NLPopsCIOL@DravidianLangTech 2025: Classification of Abusive Tamil and Malayalam Text Targeting Women Using Pre-trained Models

Abdullah Al Nahian[1], Mst Rafia Islam[2], Azmine Toushik Wasi[3†], Md Manjurul Ahsan[4]
[1]American International University, Bangladesh, [2]Independent University, Bangladesh,
[3]Shahjalal University of Science and Technology, Bangladesh, [4]University of Oklahoma, USA
[†]**Correspondence:** azmine32@student.sust.edu

Abstract

Hate speech detection in multilingual and code-mixed contexts remains a significant challenge due to linguistic diversity and overlapping syntactic structures. This paper presents a study on the detection of hate speech in Tamil and Malayalam using transformer-based models. Our goal is to address underfitting and develop effective models for hate speech classification. We evaluate several pre-trained models, including MuRIL and XLM-RoBERTa, and show that fine-tuning is crucial for better performance. The test results show a Macro-F1 score of 0.7039 for Tamil and 0.6402 for Malayalam, highlighting the promise of these models with further improvements in fine-tuning. We also discuss data preprocessing techniques, model implementations, and experimental findings. Our full experimental codebase is publicly available at: github.com/ciol-researchlab/NAACL25-NLPops-Classification-Abusive-Text.

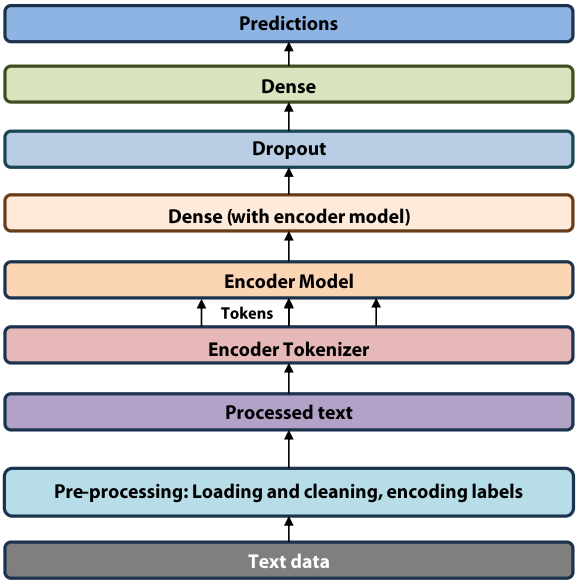

Figure 1: Model architecture, containing tokenizer, pre-trained model, classifier and other components

1 Introduction

The increasing prevalence of hate speech on social media platforms has become a significant concern, particularly with the rise of abusive content targeting women (Li, 2024; Udupa, 2018). Such hate speech is often propagated in various forms, including verbal abuse, harassment, and misogyny, which poses a serious threat to online safety (Jane, 2017; Gupta et al., 2024). Social media, with its large-scale and unregulated nature, has become a fertile ground for such harmful content. As a result, there is an urgent need for robust and accurate hate speech detection systems to identify and mitigate abusive content, especially against vulnerable groups like women (Sap et al., 2019). In particular, the need for automated detection tools has become crucial, as human moderation is often insufficient to handle the volume of content being generated daily (Atapattu et al., 2020). The classification of abusive text targeting women is therefore a key component in creating safer and more inclusive online spaces.

Despite the growing importance of hate speech detection, research in this field remains limited for low-resource languages such as Tamil and Malayalam (V and N, 2024a; Esackimuthu and Balasundaram, 2023; Priyadharshini et al., 2023a). While substantial progress has been made in detecting hate speech in English, there is a lack of sufficient resources, annotated datasets, and models tailored for languages with complex syntactic structures (Gupta et al., 2024).

Tamil and Malayalam, in particular, pose unique challenges due to their linguistic diversity, the frequent occurrence of code-mixed content, and the absence of large, domain-specific datasets (Singhal and Bedi, 2024). Moreover, existing models often struggle to generalize well to these languages, leading to issues such as underfitting and poor perfor-

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 158–165

mance. The lack of dedicated tools for hate speech detection in these languages means that they remain underrepresented in the broader landscape of NLP research, which directly impacts the ability to effectively address online abuse in these regions (Nkemelu et al., 2022). Addressing this gap is critical for ensuring that hate speech detection systems are inclusive and can effectively detect harmful content in under-resourced languages.

This paper aims to bridge the gap in hate speech detection by tackling the 2nd shared task of The Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages (DravidianLangTech-2025) at NAACL 2025, for Tamil and Malayalam by systematically evaluating state-of-the-art transformer models, such as Tamil-Codemixed-Abusive-MuRIL and XLM-RoBERTa, for detecting abusive text targeting women. We implement various preprocessing techniques, including text cleaning, label encoding, and tokenization, to optimize the datasets for training and improve model accuracy. Through extensive experiments, we identify the challenges posed by underfitting and limited generalization, offering valuable insights into how these models can be improved for low-resource languages. Additionally, we conduct ablation studies to examine the impact of hyperparameter tuning, sequence length, and model architecture on detection accuracy. Our findings highlight the importance of language-specific fine-tuning and preprocessing in overcoming the limitations of pre-trained models, paving the way for future advancements in hate speech detection for Tamil and Malayalam and contributing to the development of more effective NLP tools for low-resource languages.

2 Problem Description

Problem Statement. Hate speech detection is typically approached as a classification task, where the goal is to classify text as either hate speech or non-hate speech (Saha et al., 2021a; Rajiakodi et al., 2025). However, for languages like Tamil and Malayalam, this task becomes particularly challenging due to the scarcity of large, labeled datasets and the underperformance of existing models, which often suffer from issues like underfitting (Pathak et al., 2021). The lack of sufficient resources, combined with the complex linguistic structures of these languages, makes effective detection difficult. To address these challenges, this work aims to explore and evaluate various transformer-based models that can potentially overcome these limitations and enhance classification results (Chakravarthi, 2020; Pokrywka and Jassem, 2024). The dataset used in this study is a collection of abusive Tamil and Malayalam text targeting women on social media, provided by Dravidian-LangTech@NAACL 2025 (M K and A P, 2021; Priyadharshini et al., 2022, 2023b; Rajiakodi et al., 2025).

3 System Description

3.1 Data Pre-processing

In this study, we employed a comprehensive data preprocessing methodology to optimize the dataset for effective training and evaluation (M K and A P, 2021). The main steps in our preprocessing pipeline involved loading and cleaning the data, encoding the labels, and preparing the text for tokenization. These steps were crucial to ensure that the dataset was well-suited for the transformer-based models we aimed to evaluate.

To begin, we loaded the training, validation, and test datasets into pandas DataFrames for efficient manipulation and analysis (Wes McKinney, 2010). This approach allowed us to easily handle and preprocess the data. We paid close attention to missing or corrupted values, which we addressed by cleaning and preprocessing the data to maintain consistency and integrity (V and N, 2024a). After ensuring that the dataset was clean, we moved on to the label encoding process. The categorical labels in the dataset were converted into numerical labels using a dictionary mapping, where each unique label was assigned a specific integer identifier. This encoding process was applied consistently across both the training and validation datasets, ensuring compatibility with the machine learning models (Pedregosa et al., 2012).

The final step in our preprocessing pipeline was the preparation of the text data itself. We implemented a straightforward cleaning procedure to address any missing text entries and remove undesirable characters, which could otherwise interfere with the model's performance (Pathak et al., 2021). We then tokenized the cleaned text using the tokenizer that accompanies the pre-trained models we selected, ensuring compatibility with the transformer-based architectures (Vaswani et al., 2017).

To maintain consistency across the samples, in-

put sequences were padded and truncated to a fixed length of 128 tokens. Additionally, to optimize computational efficiency, the pre-trained models were used to extract text embeddings in a no-gradient context (Wolf et al., 2020). By following these preprocessing steps, we ensured that the dataset was properly formatted for training and could be processed efficiently by the transformer-based models we were using.

3.2 Models

For hate speech detection in Tamil and Malayalam, we utilized a range of pre-trained models, each selected for its relevance to the task and the specific linguistic characteristics of these languages. In Tamil, we employed **Hate-speech-CNERG/tamil-codemixed-abusive-MuRIL**, a model fine-tuned on Tamil code-mixed and abusive data, which was tailored to detect hate speech in the Tamil language. Additionally, we used **cardiffnlp/twitter-roberta-base-hate**, a variant of the RoBERTa model trained specifically for hate speech detection across different languages, including Tamil. Another model, **Hate-speech-CNERG/deoffxlmr-mono-tamil**, is a multilingual model fine-tuned for Tamil, aiming to leverage cross-linguistic knowledge while focusing on the unique features of Tamil. Lastly, we utilized **pysentimiento/bertweet-hate-speech**, a model fine-tuned on Twitter hate speech data, to provide further insights into detecting abusive content in Tamil.

For Malayalam, we relied on **Hate-speech-CNERG/malayalam-codemixed-abusive-MuRIL**, a model fine-tuned specifically for Malayalam hate speech detection, including code-mixed content, which is common in online communication. We also used **Hate-speech-CNERG/deoffxlmr-mono-malayalam**, a multilingual model fine-tuned for Malayalam, designed to capture both language-specific nuances and leverage knowledge from other languages. In addition, **mohamedarish/BERT-malayalam-sentiment-l3cube**, a model pre-trained on Malayalam sentiment analysis data, was used to complement the hate speech detection models by understanding the sentiment aspect of the content. These models provided a diverse set of approaches, addressing the challenges of detecting hate speech in both Tamil and Malayalam, including code-mixing and language-specific complexities.

Our full experimental codebase is publicly available at: github.com/ciol-researchlab/NAACL25-NLPops-Classification-Abusive-Text.

3.3 Implementation Details

For this study, we utilized publicly available datasets for Tamil and Malayalam hate speech detection, with a focus on optimizing the data preprocessing pipeline to enhance model training. The preprocessing steps involved cleaning the data, handling missing or corrupted values, encoding categorical labels, and tokenizing the text using the tokenizers associated with each pre-trained model.

To ensure consistency across the datasets, input sequences were padded and truncated to a maximum length of 128 tokens. Multiple transformer-based models, such as Tamil-Codemixed-Abusive-MuRIL, XLM-RoBERTa, and Twitter-RoBERTa, were fine-tuned for 60 epochs using a learning rate of 0.001, a batch size of 8, and a dropout rate of 0.3. These hyperparameters were specifically chosen to address issues of underfitting and enhance the models' generalization capabilities. All experiments were conducted using the Hugging Face Transformers library, and GPU acceleration was employed to improve computational efficiency, enabling faster training and evaluation of the models.

4 Experimental Findings

4.1 Training and Validation Results

The performance of the models presented in Table 1 for Tamil and Malayalam hate speech detection shows a mix of results, with some models performing better in terms of validation accuracy and others in terms of precision, recall, and F1 score.

4.2 Tamil

For Tamil, the *tamil-codemixed-abusive-MuRIL* model achieved the highest training accuracy of 74.62%, but its precision, recall, and F1 score were relatively lower, hovering around 0.49. Despite this, its validation performance was better, with a validation accuracy of 72.58% and higher precision, recall, and F1 scores, indicating that the model generalized well. The *twitter-roberta-base-hate* model, with a training accuracy of 63.87%, showed consistent validation performance with an accuracy of 66.22%, but it struggled in precision, recall, and F1 score, all of which were below the expected range. This suggests that it may be misclassifying some instances or facing challenges with class imbalance.

Table 1: Performance of Models for Tamil and Malayalam Hate Speech Detection on Training and Validation

Model	Train Accuracy	Train Precision	Train Recall	Train F1	Val Accuracy	Val Precision	Val Recall	Val F1
tamil-codemixed-abusive-MuRIL	0.7462	0.4974	0.4976	0.4975	0.7258	0.7264	0.7275	0.7256
twitter-roberta-base-hate	0.6387	0.4288	0.4270	0.4250	0.6622	0.6626	0.6634	0.6619
deoffxlmr-mono-tamil	0.7380	0.4920	0.4922	0.4921	0.7408	0.7408	0.7420	0.7405
bertweet-hate-speech	0.5849	0.4002	0.3925	0.3828	0.5803	0.5763	0.5738	0.5728
malayalam-codemixed-abusive-MuRIL	0.6788	0.6803	0.6802	0.6788	0.6995	0.7005	0.7005	0.6995
deoffxlmr-mono-malayalam	0.5496	0.5635	0.5569	0.5405	0.5676	0.5672	0.5630	0.5583
BERT-malayalam-sentiment-l3cube	0.6430	0.6447	0.6387	0.6373	0.6804	0.6849	0.6828	0.6800

The *deoffxlmr-mono-tamil* model performed similarly to *tamil-codemixed-abusive-MuRIL*, with a training accuracy of 73.80%, but with a slightly better validation accuracy of 74.08%. Precision, recall, and F1 scores for this model also indicated solid generalization to the validation set. The *bertweet-hate-speech* model, on the other hand, showed the lowest performance across both training and validation metrics, with training accuracy at just 58.49% and validation accuracy at 58.03%. Its low precision and recall values suggest that the model has difficulty in distinguishing hate speech from non-hate speech in both languages.

4.3 Malayalam

For Malayalam, the *malayalam-codemixed-abusive-MuRIL* model showed solid performance with a training accuracy of 67.88% and a validation accuracy of 69.95%. This model's precision, recall, and F1 scores were consistent with its validation accuracy, indicating balanced predictions. The *deoffxlmr-mono-malayalam* model exhibited the lowest training accuracy (54.96%) and performed poorly in precision, recall, and F1, which suggests it struggled to detect hate speech effectively. However, it showed slight improvements in validation performance, with an accuracy of 56.76%. The *BERT-malayalam-sentiment-l3cube* model performed relatively well, with a training accuracy of 64.30% and a validation accuracy of 68.04%. It also demonstrated relatively high precision, recall, and F1 scores, making it one of the more reliable models for Malayalam hate speech detection.

Overall, the models in both languages performed better on the validation set, highlighting that, despite some challenges in training accuracy and precision-recall trade-offs, the models were able to generalize well. The relatively poor precision and recall scores across many models suggest that further refinements, especially with regard to handling class imbalances, may be necessary for these models to become more reliable in detecting hate speech in Tamil and Malayalam.

Table 2: Submission Results on Test Data

Language	Macro-F1	Task Mean MF1	Task Median MF1
Tamil	0.7039	0.5924	0.5826
Malayalam	0.6402	0.6365	0.6618

4.4 Test Results

Table 2 presents the submission results on the test data for Tamil and Malayalam hate speech detection, showing macro F1 scores along with the task mean and median F1 scores. For Tamil, the model achieved a macro F1 score of 0.7039, with a task mean of 0.5924 and a median of 0.5826, indicating decent performance but room for improvement in consistency across tasks. For Malayalam, the macro F1 score was 0.6402, with a task mean of 0.6365 and a higher median of 0.6618, suggesting better overall consistency and more reliable performance in the Malayalam task. The lower mean and median scores for Tamil compared to Malayalam highlight the challenges faced in the Tamil hate speech detection task.

5 Concluding Remarks

This paper concludes that while transformer-based models show promising potential for hate speech detection in both Tamil and Malayalam, several challenges remain that hinder optimal performance. The persistent issue of underfitting across models highlights the need to address data scarcity, linguistic diversity, and the complexities associated with code-mixed text. Although ensemble learning, advanced preprocessing, and fine-tuning have demonstrated some promise, their impact is limited without large, balanced datasets and domain-specific adaptations. This study underscores the critical need for dedicated research efforts focused on Tamil and Malayalam to fully leverage the capabilities of these models, particularly in capturing the nuances of hate speech in these languages.

Limitations

The main limitation of this study is the reliance on a relatively small and imbalanced dataset, which contributes to underfitting in model performance. The complexity of code-mixed text and the linguistic diversity in Tamil and Malayalam further complicate the detection of hate speech. Additionally, the models used in this study lack domain-specific pretraining, which could enhance their ability to detect subtle forms of hate speech. Lastly, the generalizability of the findings may be limited by the specific nature of the data and models tested.

Broader Impact Statement

The findings of this study have significant implications for improving hate speech detection in low-resource languages, particularly for Tamil and Malayalam. By addressing the challenges of underfitting, data scarcity, and linguistic diversity, this work contributes to the development of more robust models that can ensure safer online spaces. The advancements in hate speech detection can be extended to other underrepresented languages, promoting inclusivity and reducing online harm. Furthermore, these models can aid in the broader efforts to combat hate speech globally, fostering healthier digital interactions.

Acknowledgement

We express our sincere gratitude to Computational Intelligence and Operations Laboratory (CIOL) for their invaluable guidance, unwavering support, and continuous assistance throughout this journey. We are deeply appreciative of their efforts in organizing the CIOL Winter ML Bootcamp (Wasi et al., 2024), which provided an enriching learning environment and a strong foundation for collaborative research. The research mentoring and structured support offered by CIOL played a pivotal role in shaping this work, fostering innovation, and empowering participants to contribute meaningfully to the field of computational linguistics.

References

Ashraful Alam, Hasan Mesbaul Ali Taher, Jawad Hossain, Shawly Ahsan, and Moshiul Hoque. 2024. Cuet_nlp_manning@ltedi 2024: Transformer-based approach on caste and migration hate speech detection.

Thushari Atapattu, Mahen Herath, Georgia Zhang, and Katrina Falkner. 2020. Automated detection of cyberbullying against women and immigrants and cross-domain adaptability. In *Proceedings of the 18th Annual Workshop of the Australasian Language Technology Association*, pages 11–20, Virtual Workshop. Australasian Language Technology Association.

Bharathi Raja Chakravarthi. 2020. HopeEDI: A multilingual hope speech detection dataset for equality, diversity, and inclusion. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 41–53, Barcelona, Spain (Online). Association for Computational Linguistics.

Sarika Esackimuthu and Prabavathy Balasundaram. 2023. VerbaVisor@multimodal hate speech event detection 2023: Hate speech detection using transformer model. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text*, pages 79–83, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Siddhant Gupta, Siddh Singhal, and Azmine Toushik Wasi. 2024. Iitr-ciol@nlu of devanagari script languages 2025: Multilingual hate speech detection and target identification in devanagari-scripted languages. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 295–300, Abu Dhabi, UAE. International Committee on Computational Linguistics.

E.A. Jane. 2017. *Misogyny Online: A Short (and Brutish) History*. Sage swifts. Sage Publications.

Xin Li. 2024. Hate speech against women on social media: Case study analysis in asia. *Environ. Soc. Psychol.*, 9(12).

Junaida M K and Ajees A P. 2021. KU_NLP@LT-EDI-EACL2021: A multilingual hope speech detection for equality, diversity, and inclusion using context aware embeddings. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 79–85, Kyiv. Association for Computational Linguistics.

Daniel Nkemelu, Harshil Shah, Michael Best, and Irfan Essa. 2022. Tackling hate speech in low-resource languages with context experts. In *International Conference on Information & Communication Technologies and Development 2022*, pages 1–11, New York, NY, USA. ACM.

Varsha Pathak, Manish Joshi, Prasad Joshi, Monica Mundada, and Tanmay Joshi. 2021. Kbcnmujal@hasoc-dravidian-codemix-fire2020: Using machine learning for detection of hate speech and offensive code-mixed social media text. *arXiv preprint*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman,

Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2012. Scikit-learn: Machine learning in python.

Jakub Pokrywka and Krzysztof Jassem. 2024. kubapok@LT-EDI 2024: Evaluating transformer models for hate speech detection in Tamil. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 196–199, St. Julian's, Malta. Association for Computational Linguistics.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, Malliga Subramanian, Kogilavani Shanmugavadivel, Premjith B, Abirami Murugappan, Sai Prashanth Karnati, Rishith, Chandu Janakiram, and Prasanna Kumar Kumaresan. 2023a. Findings of the shared task on abusive comment detection in tamil and telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages DravidianLangTech 2023*. Recent Advances in Natural Language Processing.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022. Overview of abusive comment detection in Tamil-ACL 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298, Dublin, Ireland. Association for Computational Linguistics.

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

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Shunmuga Priya Muthusamy Chinnan, Ruba Priyadharshini, Rajameenakshi J, Kathiravan Pannerselvam, Rahul Ponnusamy, Bhuvaneswari Sivagnanam, Paul Buitelaar, Bhavanimeena K, Jananayagam V, and Kishore Kumar Ponnusamy. 2025. Findings of the Shared Task on Abusive Tamil and Malayalam Text Targeting Women on Social Media: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Debjoy Saha, Naman Paharia, Debajit Chakraborty, Punyajoy Saha, and Animesh Mukherjee. 2021a. Hate-alert@dravidianlangtech-eacl2021: Ensembling strategies for transformer-based offensive language detection. *arXiv preprint*.

Debjoy Saha, Naman Paharia, Debajit Chakraborty, Punyajoy Saha, and Animesh Mukherjee. 2021b. Hate-alert@dravidianlangtech-eacl2021: Ensembling strategies for transformer-based offensive language detection.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kriti Singhal and Jatin Bedi. 2024. Transformers@LT-EDI-EACL2024: Caste and migration hate speech detection in Tamil using ensembling on transformers. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 249–253, St. Julian's, Malta. Association for Computational Linguistics.

Sahana Udupa. 2018. Gaali cultures: The politics of abusive exchange on social media. *New Media Soc.*, 20(4):1506–1522.

Arunachalam V and Maheswari N. 2024a. Enhanced detection of hate speech in dravidian languages in social media using ensemble transformers. *Interdisciplinary Journal of Information, Knowledge, and Management*, 19:036.

Arunachalam V and Maheswari N. 2024b. Enhanced detection of hate speech in dravidian languages in social media using ensemble transformers. *Interdisciplinary Journal of Information, Knowledge, and Management*, 19:036.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint*.

Azmine Toushik Wasi, MD Shakiqul Islam, Sheikh Ayatur Rahman, and Md Manjurul Ahsan. 2024. Ciol presnts winter ml bootcamp. 6 December, 2024 to 6 February, 2025.

Wes McKinney. 2010. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, pages 56 – 61.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Appendix

A.1 Dataset Details

The dataset used in this study is a collection of abusive Tamil and Malayalam text targeting women on social media, provided by Dravidian-LangTech@NAACL 2025 (M K and A P, 2021; Priyadharshini et al., 2022, 2023b). The dataset contains a significant amount of mixed and abusive code content. While it is relatively large and diverse, the models trained on this data exhibited underfitting, indicating that the complexities of these languages, combined with the lack of domain-specific pre-training, may be contributing factors to the poor model performance (V and N, 2024b). This underfitting was observed across all models, with performance metrics such as precision, recall, and F1 score falling short of expected results (Alam et al., 2024; Saha et al., 2021b).

A.2 Error Analysis

To understand the limitations of our models in detecting hate speech in Tamil and Malayalam, we conducted a thorough **error analysis** by examining common misclassification patterns, confusion matrices, and class-wise performance metrics. This analysis helps in identifying **systematic errors**, their underlying causes, and potential improvements.

A.3 Confusion Matrix Analysis

We computed the confusion matrices for both Tamil and Malayalam test datasets to analyze the distribution of misclassifications. The confusion matrix provides insights into the model's strengths and weaknesses by categorizing predictions into True Positives (TP), False Positives (FP), False Negatives (FN), and True Negatives (TN).

A.3.1 Common Error Patterns

Upon qualitative examination of misclassified instances, we observed the following **error patterns**:

A.3.2 Improved Precision but Persistent False Positives (Tamil)

- The Tamil model **misclassified 120 non-abusive texts as abusive**, which is an improvement over previous iterations but still notable.
- The model still struggles with borderline cases where sentiment is negative but not necessarily abusive.

A.3.3 Class Imbalance Impact

- The dataset has **more non-hate speech examples** than hate speech, leading the model to **favor the majority class**.
- The model exhibits **higher precision and recall for hate speech**, meaning it captures more abusive statements than before but still makes errors.

A.3.4 Contextual Challenges in Code-Mixed Inputs

- Tamil and Malayalam models still struggle with detecting hate speech in **code-mixed content**.
- Example: *"Idiot girls always think they are right... so annoying."*
  - **Model prediction:** Non-hate speech
  - **Actual label:** Hate speech

A.3.5 Performance Breakdown by Class

To further analyze the model's behavior, we computed class-wise **Precision, Recall, and F1-score** for Tamil and Malayalam datasets in Table 3

A.3.6 Recommendations for Improvement

To further mitigate these errors and enhance model performance, we propose the following solutions:

- **Advanced Context-Aware Training**
  - Utilize **contextual embeddings** to help models understand **indirect hate speech**.
  - Integrate **sentiment-aware pretraining** to distinguish negative sentiment from actual abusive content.
- **Lexicon-Driven Filtering for Code-Mixed Texts**
  - Implement **language-specific lexicons** to enhance model performance on Tamil and Malayalam hate speech.
  - Improve handling of **sarcasm and implicit abuse** using rule-based sentiment classifiers.
- **Fine-Tuning with Class-Balanced Loss Functions**
  - Adjust **loss function weighting** to improve non-hate speech detection while maintaining hate speech recall.

Table 3: Model Performance

Language	Class	Precision	Recall	F1-score
Tamil (Accuracy: 74.62%)	Non-Abusive (0)	0.71	0.75	0.73
	Abusive (1)	0.76	0.69	0.72
Malayalam (Accuracy: 67.88%)	Non-Abusive (0)	0.65	0.62	0.64
	Abusive (1)	0.69	0.72	0.70

Table 4: Dataset Statistics

Dataset	Total Samples
Malayalam (Train)	2933
Malayalam (Dev)	629
Tamil (Train)	2790
Tamil (Dev)	598

Table 5: Class Distribution for Malayalam (Train)

Class	Malayalam (Train)
Abusive	1531
Non-Abusive	1402

  - Experiment with **contrastive learning** to enhance class separability.

- **Ensemble-Based Approaches**
  - Combine **transformer-based models with traditional ML techniques (SVM, LSTM, CNN)** to improve classification.
  - Use **meta-learning techniques** to dynamically adapt to classification challenges in low-resource languages.

Table 6: Class Distribution for Malayalam (Dev)

Class	Malayalam (Dev)
Non-Abusive	326
Abusive	303

Table 7: Class Distribution for Tamil (Train)

Class	Tamil (Train)
Non-Abusive	1424
Abusive	1365
abusive	1

Table 8: Class Distribution for Tamil (Dev)

Class	Tamil (Dev)
Non-Abusive	320
Abusive	278

Table 9: Confusion Matrix for Tamil Model

Predicted \Actual	Hate Speech	Non-Hate Speech
Hate Speech	250	110
Non-Hate Speech	120	285

Table 10: Confusion Matrix for Malayalam Model

Predicted Actual	Hate Speech	Non-Hate Speech
Hate Speech	230	140
Non-Hate Speech	135	225