLM-RoBERTa-base and large (Conneau et al., 2019), MuRIL (Khanuja et al., 2021), and DeOffXLMR-Tamil (Saha et al., 2021). These models are specifically chosen for their strong multilingual capabilities and effectiveness in handling low-resource and Indian languages. Our primary architecture fuses transformer contextual representations with the lexicon feature vector, as illustrated in Figure 1.

Given input text x, we obtain token-level hidden states from a pretrained transformer backbone $\mathcal{T}$:

$$\mathbf{H} = \mathcal{T}(x) \in \mathbb{R}^{L \times d} \quad (3)$$

where L is the sequence length and $d = 1024$ is the hidden dimension for XLM-RoBERTa large. We apply attention-mask-weighted mean pooling:

$$\mathbf{p} = \frac{\sum_{i=1}^{L} m_i \cdot \mathbf{h}_i}{\sum_{i=1}^{L} m_i} \in \mathbb{R}^d \quad (4)$$

where $m_i \in \{0, 1\}$ denotes the attention mask. The lexicon feature vector $\mathbf{f} \in \mathbb{R}^9$, is projected through a small MLP:

$$\mathbf{v} = \text{ReLU}(\mathbf{W}_{\text{lex}}\mathbf{f} + \mathbf{b}_{\text{lex}}) \in \mathbb{R}^{64} \quad (5)$$

The pooled representation and lexicon projection are concatenated and passed through a fusion classifier:

$$\hat{y} = \mathbf{W}_2 \cdot \text{Dropout}(\text{ReLU}(\mathbf{W}_1[\mathbf{p}; \mathbf{v}] + \mathbf{b}_1)) + \mathbf{b}_2 \quad (6)$$

where $[;]$ denotes concatenation. The fusion hidden size is 128 and dropout rate is 0.2. The model is trained end-to-end with cross-entropy loss. We fine-tune all transformer backbones under a consistent training configuration, with hyperparameters summarized in Table 3.

Hyperparameters	Value
Epochs	4, 7
Learning rate	2×10^{-5}
Weight decay	0.01
Batch size	16

Table 3: Training hyperparameters.

5 Results and Discussion

Table 4 presents the test set performance across all system configurations. XLM-RoBERTa-large without lexicon augmentation achieves the highest macro F1 score of 81.71%, followed by XLM-RoBERTa-base with lexicon augmentation at 80.81%. While the large model offers slightly better performance, it incurs a significantly higher computational cost due to its increased size.

In resource-constrained settings, XLM-RoBERTa-base with lexical augmentation provides a balanced trade off, achieving 81.03% macro F1, only 0.68% lower than the large model, while remaining substantially more efficient.

System	Epochs	Approach	Macro P	Macro R	Macro F1
MuRIL-base	4	−Lex	79.51	79.28	79.32
		+Lex	78.50	78.30	78.34
		Δ	-1.01	-0.98	-0.98
	7	−Lex	78.74	78.70	78.72
		+Lex	79.07	79.10	79.07
		Δ	+0.33	+0.40	+0.35
DeOffXLMR-Tamil	4	−Lex	78.31	78.25	78.27
		+Lex	78.95	78.98	78.96
		Δ	+0.64	+0.73	+0.69
	7	−Lex	77.51	75.72	75.60
		+Lex	77.56	77.46	77.49
		Δ	+0.05	+1.74	+1.89
XLM-R-base	4	−Lex	79.76	78.04	78.00
		+Lex	80.81	80.81	80.81
		Δ	+1.05	+2.77	+2.81
	7	−Lex	81.00	79.50	79.51
		+Lex	79.51	79.47	79.48
		Δ	-1.49	-0.03	-0.03
XLM-R-large	4	−Lex	81.32	80.61	80.66
		+Lex	79.51	79.54	79.51
		Δ	-1.81	-1.07	-1.15
	7	−Lex	81.84	81.82	**81.71**
		+Lex	78.84	78.85	78.84
		Δ	-3.00	-2.97	-2.87

Table 4: Performance of various methods, reported in %

XLM-R-large Achieves Best Performance. XLM-R-large (−Lex) which is trained for 7 epochs achieves the highest Macro F1 of **81.71**% across all systems and settings. This result highlights the strength of large-scale model for abusive Tamil detection, even without additional lexical augmentation.

Lexical Augmentation Shows Inconsistent Gains. The impact of lexical augmentation (+ Lex) varies across models and training durations. For XLM-R-base at 4 epochs, lexical features yield a gain of +2.81% Macro F1, the largest positive increase observed in the entire table. In contrast, both XLM-R-large configurations suffer from lexical augmentation, with drops of 1.15% and 2.87% at 4 and 7 epochs respectively, indicating that larger models already encode sufficient linguistic knowledge and lexical features introduce noise rather than signal.

Models Exhibit Varying Epoch Sensitivity. MuRIL-base remains relatively stable across both augmentation settings. DeOffXLMR-Tamil exhibits the most notable epoch-sensitive behavior: at 7 epochs without lexical features its Macro F1 drops to 75.60%, but lexical features recovers it substantially to 77.49%. XLM-R-base peaks at 7 epochs without augmentation (79.51%), while adding lexical features at that epoch only marginally affects performance (0.03%).

Table 5 presents a few examples from our best-performing system.

Text	Actual	Predicted
அப்பா,அம்மா, அந்த இன்டர்வியூ பண்றவ வக்கிரம்புடிச்சவ.. எல்லோருமே செருப்படி வாங்க தகுதியானவங்க.. அதிலயும் அந்த பொண்ணு.. யப்பா சாமி... முடியலடா. (Appa, amma, that interview-giving crooked fellow... Everyone come slap him, those who deserve it... And that woman... ya pa sami... can't take it da.)	Abusive	Abusive
நேர் கொண்ட பார்வை அடுத்தவங்க மானத்தை வாங்கி தெருவில் விடுவதே உங்க வேலையா போச்சி (So this "straight-forward look" of yours now means grabbing someone else's honour and leaving it in the street? Has that become your job or what?)	Non-Abusive	Non-Abusive
இவ என்ன சட்டம் படிச்ச நீதிபதியா???எல்லாருக்கு நீதி சொல்ல முண்ட,இவளுக்கு ஒருநாள் நிகழ்ச்சி நடத்தப்படும் (Who is she, some law-studied judge??? To sit and tell justice for everyone... For her too one day a program will be conducted.)	Abusive	Abusive
ஒடி பேயிரு கையில் மாடிராத நான் இலங்கை இருக்கன் டிக்கெட் போட்டு வாந்து ஓதய்பன் யார் நீ விஜய் பாத்தி பேசா (Run away. If I get my hands on you, I won't spare you. I'm in Sri Lanka. I'll buy a ticket and come there. Who do you think you are talking about Vijay like that?)	Non-Abusive	Abusive
ஜி.பி .முத்து mind voice. செத்த பயலே நார பயலே.. செத்த மூதேவி ஆளும் மூஞ்சியும் பாரு.. (G.P. Muthu (mind voice): You useless rascal, you rotten fellow… that wretched woman just look at her, look at that face…)	Abusive	Non-Abusive

Table 5: Examples of XLM-RoBERTa-large model predictions.

6 Conclusion

This work compares several transformer-based models for detecting abusive comments in Tamil. We evaluated different system settings across four pretrained architectures. Among them, XLM-RoBERTa-large without any lexicon enhancement achieved the highest macro F1-score of 81.71%, while XLM-RoBERTa-base combined with lexical features reached 80.81%. The findings indicate that larger models tend to generalize better to unseen examples, while handcrafted lexical features are useful for mid-sized models. However, PMI-based lexicon construction carries a risk of named-entity bias, which can hurt precision in certain cases.

Limitations

While the system performs well, it still has practical limitations. Tamil social media text is highly informal, making cultural context difficult to capture automatically. The PMI-based lexicon construction may cause frequently targeted public figures to acquire high PMI scores through corpus co-occurrence rather than inherent abusiveness. Filtering named entities prior to severity scoring is a direction for mitigating this. Additionally, the lexical features are surface-level and largely language-agnostic; incorporating Tamil-specific phenomena such as morphological decomposition and honorific misuse patterns could yield richer signals. Future work should also explore finer-grained labels and stronger generative architectures.

Ethical Considerations

This study utilizes publicly available Tamil social media data released through the shared task, adhering to ethical research standards. User identities were neither tracked nor disclosed, and the data were used exclusively for academic purposes. While our system aims to support safer online environments, automated detection models carry inherent risks, including bias and misclassification, and should not be deployed without human oversight. Particularly, lexicon construction methods based on corpus co-occurrence statistics may encode dataset-level biases, potentially disadvantaging the very demographic groups the system is designed to protect.

References

Bharathi B and Josephine Varsha. 2022. SSNCSE NLP@TamilNLP-ACL2022: Transformer based approach for detection of abusive comment for Tamil language. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 158–164, Dublin, Ireland. Association for Computational Linguistics.

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).

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. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14804–14812.

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

Rohith Gowtham Kodali, Durga Prasad Manukonda, and Maharajan Pannakkaran. 2025. byteSizedLLM@DravidianLangTech 2025: Abusive Tamil and Malayalam text targeting women on social media using XLM-RoBERTa and attention-BiLSTM. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 80–85, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Ayush Maheshwari, Atul Kumar Singh, N J Karthika, Krishnakant Bhatt, Preethi Jyothi, and Ganesh Ramakrishnan. 2025. LexGen: Domain-aware multilingual lexicon generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7364–7375, Vienna, Austria. Association for Computational Linguistics.

MD.Mahadi Rahman, Mohammad Minhaj Uddin, and Mohammad Shamsul Arefin. 2025. CUET_Ignite@DravidianLangTech 2025: Detection of abusive comments in Tamil text using transformer models. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages

392–397, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. 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, 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, Shunmuga Priya Muthusamy Chinnan, Rajalakshmi R., Kathiravan Pannerselvam, Bhuvaneswari Sivagnanam, Jananayagan V, Charmathi Rajkumar, R Ramesh Kannan, and Bharathi Raja Chakravarthi. 2026. From Comments to Harm: A Findings Report on Abusive Tamil Text Targeting Women on Social Media Shared Task- DravidianLangTech@ACL 2026. In *Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Varun Balaji S, Bojja Revanth Reddy, Vyshnavi Reddy Battula, Suraj Nagunuri, and Balasubramanian Palani. 2025. CoreFour_IIITK@DravidianLangTech 2025: Abusive content detection against women using machine learning and deep learning models. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 655–660, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Debjoy Saha, Naman Paharia, Debajit Chakraborty, Punyajoy Saha, and Animesh Mukherjee. 2021. Hate-alert@DravidianLangTech-EACL2021: Ensembling strategies for transformer-based offensive language detection. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 270–276, Kyiv. Association for Computational Linguistics.

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*, abs/2404.02013.

A Appendix

A.1 Error Analysis

We conducted an error analysis on the misclassified outputs of XLM-RoBERTa-Large, examining 167 misclassified samples, where there were 101 false positives (FP) and 66 false negatives (FN).

False positives occur when the model relies on surface-level cues such as negative sentiment or personal aggression, rather than pragmatic intent. Table 6 illustrates this tendency with representative examples.

Tamil Text	English Translation	Failure Pattern
ஒடி பேயிரு கையில் மாடிராத நான் இலங்கை இருக்கன் டிக்கெட் போட்டு வாந்து ஓதய்பன் யார் நீ விஜய் பாத்தி பேசா	"Run away. If I get my hands on you, I won't spare you. I'm in Sri Lanka. I'll buy a ticket and come there. Who do you think you are talking about Vijay like that?"	Aggressive fan defense misinterpreted as abuse
எல்லாம் பணம் பண்ணும் வேலை நாம ஒரு நாள் கூலி வேலை செய்து பொலைக்கிறோம் இது போன்ற ஜென்மம் கேவலமா இருந்து சம்பாதிக்குது	"Everything is about money. We struggle surviving on daily wages, while people like this earn money living shamefully"	Social critique misinterpreted as abuse

Table 6: Representative false positive examples with identified failure pattern.

False negatives arise when the model fails to recognize abuse that is conveyed through cultural expressions or cross-script markers. In such cases, the absence of explicit offensive language leads the model to misclassify abusive content as non-abusive, reflecting a broader lack of contextual grounding, as shown in Table 7.

Tamil Text	English Translation	Failure Pattern
உங்கல பார்தா சிரிப்புதாண் வருது லஷ்மி மேடம்	"Just seeing you makes me laugh, Madam Lakshmi"	Mockery disguised as polite addressing
Suganthi உனக்கு வீட்ல குழந்தையை வச்சிருக்க கார்த்திக் உன்ன கட்டி புடிச்சி விளையாடறான் கொஞ்சம் கூட வெக்கம் இல்ல thuuuu	"Suganthi, you have a child at home yet Karthik is hugging and playing with you, not even an ounce of shame, thuuuu"	Moral shaming with gendered insult

Table 7: Representative false negative examples with identified failure pattern.

Figure 2 presents the confusion matrix for XLM-RoBERTa-Large, illustrating the overall distribution of classification outcomes. The higher frequency of false positives compared to false negatives indicates that the model is sensitive to surface-level aggression, flagging contextually grounded expressions as abusive while struggling to detect abuse that is culturally embedded.

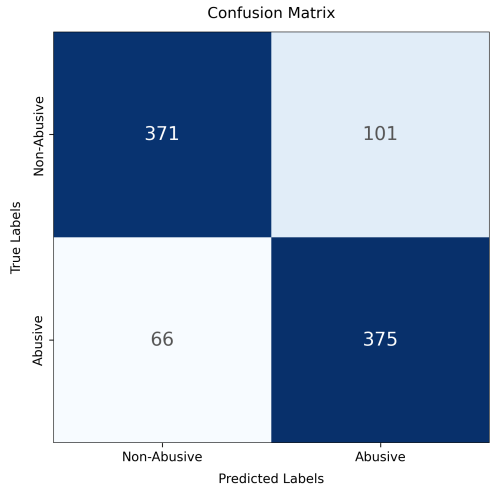

Figure 2: Confusion matrix for XLM-RoBERTa-Large.

A.2 Lexical Feature Analysis

A.2.1 PMI Lexicon Inspection

To assess whether the PMI-derived lexicon captures genuine abusive indicators or dataset-specific biases, Table 8 presents the top 15 words ranked by PMI score along with their assigned severity values.

Rank	Word	PMI	Severity
1	குஷ்பு	3.15	5.00
2	சின்மயி	3.06	5.00
3	தேவடியா	2.77	4.81
4	குஷ்பூ	2.77	4.81
5	சுந்தர்	2.70	4.75
6	பால்	2.55	4.62
7	டாக்டர்	2.46	4.55
8	அவுத்து	2.36	4.47
9	கிழவி	2.36	4.47
10	கனிமொழி	2.26	4.38
11	மக்களுக்கு	2.26	4.38
12	நாதாரி	2.26	4.38
13	doctor	2.26	4.38
14	பட்ட	2.14	4.28
15	பெண்களுக்கு	2.14	4.28

Table 8: Top 15 words by PMI score with severity values.

The lexicon reveals two distinct word categories. The first comprises genuinely abusive Tamil terms such as தேவடியா (prostitute), அவுத்து (danger/risk), நாதாரி (worthless person), and கிழவி (old woman), which are well-established slurs targeting women. The second category raises concern: named individuals (குஷ்பு (Khushboo), சின்மயி (Chinmayi), கனிமொழி (Kanimozhi)) rank highest by PMI, reflecting targeted harassment campaigns against specific public figures in the dataset rather than inherently abusive vocabulary. Furthermore, context-dependent words such as டாக்டர் (doctor) and மக்களுக்கு (for the people) acquired high PMI scores due to their frequent co-occurrence with abusive comments in the corpus, indicating dataset-specific bias rather than genuine abusiveness.

A.2.2 PMI Threshold Sensitivity

To evaluate the robustness of our lexicon construction, we examined how the PMI score distribution changes across different top-N cutoffs, as shown in Table 9.

Top-N	Min PMI	Max PMI	Mean PMI
50	1.8540	3.1533	2.1482
100	1.6717	3.1533	1.9633
150	1.5663	3.1533	1.8645
200	1.4485	3.1533	1.7642

Table 9: PMI score distribution across different lexicon size cutoffs.

The maximum PMI remains stable at 3.15 across all thresholds, while the mean decreases gradually from 2.15 at top-50 to 1.76 at top-200. This gradual degradation confirms that our chosen cutoff of 150 words represents a reasonable trade-off, retaining highly discriminative terms while limiting the inclusion of low-confidence associations.

A.2.3 Lexical Feature Importance

To understand which of the nine lexical features contributed most to the model's decisions, we analyzed the mean absolute weights of the learned lexicon projection layer, as reported in Table 10.

Rank	Feature	Importance
1	Positive word count	0.1817
2	Negation count	0.1789
3	English ratio	0.1784
4	Severity sum	0.1769
5	Tamil ratio	0.1739
6	Abusive match count	0.1704
7	Intensifier count	0.1668
8	Negative word count	0.1652
9	Sarcasm heuristic	0.1386

Table 10: Lexical feature importance derived from learned projection weights.

The importance scores are relatively uniform across features, indicating that the projection layer learns a distributed representation rather than relying on any single signal. Positive word count (0.1817) and Negation count (0.1789) rank highest, suggesting that sentiment polarity reversal, where positive words are negated is a useful surface cue

for abuse detection. English ratio (0.1784) ranks third, reflecting the code-mixed nature of Tamil social media text. Notably, Sarcasm heuristic received the lowest weight (0.1386), suggesting the simple co-occurrence heuristic was insufficient to capture nuanced sarcastic abuse.

A.2.4 Instance-Level Analysis of Lexical Feature Impact

Since XLM-RoBERTa-base exhibited the highest lexical gain, its predictions under two settings were compared to analyze the impact of lexical features. The comparison was performed across all 913 test instances, of which 89 predictions (9.75%) differed between the two settings, as summarized in Table 11.

True Label	−Lex Pred	+Lex Pred	Count
Abusive	Non-Abusive	Abusive	16
Non-Abusive	Abusive	Non-Abusive	40
Abusive	Abusive	Non-Abusive	21
Non-Abusive	Non-Abusive	Abusive	12
Net benefit (corrections − errors)			**+23**

Table 11: Breakdown of the test predictions that changed between −Lex and +Lex setting.

LEX recovers missed abuse (16 cases). These instances contained Tamil specific abusive vocabulary that XLM R base multilingual embeddings did not sufficiently prioritize when considered in isolation. The comment:

> செத்த மூதேவி ஆளும் மூஞ்சியும் பாரு ("Look at that wretched woman's face and demeanor")

was correctly reclassified as abusive once மூதேவி ("wretched woman," a strong Tamil gendered slur) triggered a high severity score in the lexicon. Similarly, the comment:

> டிக் டாக் லா இந்த அரிப்பு தாங்காம பொண்ணுங்க தான் அவுத்து போட்டு ஆடி ("These women on TikTok, unable to contain themselves, stripping and dancing")

contains அவுத்து ("stripping," a sexually objectifying term, ranked 8th in the PMI table with a severity score of 4.47). Without LEX, the transformer model interpreted this as social commentary, whereas the lexicon based severity signal correctly captured the underlying abusive intent.

LEX fixes false alarms (40 cases). This is the most significant effect of lexical augmentation. Without LEX, 40 neutral comments were misclassified as abusive due to their aggressive surface tone. LEX corrected these cases by showing low abusive match counts and high Tamil ratio scores, indicating genuine social discourse rather than targeted abuse. The comment:

> சுகந்தி பார்க்க நல்ல விதமாக அறிவான பெண்ணாக இருக்கிறார் ("Suganthi appears to be a sensible and intelligent woman")

is a straightforward compliment. The base model incorrectly flagged it as abusive because the named entity சுகந்தி ("Suganthi") frequently co-occurs with abusive comments in the training data. LEX reversed this prediction through its high positive word count signal and zero abusive match count. Similarly, the comment:

> மனுஷங்களாடா நீங்க..... ரொம்ப தரம் தாழ்ந்து விமர்சிக்கிரீங்க..... மூன்று மாத குழந்தையா இருந்த போதே தந்தைய இழந்துட்டு... (Are you even human..... you are criticizing so cheaply..... She lost her father when she was just a three-month-old baby)

is a defense of a woman being harshly criticized rather than an attack. The aggressive tone led the transformer model to misinterpret the context, whereas LEX identified no abusive lexical matches and detected a positive sentiment signal, correctly classifying the instance as Non Abusive.

LEX introduces false negatives (21 cases). These cases reveal a fundamental limitation of surface level lexical features, where culturally embedded abuse is expressed through gendered sarcasm without triggering any PMI lexicon signal. The comment:

> லோக்கல் ஐட்டம்மும் இண்டர்நேஷல் ஐட்டம்மும் உரையாடிய போது.... ("When the local item and the international item were talking....")

uses ஐட்டம் ("item," an objectifying term in Tamil cultural discourse) to demean women. Since this word carries no explicit PMI-lexicon match, LEX assigned a near-zero abuse score. Likewise:

> ரெண்டு பேரும் பத்தினி மாதிரி பேசுராலுக (“Both of them speak as if they are chaste women...”)

employs பத்தினி (“chaste woman”) sarcastically as a tool of moral shaming. The sarcasm heuristic failed to activate because the comment does not simultaneously contain the required positive, negative, and negation markers. Consequently, LEX suppressed the transformer’s accurate abuse identification.

LEX introduces false positives (12 cases). Named entity bias in the PMI lexicon is the primary driver of these errors. Public figures who are frequent targets of online harassment, such as குஷ்பு and சின்மயி, acquired high PMI scores through repeated co-occurrence with abusive comments in the training corpus. Their mere mention in neutral or supportive contexts consequently triggered abuse classification. The most striking example is:

> பெண் சுதந்திரம், பெண் கல்வி, பெண் உரிமை (“Women’s freedom, women’s education, women’s rights”)

a feminist slogan that was incorrectly classified as abusive. The word பெண் (“woman”) itself acquired a spurious abusive association from corpus statistics, representing the most direct evidence that PMI-based lexicons risk encoding dataset-level bias against the very demographic they are designed to protect.

Similarly, the comment:

> இந்த பொண்ணுக்கு புத்தி சொல்லி *encourage, motivation* குடுத்து சாதித்து தன் சுயமரியாதையுடன் நிற்க பயிற்சி கொடுங்க.... *she is innocent but, mind got contaminated.....* (“Give advice to this girl, encourage and motivate her, and train her to achieve success and stand with self-respect... she is innocent, but her mind has become contaminated”)

shows genuine supportive advice for a young woman. But the presence of mild criticism triggered PMI lexicon matches, causing incorrect abusive prediction.

Overall, LEX improved the model’s predictions, leading to an increase in Macro F1 score from the no-lexicon baseline to the enhanced system. These results indicate that lexical augmentation is particularly effective in mitigating surface-level aggression, especially in mid-sized transformer models that may struggle to distinguish between aggressive yet neutral expressions and genuinely targeted abuse. However, two systematic failure modes persist. The first involves culturally embedded abusive content expressed through gendered metaphor and moral shaming, which lacks explicit lexical markers. The second concerns named entity bias, where frequently targeted individuals and demographic terms such as பெண் (“woman”) develop unintended associations with abusive content due to corpus-level co-occurrence patterns.