International Journal of Engineering Research and Sustainable Technologies
Volume II, No.1, March 2024, 29 - 35
ISSN: 2584-1394 (Online version)

MITIGATING CLASS IMBALANCE IN OFFENSIVE LANGUAGE DETECTION IN MALAYALAM THROUGH NLPAUG

Munawwar K V[1] and Nandhini K[2]
[1]Research Scholar, Central University of Tamil Nadu, India
[2]Assistant Professor, Central University of Tamil Nadu, Thiruvarur, Tamil Nadu
* **Corresponding author email address**: munumunawwar8189@gmail.com , nandhinikumaresh@cutn.ac.in

DOI: https://doi.org/10.63458/ijerst.v2i1.73 | ARK: https://n2t.net/ark:/61909/IJERST.v2i1.73

Abstract

The rise of technology alongside the prevalence of social media and the promotion of free speech has resulted in an increased presence of vulnerable content in the public sphere. Currently, various researches demonstrate that the identification of offensive language plays a crucial role in preventing or protecting vulnerable groups. Our attention is directed towards the detection of offensive language in Malayalam, recognizing the scarcity of existing research in this area for the Malayalam language. mBERT demonstrates effectiveness across Indian languages. To address class imbalances within datasets, we employed NlpAug for word-level augmentation and achieved a significant improvement in macro F1 score of 0.31.

Keywords: *Offensive language, mBERT, NlpAug, stratified K-fold, Data Augmentation*

1. Introduction

Offensive language identification in the Malayalam language presents distinctive obstacles due to the intricate nature of the language and the assorted varieties of offensive material prevalent in online communication [1,18,19]. With the rapid expansion of social media platforms and online forums, the necessity to identify and alleviate offensive language in Malayalam text has become increasingly crucial to uphold a secure and respectful online environment. Malayalam, a Dravidian language predominantly spoken in the Indian state of Kerala and the Union Territory of Lakshadweep, is characterized by its extensive lexicon, intricate syntax, and diverse regional dialects. Nevertheless, the widespread presence of offensive content, which includes hate speech, cyberbullying, and profanity, in Malayalam text across various online platforms poses significant challenges for initiatives in content moderation and community management. The manual moderation of such content is often time-consuming, resource-intensive, and subject to subjectivity, thus rendering automated offensive language identification an indispensable solution.

In this particular context, the utilization of Natural Language Processing (NLP) techniques presents promising opportunities for the automation of offensive language detection in Malayalam text. By harnessing the power of machine learning algorithms, linguistic analysis, and semantic comprehension, NLP models can effectively discern and categorize instances of offensive language with a considerable degree of accuracy [12,14]. Moreover, the incorporation of data augmentation techniques such as NlpAug empowers researchers to amplify the training data, thereby reinforcing the robustness and efficacy of the model in detecting subtle variations of offensive language.

This research paper aims to examine the utilization of NlpAug to identify offensive language in Malayalam text. We will explore the effectiveness of NlpAug in augmenting training data for NLP models, its impact on the performance of these models, and its ability to improve the detection of offensive language nuances specific to the Malayalam language. By combining NLP techniques with data augmentation strategies, our goal is to develop reliable and efficient systems for detecting offensive language, specifically tailored for the Malayalam language. This research will contribute to the creation of safer and more inclusive online communities

2. Related Works

Several investigations have been conducted on the detection of offensive comments and the classification of hate speech in a variety of languages, such as English, Arabic, and Chinese. However, there has been a lack of specific research devoted to Malayalam, which poses unique challenges due to its native script. Nonetheless, valuable insights can be gleaned from relevant studies in the wider scope of multilingual offensive content detection and research on offensive comment detection in different languages.

Patankar et al. put an adaptable ensemble method to identify offensive comments in a language with limited resources, with a specific emphasis on the necessity of meticulous detection [13]. Gupta et al. introduced a collection of offensive comments in Indic languages, including Malayalam, and introduced a multilingual model named abusexlmr for detecting offensive comments, achieving the most advanced results [12]. The study on sentiment analysis of code-mixed text in Dravidian languages introduced techniques for sentiment analysis of code-mixed text, which could potentially be applied to Malayalam-English code-mixed data [15]. Furthermore, research papers such as "Understanding Emojis for Sentiment Analysis" [25] and "Pre-Processing and Emoji Classification of WhatsApp Chats for Sentiment Analysis" [23] dealt with the role of emojis in sentiment analysis, which is a relevant factor in analysing Malayalam text. While most of the related works in Malayalam that utilize transformer models focus solely on text written in pure native Malayalam script, some studies explore code-mixed text limited to the English/Latin script [1,2,3]. However, the dataset employed in this study includes not only text written in pure native Malayalam script but also code-mixed text written in English script and a combination of both native Malayalam script and code-mixed script. Promising outcomes for Malayalam offensive detection could be obtained by fine-tuning models like XLMRoBERTa and mBERT using this dataset [12,13].

Research in Natural Language Processing (NLP) is experiencing rapid growth in the exploration of data-balancing techniques for low-resource languages, with numerous notable contributions. Other work focused on addressing data sparsity issues in low-resource dependency parsing, investigating approaches such as cross-lingual transfer learning, semi-supervised learning, and synthetic data generation [4,5,6]. Das et al. a low-resource language, proposed effective techniques for dependency parsing under data scarcity, including cross-lingual transfer learning and semi-supervised learning. Iqbal et al. conducted research on data augmentation strategies for low-resource Named Entity Recognition (NER) tasks, with the aim of improving performance across different text domains through techniques such as paraphrasing and synonym replacement [8].

Alakrot et al. proposed cross-lingual parameter-sharing techniques in neural network parsers for low-resource settings in 2018, leveraging data from resource-rich languages to enhance parsing accuracy in underrepresented languages [9]. Zhou et al. addressed data scarcity in low-resource part-of-speech tagging tasks by utilizing transfer learning techniques, including fine-tuning pre-trained models and domain adaptation [10]. Collectively, these papers contribute valuable insights and methodologies for effectively managing data scarcity and imbalance in low-resource language settings, thus making significant contributions to advancements in NLP research and applications. As a research gap, this study put forward augmentation techniques such as replacing synonyms, inserting randomly, and swapping randomly to create augmented data to enhance the performance of text classification in situations where resources are limited. Finally, these studies collectively contribute valuable insights and methodologies to tackle the scarcity and imbalance of data in low-resource NLP scenarios, thereby presenting promising avenues for enhancing performance in such contexts [7,17,18,19,20].

3. Methodology

An explanation of the steps involved in the proposed approach is provided in Figure 1. Before the fine-tuning process of the models, we execute preprocessing steps on the Malayalam dataset. These steps encompass the conversion of the labels Not offensive (NF), Offensive-Targeted-Insult-Individual (OTII), Offensive-Targeted-Insult-Group (OTIG), and Offensive-Untargeted (OU) to 0, 1, 3, and 4 respectively for classification purposes.

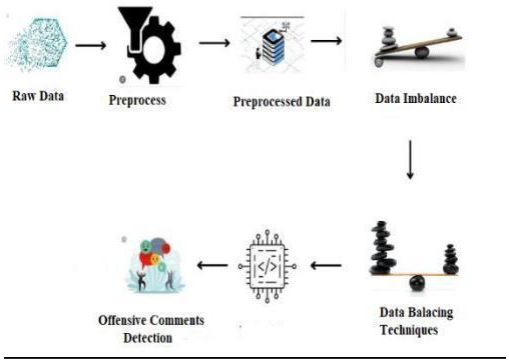

Fig. 1. Methodology

Additionally, the text is cleaned by eliminating irrelevant characters, converting comments in English and codemix script to Malayalam, and tokenizing the comments into subword units that are compatible with the language models. Following this, instances of the mBERT model are initialized. These models have been pre-trained on extensive monolingual and multilingual corpora, encompassing approximately 104 languages. They possess the capability to capture comprehensive linguistic representations and have undergone training for various natural language understanding tasks, thereby providing a solid foundation for fine-tuning.

The fine-tuning process involves a series of steps to adapt the models to the offensive comment detection task in the Malayalam language. The modification of the classification head of the mBERT models is performed to accommodate the multi-class classification task of identifying categories. This entails the addition of a SoftMax layer on top of the model's final hidden states, which generates probability scores for the classes. The subsequent step involves mapping the pre-processed tokenized comments to their corresponding embeddings, thereby serving as input to the models. This ensures that the models capture the semantic and syntactic information present in the Malayalam text.

During the training procedure, the dataset is divided into training, validation, and test sets. The optimization of the models is achieved through the utilization of a cross-entropy loss function, which compares the predicted probabilities with the ground truth labels. Hyper-parameter tuning is then conducted to determine the optimal configuration for the models. This involves adjusting learning rates, batch sizes, and the number of training epochs. The performance of the models is monitored using the validation set during training, enabling the selection of the best hyper-parameter settings. Once the fine-tuning process is completed, the performance of the mBERT models is evaluated on the test set. Various evaluation metrics, such as accuracy, weighted and macro scores (precision, recall, and F1 score), are employed to assess the effectiveness of the models in detecting offensive comments in the Malayalam dataset.The objective of fine-tuning the mBERT model on our Malayalam dataset is to leverage its pre-trained knowledge and adapt it to effectively detect offensive language while considering the intricacies of different writing scripts. This proposed approach enables us to harness the capabilities of these cutting-edge language models in addressing the task of Malayalam Offensive comment detection.

3.1 Dataset description

The objective of this study is to develop a system capable of accurately identifying offensive text in social media. To accomplish this, we utilize a standardized corpus provided by NLP-CUET@DravidianLangTech-EACL2021 [21,22]. Since social media texts often contain code-mixed content, traditional systems trained on monolingual data struggle to classify such data due to the complexities of code-switching at different linguistic levels. To address this issue, we transform the text into a pure form. Before system development, each instance in the dataset is converted into pure Malayalam text using both translation and transliteration techniques.

The main goal of this project is to create a system that can effectively detect offensive texts. To achieve this, we make use of the corpus provided by NLP-CUET@DravidianLangTech-EACL2021. The task assigned to us is a multi-class classification problem where the model is required to classify a given text into one of four

predetermined classes. These classes are Not offensive (NF), Offensive-Targeted-Insult-Individual (OTII), Offensive-Targeted-Insult-Group (OTIG), and Offensive-Untargeted (OU). The datasets used for training and testing are highly imbalanced, with a significantly larger number of instances in the NF class compared to the other classes. The train/test split is done in an 85/15 ratio, resulting in 12,514 instances for training and 2,209 instances for testing. The dataset description is given in Table 1.

During the pre-processing stage, the dataset undergoes a cleaning process. This involves removing unwanted characters, symbols, punctuation, emojis, and numbers from the texts, resulting in a cleaned dataset specifically tailored for the Malayalam language.

Table 1. Dataset Description.

Classes	No. of instants
Not offensive (NF)	12019
Offensive-Targeted-Insult-Individual (OTII)	120
Offensive-Targeted-Insult-Group (OTIG)	205
Offensive-Untargeted (OU)	170

3.2 Transformers

Transformers are innovative deep-learning models designed for natural language processing tasks. They use self-attention mechanisms to capture contextual relationships between words. This allows them to handle long-range dependencies effectively [17,21,25]. The self-attention mechanism weighs the importance of each word in the input sequence. Transformers can capture complex linguistic patterns and semantic relationships accurately. They consist of multiple layers of self-attention and feedforward neural networks. Each block processes the input sequence independently and passes the output to the next block. Transformers outperform previous models in various NLP tasks. They are highly parallelizable and scalable, making them suitable for training on large datasets. They can be fine-tuned for domain-specific tasks with small amounts of data. Transformer-based architectures like BERT, GPT, and T5 have advanced the state-of-the-art performance in NLP. Transformers are a cornerstone of modern NLP, offering exceptional performance across a range of tasks and applications. They capture rich contextual information and learn powerful representations from text data, propelling them to the forefront of AI research and development.

3.3 mBERT

Utilizing Multilingual BERT (mBERT) for Malayalam offensive language detection presents several compelling advantages. Firstly, mBERT's pre-training on a vast corpus covering various languages, including Malayalam, facilitates cross-lingual transfer learning [12,13]. This capability allows the model to grasp general language patterns and semantics, proving beneficial for offensive language detection in Malayalam, particularly when labeled data is scarce. Secondly, the adoption of mBERT avoids the resource-intensive process of building a language-specific BERT model from scratch, circumventing the need for extensive language-specific data or computational resources. Moreover, mBERT's shared representations across languages enhance its generalization to unseen data and language usage variations, thereby improving its effectiveness in detecting offensive content in Malayalam text. Additionally, fine-tuning mBERT on offensive language detection data in Malayalam further adapts the model to the target domain, enhancing its discriminatory abilities between offensive and non-offensive language. Lastly, by leveraging pre-existing multilingual resources, mBERT reduces the annotation effort and cost associated with collecting and labelling large-scale training data specific to Malayalam offensive language detection. This pragmatic approach allows for the development of robust and accurate offensive language detection systems for Malayalam, even in low-resource settings.

3.4 NlpAug

NlpAug, also known as Natural Language Processing Augmentation, is a Python library that has been meticulously designed to facilitate the process of data augmentation for text-based tasks in the field of Natural Language Processing (NLP). This library offers a diverse range of augmentation techniques that can be applied to textual data, thereby enhancing the performance, robustness, and generalization capabilities of models. At its core, NlpAug employs a broad set of strategies that encompass various operations, including synonym replacement, random insertion or deletion of words, character-level perturbations, and augmentation based on contextual word embeddings. One of the most notable features of NlpAug is its flexibility and extensibility, as it supports augmentation techniques that are specifically tailored for different NLP tasks, such as text classification, named entity recognition, machine translation, and sentiment analysis. Moreover, NlpAug allows users to customize augmentation parameters, such as the degree of augmentation, the probability of applying each augmentation operation, and the selection of augmentation techniques, thereby providing fine-grained control over the augmentation process.

NlpAug has been designed to be highly efficient and lightweight, which makes it suitable for seamless integration into existing NLP pipelines and workflows. Another noteworthy aspect of NlpAug is its ability to support text augmentation in multiple languages, thereby enabling the augmentation of text data in diverse linguistic contexts. This particular feature is particularly beneficial for addressing challenges related to data scarcity and improving the generalization capabilities of models across different linguistic domains.

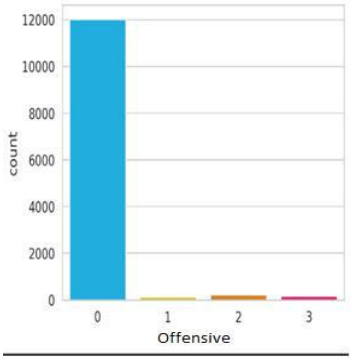

Figure 2.Before Augmentation

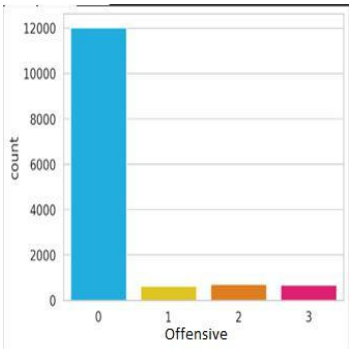

Figure 3. After Augmentation.

In our work, we augmented 20% of data through NlpAug to improve model performance. The dataset distribution in various classes before and after augmentation is given in Figures 2 and Figure 3.

4. Result and Discussion

In this section, the results of the experiments on Malayalam offensive comment detection using the mBERT, mBERT with Stratified K Fold and mBERT with NlpAug models are given in the tables. The model type used for mBERT is "bert-base-multilingual-cased" from the Hugging face transformers library in Python. The experimental setup for executing the code is Google Colab's Tesla T4 GPU. For hyper-parameter tuning the learning rate of 2e-5, AdamW optimizer, batch size of 16, and 4 training epochs are used. Both models use the same hyperparameter tunings for experiments

Table 2.Comparison of various Models

	W. P	W. R	W. F1	M.P	M.R	M.F1	ACC
Base	.97	.96	.98	.41	.42	.40	96.5%
Stratified K Fold	.98	.95	.97	.35	.32	.34	96.1%
NlpAug	.98	.98	.98	.66	.79	.71	98.1%

As specified in Table 2, we can see that mBERT+NlpAug has a higher score for Macro Precision, Recall and F1-score. The score was similar for the other two occasions. Also, Accuracy improved by ~2% in mBERT+NlpAug.

The experimental results highlight that the mBERT+NlpAug performs better than the other two methods in detecting offensive comments in the Malayalam language

5. Conclusion

In our study, we utilized mBERT for identifying offensive language in Malayalam, and further enhanced its performance through data augmentation using NlpAug. The results demonstrated a notable improvement in macro F1 scores across different approaches, with base, Stratified K Fold, and NlpAug achieving scores of 0.40, 0.34, and 0.71 respectively. This highlights the effectiveness of incorporating NlpAug for data augmentation, significantly enhancing the performance of mBERT in detecting offensive language in Malayalam text.

References

1. Akhtar, M. S., Sawant, P., Sen, S., Ekbal, A., & Bhattacharyya, P. , Solving data sparsity for aspect-based sentiment analysis using cross-linguality and multi-linguality. Association for Computational Linguistics.2018
2. Fadaee, M., Bisazza, A., & Monz, C., Data augmentation for low-resource neural machine translation. arXiv preprint arXiv:1705.00440.2017.
3. Feng, X., Feng, X., Qin, B., Feng, Z., & Liu, T. , Improving Low Resource Named Entity Recognition using Cross-lingual Knowledge Transfer. In IJCAI (Vol. 1, pp. 4071-4077).2018
4. Wei, J., & Zou, K, 'Eda: Easy data augmentation techniques for boosting performance on text classification tasks'. arXiv preprint arXiv:1901.11196.2019
5. Mo, Y., Yang, J., Liu, J., Wang, Q., Chen, R., Wang, J., & Li, Z, 'mCL-NER: Cross-Lingual Named Entity Recognition via Multi-view Contrastive Learning. arXiv preprint arXiv:2308.09073. 2023.
6. Ojha, A. K., & Zeman, D. , 'Universal Dependency treebanks for low-resource Indian languages: The case of Bhojpuri. In Proceedings of the WILDRE5–5th workshop on Indian language data: resources and evaluation (pp. 33-38).2020.
7. Das, M., Banerjee, S., Saha, P., & Mukherjee, A., 'Hate Speech and Offensive Language Detection in Bengali. https://doi.org/10.48550/arXiv.2210.03479, 2022.
8. Iqbal, M., Nisar, S., & Iqbal, W. ' Offensive Language Detection for Low Resource Language Using Deep Sequence Model. https://doi.org/10.1109/tcss.2023.3280952, 2023
9. Alakrot, A., Murray, L., & Nikolov, N. S., ' Towards Accurate Detection of Offensive Language in Online Communication in Arabic. 142. https://doi.org/10.1016/J.PROCS.2018.10.491, 2018.
10. Zhou, L., Cabello, L., Cao, Y., & Hershcovich, D. 'Cross-Cultural Transfer Learning for Chinese Offensive Language Detection. abs/2303.17927. https://doi.org/10.48550/arXiv.2303.17927, 2023.
11. Chakravarthi, B. R., Priyadharshini, R., Banerjee, S., Jagadeeshan, M. B., Kumaresan, P. K., Ponnusamy, R., ... & McCrae, J. P,'Detecting abusive comments at a fine-grained level in a low-resource language. Natural Language Processing Journal, 3, 100006., 2023
12. Gupta, V., Roychowdhury, S., Das, M., Banerjee, S., Saha, P., Mathew, B., & Mukherjee, A., 'Multilingual Abusive Comment Detection at Scale for Indic Languages. Advances in Neural Information Processing Systems, 35, 26176-26191., 2022.
13. Patankar, S., Gokhale, O., Litake, O., Mandke, A., & Kadam, D., 'Optimize_Prime@ DravidianLangTech-ACL2022: Abusive Comment Detection in Tamil. arXiv preprint arXiv:2204.09675. 2022
14. Bigoulaeva, I., Hangya, V., & Fraser, A. 'Cross-lingual transfer learning for hate speech detection. In Proceedings of the first workshop on language technology for equality, diversity and inclusion (pp. 15-25).2021.
15. Puranik, K, 'IIITT@ Dravidian-CodeMix-FIRE2021: Transliterate or translate? Sentiment analysis of code-mixed text in Dravidian languages. arXiv preprint arXiv:2111.07906. 2021
16. Sultan, A., Salim, M., Gaber, A., & Hosary, I. E, ' WESSA at SemEval-2020 Task 9: Code-mixed sentiment analysis using transformers. arXiv preprint arXiv:2009.09879. 2020
17. Bhowmick, A., & Jana, A., 'Sentiment Analysis for Bengali Using Transformer Based Models. In Proceedings of the 18th International Conference on Natural Language Processing (ICON) (pp. 481-486).2021

18. Priyadharshini, R., Chakravarthi, B. R., Malliga, S., Cn, S., Kogilavani, S. V., Premjith, B., ... & Kumaresan, P. K. 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 (pp. 80-87). (2023, September).
19. Priyadharshini, R., Chakravarthi, B. R., Cn, S., Durairaj, T., Subramanian, M., Shanmugavadivel, K., ... & Kumaresan, P. 'Overview of abusive comment detection in Tamil-ACL 2022. In Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages (pp. 292-298).2022
20. Shanmugavadivel, K., Hegde, S. U., & Kumaresan, P. K, 'Overview of Abusive Comment Detection in Tamil-ACL 2022. DravidianLangTech, 2022, 292. 2022
21. Sharif, O., Hossain, E., & Hoque, M. M.,' Nlp-cuet@ dravidianlangtech-eacl2021: Offensive language detection from multilingual code-mixed text using transformers. arXiv preprint arXiv:2103.00455. 2021
22. Hossain, E., Sharif, O., & Hoque, M. M., ' NLP-CUET@ LT-EDI-EACL2021: multilingual code-mixed hope speech detection using cross-lingual representation learner. arXiv preprint arXiv:2103.00464.21. 2021
23. Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).
24. Yoo, B., & Rayz, J. T., 'Understanding emojis for sentiment analysis. In The International FLAIRS Conference Proceedings (Vol. 34).2021.
25. Effect of Emojis in Classifying Telugu Code Mixed Movie Reviews, 3rd International Conference on Mathematical Modeling & Computational Science ICMMCS'23.
26. Mohta, Astha, Atishay Jain, Aditi Saluja, and Sonika Dahiya. "Pre-Processing and Emoji Classification of WhatsApp Chats for Sentiment Analysis." In 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), pp. 514-519. IEEE, 2020.
27. Offensive Language Detection from Multilingual Code-Mixed Text using Transformers by Omar Sharif, Eftekhar Hossain, and Mohammed Moshiul Hoque, 2021