From Comments to Harm: A Findings Report on Abusive Tamil Text Targeting Women on Social Media Shared Task

Bhuvaneswari Sivagnanam[1], Kathiravan Pannerselvam[1], Jananayagan V[1], Charmathi Rajkumar[2], Ramesh Kannan R[3], Ratnavel Rajalakshmi[3], Shunmuga Priya Muthusamy Chinnan[4], Saranya Rajiakodi[1], Bharathi Raja Chakravarthi[4]

[1]Central University of Tamil Nadu, India, [2]Kongu Engineering College, Erode, India
[3]Vellore Institute of Technology, Chennai, India,
[4]Unit for Inclusive AI, Data Science Institute, University of Galway, Ireland

Abstract

This paper presents an overview of the second shared task on Abusive Tamil Text Targeting Women on Social Media as a binary classification problem (abusive vs. non-abusive). We release a dataset of Tamil YouTube comments and evaluate submissions using macro-F1 to encourage balanced performance in a noisy, low-resource setting. There are 89 teams registered for this task and 24 teams submitted the results. The approaches used by the teams includes transformer fine-tuning, heterogeneous ensembles, classical baselines, and large language models using prompting and LoRA. Results show that the best-performing system scored 0.8297 macro-F1 and many submissions are around 0.79-0.81. Across submissions, transformer fine-tuning with domain-aligned encoders is consistently strong, while additional gains are frequently associated with Tamil-aware normalization and macro-F1-oriented calibration such as class-weighted learning and validation-based threshold tuning. Overall, the findings highlights the importance of language-aware preprocessing and careful decision calibration for reliable moderation of women-targeted abusive Tamil social media text.

Disclaimer: This paper (including figures and examples) may contain offensive or harmful language, including abusive content targeting women. All such text is presented solely for research and educational purposes and it does not reflect the author's views. Reader discretion is advised.

1 Introduction

In recent years, the social media platforms has grown significantly and reshaped how people communicate, share their views, and engage in public discussions (Ratkiewicz et al., 2011; Stephenson et al., 2018). While these digital spaces provide opportunities for connection and information sharing, they have also become platforms where harmful content, especially gender-based abuse, spreads (Suzor et al., 2019). Particularly, women are the most frequent targets of harassment, misogynistic comments and gender-based hostility within such online environments[1,2]. This not only affects their mental well-being and participation in digital platforms but also reflects broader social inequalities and discriminatory attitudes towards them (Kavanagh et al., 2019; Miranda, 2023).

In such settings, *offensive language* broadly refers to text that violates norms of respectful communication which includes insults, profanity, harassment, and other hostile expressions that can harm individuals or groups (Vidgen et al., 2019; Chakravarthi et al., 2021). *Abusive language* is more specific of offensive language in which the intent and effect are explicitly degrading, threatening, or humiliating toward a target (Vidgen et al., 2019) (Caselli et al., 2020). In this shared task, the target is women in which the abusive content frequently appears as misogynistic harassment, sexualised insults, or gendered ridicule aimed at policing women's presence and voice in public online spaces (Rajiakodi et al., 2025).

Automatically detecting such abuse is difficult even in high-resource languages because many abusive comments are indirect as they rely on sarcasm, insinuation, stereotypes, or context-dependent meanings so they cannot be captured reliably by simple lexical matching (Ocampo et al., 2023; Vidgen et al., 2019). The challenge is increased for Tamil, a low-resource language, where social-media writing introduces additional variability (Pannerselvam and Rajiakodi, 2026; Rajiakodi et al., 2025) such as informal and non-standard spellings, spoken forms, and regional variation, along with pervasive Tamil-English code-mixing

[1] https://www.amnesty.org/en/documents/act30/8070/2018/en/

[2] https://www.unwomen.org/en/articles/faqs/digital-abuse-trolling-stalking-and-other-forms-of-technology-facilitated-violence-against-women

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 88–100

and Romanised writing (Chakravarthi et al., 2021; Saumya et al., 2021; Prasanna and Arora, 2024). These characteristics increase lexical sparsity and complicate tokenisation and feature learning, reducing the robustness of conventional text-processing pipelines (Saumya et al., 2021; Prasanna and Arora, 2024). As a result, building reliable models for women-targeted abusive Tamil detection remains an active research problem, motivating shared-task benchmarks that enable systematic comparison of approaches under a unified evaluation protocol (Rajiakodi et al., 2025; Seemann et al., 2023).

Content moderation increasingly combines encoder-based classifiers with large language models (LLMs). DetoxBench reports that LLM (e.g., GPT-style models, Claude, Mistral/Mixtral, and Jurassic) performance varies widely across abuse categories (Chakraborty et al., 2024), and studies on implicit abuse show that zero-/few-shot prompting can help but remains inconsistent for context-dependent cases (Jaremko et al., 2025). This shared task builds on the first shared task on abusive Tamil and Malayalam text targeting women (Rajiakodi et al., 2025). We organize the *second* edition, *Abusive Tamil Text Targeting Women on Social Media*, as part of DravidianLangTech@ACL 2026[3], with a slightly modified setup to support clearer benchmarking and analysis. The objective of this shared task was to encourage the participants to develop a models capable of detecting abusive text targeting women in social media.

This paper is structured as follows: we describe the dataset construction process and annotation guidelines, then summarize the participating systems and evaluation setup. Next, we report the official results and highlight key observations and trends across submissions. Finally, we discuss common challenges and error patterns, and outline directions for future work toward more reliable and inclusive moderation for Tamil-speaking online communities.

2 Related work

Research on abusive text classification advanced by the several shared tasks and datasets. Hate Speech (Satapara et al., 2023; Mubarak et al., 2022) and offensive content detection (Chakravarthi et al., 2022) shared task on English and other Indian languages advanced the deeper research on abusive text classification. The Hasoc (Satapara et al., 2023; Mubarak et al., 2022) shared task adopted twitter dataset for English, Arabic and Indo-Aryan Languages for hate speech detection. The DravidianCodeMix (Chakravarthi et al., 2022) dataset provided 60000 instances for offensive language detection on code-mixed languages of Tamil-English, Malayalam-English, and Kannada-English. Recently Hate speech identification on Bangala language (Hossan et al., 2025) adopted multi-task setup to include the severity and target group identification. These shared tasks established the benchmarking procedures and standard datasets for abusive language detections.

English is a high-resource language, extensive studies have been carried (Fortuna et al., 2021) for identifying Abusive language, but limited research on low-resource languages such as Tamil. Research on languages such as Tigrinya (Gaim et al., 2025), German (Satapara et al., 2023) and various Indian languages underscores the importance of developing annotated datasets and multilingual models to address the resource scarcity. Recent research on code-mixed datasets for languages such as Telugu-English and Nepali-English (Pandey et al., 2025; Zia Ur Rehman et al., 2023) introduced to support multilingual abusive language detection. These works highlights the significance of multilingual embeddings and cross-lingual transfer learning in developing abusive language detection in low-resource settings.

Abusive text classification evaluation practices rely on standard metrics such as accuracy, precision, recall, and F1-score (Davidson et al., 2017; Ganganwar and Rajalakshmi, 2022). Shared tasks on abusive comment detection in various languages has commonly used macro F1-score (Fortuna et al., 2021; Satapara et al., 2024) as the primary evaluation metric due to class imbalance issues. Binary classification schemes (Priyadharshini et al., 2023) and Cross-dataset evaluation (Fortuna et al., 2021) have been adopted across domains to test the model generalization. Further studies (Chiril et al., 2022) shows statistical validation techniques such as validation and significance testing to ensure reliable performance comparisons.

The study of gender-targeted abuse has emerged as a key area within abusive language detection. Shared tasks and various datasets focusing on regional languages and English have been developed to identify the targeted women using abusive contents and targeted on marginalized groups.

[3] https://sites.google.com/view/dravidianlangtech-2026/home?authuser=0

Hate speech datasets such as HatEval (Basile et al., 2019), (Chiril et al., 2022) explicitly added women as a target group for hate speech identification. Another major challenge is the detection of abusive content in code-mixed and Romanized text such as Tanglish (Benhur and Sivanraju, 2021). Transformer-based models have been successfully applied to Tanglish offensive language detection tasks using multilingual pretrained representations. Code-mixed datasets such as DravidianCodeMix (Chakravarthi et al., 2022) further demonstrate the linguistic complexity of Romanized and multilingual social media text. Intimate Partner Violence (IPV) related posts were detected using a RoBERTa based transformer model approach on manually annotated English dataset collected from Twitter and Reddit (Guo et al., 2023). The model able to distinguish abusive self reports from general discussions with an accuracy of 78 %. The dataset is small and it is limited to detect explicit English comments. These challenges highlight the urgent need for robust models capable of handling low resource, multilingual and noisy texts in social media. All the above discussed articles explore various implementations and performance metrics for abusive text detection, hate speech detection and offensive language detection. They provide valuable insights into implementation strategies and deep learning models. However, there remains a critical gap in targeting gender based abuse towards women in particularly Dravidian language like Tamil.

3 Task description

This second edition of the shared task is organized on the Codabench platform[4] and focuses on binary classification of Tamil YouTube comments. In this edition, we only detect abusive comments that target women. Each comment is classified as either Abusive or Non-Abusive.

A comment is labeled as Abusive only if the abusive content is targeted at women. Comments that contain bad words but do not target women are labeled as Non-Abusive. For example, if a man is insulted using offensive words, even if those words mention his mother or other female family members, the comment is still considered Non-Abusive because the main target is not a woman. Most abusive comments targeting women include references to female body parts or insults about women's character, behavior, or morality.

Further, the abusive category includes comments that contain offensive, insulting, threatening, harassing or otherwise harmful language. The abuse may be explicit or implicit, and may appear in direct, sarcastic, or coded forms, but it reflects an intention to demean or harm the target, and the Non-Abusive category includes comments that are neutral, positive, informative, or express general opinions without harmful intent. To support system development and evaluation, annotated datasets are provided in Tamil, including training, and test splits. Participants are required to build models that automatically classify each test comment into one of the two predefined categories. The task aims to advance research on the detection of abusive Tamil language and to contribute to safer online communication in low-resource language settings.

4 Dataset

We created a Tamil YouTube comment dataset with annotations for abusive text targeting women. The comments were collected using the automated scraping tools combined with manual filtering from publicly available comment sections of popular Tamil YouTube channels. Comments were gathered from publicly available videos related to entertainment, politics, and social discussions to make sure diversity in abusive and non-abusive expressions. The dataset contains abusive comments specifically directed at women, as well as comments that are not targeted at women, even if they target men or are neutral. It was prepared to reflect normal social media Tamil language usage, including informal spelling, slang words, and platform-style punctuation. The classification task is binary, with two labels: Abusive (targeting women) and Non-Abusive (not targeting women).

4.1 Annotation process

Annotator background. The annotation process was done by three annotators with different academic backgrounds to improve labeling quality. As shown in Table 1, two annotators were Computer Science students and one was a Social Work student. Among them, two were Ph.D. students and one was a postgraduate student. This diversity helped reduce bias when identifying abusive content, since interpreting abusive language can sometimes be subjective.

All annotators were native Tamil speakers and

[4] https://www.codabench.org/competitions/11326/#/pages-tab

were familiar with informal social media language, including slang words and casual writing styles. Before annotation, they were given clear guidelines explaining abusive content targeting women, including both direct and indirect abusive expressions. When there were differences in opinions, the annotators discussed and agreed on a final label. This helped improve the consistency and quality of the dataset labels.

Characteristic	Details
# Annotators	3
Gender Distribution	2 Female, 1 Male
Educational Background	2 CS Students 1 SW Student
Level of Study	2 Ph.D. and 1 PG
Native Language	Tamil
Social Media Language Familiarity	Yes

Table 1: Annotator Demographic Details. CS- Computer Science, SW- Social Work

Guidelines and labeling. Annotators labeled each comment based on whether it contained abusive content targeting women. The interpretation of Abusive and Non-Abusive content can sometimes be subjective, as it depends on social and cultural context. Therefore, multiple annotators evaluated each comment to improve labeling reliability. Direct abusive content included insults, derogatory words, or explicit attacks targeting women. Contextual meaning was also considered during annotation. Figure 1 and Figure 2 present sample annotated comments.

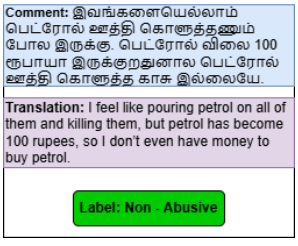

Figure 1: Example of Non-Abusive Comment Expressing Anger. Although the comment expresses emotional anger, it does not directly contain abusive language targeting women.

Each comment was annotated by multiple annotators, and the final label was determined using majority voting. This approach helped reduce individual bias and improved labeling consistency, especially for borderline cases where subjective interpretation may vary.

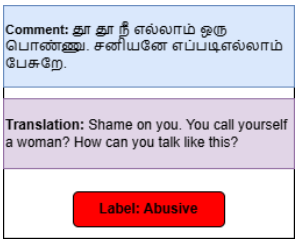

Figure 2: Example of Abusive Comment. The comment contains direct insulting expressions targeting women and was therefore labeled as abusive.

Split	#Abusive	#Non-Abusive	Total
Train	1769	1883	3652
Test	441	472	913
Total	2210	2355	4565

Table 2: Label distribution of the dataset.

Inter-annotator agreement. For this task, the Krippendorff's Alpha (Marzi et al., 2024) is 0.6474, indicating moderate agreement. This level of agreement is common in subjective social-media moderation tasks where sarcasm, implicit harassment, and borderline insults can lead to genuine annotator disagreement.

Dataset splits and evaluation setup. The dataset is split into a training set (3652 instances) and a test set (913 instances) (Table 2). The test set is released in two forms: an unlabeled version for evaluation and a labeled version for subsequent error analysis.

5 Evaluation setup

The submitted models are evaluated using standard classification metrics computed over the two classes: *Abusive* and *Non-Abusive*. The performance of the classification system is measured using Accuracy, Precision, Recall, macro F1-score, and Weighted F1-score (Sharma et al., 2024).

Accuracy measures the proportion of correctly predicted instances among the total number of instances. Precision measures the proportion of correctly predicted instances of a class among all instances predicted for that class. Recall measures the proportion of correctly identified instances of a class among all actual instances of that class. The F1-score combines Precision and Recall into a single measure by taking their balanced mean.

The primary evaluation metric for ranking is the macro-averaged F1-score. The F1-score is first computed independently for each class. The macro F1 score (Eq. (1)) is then obtained by averaging the

F1-scores of all classes, giving equal importance to each class regardless of its size (Hinojosa Lee et al., 2024).

$$\text{macro F1} = \frac{1}{N}\sum_{i=1}^{N} F1_i \tag{1}$$

Since abusive language datasets are often imbalanced, the macro F1-score ensures that performance on the minority class is properly reflected. The final ranking of systems is based on the macro-averaged F1-score computed on the test set.

Weighted F1-score (Eq. (2)) computes the F1-score for each class and averages them according to the proportion of instances in each class.

$$Weighted\ F1 = \sum_{i=1}^{N} \left(\frac{n_i}{N} \times F1_i\right) \tag{2}$$

where N represents the total number of classes, and n_i denotes the number of instances in class i.

Participants are allowed to submit multiple runs. All valid submissions are considered and the highest macro-F1 score for each participant is used to determine the final ranking. Only the best-performing submission from each participant is included in the official results. The use of external data and pre-trained models is permitted. However, participants must clearly describe any additional resources used in their system description paper to ensure transparency. Participants must submit their predictions for the test set. Each line in the submission file should correspond to a test instance and contain the predicted label (Abusive or Non-Abusive) in the same order as the test data. Submissions that do not follow the required format may not be evaluated. This evaluation setup ensures fairness and clarity in comparing different models.

6 Overview of participant systems

The submitted systems span diverse approaches for detecting abusive Tamil text targeting women on social media, ranging from fine-tuned transformer encoders to ensembles, LLM-based prompting/LoRA, and lightweight classical baselines. Several submissions also reported robustness-oriented preprocessing for noisy or obfuscated text, and a small subset used external data.

6.1 Top-ranked systems (by macro-F1)

The top of the leaderboard is dominated by transformer-based systems with either domain adaptation, Indian language pretraining, or Tamil-aware normalization. nitc-hsr (Rank 1) submitted an ensemble of Indian-language/domain-adapted encoders (IndicBERT and AbuseXLMR). primeline_abusive_Tamil (V et al., 2026) (Rank 2) fine-tuned MuRIL using a simple and reproducible pipeline without handcrafted features. HNK (Rank 3) emphasized grapheme-aware preprocessing before tokenization and reported consistent gains across multiple transformer backbones. CUET_Synthetica (Rank 4) proposed a large heterogeneous ensemble combining Tamil-specialized models (e.g., MuRIL, IndicBERTv2) with multilingual transformers (e.g., XLM-RoBERTa-Large), together with duplicate removal and Tamil-specific preprocessing. DPR (Prakash et al., 2026) (Rank 5) conducted a controlled comparison of multilingual transformers (mBERT, MuRIL, XLM-RoBERTa, IndicBERT) under consistent settings to study transfer robustness for Tamil social-media abuse detection.

6.2 Transformer-based fine-tuning

Many teams contributed strong baselines by varying pretraining choice and fine-tuning configuration. CHMODE_777 (Karunanidhi and Arumugam, 2026) compared MuRIL and XLM-RoBERTa, highlighting the importance of model selection and reporting gains from longer context windows (up to 256 tokens) and Indian language pretraining. DLRG (Sankar and Rajalakshmi, 2026) fine-tuned MuRIL with label normalization and HTML cleaning, using stratified splits, mixed-precision training, weight decay, and warmup scheduling; they emphasised that carefully optimised fine-tuning can be a strong baseline without external data or ensembles. IndiLangTech fine-tuned IndicBERTv2 with a standard classification head. TriVector (Rebayet et al., 2026) used a Tamil hate-speech domain-adapted XLM-R model and incorporated statistical lexicon features. ADAPTIVEMINDS fine-tuned XLM-RoBERTa directly on the provided labels with a lightweight setup, without manual relabeling/normalization. Medhastra_abusive also fine-tuned XLM-RoBERTa and highlighted a straightforward end-to-end pipeline without rules, lexicons, or external resources.

6.3 Ensembles submissions

A subset of teams explored ensembles and hybrid pipelines to improve robustness. Lannisters reported Tamil-aware normalization (Unicode/HTML handling), explicit handling of obfuscated/censored abuse, lightweight augmentation/noise strategies, and confidence-aware prediction for flagging low-confidence cases. SuperNova (K et al., 2026) evaluated multiple paradigms, including end-to-end MuRIL fine-tuning and embedding-based classifiers (MuRIL or SBERT embeddings with tree/boosting models), highlighting trade-offs between accuracy and simplicity.

6.4 macro-F1-based optimization

Several submissions explicitly targeted balanced performance using macro-F1-aware optimization choices. N-cuboid-MP applied class-weighted training and validation-based threshold optimization rather than using a fixed 0.5 decision boundary. DynamicDuo emphasized cost-sensitive fine-tuning and probability-based decision thresholding to reduce false negatives.

6.5 LLM-based approaches and deployment-oriented framing

Two submissions explicitly connected the task to modern moderation workflows using LLMs or human-in-the-loop design. VITECH (Krubhakaran et al., 2026) evaluated multiple LLM families (Gemma, LLaMA, Qwen, and DeepSeek-Distilled) under zero-shot/few-shot prompting and LoRA fine-tuning, reporting that LoRA and larger model sizes (e.g., 8B) consistently improved performance over smaller models. Infinity_Preetha proposed a deployable moderation workflow layered on top of transformer classification, using a progressive three-warning escalation mechanism that forwards persistent offenders to human moderators for verification and action.

6.6 Lightweight baselines and external data

A few teams reported lightweight pipelines that remain attractive for efficiency and interpretability. GauriWarriors used a FastText-based approach with subword modeling and auto-tuning to handle spelling variation and informal text. KEC's Code Crafters used an external Kaggle dataset[5] and combined word- and character-level features in a simple ML classifier to capture noisy spelling patterns. KECInference submitted two contrasting runs like MuRIL fine-tuning and a TF-IDF+SVM baseline to illustrate the trade-off between contextual encoders and lexical models. Finally, tamilgoodbadtxt presented a model-agnostic transformer pipeline enabling consistent preprocessing and easy switching across pretrained backbones, while VISION (Subramanian et al., 2026) and CodeTamizh described practical pipelines emphasizing preprocessing and scalability, although the exact backbone and settings were not fully specified in their short submission notes.

[5] https://www.kaggle.com/datasets/prfsr007/abusive-tamil-dataset

7 Results and Discussion

There are total of 89 teams registered for this shared task, and among them 24 teams submitted their results for ranking and evaluation. This gap may due to the fact of practical challenges associated with the low-resource Tamil abusive language detection which includes the noisy social media text and computational constraints.Table 3 reports the official leaderboard for the shared task which ranks the systems by macro-F1 (our primary evaluation metric). Alongside macro-F1, we also reported Accuracy, Precision, Recall, and Weighted-F1 for completeness.

7.1 Overall leaderboard

The top system, nitc-hsr (A and Kumar, 2026) (Run 2), achieved the best overall performance with a macro-F1 of 0.8297, followed by prime-line_abusive_Tamil (V et al., 2026) (Run 1, macro-F1 0.8133) and HNK (R et al., 2026) (Run 1, macro-F1 0.8103). The next positions were occupied by CUET_Synthetica (Rishta et al., 2026) (Run 2, 0.8086) and DPR (Prakash et al., 2026) (Run 4, 0.8072), indicating a competitive top tier where multiple transformer-based systems performed within a narrow margin. Several teams achieved macro-F1 values are around 0.79-0.80 which demonstrates that strong performance is achievable under a range of design choices.

7.2 Statistical Significance Analysis

To examine whether the differences among the top-ranked systems were statistically meaningful, we performed bootstrap significance testing with 10,000 resampling iterations using Macro-F1 as the evaluation metric. The analysis was conducted on the prediction outputs of the top three systems. The

Team	RUN	Accuracy	Precision	Recall	macro-F1	Weighted-F1	Rank
nitc-hsr (A and Kumar, 2026)	RUN 2	0.8302	0.8310	0.8293	0.8297	0.8300	1
primeline_abusive_Tamil (V et al., 2026)	RUN 1	0.8138	0.8140	0.8131	0.8133	0.8100	2
HNK (R et al., 2026)	RUN 1	0.8105	0.8103	0.8104	0.8103	0.8000	3
CUET_Synthetica (Rishta et al., 2026)	RUN 2	0.8094	0.8106	0.8082	0.8086	0.8100	4
DPR (Prakash et al., 2026)	RUN 4	0.8072	0.8072	0.8076	0.8072	0.8100	5
CHMODE_777 (Karunanidhi and Arumugam, 2026)	RUN 2	0.8061	0.8061	0.8065	0.8061	0.8000	6
Infinity_Preetha	RUN 4	0.8061	0.8059	0.8059	0.8059	0.8100	7
Lannisters	RUN 1	0.8028	0.8032	0.8020	0.8023	0.8000	8
VISION (Subramanian et al., 2026)	RUN 10	0.8018	0.8018	0.8022	0.8017	0.6800	9
N-Cuboid-MP	RUN 2	0.8007	0.8010	0.7998	0.8001	0.8000	10
SuperNova (K et al., 2026)	RUN 1	0.8007	0.8031	0.7989	0.7994	0.6800	11
VITECH (Krubhakaran et al., 2026)	RUN 1	0.7963	0.7963	0.7957	0.7959	0.8000	12
IndiLangTech	RUN 1	0.7941	0.7938	0.7940	0.7939	0.7900	13
DLRG (Sankar and Rajalakshmi, 2026)	RUN 1	0.7919	0.7919	0.7913	0.7915	0.7900	14
TriVector (Rebayet et al., 2026)	RUN 1	0.7897	0.7895	0.7898	0.7896	0.7900	15
tamilgoodbadtxt (K and B, 2026)	RUN 1	0.7864	0.7881	0.7876	0.7864	0.7800	16
ADAPTIVEMINDS	RUN 1	0.7700	0.7735	0.7678	0.7681	0.7700	17
GauriWarriors	RUN 1	0.7338	0.7333	0.7010	0.7117	0.7335	18
CodeTamizh	RUN 1	0.5268	0.5246	0.5181	0.4881	0.6800	19
Medhastra_abusive	RUN 1	0.5170	0.2585	0.5000	0.3408	0.6800	20
DynamicDuo	RUN 1	0.5170	0.2585	0.5000	0.3408	0.6800	20
DravidianNLP	RUN 1	0.5170	0.2585	0.5000	0.3408	0.6800	20
KEC's Code Crafters	RUN 1	0.5170	0.2585	0.5000	0.3408	0.6800	20
KECInference	RUN 1	0.5170	0.2585	0.5000	0.3408	0.6800	20

Table 3: Leaderboard of Participating Systems Ranked by macro-F1

Team	Model(s) reported	Pre proc	Indian-PT	Ens-emble	Ext. Data	Imb.	Thr.	Aug.	HITL	LLM/ LoRA
nitc-hsr (A and Kumar, 2026)	IndicBERT + AbuseXLMR	✗	✓	✓	✗	✗	✗	✗	✗	✗
primeline_abusive_Tamil (V et al., 2026)	MuRIL	✗	✓	✗	✗	✗	✗	✗	✗	✗
HNK (R et al., 2026)	Multiple transformers + grapheme-aware normalization	✓	✓	✗	✗	✗	✗	✗	✗	✗
CUET_Synthetica (Rishta et al., 2026)	MuRIL, IndicBERTv2, XLM-R	✓	✓	✓	✗	✗	✗	✗	✗	✗
DPR (Prakash et al., 2026)	mBERT, MuRIL, XLM-R, IndicBERT	✗	✓	✗	✗	✗	✗	✗	✗	✗
CHMODE_ 777	MuRIL vs XLM-R	✗	✓	✗	✗	✗	✗	✗	✗	✗
Infinity_Preetha	XLM-R + progressive 3-warning escalation workflow	✗	✗	✗	✗	✗	✗	✗	✓	✗
Lannisters	XLM-R	✓	✗	✗	✗	✗	✓	✓	✓	✗
VISION (Subramanian et al., 2026)	(Not specified)	✓	✗	✗	✗	✓	✗	✗	✗	✗
N-Cuboid-MP	XLM-R fine-tuning	✗	✗	✗	✗	✓	✓	✗	✗	✗
SuperNova (K et al., 2026)	MuRIL FT	✗	✓	✓	✗	✗	✗	✗	✗	✗
VITECH (Krubhakaran et al., 2026)	Gemma / LLaMA / Qwen / DeepSeek	✗	✗	✗	✗	✗	✗	✗	✗	✓
IndiLangTech	IndicBERTv2 fine-tuning	✗	✓	✗	✗	✗	✗	✗	✗	✗
DLRG (Sankar and Rajalakshmi, 2026)	MuRIL fine-tuning	✓	✓	✗	✗	✗	✗	✗	✗	✗
TriVector (Rebayet et al., 2026)	Hate-speech-CNERG/deoffxlmr-mono-tamil	✗	✗	✗	✗	✗	✗	✗	✗	✗
tamilgoodbadtxt (K and B, 2026)	Model-agnostic transformer pipeline	✓	✓	✗	✗	✗	✗	✗	✗	✗
ADAPTIVEMINDS	XLM-R fine-tuning	✗	✗	✗	✗	✗	✗	✗	✗	✗
GauriWarriors	FastText	✗	✗	✗	✗	✗	✗	✗	✗	✗
CodeTamizh	Model not specified	✗	✗	✗	✗	✗	✗	✗	✗	✗
Medhastra_abusive	XLM-R fine-tuning	✗	✗	✗	✗	✗	✗	✗	✗	✗
DynamicDuo	XLM-R	✗	✗	✗	✗	✓	✓	✗	✗	✗
DravidianNLP	(Not specified)	✓	✗	✗	✗	✗	✗	✗	✗	✗
KEC's Code Crafters	Traditional ML	✗	✗	✗	✓	✗	✗	✗	✗	✗
KECInference	MuRIL and TF-IDF + SVM	✗	✓	✗	✗	✗	✗	✗	✗	✗

Table 4: Summary of Descriptive trends of Participating Systems: Reported models and Techniques (✓ = reported, ✗ = not reported in the team description). **Preproc**: cleaning/normalization; **Indian-PT**: Indian-language pretraining (e.g., MuRIL/IndicBERT/AbuseXLMR); **Imb.**: class-imbalance handling (e.g., weighted loss/cost-sensitive); **Thr.**: validation-based threshold tuning; **Aug.**: data augmentation; **HITL**: human-in-the-loop moderation workflow; **LLM/LoRA**: prompting or LoRA fine-tuning of LLMs.

results (5) show that although nitc-hsr (A and Kumar, 2026) achieved the highest Macro-F1 score, its improvement over primeline_abusive_Tamil (V et al., 2026) was not statistically significant. Similarly, primeline_abusive_Tamil (V et al., 2026) and HNK (R et al., 2026) showed highly overlapping performance. nitc-hsr (A and Kumar, 2026) showed only a marginal advantage over HNK (R et al., 2026). These findings suggest that the top-ranked systems achieved broadly comparable performance, and small leaderboard differences should be interpreted cautiously.

Comparison	Mean Δ F1	95% CI	p-value
nitc-hsr (A and Kumar, 2026) vs PRIMELINE	0.0164	[-0.0047, 0.0376]	0.0676
nitc-hsr (A and Kumar, 2026) vs HNK (R et al., 2026)	0.0192	[-0.0015, 0.0400]	0.0365
PRIMELINE vs HNK (R et al., 2026)	0.0028	[-0.0178, 0.0237]	0.3992

Table 5: Bootstrap significance testing among the top-ranked systems using Macro-F1.

7.3 Trends across submissions

To avoid relying only on qualitative descriptions, we summarize reported techniques using Table 4. Overall, 10/24 submissions used Indian language or domain-adapted pretrained encoders (e.g., MuRIL/IndicBERT/AbuseXLMR), and 6/24 explicitly mentioned Tamil-aware preprocessing or

normalization (e.g., grapheme/Unicode handling, HTML/obfuscation cleaning). This pattern aligns with the leaderboard, where higher-ranked systems frequently adopted language/domain-aligned pretraining and, in some cases, Tamil-specific normalization.

A smaller subset of teams explicitly targeted macro-F1 through imbalance-aware optimization, such as class-weighted training, cost-sensitive loss functions, or validation-based threshold tuning. While the class distribution is relatively balanced in our dataset, such techniques are commonly used to calibrate decision boundaries and improve class-balanced metrics. Ensembles were explored by several teams.

LLM-based approaches were evaluated as moderation-relevant baselines using prompting (zero-shot/few-shot) and parameter-efficient adaptation (LoRA). While such methods offer flexibility, encoder-based fine-tuning remains a strong and computationally efficient choice for this binary classification setting.

7.4 Discussions

From the leaderboard and the reported system designs, we highlight three observations. First, reported gains frequently came from language/domain-aligned pretraining and Tamil-aware normalization, which help models handle noisy social-media text. Second, calibration choices including class weighting and validation-based threshold tuning can improve macro-F1 by producing more balanced predictions. Third, robustness-oriented preprocessing (e.g., handling obfuscated or masked abuse) is better aligned with real-world moderation needs.

Finally, multiple teams obtained identical macro-F1 values at the bottom of the leaderboard (Rank 20). To better understand this pattern, we examined the prediction distributions of several lower-ranked systems. The analysis showed that these systems did not uniformly collapse to a single class prediction; instead, they predicted both *Abusive* and *Non-Abusive* labels, although with varying degrees of bias toward the majority class. These observations suggest that identical macro-F1 scores at the bottom of the leaderboard may arise from different failure modes, including majority-class bias, skewed decision thresholds, difficulty handling minority-class instances, and inconsistent prediction behavior, rather than complete majority-class collapse alone.

8 Conclusion

This shared task introduced a benchmark for detecting abusive Tamil text targeting women in social-media comments that often contain code-mixing, informal spelling, and non-standard orthography. A total of participating teams explored a range of approaches, including Tamil-aware preprocessing, Indic-language pretrained transformers, and imbalance-aware training strategies. The best-performing system achieved a macro-F1 score of 0.8297, while several systems obtained competitive results in the 0.7900-0.8100 range. These results demonstrate that language-specific preprocessing and robust training techniques can improve abusive language detection in noisy online environments. However, several challenging cases remain, particularly those involving masked abusive words, ambiguous targets, quoted harmful language, and sarcasm. The shared task dataset and benchmark are expected to support further research on abusive language detection for low-resource Dravidian languages. Future work can explore finer-grained annotation schemes, contextual modeling, and cross-platform evaluation to improve robustness and generalization.

9 Limitations

The dataset used in this shared task is collected from YouTube comments, which may introduce platform-specific communication patterns and topical biases. As a result, models trained on this dataset may not directly generalize to other social media platforms without additional adaptation (Vidgen and Derczynski, 2020). Tamil social-media text also contains dialectal variation, transliteration, creative spelling, and frequent English code-mixing, which makes normalization and interpretation challenging. Another limitation concerns the binary labeling scheme, which compresses multiple forms of harmful language into a single category and may fail to capture implicit abuse, sarcasm, or counterspeech. In addition, the distinction between abusive and non-abusive comments can sometimes be subjective. During the shared task, some participants over mail sought clarification regarding this boundary, particularly in cases involving sarcasm, indirect criticism, or quoted language. This limitation may be addressed in future shared tasks through finer-grained annotation schemes. Finally, the limited availability of large-scale Tamil resources, including domain-pretrained models and

extensive annotated corpora, may restrict achievable system performance.

Ethical Statement

This shared task dataset contains abusive Tamil social media text targeting women and may include offensive or distressing language, it is provided strictly for research purposes and does not reflect the author's views. We sourced text from publicly available content, removed direct personal identifiers where possible, and released only the minimum information needed for the task, with annotation conducted by trained Tamil-proficient annotators under content warnings and well-being safeguards. The dataset is intended to support research on responsible moderation, but it may carry risks of bias and misuse, we therefore recommend careful, transparent use with human oversight and evaluation of potential harms.

Usage of AI Declaration

We used AI-based tools to support language editing and formatting of this manuscript (e.g., improving clarity, grammar, and LaTeX presentation). All technical content, experimental results, interpretations, and final wording were reviewed and verified by the authors. No AI system was used to generate or alter the dataset, annotations, or official leaderboard results.

Acknowledgements

The authors, Bharathi Raja Chakravarthi was funded by a research grant from Research Ireland under grant number SFI/12/RC/2289_P2 (Insight_2).

References

Rameez Mohammed A and S D Madhu Kumar. 2026. NITC-HSR@DravidianLangTech 2026: Ensembling Multilingual Transformer Models for Detecting Abusive Tamil Text Targeting Women on Social Media. In *Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019)*, pages 54–63.

Sean Benhur and Kanchana Sivanraju. 2021. Pretrained transformers for offensive language identification in tanglish. *arXiv preprint arXiv:2110.02852*.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6193–6202, Marseille, France. European Language Resources Association.

Joymallya Chakraborty, Wei Xia, Anirban Majumder, Dan Ma, Walid Chaabene, and Naveed Janvekar. 2024. Detoxbench: Benchmarking large language models for multitask fraud & abuse detection. *arXiv preprint arXiv:2409.06072*.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Navya Jose, Thomas Mandl, Prasanna Kumar Kumaresan, Rahul Ponnusamy, John Philip McCrae, Elizabeth Sherly, and 1 others. 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.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Navya Jose, Elizabeth Sherly, and John McCrae. 2022. Dravidiancodemix: Sentiment analysis and offensive language identification dataset for dravidian languages in code-mixed text. *Language Resources and Evaluation*.

Patricia Chiril, Endang Pamungkas, Farah Benamara, Véronique Moriceau, and Viviana Patti. 2022. Emotionally informed hate speech detection: A multi-target perspective. *Cognitive Computation*, 14(3):1322–1336.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Preprint*, arXiv:1703.04009.

Paula Fortuna, Juan Soler-Company, and Leo Wanner. 2021. How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management*, 58(3):102524.

Fitsum Gaim, Hoyun Song, Huije Lee, Changgeon Ko, Eui Jun Hwang, and Jong C. Park. 2025. A multitask benchmark for abusive language detection in low-resource settings. *Preprint*, arXiv:2505.12116.

Vaishali Ganganwar and Ratnavel Rajalakshmi. 2022. Mtdot: A multilingual translation-based data augmentation technique for offensive content identification in tamil text data. *Electronics*, 11(21).

Yuting Guo, Sangmi Kim, Elise Warren, Yuan-Chi Yang, Sahithi Lakamana, and Abeed Sarker. 2023. Automatic detection of intimate partner violence victims from social media for proactive delivery of support.

In *AMIA Joint Summits on Translational Science Proceedings*, pages 254–260. American Medical Informatics Association.

Maria Cristina Hinojosa Lee, Johan Braet, and Johan Springael. 2024. Performance metrics for multilabel emotion classification: comparing micro, macro, and weighted f1-scores. *Applied Sciences*, 14(21):9863.

Md. Refaj Hossan, Kawsar Ahmed, and Mohammed Moshiul Hoque. 2025. CUET-NLP_Zenith at BLP-2025 task 1: A multi-task ensemble approach for detecting hate speech in Bengali YouTube comments. In *Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025)*, pages 443–452, Mumbai, India. Association for Computational Linguistics.

Julia Jaremko, Dagmar Gromann, and Michael Wiegand. 2025. Revisiting implicitly abusive language detection: Evaluating llms in zero-shot and few-shot settings. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3879–3898.

Kiruthika K, Roahiyaa T, and Premjith B. 2026. SUPERNOVA@DravidianLangTech 2026: Transformer and Ensemble Approaches for Abusive Tamil Text Detection Targeting Women. In *Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Varalakshmi K and Bharathi B. 2026. TAMILGOODBADTXT@DravidianLangTech 2026:A Multilingual Transformer-Based Approach for Abusive Language Identification in Tamil Social Media. In *Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Arunaggiri Pandian Karunanidhi and Prabalakshmi Arumugam. 2026. CHMOD_777@DravidianLangTech 2026: Context-Aware Fine-tuned MuRIL for Abusive Tamil Text Detection on Social Media. In *Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Emma Kavanagh, Chelsea Litchfield, and Jaquelyn Osborne. 2019. Sporting women and social media: Sexualization, misogyny, and gender-based violence in online spaces. *International Journal of Sport Communication*, 12(4):552–572.

Triambiga Krubhakaran, Senthil Kumar B, Kaviya Nagarajan, and Balaji N. 2026. VITECH@DravidianLangTech2026: Prompting and LoRA Adaptation for Tamil Abusive Language Detection - A Comparative Study of Open LLMs. In *Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Giacomo Marzi, Marco Balzano, and Davide Marchiori. 2024. K-alpha calculator–krippendorff's alpha calculator: a user-friendly tool for computing krippendorff's alpha inter-rater reliability coefficient. *MethodsX*, 12:102545.

Sandra Lopes Miranda. 2023. Analyzing hate speech against women on instagram. *Open Information Science*, 7(1):20220161.

Hamdy Mubarak, Hend Al-Khalifa, and Abdulmohsen Al-Thubaity. 2022. Overview of osact5 shared task on arabic offensive language and hate speech detection. In *OSACT*.

Nicolás Benjamín Ocampo, Ekaterina Sviridova, Elena Cabrio, and Serena Villata. 2023. An in-depth analysis of implicit and subtle hate speech messages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1997–2013.

Manish Pandey, Nageshwar Prasad Yadav, Mokshada Adduru, and Sawan Rai. 2025. Creating and evaluating code-mixed nepali-english and telugu-english datasets for abusive language detection using traditional and deep learning models. *Preprint*, arXiv:2504.21026.

Kathiravan Pannerselvam and Saranya Rajiakodi. 2026. Systematic literature review on hate speech detection in indian low-resource languages. *Journal of Computational Social Science*, 9(1):5.

Diya Prakash, Praveen Kumar S, Ranjith Kumar R, Siranjeevi Rajamanickam, Balasubramanian Palani, and Jobin Jose. 2026. DPR@DravidianLangTech 2026: Transformer-Based Abusive Content Detection for Tamil Text Targeting Women on Social Media. In *Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Kabilan Prasanna and Aryaman Arora. 2024. Irumozhi: Automatically classifying diglossia in tamil. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3096–3103.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, M Subramanian, and 1 others. 2023. Overview of shared-task on abusive comment detection in tamil and telugu. In *DravidianLangTech*.

Hanish Vigneshwar R, Nahul Alaguraj, Karthikeyan M, and Ratnavel Rajalakshmi. 2026. HNK@DravidianLangTech 2026: Investigating Grapheme-Level Normalization for Abusive Tamil Text Classification. In *Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Shunmuga Priya Muthusamy Chinnan, Ruba Priyadharshini, Kathiravan Pannerselvam, Rahul Ponnusamy, Bhuvaneswari Sivagnanam, Paul Buite-

laar, Jananayagan Jananayagan, Kishore Kumar Ponnusamy, and 1 others. 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*, pages 671–681.

Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer. 2011. Detecting and tracking political abuse in social media. In *Proceedings of the International AAAI Conference on Web and social media*, volume 5, pages 297–304.

Oarisa Rebayet, Tahmima Hoque Eid, Fawzia Tabassum, and Hasan Murad. 2026. TriVector@DravidianLangTech 2026: Abusive Tamil Text Detection on Social Media Using Lexicon-Augmented Transformers. In *Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Miftahul Jannat Rishta, Sumaiya Zaman, Shiti Chowdhury, and Hasan Murad. 2026. CUET_SYNTHETICA@DravidianLangTech 2026: Multi Architecture Transformer Ensemble for Detecting Abusive Tamil Text Targeting Women. In *Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Mirudhula Sankar and Ratnavel Rajalakshmi. 2026. DLRG@DravidianLangTech 2026: Explainable Transformer-Based Detection of Abusive Tamil Text Targeting Women on Social Media. In *Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Shrey Satapara, Prasenjit Majumder, Thomas Mandl, Sandip Modha, Hiren Madhu, Tharindu Ranasinghe, Marcos Zampieri, Kai North, and Damith Premasiri. 2023. Overview of the hasoc subtrack at fire 2022: Hate speech and offensive content identification in english and indo-aryan languages. In *Proceedings of the 14th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '22, page 4–7, New York, NY, USA. Association for Computing Machinery.

Shrey Satapara, Sarah Masud, Hiren Madhu, Md. Aflah Khan, Md. Shad Akhtar, Tanmoy Chakraborty, Sandip Modha, and Thomas Mandl. 2024. Overview of the hasoc subtracks at fire 2023: Detection of hate spans and conversational hate-speech. In *Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '23, page 10–12, New York, NY, USA. Association for Computing Machinery.

Sunil Saumya, Abhinav Kumar, and Jyoti Prakash Singh. 2021. Offensive language identification in dravidian code mixed social media text. In *Proceedings of the first workshop on speech and language technologies for Dravidian languages*, pages 36–45.

Nina Seemann, Yeong Su Lee, Julian Höllig, and Michaela Geierhos. 2023. The problem of varying annotations to identify abusive language in social media content. *Natural Language Engineering*, 29(6):1561–1585.

Deepawali Sharma, Vivek Kumar Singh, and Vedika Gupta. 2024. Tab hate: a target-based hate speech detection dataset in hindi. *Social Network Analysis and Mining*, 14(1):190.

Victoria L Stephenson, Brittany M Wickham, and Nicole M Capezza. 2018. Psychological abuse in the context of social media. *Violence and Gender*, 5(3):129–134.

Malliga Subramanian, Kogilavani Shanmugavadivel, Samyuktha K S, Santhiya D, Saranya P, Samritha A R, and Sivapriyan P. 2026. VISION@DravidianLangTech 2026: Abusive Tamil Text Detection Targeting Women on Social Media. In *Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Nicolas Suzor, Molly Dragiewicz, Bridget Harris, Rosalie Gillett, Jean Burgess, and Tess Van Geelen. 2019. Human rights by design: The responsibilities of social media platforms to address gender-based violence online. *Policy & internet*, 11(1):84–103.

Rithikaa V, Sanjay Krishnan K, Nithya Varshini C N, and R S. Sumathi. 2026. PrimeLine@DravidianLangTech 2026: Abusive Tamil Comment Detection Using MuRIL. In *Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12):e0243300.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott A Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the third workshop on abusive language online*, pages 80–93.

Mohammad Zia Ur Rehman, Somya Mehta, Kuldeep Singh, Kunal Kaushik, and Nagendra Kumar. 2023. User-aware multilingual abusive content detection in social media. *Information Processing & Management*, 60(5):103450.

10 Appendix

10.1 Error Analysis

To better understand the failure modes of abusive Tamil text detection targeting women, we

conducted a qualitative error analysis on the predictions of the top five systems: nitc-hsr (A and Kumar, 2026), PRIMELINE_ABUSIVE, HNK (R et al., 2026), CUET_Synthetica (Rishta et al., 2026), and DPR (Prakash et al., 2026). Rather than estimating system-level performance, this analysis aims to identify repeated linguistic and contextual occurrences that contributed to misclassification. -media language patterns, target ambiguity, and annotation boundary conditions.

10.1.1 Setup

For error analysis, we examined the predictions from the top-5 systems (NIT-HSR, PRIMELINE_ABUSIVE, HNK (R et al., 2026), CUET_Synthetica (Rishta et al., 2026), DPR (Prakash et al., 2026)) and manually inspect a curated set of mispredicted instances from the first 250 samples of test set and we use it for qualitative analysis rather than estimating overall performance. We qualitatively analyze false positives (FP- gold Non-Abusive but predicted Abusive) and false negatives (FN- gold Abusive but predicted Non-Abusive) and grouping them into categories relevant to Tamil social-media abuse targeting women.

10.1.2 Common error categories

Masked or Obfuscated Abuse (False Negatives). Users often evade moderation by masking abusive terms using partial redaction or character substitution. Models that rely on surface lexical cues struggle when the key abusive word is hidden. In several such cases, systems including nitc-hsr (A and Kumar, 2026), HNK (R et al., 2026), and CUET_Synthetica (Rishta et al., 2026) predicted Non-Abusive despite the presence of masked offensive language. These errors suggest that current transformer models lack robustness to obfuscated abusive expressions and would benefit from normalization strategies capable of handling masked tokens.

Quoting, Reporting, and Counterspeech (False Positives). False positives also occur when users criticize or question abusive behavior using strong language. In such instances, the intent of the comment is corrective rather than abusive. Systems such as primeline_abusive_Tamil (V et al., 2026), HNK (R et al., 2026), and DPR (Prakash et al., 2026) occasionally predicted Abusive because the models reacted to emotionally expressive tokens without capturing the broader stance of the comment.

Profanity or Aggressive Tone Without Abusive Intent (False Positives). Several comments contain strong insults or aggressive tone but are labeled Non-Abusive under the task definition because they are not directed toward women. Systems including primeline_abusive_Tamil (V et al., 2026), HNK (R et al., 2026), CUET_Synthetica (Rishta et al., 2026), and DPR (Prakash et al., 2026) frequently predicted such instances as Abusive. This indicates that many models equate lexical signals of toxicity with abusive language, leading to overprediction of the abusive class.

Threat or Violent Expressions Without Clear Target (False Positives). Expressions containing violent or threatening verbs often trigger abusive predictions even when the target of the comment is unclear or when the statement is used rhetorically. In several examples, all top systems primeline_abusive_Tamil (V et al., 2026), HNK (R et al., 2026), CUET_Synthetica (Rishta et al., 2026), and DPR (Prakash et al., 2026) classified such comments as Abusive, despite the ground truth annotation being Non-Abusive. This suggests that transformer models strongly associate threat-related vocabulary with abusive intent regardless of contextual interpretation.

Code-Mixed and Noisy Social Media Language. Tamil social media comments frequently contain a mixture of Tamil and English words, transliterated expressions, and spelling variations. Although multilingual transformer models such as MuRIL and IndicBERT are designed to handle multilingual data, inconsistencies in spelling and informal writing styles still lead to errors. In several cases, CUET_Synthetica (Rishta et al., 2026) and nitc-hsr (A and Kumar, 2026) predicted Non-Abusive when abusive intent was present within code-mixed expressions.

Annotation Boundary Cases and Ambiguity. Finally, certain comments fall near the boundary of the annotation guidelines, where offensiveness depends on whether the insult is explicitly directed toward women or whether it represents general criticism. Such borderline cases produced inconsistent predictions across multiple systems including primeline_abusive_Tamil (V et al., 2026), HNK (R et al., 2026), and DPR (Prakash et al., 2026). These observations highlight how dataset definitions and annotation policies influence automated classification outcomes.

Overall, most misclassifications arise from re-

liance on surface lexical cues, limited ability to interpret discourse context, and difficulty handling noisy or obfuscated social media language.