NITC-HSR@DravidianLangTech 2026: Ensembling Multilingual Transformer Models for Detecting Abusive Tamil Text Targeting Women on Social Media

Rameez Mohammed A[1] **and S D Madhu Kumar**[2]
Department of Computer Science and Engineering
National Institute of Technology Calicut
Kozhikode, India
[1]rameez_p200051cs@nitc.ac.in, [2]madhu@nitc.ac.in

Abstract

The proliferation of misogynistic content on social media platforms is a serious problem that requires the development of automated detection systems, which is a challenging task for low-resource languages like Tamil. This study investigates the effectiveness of multilingual transformer models for identifying abusive Tamil text targeting women in social media. Results indicate that such models achieve strong baseline performance on this task. Furthermore, an ensemble of two best performing models was found to improve the classification performance further. The results also highlighted the significance of domain-specific pre-training for improving classifier performance. The best performing ensemble model achieved a weighted F1 score of 0.83 on the test set, placing our approach in first position in the shared task.

1 Introduction

Of late, social media has become an overwhelmingly accepted medium for dissemination of ideas, communication and entertainment. However, its widespread prevalence has also led to harmful behaviours like cyber bullying, hate speech, harassment etc. This has led to many adverse effects like psychological distress, marginalization of communities based on gender, religion, race etc. These challenges emphasize the need for developing robust detection and mitigation mechanisms to identify, prevent and reduce the harmful impacts of abusive and toxic content in social media environments.

Hate speech directed against women has become a pervasive issue in contemporary online spaces, reflecting deep-seated gender biases and discriminatory attitudes. Such expressions often manifest in various forms, including overt sexism, the use of vulgar or abusive language, derogatory remarks aimed at undermining women's opinions and achievements, and subtle forms of belittlement or stereotyping. The persistence of such content poses significant challenges to ensuring safe, inclusive, and respectful digital environments (Madhukumar and Mohammed A, 2023). Systems for detecting abusive content in social media have been widely researched and deployed for English, but there is a significant lack of resources and models for the same in regional Indian languages. As such, developing reliable mechanisms to identify abusive content targeted at women in regional Indian languages is a significant research problem.

A shared task of DravidianLangTech 2026 focuses on detecting abusive Tamil text aimed at women in social media, and the organizers have released a dataset of comments in Tamil towards the same. The dataset contains comments sourced from social media which are classified as either Abusive of Non-Abusive (Sivagnanam et al., 2026). A number of multilingual transformers like mBERT, IndicBERT etc., were fine-tuned on this dataset and their performance were evaluated. The effectiveness of AbuseXLMR (Gupta et al., 2022), a domain-adapted transformer model pre-trained on abusive speech corpora in several Indian languages, was also assessed on this dataset. Finally, experiments were conducted with an ensemble of the two best performing transformer models to investigate potential performance improvements through model combination. The python code for the experiments can be found in GitHub. [1]

2 Related Work

Several works for detecting hate and abusive speech have been carried out for English texts, but the suitability of these methods for regional Indian languages like Tamil is under-researched. This can largely be attributed to the paucity of high quality labeled datasets in Indian languages.

[1]https://github.com/rameezp200051cs/DravidianLangTech-2026

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 316–320
July 3, 2026 ©2026 Association for Computational Linguistics

Shared tasks associated with HASOC Chakravarthi et al., 2020, 2021, DravidianLangTech Priyadharshini et al., 2022, 2023; Premjith et al., 2024; Rajiakodi et al., 2025 etc. have provided datasets in multiple Indian languages for hate and abusive speech detection. These datasets have contributed significantly towards research related to hate speech detection mechanisms for monolingual, as well as code-mixed low resource Indian languages.

Saha et al. (2021) used multilingual transformers like XLM-RoBERTA, MuRIL and IndicBERT as well a BERT-CNN fusion model to classify code-mixed offensive comments in three Dravidian languages - Tamil, Kannada and Malayalam. They observed that a Genetic Algorithm based ensembling helped improve the performance of the models.

Subramanian et al. (2022) evaluated the performance of various ML models and transformer models on a dataset of Tamil Youtube comments and found that transformer models performed better in the task.

Gupta et al. (2022) released the MACD dataset, which is a large scale annotated dataset of abusive comments in five Indian languages. They further released AbuseXLMR, a pretrained abuse detection model based on XLM-RoBERTa model, for social media content in Indian languages, including Tamil.

Sreelakshmi et al. (2024) studied different multilingual transformer models with the objective of finding a single pre-trained embedding model which can be effective in detecting hateful and offensive language in code-mixed Kannada, Tamil and Malayalam.

Habiba and Aghila (2025) used an LSTM based approach for detecting abusive language in code-mixed Malayalam and Tamil languages as part of a DravidianLangTech 2025 shared task. Hanif et al. (Hanif and Rahman, 2025) compared the performance of multilingual transformer models for the same task and observed that IndicBERT gave the best results for both the languages.

3 Dataset Description

The datasets provided for the task contained Tamil language text sourced from social media. The texts were classified as either Abusive or Non-Abusive, depending on whether the text implied misogyny or not. Train and test datasets were made available separately by the organizers of the task. Both datasets are nearly balanced. The dataset statistics are as shown in Table 1.

Dataset	Abusive	Non-Abusive	Total
Train	1769	1883	3652
Test	441	472	913

Table 1: Dataset statistics

4 Preprocessing Data

An exploratory analysis of the text provided in the datasets indicated low code-mixing. Preprocessing of the text was done to remove URLs, hashtags, email addresses, and extra whitespaces. Hashtag symbols were also removed, but the associated keyword text were retained since it may contribute useful signals for classification.

5 Methodology

Multilingual transformer models have been found to be effective for offensive speech detection tasks in low-resource languages like Tamil. Five multilingual transformer models - XLM-RoBERTa (Conneau et al., 2020), mBERT (Devlin et al., 2019), MuRIL (Khanuja et al., 2021), IndicBERTv2 (Doddapaneni et al., 2023) and AbuseXLMR (Gupta et al., 2022) were used for the classification task. Of these models, XLM-RoBERTa and mBERT are pre-trained on a wide variety of languages spoken worldwide. MuRIL and IndicBERTv2 are pre-trained in Indian languages. AbuseXLMR is pre-trained exclusively on abusive social media comments in five Indian languages. The models have all been pre-trained in Tamil, which makes them suitable to be employed for the given task. All models were fine-tuned on 80% of the training dataset. 20% of the training set was designated as the validation set and the model performance was evaluated on the same. We fine-tuned models using the Hugging Face Transformers library with PyTorch. AdamW was used as the optimizer. The hyperparameters used for training the models are as shown in Table 2. They were obtained by

Hyperparameters	Values
Learning Rate	2e-5
Epochs	7
Weight Decay	0.01
Batch Size	16

Table 2: Hyperparameters used for training

Model	Accuracy	Precision	Recall	F1 Score
mBERT	0.7836	0.7438	0.8446	0.7910
XLM-RoBERTa	0.8098	0.7961	0.8164	0.8061
AbuseXLMR	0.8194	0.7761	0.8816	**0.8254**
MuRIL	0.8276	0.8295	0.8107	0.82
IndicBERTv2	0.8304	0.8267	0.8220	0.8244

Table 3: Performance of multilingual transformers on the train set

Ensembling Strategy	Accuracy	Precision	Recall	F1 Score
Soft Voting	0.8386	0.8172	0.8588	0.8375
Hard Voting	0.8249	0.8087	0.8362	0.82
Weighted Voting	0.8386	0.8172	0.8586	0.8375
Stacking - Logistic Regression	0.8497	0.8588	0.8249	**0.8415**
Stacking - Random Forest	0.83	0.8452	0.8022	0.8232

Table 4: Performance of AbuseXLMR - IndicBERTv2 ensemble on train set

manual tuning, following the commonly adopted settings for fine-tuning transformer models for text classification tasks.

5.1 Ensembling Models

Model ensembling is often employed as an effective strategy for improving classification performance, by leveraging the complementary strengths of the individual models. As such, we explored whether combining the predictions of the models under consideration could lead to further gains in classifier performance. As can be observed from Table 3, AbuseXLMR and IndicBERTv2 achieved the highest individual F1 scores. Accordingly, an ensemble of these two best performing models were evaluated with four different strategies:

- Soft voting - averages the predicted class probabilities of the models to determine the final classification
- Hard voting - takes a majority vote of the class labels predicted by the models to determine the final classification
- Weighted voting - combines predictions by giving higher weights to the predictions of the better performing model
- Stacking - trains meta-learners (linear regression, random forest) on the predictions of the models

6 Results

Table 3 reports the performance metrics obtained for the multilingual models on the classification task. The metrics listed are accuracy, precision, recall and macro F1 score. All the models perform strongly on the task, as evidenced by the metrics obtained. AbuseXLMR achieves the highest F1 score of 0.8254, followed by IndicBERTv2 with an F1 score of 0.8244. Since AbuseXLMR has been pretrained on abusive texts in Tamil (among other languages), it is able to perform marginally better than IndicBERTv2, which itself has been trained extensively on Indian languages. However, AbuseXLMR performs significantly better than its base model, XLM-RoBERTa, underscoring the fact that domain specific pretraining can lead to improvements over general-purpose multilingual models for abusive speech detection tasks.

Table 4 reports the performance of ensembled configurations of the two best performing multilingual models, AbuseXLMR and IndicBERTv2. Ensemble models exhibit stronger performance on the classification task as compared to the various multilingual models evaluated individually, as is evident from the higher values for the various metrics. Stacking with logistic regression strategy achieved the highest F1 score of 0.8415, combining the predictions from AbuseXLMR and IndicBERTv2 models. The better performance of this ensemble may be due to the complementary strengths of the two models - domain-specific pre-training (AbuseXLMR) and language-specific representations (IndicBERTv2), which the meta-learner combines adaptively based on prediction confidence and instance characteristics.

Table 5 presents the evaluation metrics achieved by the best performing ensemble model on the test

Model	Accuracy	Precision	Recall	Macro F1 Score	Weighted F1 Score
AbuseXLMR + IndicBERTv2 Ensemble (Stacking - Logistic Regression)	0.8302	0.8310	0.8293	0.8297	0.83

Table 5: Performance of AbuseXLMR - IndicBERTv2 ensemble on test set

set provided by the organizers. The model obtained a macro F1 of 0.8297 and weighted F1 of 0.83, placing our approach in first position in the shared task. The second placed team obtained a macro F1 of 0.8133 and weighted F1 of 0.81, whereas the third placed team obtained a macro F1 of 0.8103 and weighted F1 of 0.80.

7 Conclusion

This work investigates the use of multilingual transformer models for the task of detecting abusive Tamil text aimed at women in social media. Our experiments demonstrate that pre-trained multilingual models achieve strong baseline performance on this task. Moreover, it was observed that a domain-adapted model, AbuseXLMR, outperformed other models, highlighting the importance of task-specific pre-training. Ensembling AbuseXLMR and IndicBERTv2 models led to additional performance gains, highlighting the benefits of combining complementary model strengths. As future work, we plan to extend this approach for misogyny detection in other low-resource Indian languages as well.

8 Limitations

Dataset scarcity is the main hurdle to creating effective abuse detection models for low-resource languages like Tamil. The dataset provided for the shared task was used as is, with minimal preprocessing, for fine-tuning the models. Augmenting the dataset, like using back-translation, synonym replacements etc. may have improved the classification performance of the models. The models discussed in the paper are variants of multilingual transformers, fine tuned on the provided dataset. An evaluation of classifiers based on conventional machine learning models like SVM, Random Forest etc. was not conducted. A relative comparison of the performance of the classifiers based on these two approaches could have been informative.

9 GenAI Usage Disclosure

Anthropic's Claude was used exclusively for language refinement of the manuscript and debugging the code scripts associated with our methodology. No original text, arguments, or experimental conclusions were generated by the AI.

References

Bharathi Raja Chakravarthi, Prasanna Kumar Kumaresan, Ratnasingam Sakuntharaj, Anand Kumar Madasamy, Sajeetha Thavareesan, B Premjith, K Sreelakshmi, Subalalitha Chinnaudayar Navaneethakrishnan, John P McCrae, and Thomas Mandl. 2021. Overview of the hasoc-dravidiancodemix shared task on offensive language detection in tamil and malayalam. In *FIRE (Working Notes)*, pages 589–602.

Bharathi Raja Chakravarthi, Anand Kumar M, John P McCrae, Bhavukam Premjith, KP Soman, and Thomas Mandl. 2020. Overview of the track on hasoc-offensive language identification-dravidiancodemix. In *FIRE (Working notes)*, pages 112–120.

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.

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.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no indic language behind: Building monolingual corpora, benchmark and models for indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426.

Vikram Gupta, Sumegh Roychowdhury, Mithun Das, Somnath Banerjee, Punyajoy Saha, Binny Mathew, Animesh Mukherjee, and 1 others. 2022. Multilingual abusive comment detection at scale for indic languages. *Advances in Neural Information Processing Systems*, 35:26176–26191.

A Habiba and G Aghila. 2025. Dltcnitpy@ dravidianlangtech 2025 abusive code-mixed text detection system targeting women for tamil and malayalam languages using deep learning technique. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 567–572.

Tareque Md Hanif and Md Rashadur Rahman. 2025. Cuet_agile@ dravidianlangtech 2025: Fine-tuning transformers for detecting abusive text targeting women from tamil and malayalam texts. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 315–319.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, and 1 others. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

SD Madhukumar and Rameez Mohammed A. 2023. Hate speech detection in indian languages: A brief survey. In *2023 IEEE 2nd International Conference on Data, Decision and Systems (ICDDS)*, pages 1–5. IEEE.

B Premjith, Bharathi Raja Chakravarthi, Prasanna Kumar Kumaresan, Saranya Rajiakodi, Sai Karnati, Sai Mangamuru, and Chandu Janakiram. 2024. Findings of the shared task on hate and offensive language detection in telugu codemixed text (hold-telugu)@ dravidianlangtech 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 49–55.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Kogilavani SV, Abirami Murugappan, Prasanna Kumar Kumaresan, and 1 others. 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.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumar Kumaresan. 2022. Findings of the shared task on abusive comment detection in tamil. In *Proceedings of the second workshop on speech 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 Buitelaar, 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.

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.

Bhuvaneswari Sivagnanam, Kathiravan Pannerselvam, Jananayagan V, Charmathi Rajkumar, Ramesh Kannan R, Ratnavel Rajalakshmi, Shunmuga Priya Muthusamy Chinnan, Saranya Rajiakodi, and Bharathi Raja Chakravarthi. 2026. From Comments to Harm: A Findings Report on 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.

K Sreelakshmi, B Premjith, Bharathi Raja Chakravarthi, and KP Soman. 2024. Detection of hate speech and offensive language codemix text in dravidian languages using cost-sensitive learning approach. *IEEE Access*, 12:20064–20090.

Malliga Subramanian, Rahul Ponnusamy, Sean Benhur, Kogilavani Shanmugavadivel, Adhithiya Ganesan, Deepti Ravi, Gowtham Krishnan Shanmugasundaram, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2022. Offensive language detection in tamil youtube comments by adapters and cross-domain knowledge transfer. *Computer Speech & Language*, 76:101404.