Mitigating Abusive Comment Detection in Tamil Text: A Data Augmentation Approach with Transformer Model

Reshma Sheik and **Raghavan Balanathan** and **S. Jaya Nirmala**
Department of Computer Science and Engineering,
National Institute of Technology, Tiruchirapalli,
Tamil Nadu, India
rezmasheik@gmail.com, raghavanb21@gmail.com, sjaya@nitt.edu

Abstract

With the increasing number of users on social media platforms, the detection and categorization of abusive comments have become crucial, necessitating effective strategies to mitigate their impact on online discussions. However, the intricate and diverse nature of low-resource Indic languages presents a challenge in developing reliable detection methodologies. This research focuses on the task of classifying YouTube comments written in Tamil language into various categories. To achieve this, our research conducted experiments utilizing various multi-lingual transformer-based models along with data augmentation approaches involving back translation approaches and other pre-processing techniques. Our work provides valuable insights into the effectiveness of various preprocessing methods for this classification task. Our experiments showed that the Multilingual Representations for Indian Languages (MURIL) transformer model, coupled with round-trip translation and lexical replacement, yielded the most promising results, showcasing a significant improvement of over 15 units in macro F1-score compared to existing baselines. This contribution adds to the ongoing research to mitigate the adverse impact of abusive content on online platforms, emphasizing the utilization of diverse preprocessing strategies and state-of-the-art language models.

1 Introduction

Abusive speech, a form of communication intending to harm or promote hatred against vulnerable individuals or groups based on factors like gender, racial background, religious beliefs, complexion, or physical capabilities, often employs offensive language. It can lead to severe psychological effects on the targeted people, potentially pushing them towards wrongful actions (Prasanth et al., 2022). With the widespread expansion of social networking sites such as Facebook and Twitter, the internet has become a hub for extensive information exchange. However, this digital boom has introduced challenges, particularly the rise in issues like hate speech and cyberbullying. Among these, abusive comments stand out as a troubling concern, characterized by the use of offensive language against individuals or groups. Online abuse has far-reaching effects, including diminished self-esteem, depression, widespread harassment, etc. Identifying and addressing these comments is of utmost importance. Effectively categorizing these comments not only aids in assessing their seriousness but also equips authorities with the means to undertake suitable measures against those responsible.

Our study concentrates on detecting abusive comments in Tamil, a Dravidian language spoken by the Tamil community in South Asia, which holds a special place as one of India's 22 scheduled languages (Priyadharshini et al., 2022a). The limited linguistic resources of Tamil present a hurdle for natural language processing, creating challenges in the acquisition of sufficient datasets. Detecting abusive comments essentially involves text classification, which aims to categorize text into predefined classes or categories. While previous research has used transformer models (Vaswani et al., 2017) for this purpose, data augmentation in Tamil has been unexplored due to its status as a low-resource language and its inherent challenges.

This paper focuses on conducting experiments employing various transformer-based models and exploring different data preprocessing and augmentation techniques. Back translation methods such as round-trip translation and lexical replacement have also been explored. Our primary aim was to identify the most effective model by analyzing the outcomes, focusing on addressing the critical challenge of combating abusive comments in the digital landscape of Tamil-speaking communities.

2 Related Work

Identifying online abusive content is a challenging task, particularly due to the hairline boundary between abusive language and free expression (Khairy et al., 2021). The paper (Ziehe et al., 2021) outlines methods for detecting Hope Speech in short, casual written content in English, Malayalam, and Tamil using a range of machine learning methods. It emphasizes that even simple algorithms can produce satisfactory outcomes when provided with sufficient training data. The most successful approach is cross-lingual transfer learning through fine-tuning XLM-RoBERTa.

The paper (Kumar et al., 2022) highlights the pressing need to swiftly detect and remove hate speech and offensive content from social platforms due to their rapid spread and detrimental effects. Detecting hate speech poses a significant challenge, particularly in code-mixed languages like Hindi–English, Tamil–English, Malayalam–English, and Telugu–English. This comprehensive study investigates and contrasts various machine learning and deep learning techniques for addressing this issue.

Team CENTamil (Prasanth et al., 2022) employed TF-IDF with character-level word boundary analysis and the Random Kitchen Sink (RKS) algorithm to generate feature vectors for abusive comment detection in the Tamil language. Classification was performed using a Support Vector Machine (SVM) with a polynomial kernel. This method was applied to the Tamil dataset, achieving the first rank with a macro f1-score of 0.32 (Priyadharshini et al., 2022a). Subsequently, a comprehensive comparison was conducted by (Chakravarthi et al., 2023) to evaluate various machine learning algorithms with diverse feature extractors for detecting abusive comments in Tamil and code-mixed Tamil–English.

The study by (Rajalakshmi et al., 2023) focuses on enhancing Tamil text representation for hate speech detection. It explores various embedding techniques, including TF-IDF and pre-trained transformer models. Stemming algorithms are applied to address the linguistic complexity of Tamil. Experiments with classifiers like logistic regression, SVM, Stochastic Gradient Descent (SGD), decision trees, and ensemble models reveal that stemming improves performance, and MuRIL, combined with majority voting, achieved the best results in classifying offensive content, considering data imbalance.

3 Dataset

The Abusive Comment Detection Dataset by (Priyadharshini et al., 2022b) comprises comments written in the Tamil language, sourced from YouTube's comment section. This dataset includes comments and their respective labels categorized into nine distinct classes: *Misandry, Misogyny, Counter-Speech, Xenophobia, Hope-Speech, Transphobic, Homophobia, Not-Tamil, and None-of-the-above*. The class Counter-speech involves delivering feedback through fact-based arguments in a non-aggressive way, while hope speech encompasses optimistic narratives detailing how individuals cope with and overcome adversity. Misogynistic class is aimed at women or particular gendered groups, while homophobic content constitutes a form of gender-based harassment utilizing disparaging labels. Transphobic statements are phrases that are used to belittle vulnerable transgender individuals. Xenophobia is characterized by fear of or hate for what is thought to be alien or odd is known as xenophobia (Chakravarthi et al., 2023). Figure 1 illustrates the dataset split showcasing distinct classes and their percentages.

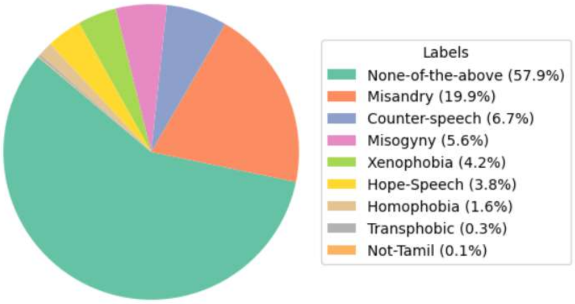

Figure 1: Data visualization

The dataset has been divided into three sets: Training, Validation, and Test, with 2240, 560, and 699 samples, respectively. Each entry in the training dataset consists of a Tamil text comment and its associated label.

4 Methodology

4.1 Data Pre-processing

Our work implemented and tested a series of data preprocessing steps to enhance the performance of our classification model. Initially, stop words and emojis were removed from the text. Subsequently, lexical replacement of English words with their equivalent Tamil counterparts within the text was performed. This step was crucial for ensuring the model could effectively process and classify text in Tamil. To address class imbalance in the augmented dataset, oversampling techniques were employed, involving duplicating instances of minority classes to create a more balanced training set. These pre-processing steps aimed to enhance the dataset's quality and representation for the subsequent classification task, ensuring the model's improved generalization and effective classification of Tamil text.

4.2 Model Classification

In the context of model classification, our paper used a range of multilingual transformer models, as addressed in the subsequent paragraphs, and these models underwent optimization with hyperparameter tuning.

MuRIL (Khanuja et al., 2021) is a BERT model pre-trained on 17 Indian languages, including transliterated counterparts is specifically designed for Indian languages and has been trained on extensive Indic text corpora.

XLM-RoBERTa (Conneau et al., 2020) is an extension of RoBERTa (Liu et al., 2019) designed to support multiple languages. It undergoes pre-training on a substantial dataset of 2.5TB from filtered CommonCrawl[1] data, encompassing content from 100 different languages.

M-BERT (Kenton and Toutanova, 2019), short for Multilingual BERT, is a pre-trained model that has been trained on a diverse corpus of text from the most prominent 104 languages, employing a masked language modeling (MLM) objective.

IndicBERT (Kakwani et al., 2020) is an ALBERT-based (Lan et al., 2019) multilingual model exclusively trained on 12 prominent Indian languages, including Tamil. Its pre-training process utilizes a novel monolingual corpus of around 9 billion tokens.

For the purpose of model selection on the training dataset, HuggingFace's (Wolf et al., 2020) SimpleTransformers framework was used to train transformer models for text classification. This allowed us to compare different models and determine the best-performing one before implementing data augmentation techniques.

Hyperparameter tuning is a crucial stage in model development, as these settings strongly influence the performance of a model. The hyperparameters were chosen by exploring various values that yielded superior results during model validation, focusing on optimizing the macro-averaged F1 score. The optimizer used is AdamW (Kingma and Ba, 2014), and Table 1 shows the other training parameters for model classification.

Training Parameters	Value
No: of epochs	3
Learning Rate	4e-5
Maximum Sequence Length	128
Batch Size	32

Table 1: Training parameters for model classification

4.3 Data Augmentation

We performed data augmentation utilizing a round-trip translation method with English and Hindi as intermediate languages to generate additional training data for Tamil, a low-resource language (Sugiyama and Yoshinaga, 2019). This process involves using Google Translate, a multilingual neural machine translation service, to first translate the original text from Tamil to English and Hindi and then re-translating it back into Tamil. This produced a diverse synthetic dataset for Tamil, which was used to train the MuRIL transformer model. Figure 2 shows the round-trip translation for the data augmentation process.

5 Results and Discussion

The results obtained by various transformer models trained on raw data is shown in Table 3. MuRIL attained the top macro F1 score of 0.44 compared to other transformer models, which is 12 units higher than the baseline model's performance. Table 2 displays the F1 scores for each class when comparing the baseline model to our best-performing model. The Transphobic class attains the lowest individual F1-score of zero due to its limited representation, consisting of only six data points in the test dataset. Subsequently, we explored data pre-processing and augmentation techniques with MuRIL to enhance its performance.

5.1 Effect of Preprocessing

The effects of different data pre-processing methods, including stop word removal, emoji removal, and lexical replacement, on MuRIL are presented

[1] https://github.com/ThilinaRajapakse/simpletransformers

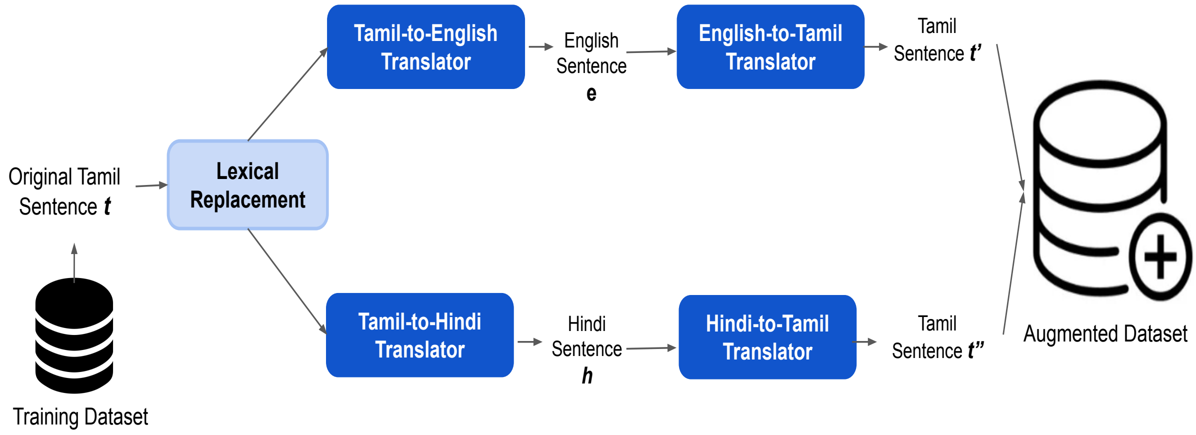
Original Tamil Sentence *t*
Training Dataset
Lexical Replacement
Tamil-to-English Translator
English Sentence **e**
English-to-Tamil Translator
Tamil Sentence *t'*
Tamil-to-Hindi Translator
Hindi Sentence *h*
Hindi-to-Tamil Translator
Tamil Sentence *t''*
Augmented Dataset

Figure 2: Round-trip Translation for Data Augmentation

Class	F1-score (Base)	F1-score (Ours)
Counter-speech	0.35	0.32
Homophobia	0.67	0.71
Hope-Speech	0.26	0.41
Misandry	0.71	0.64
Misogyny	0.54	0.50
None-of-the-above	0.83	0.76
Transphobic	0.00	0.00
Xenophobia	0.18	0.43

Table 2: Class-wise F1-score

in Table 4. It's important to highlight that eliminating stop words and emojis did not enhance the model's performance and excluded this step from the pre-processing pipeline. On the other hand, the lexical replacement of English words with Tamil words showed a slight improvement in the macro F1 score. As a result, our work incorporated this approach along with the round-trip translation technique.

5.2 Effect of Data Augmentation

The results of data augmentation with various combinations of intermediate languages are shown in Table 5. Our findings showed a significant increase in accuracy by using double round-trip translations with English and Hindi as intermediary languages. However, the macro-averaged F1 score reached its highest point at 0.47 using MuRIL when employing a single round-trip translation with English as the intermediary language, while the weighted-averaged F1 score remained consistent throughout the experiments. This indicates that data augmentation, particularly through round-trip translation, can be a valuable method to enhance the performance of NLP tasks in languages with limited linguistic resources, such as Tamil, with the choice of intermediary language being a critical factor in achieving optimal results.

6 Conclusion and Future Work

Our study presented promising insights into the effectiveness of data augmentation for abusive comment detection in a low-resource language like Tamil. Having observed an improvement in accuracy when employing two round-trip translations, the macro-averaged F1 score achieved its highest score with a single round-trip translation using English as an intermediate language. These results highlight data augmentation's potential to improve transformer-based models' performance in detecting abusive comments in low-resource languages, emphasizing the importance of selecting the right intermediate language for optimal results. The implementation details are shared in the Github repository. [2]

In the future, we would like to expand our investigation by incorporating a wider selection of intermediate languages in the round-trip translation process, allowing us to assess how these language choices influence the outcomes in diverse linguistic contexts on ensemble-based models.

[2] https://github.com/ICON-ML/ACDT

Model/ Metrics	Acc	Macro F1	Macro Pr	Macro Rec	Weight F1	Weight Pr	Weight Re
CEN-Tamil (Prasanth et al., 2022)	-	0.32	0.38	0.29	-	-	-
Indic BERT	0.66	0.35	0.43	0.31	0.62	0.62	0.66
M-BERT	0.67	0.43	0.45	0.43	0.67	0.67	0.67
XLM-RoBERTa-base	0.65	0.42	0.42	0.44	0.65	0.67	0.65
MuRIL	**0.67**	**0.44**	0.48	0.44	**0.67**	0.68	0.67

Table 3: Comparison of multi-lingual transformer models with existing baseline

Criteria	Accuracy	Macro F1	Macro Pr	Macro Rec	Weight F1	Weight Pr	Weight Rec
Without Pre-processing	0.67	0.44	0.48	0.44	0.67	0.68	0.67
Stop word Removal	0.65	0.41	0.43	0.4	0.65	0.64	0.65
Emoji Removal	0.66	0.43	0.47	0.41	0.65	0.65	0.66
Lexical Replacement	**0.67**	**0.45**	0.47	0.45	0.66	0.67	0.67

Table 4: Effect of Pre-processing Techniques

Criteria	Accuracy	Macro F1	Macro Pr	Macro Rec	Weight F1	Weight Pr	Weight Rec
Without Augmentation	0.67	0.45	0.47	0.45	0.66	0.67	0.67
Round-trip translation (English)	0.66	**0.47**	0.47	0.48	0.66	0.68	0.66
Round-trip translation (Hindi)	0.65	0.44	0.45	0.44	0.65	0.66	0.65
Round-trip translation (Hindi and English)	**0.69**	0.42	0.46	0.4	**0.67**	0.66	0.69

Table 5: Effect of Data Augmentation

References

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Shubanker Banerjee, Manoj Balaji Jagadeeshan, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Sean Benhur, and John Philip McCrae. 2023. Detecting abusive comments at a fine-grained level in a low-resource language. *Natural Language Processing Journal*, 3:100006.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard 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.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

Marwa Khairy, Tarek M Mahmoud, and Tarek Abd-El-Hafeez. 2021. Automatic detection of cyberbullying and abusive language in arabic content on social networks: a survey. *Procedia Computer Science*, 189:156–166.

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

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Roy Pradeep Kumar, Bhawal Snehaan, Subalalitha Chinnaudayar, et al. 2022. Hate speech and offensive language detection in dravidian languages using deep ensemble framework. In *Computer Speech & Language*, volume 75, page 101386. Elsevier.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

SN Prasanth, R Aswin Raj, P Adhithan, B Premjith, and Soman Kp. 2022. Cen-tamil@ dravidianlangtech-acl2022: Abusive comment detection in tamil using tf-idf and random kitchen sink algorithm. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 70–74.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022a. 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.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumar Kumaresan. 2022b. 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*.

Ratnavel Rajalakshmi, Srivarshan Selvaraj, Pavitra Vasudevan, et al. 2023. Hottest: Hate and offensive content identification in tamil using transformers and enhanced stemming. *Computer Speech & Language*, 78:101464.

Amane Sugiyama and Naoki Yoshinaga. 2019. Data augmentation using back-translation for context-aware neural machine translation. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 35–44, Hong Kong, China. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 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.

Stefan Ziehe, Franziska Pannach, and Aravind Krishnan. 2021. Gcdh@ lt-edi-eacl2021: Xlm-roberta for hope speech detection in english, malayalam, and tamil. In *proceedings of the first workshop on language Technology for Equality, diversity and inclusion*, pages 132–135.