Gradient-Guided Data Augmentation with mBERT and MuRIL for Malayalam Offensive Language Detection

Munawwar K V[1], Nandhini K[2]
Research Scholar, Central University of Tamil Nadu, Thiruvarur, Tamil Nadu – 610005, India[1]
Assistant Professor, Central University of Tamil Nadu, Thiruvarur, Tamil Nadu – 610005, India[2]

Abstract—**The widespread adoption of social media platforms has facilitated increased usage of offensive content, particularly in native languages where users express themselves more freely. Automated offensive language detection in low-resource languages such as Malayalam faces significant challenges due to severe class imbalance, where non-offensive samples substantially outnumber offensive instances, resulting in biased model performance and diminished detection accuracy for underrepresented classes. This study addresses the critical challenge of class imbalance in Malayalam offensive language identification through a comprehensive data augmentation framework. We propose a novel gradient-guided augmentation technique specifically designed to mitigate minority class imbalance by selectively enhancing underrepresented class samples through the identification and synthesis of challenging instances that improve model robustness. The effectiveness of various augmentation strategies is systematically evaluated, including back-translation, paraphrasing, and NLPAUG techniques, integrated with mBERT and MuRIL models. Our gradient-guided augmentation approach demonstrates substantial performance improvements, achieving a notable 0.09 increase in recall score compared to the baseline model's 0.74 recall, while preserving overall model performance on imbalanced Malayalam offensive language datasets. The proposed methodology offers a promising solution for addressing class imbalance challenges in offensive content detection for low-resource languages. The results highlight that integrating augmentation with explainability not only improves classification performance but also helps in overcoming certain limitations associated with the previous methods, while also contributing more to the study.**

Keywords—Offensive comment detection; gradient guided augmentation; NLPAUG; back translation; paraphrasing with MultiIndicParaphraseGeneration

I. Introduction

Exponential growth in online information media of messages and social networks has altered in an unprecedented manner the way people interact or communicate with each other; nonetheless, the technology provides unprecedented possibilities of self-expression, information communication, and communication with people. Nonetheless, this technology has resulted in an unprecedented growth in the propagation of information in the form of hate speech, cyberbullying, offensive messages and texts, and information itself. This information can have significant social and psychological ramifications in intimidating people and even inciting society to hostility against itself. In the realm of artificial intelligence (AI) and natural processing languages (NLP), the detection of offending language automatically has become an essential duty; the objectives of the offence detection tool are said to include the detection and flagging of information of harm in advance of the actual indulgence in the harm itself [1,2].

Although significant progress has been observed in languages like English, where large annotated data and benchmark corpora are available [3], languages like Malayalam remain understudied. Malayalam belongs to the Dravidian family of languages and has about 35 million speakers. It is spoken mainly in the state of Kerala. The complexity of the Malayalam language lies in its inflexibility and the complexity of its morphological patterns and code mixing from other languages like English and other regional languages spoken in India. This linguistic complexity complicates both tokenization and semantic representation, making offensive content detection particularly challenging [1, 4]. Moreover, social media text is often informal, containing colloquial expressions, transliterated words, and dialectal variations, which further exacerbate the difficulty of automated understanding.

Initial approaches of offence detection in Malayalam relied mainly on traditional machine learning classifiers such as Support Vector Machines, Random Forests, and logistic regression. These broadly used manually constructed features, such as bag-of-words, TF-IDF, and surface-level embeddings [2, 5]. Though the initial approaches established the foundation of Malayalam NLP, they did not perform well in dealing with rich morphology, subtle semantics, and code-mixed expressions that are typical in user-generated materials. Class imbalance is an underlying and large problem in the field, with the vast majority of offending content only being in the minority of the overall material. Traditional approaches of solving this, such as resampling, cost-sensitive learning, and feature-level augmentation, yield incremental gain that often is overfitting in some sets [3,6,27].

Recent breakthroughs in deep learning, specifically in transformer-based architectures like multilingual BERT (mBERT) and Google's MuRIL, have exhibited extraordinary prospects in low-resource languages through the use of cross-lingual and contextual embeddings [4,7,26]. Through such models, contextual relations and semantic subtleties can be effectively grasped even in morphologically complex languages like Malayalam. However, they continue to be limited in their effectiveness due to the availability of sparse annotated data and their inclination toward being biased against majority classes. Therefore, the necessity of data augmentation strategies that will be able to augment minority classes and compensate for class imbalance continues to be critical. Methods like contextual word embedding augmentation (NLPaug), back translation, paraphrasing, and gradient-guided token-level augmentation have promised the increase of model resilience, the expansion of recall, as well as the decrease in bias against non-offensive classes [5,6,7].

Based on these observations, this work introduces an end-to-end Malayalam offensive language detection framework that combines various augmentation techniques with transformer-based architectures. NLPaug is used to create semantically similar variations of minority-class instances, back-translation injects paraphrased sentences in the form of intermediate languages, and transformer-based paraphrasing creates further expansions of low-frequency classes. IG-guided augmentation allows selecting the tokens most responsible for the model's prediction and targets them for focused augmentation, such that the generated instances are informative and add meaningfully to minority-class learning [7].

The constructed framework integrates NLPaug, Back translation, paraphrasing, and IG-guided augmentation in tandem with transformer-based models like mBERT and MuRIL, yielding a strong solution against the issues of class imbalance and morphological richness. This research aims to develop a robust, scalable, and interpretable framework that addresses challenges like Morphological richness, complex inflexion patterns, Informal expressions, dialectal variations in social media text and Class imbalance by applying data augmentation methods and explainable AI techniques with transformer models.

A. Research Questions

1) How can class imbalance in pure Malayalam offensive language datasets be effectively mitigated to improve minority-class detection?

2) What is the impact of applying data augmentation techniques on transformer-based model performance?

3) How can explainable AI methods guide augmentation to enhance interpretability while preserving semantic coherence in augmented examples?

4) How effective are these approaches in improving robustness and performance on morphologically rich pure Malayalam text?

B. Contributions

This work makes the following key contributions:

1) Uses explainable AI-guided augmentation to generate contextually meaningful examples while preserving semantic integrity.

2) Provides a comprehensive evaluation of the effectiveness of augmentation in reducing class imbalance and improving minority-class detection.

3) Display improved performance and strength of transformer-based models on morphologically complex low-resource languages like Malayalam.

II. Related Work

The detection of offensive language in Malayalam texts has progressed from traditional machine learning methods to more powerful transformer-based models, following the broader trend in natural language processing research. Initial works in this area used classifiers such as Support Vector Machines, Random Forest, and Logistic Regression, using features such as bag-of-words and Term Frequency-Inverse Document Frequency [1,2]. Although these methods are computationally efficient and provide good intuitions for understanding, they are fundamentally limited in their inability to understand context and linguistic nuances. This is particularly relevant for Malayalam, which is a morphologically complex language with intricate inflectional and code-mixing features. Shallow representations and poor generalization are common issues in traditional machine learning-based methods.

Another important limitation of earlier works is the fact that the majority of the works rely on the idea that conventional preprocessing and feature engineering techniques are adequate for dealing with informal and code-mixed texts. In the case of Malayalam social media texts, there will be transliterations, non-standard spellings, and usage of localized words, which will negatively impact the performance of conventional models. Moreover, the majority of the earlier works do not adequately address the class imbalance problem, considering the fact that there will be fewer offensive texts. Although oversampling, undersampling, and cost-sensitive learning have been used in earlier works [3, 6, 8], it is found that these methods do not significantly contribute to the solution and may result in overfitting and distortion of the original data.

The introduction of transformer-based models, including multilingual BERT and MuRIL, is seen as a significant advancement in this field [4,7]. These models incorporate contextual embeddings and cross-lingual transfer learning, which help in better capturing semantic nuances, morphological variations, and code-mixed texts. These models are also seen to perform better in practice compared to traditional models using multiple metrics. However, these models are also limited by two main conditions: first, that pre-training on multilingual corpora captures linguistic phenomena well for the Malayalam language, and second, that fine-tuning on limited annotated data is sufficient. However, in practice, this is not always true and leads to suboptimal performance, especially for minority classes such as offensive texts.

To address this, recent research in this area includes using data augmentation [13] techniques such as back-translation, synonym replacement, contextual word embeddings such as NLPaug, and transformer-based paraphrasing [4,5,6]. These techniques artificially augment the training data, especially for minority classes, and are seen to perform better in recall and F1-scores for tasks such as hate speech detection and sentiment analysis [9,10].

However, most data augmentation techniques are seen to work in a heuristic manner without considering which part of the input is more relevant for better predictions.

A more recent and promising approach has been the integration of techniques based on Explainable AI (XAI). Techniques such as Integrated Gradients (IG) have the ability to identify the specific words that have the greatest influence on the output of the model [3,6,7]. Compared to other techniques such as SHAP and LIME, the IG approach is computationally efficient, especially when working with the transformer model, as well as longer texts [3,12]. However, there has been limited work conducted on the integration of these techniques, as well as the specific language of interest, namely the Malayalam language.

Building upon these gaps, this study seeks to place itself at the crossroads of transformer-based modeling, data augmentation, and explainability in AI. Unlike other studies that focus on each of these components in isolation, this study proposes an IG-guided data augmentation mechanism for Malayalam offensive language detection. Specifically, this study seeks to leverage attribution scores for more semantically meaningful data augmentation, thus handling class imbalance more adequately. Furthermore, this study also seeks to leverage transformer-based models such as mBERT and MuRIL to enhance both effectiveness and explainability in handling code-mixing, morphological complexity, and limited annotated data.

III. Methodology

The Architecture shown in Fig. 1 demonstrates how to use transformer models and data augmentation to create an offensive language detection system for Malayalam. To label the data, it first scrapes comments from online sources and then annotates them. After that, the unbalanced dataset is either used straight away or put through several data augmentation techniques, including back translation, gradient-guided methods, paraphrasing, and NLPaug. These techniques produce an enhanced dataset with the goal of enhancing diversity and class balance. MBERT and MuRIL models are trained on both the original and augmented datasets. The last phase entails assessing these models' performance to ascertain how well each augmentation technique improves classification robustness and accuracy.

A. Scrapping and Annotation

The data for the research was collected through web scraping on YouTube comments, with the help of the official YouTube Data API. The data is a collection of various kinds of Malabar videos with different linguistic varieties, involving issues like the political scene, entertainment, news, culture, etc. The comments on these were scrapped for text classification, ensuring that only high-quality data is cleaned, with no spam, links, emojis, redundant comments, etc.

All comments were manually labeled into one of two classes: offensive (label 0) or non-offensive (label 1). Offensive comments contained abusive, hateful, or derogatory language, whereas non-offensive comments were neutral or constructive. The final dataset contained a total of 30,231 comments, of which 22,678 (75%) were classified as non-offensive and 7,553 (25%) classified as offensive, as shown in Fig. 2. This heavily skewed distribution emphasizes an extreme class imbalance, which is problematic for training robust classifiers.

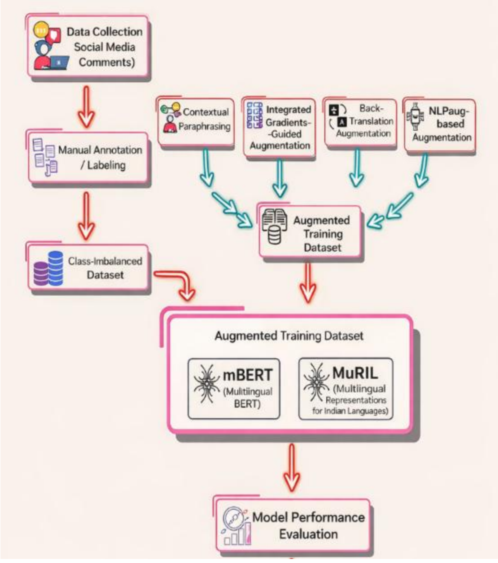
Data Collection Social Media Comments)
Manual Annotation / Labeling
Class-Imbalanced Dataset
Contextual Paraphrasing
Integrated Gradients--Guided Augmentation
Back-Translation Augmentation
NLPaug-based Augmentation
Augmented Training Dataset
Augmented Training Dataset
mBERT (Mulitiingual BERT)
MuRIL (Multilingual Representations for Indian Languages)
Model Performance Evaluation

Fig. 1. Proposed architecture for data augmentation using targeted augmentation.

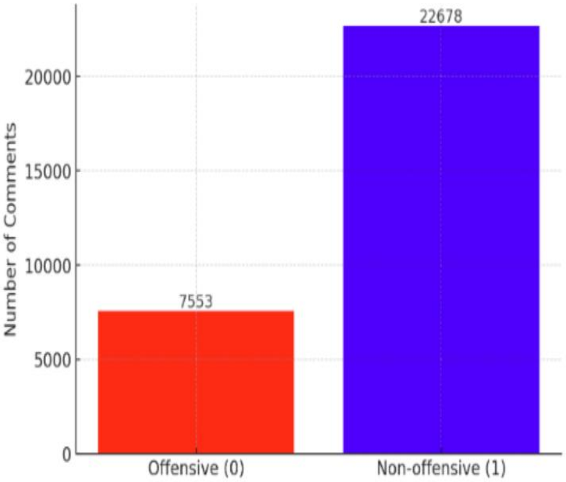

Fig. 2. Dataset statistics.

In order to achieve annotation reliability, multiple annotators were involved in the process of marking under established annotation instructions. The agreement of the annotations was measured via Cohen's Kappa coefficient (κ), which accounts for chance agreement.

$$\kappa = \frac{\mathrm{Po}-\mathrm{Pe}}{1-\mathrm{Pe}} \tag{1}$$

Where Po is the observed agreement that is being considered, and Pe is the expected agreement from chance. The observed agreement was 0.87, and the expected agreement was 0.26. After substituting the values in Eq. 1,

$$\kappa = \frac{.87-0.26}{1-0.260} = 0.82 \tag{2}$$

A kappa value of 0.82 [Eq. (2)], demonstrates strong agreement [11], which confirms that the annotation process was uniform and trustworthy. The high level of inter-annotator agreement and the large-scale dataset create a strong foundation that is needed to train and test Malayalam-based offensive language detection models.

B. mBERT (Multilingual BERT)

The Multilingual BERT model embodies an advanced transformer concept that revealed advancements in detecting offensive text in the Telugu script by Reddy et al. and Devlin et al. [14,15]. It is pre-trained on information gathered from a broad range of sources, such as Wikipedia articles in 104 languages worldwide, such as Malayalam script. mBERT reveals the immense possibilities of transferring knowledge to languages with limited resources.

The Multilingual BERT applies bidirectional encoding that efficiently interprets the Malayalam script by detecting contextual clues from both sides of the text, where this is necessary due to the complex permutations of the Malayalam script with respect to word order. In the case of raw Malayalam text, mBERT utilizes WordPiece tokenization to adapt to the complexities attached to the use and understanding of the Malayalam alphabet. It adequately manages special characteristics attached to the use and understanding of the Malayalam alphabet, such as the need for conjunct consonants and vowels in word formation. The bidirectional attention mechanism is especially helpful in Malayalam because the word order is variable and context is critical in disambiguating meanings, particularly in determining subtle offensive content that is conditioned on cultural and contextual information. The training strategy adopted in mBERT is masked language modeling in which random tokens are blanked out to induce the model to generate the blanked-out tokens from the surrounding context. The strategy helps improve the model's capability in detecting Malayalam offensive language because the model can better recognize the tacit relationship among words and word sequences that may be essentially conveying the offensive language. The attention mechanisms that capture the long-range dependencies are essential in deciphering Malayalam complex sentences in which the offensive content is widely spread across the sentence structure.

C. MuRIL (Multilingual Representations for Indian Languages)

Developed at Google Research India, MuRIL is a notable breakthrough in transformer-based models specifically for Indian languages, especially Malayalam. The model was pre-trained on 17 Indian languages and their transliterations, overcoming numerous Malayalam text analysis constraints of mBERT [16,17]. The architecture involves the same basic design as that of BERT, but with a well-curated lexicon and an Indian language dataset that allows the model to better recognize offending language use in Malayalam. The model comprises 12 transformer layers with 768 hidden dimensions and 12 attention heads, like BERT-base, but with a vocabulary of 200,000 subword tokens, much better at covering Malayalam compared to mBERT. The large vocabulary reduces the occurrences of intense subword tokenization and maintains semantic information during processing. Pre-training was done with mixed masked language modeling and next sentence prediction on a multilingual corpus, including Wikipedia, CommonCrawl, and Malayalam news articles.

The training data of MuRIL covers formal linguistic forms from news and encyclopedic texts, along with informal forms from web pages, familiarizing the model with diverse linguistic patterns typical of social media sites where offending language is found. The pre-training also involved the use of transliterated Malayalam text, critical when familiarizing the model with informal orthographic forms and phonetic variations in online text. The exposure is critical when detecting offending language since abusive text may utilize non-standard spellings and informal forms in an effort to escape detection. The importance of the attention mechanism in MuRIL was particularly increased with an objective to address Indian language nuances and especially their capacity to manage intricate morphological structures, such as those of the Malayalam language. Indeed, it proves to be an incredibly effective tool in addressing the intricate structure of Malayalam, where morphemes flow smoothly into each other to generate differentiated words of understated nuances.

D. NLPAug

NLPAug is another augmentation library that is complete and has been found to possess several unique methodology options that were found to be applicable and relevant when handling the Malayalam language [18,19]. The library allows one to augment characters, words, and sentences, and it has been found that the unique characteristics of the Malayalam language can be adequately addressed through the module's development and design, which is significant when handling an offence-detecting task.

Character-level augmentation in NLPAug involves complex operations under Character-level augmentation [20], in which consideration of a complex Malayalam script system consisting of 55 basic script characters, a high number of conjunct consonants, many vowel signs, combining characters, etc., must be augmented to maintain a valid script and phonetic consistency between them. Thus, in NLPAug for Malayalam, a deep knowledge of Malayalam orthographical rules has to be achieved during a function implementing a valid insertion of Malayalam characters in strategic places throughout a text to maintain a valid Malayalam script without compromising readability of text in Malayalam, as well as maintaining a non-offensive nature of text content after deleting characters in any sequence in a text without compromising readability of a word comprised of a group of characters being deleted in a text. Such a process in NLPAug avoids a text from becoming unreadable due to invalid schemes of disuniting essential pairs of characters in text written in Malayalam.

Character Substitution behaves similarly to the normal evolution of the written word and the differences in dialects of the same spoken word; characters are substituted according to original characters in the word and similar characters in the Malayalam alphabet. Using the Malayalam alphabet's pronunciation mapping as input. Visual Similarity Substitution replaces characters in the word with similar-looking characters in the original word. This behaves similarly to errors made during typing and the various forms of obfuscation done on the original material in order to avoid detection mechanisms. Character swapping operations swap nearby characters to mimic typing errors and natural variation patterns, with Malayalam script constraints to preserve phonetic viability and script appropriateness, with sophisticated swapping techniques including character dependencies wherein some characters have the requirement that they will be required in certain sequences, invalidating invalid character sets that would degrade word structure or semantic meaning.

Word-level augmentation [9] via NLPAug offers high-level operations optimized specifically for Malayalam's agglutinative morphological structure and rich vocabulary variation. The synonym replacement employs Malayalam wordnets, semantic similarity models, and contextual embeddings to find the relevant synonyms that will sustain the original text's offensive or non-offensive characteristics, with the selection of synonyms undergoing semantic filtering so that the replacement words will sustain emotional intensity, cultural nuance, and the offending purpose when present. The system is required to keep large Malayalam synonym databases organized in terms of semantic fields, emotional valence, and cultural context, with the synonym replacement exercising care so that the noxious nature of offending terms is maintained while offering linguistic variation critical to the training of strong models, including cultural sensibility filtering so that the synonyms will sustain the correct regional and social context, the latter critical in the case of the offending expressions that draw meaning from reference culture in use or community-related language patterns. Deletion operations eliminate random words with controlled likelihood while keeping sentences semantically intact and syntactically coherent, using Malayalam syntax awareness through controlled elimination of relatively insignificant words like particles, auxiliaries, and conjunctions while leaving core semantic information intact. The advanced deletion methodology utilizes dependency parser techniques in order to delete words that result in increased accuracy and efficiency. The deletion methodology does not decrease the efficiency of sentence classification and comprehension. The random insertion of words is carried out to maintain sentence length and meaning. The random insertion of words is done using semantically accurate Malayalam words. It is carried out based on an adequate vocabulary list specific to particular sentence semantics, emotional content, and cultural reference. Word replacement in a specific context utilizes pre-trained language models of the Malayalam language. It utilizes transformer models, showing increased accuracy and efficiency in replacing words with increased accuracy and pertinent consideration of the context. The models show higher efficiency and accuracy, wherein they introduce variation in profanity without compromising on their meaning. The mathematical representation is given as:

$$nlpaug(xi) = x_i^{\sim NA} \tag{3}$$

where, $x_i^{\sim NA}$ is a syntactically or lexically perturbed version of xi using substitution, insertion, or deletion operations in Eq. (3).

E. Back Translation

The efficient augmentation scheme specially calibrated for Malayalam's unique linguistic properties and cultural context is back translation [10,21,23]. For the purpose of generating paraphrastic versions while preserving the semantic content necessary for identification of offensive language, the process consists of translating the Malayalam text into one or several zeros, followed by back-translation into Malayalam. Through the use of a translation system highly variant in terms of language, the process alleviates the scarcity of parallel corpus data for Malayalam.

The Malayalam back-translation pipeline assimilates advanced neural machine translation architectures specifically trained on Malayalam–English/Malayalam–Hindi parallel texts with the main translation direction using Malayalam-to-English translation, then English-to-Malayalam back-translation using the relatively plentiful Malayalam–English parallel data available from numerous sources, including official documents, news publications, and textbooks. The translation models utilize transformer-based architecture specifically fine-tuned on Malayalam-to-Malayalam and Malayalam-to-English language pairs that utilize Malayalam-specialized tokenization techniques that account for complex script characteristics such as conjunct consonants, vowel signs, and morphological variation, with the morphological information crucial for proper translation being maintained through the tokenization while subword efficiency needs of neural translation models are also met. Quality control measures validate the translation effectively with confidence scoring, attention analysis, and semantic similarity checking, with the system tracking translation quality using automated measures such as BLEU scores, measures of semantic similarity, and linguistic validity checks specific to Malayalam being used, while poor-quality translations are automatically filtered out to avoid adding corrupted training data that may negatively affect model quality.

Malayalam offensive language detection using back translation involves detailed preservation of cultural context as well as community-based terms that yield offensive meaning, with the translation being carried out using cultural awareness-based mechanisms that identify Malayalam-based cultural terms, religious terms, and social terms that wouldn't translate directly to the intermediate language, dealt with through special translation dictionaries as well as cultural context preservation methodologies. The system is equipped with Malayalam cultural lexicons that keep track of culture-based terms that demand special care when being translated, with religion-based terms, caste terms, local cultural norms, and traditional terms being given special translation treatment so that the back-translated material is culturally correct, important in offense detection because many harmful terms get their meaning from the cultural context they're used in that generic translation-based systems may fail accurately. The mathematical representation is given as:

$$A_{bt}(xi) = T^{-1}(T(xi; Ls \rightarrow Lt); Lt \rightarrow Ls) = x_i^{\sim BT} \tag{4}$$

where, *T(.)* is a translation function from source language *Ls* pivot language *Lt*, and back in Eq. (4).

F. Paraphrasing With MultiIndicParaphraseGeneration

MultiIndicParaphraseGeneration is a sophisticated paraphrasing tool designed especially for Malayalam and other Indian languages [22,23,25]. The project maintains the linguistic semantics, cultural semantics, as well as the classification labels that play an important role in the detection of offensive language while dealing with the challenges that come with the generation of the data for the process of paraphrasing through morphological density. This process makes use of highly advanced transformer networks that undergo accurate customization during the process of creating Malayalam data.

As with the inclusion of the layers of cross-attention between the representation of the encoder and the decoder, keeping the characteristics of the agglutinative nature of the Malayalam language due to the presence of individual words with multiple units of semantics, the ability to attend to the more influential semantics of the word while being liberal towards the other units of language with minimal semantics or significance, the attention within the linguistic architecture is custom-made to conform to the characteristics of the Malayalam language. Because the variations in morphological type often provide natural paraphrastic forms while themselves being semantically similar, the architecture embodies positional encoding schemes that are tailored to the morphological attributes and script properties of Malayalam. Typical positional codes are enhanced with morphological insight that is aligned with word internal structure and morpheme edges.

MultiIndicParaphraseGeneration utilizes state-of-the-art controlled generation techniques that provide fine-grained control over paraphrase characteristics of interest in offense detection in language, providing regulation over linguistic register, allowing generation of paraphrases that are suitably formal or informal depending on the type of material, with the application of register control being particularly useful in offense language detection because harmful material is realized in various linguistic forms from the formal style of hate speech through informal abusive speech. Semantic intensity management allows the generation of paraphrases that maintain, increase, or decrease the emotional intensity of original material while communicating primary semantic meaning, allowing the generation of paraphrases that contain levels of offensiveness different from the original material while communicating correct classification labels, with the intensity management system utilizing Malayalam-limited emotional lexicons as well as cultural understanding to appropriately modulate the intensity. The mathematical equation is given by:

$$A_{pp}(xi) = x_i^{\sim PP} \tag{5}$$

where, paraphrase generation maps $xi \mapsto x_i^{\sim PP}$ while preserving semantics in Eq. (5).

G. Gradient-Guided Augmentation with Integrated Gradients

Gradient-driven augmentation through Integrated Gradients (IG) is one of the sophisticated, smart augmentation techniques that detects the highest-impact training samples regarding their role in learning from the model and uses such high-impact training samples selectively for focused augmentation [3,24]. Such an augmentation strategy changes the conventional random augmentation in that the computational efforts are applied selectively on the highest-potential training examples so that augmentation efforts get maximized in Malayalam offensive language detection tasks.

Integrated Gradients is an attribution algorithm (Algorithm 1) that calculates the role of each input feature in model outputs through integration along the path from a baseline input to the true input, with this involving high-level management of Malayalam text tokenization, embedding, and computing gradients that take account of the language's distinctive features including rich morphology, script characteristics, and cultural context dependencies. Baseline selection in computing IG in Malayalam is done with care in selecting appropriate neutral reference points that adhere to Malayalam linguistic structure but offer informative baselines for attribution, with typical baseline techniques including using neutral Malayalam text instances, zero embeddings, or average token embeddings from large Malayalam corpora, whereas advanced baseline selection involves Malayalam-specific whereby morphological neutrality, cultural context elimination, and syntactic structure preservation are considered.

Malayalam IG computation employs high-level interpolation techniques in such a way that they are linguistically valid according to Malayalam linguistics along the path of union, in which the points along the path of union are linguistically valid Malayalam text or embedding representations, in such a way that vigilance in avoiding such attribution artifact formation resulting from linguistically invalid intermediate points in the path of union has critical importance, especially in languages such as Malayalam that are linguistically complex, with high-level tokenization techniques that use morphological analysis in computing semantic coherence in the process of attribution computation, as well as gradient computation of attribution that employs linguistically valid tokenization in Malayalam that retains morphological boundaries along with semantic coherence crucial to accurate attribution, in which tokenization strategy in subwords has critical importance in balancing efficiency and semantic coherence to ensure that the attribution score formation computes semantic units as opposed to unnecessary fragments in Malayalam linguistics.

These samples will be identified by conducting a statistical analysis to judge their attribution scores. The process for selecting the samples for gradient-guided augmentation will include full scoring mechanisms for the samples present in the training data set based on their IG values and their patterns of distribution. Additionally, during the selection process for sentence samples for focused augmentation, several aspects will be included to validate a sentence's scoring mechanisms, such as sentence lengths and their complexity for Malayalam samples and their cultural context. As a result, statistical significance testing will ensure the selection of the most influential samples through the use of several statistical mechanisms, ensuring a better and more efficient selection process for samples through the identification and elimination of computation artifacts. The focused augmentation process will include several mechanisms for data augmentation for samples identified through their very high values for IG. This allows for the selection of augmentation techniques that preserve the significant characteristics while introducing useful variation. In order to create a variety of augmented variants that explore different facets of linguistic variation while fully covering potential variations and preserving the influential characteristics identified through IG analysis, multi-technique augmentation applies collections of back translation, paraphrase, and NLPAug operations on single high-influence samples. With judicious coordination of augmentation procedure across techniques to avoid conflicting alterations that would endanger sample quality or patterns of attribution. The mathematical representation is:

Let $IG\ (xi, yi)\ \in R^d$ denote the attribution scores from Integrated Gradients for each token in xi . Then targeted augmentation modifies only the high-saliency tokens:

$$A_{ig}(xi) = f\left(xi, Top - k\left(IG(xi, yi)\right)\right) = x_i^{\sim IG} \qquad (6)$$

where, $Top - k(\cdot)$ selects the k most influential tokens, and $f(\cdot)$ applies augmentation (substitution, backtranslation, or rephrasing) only on them in Eq. (6).

Algorithm 1: IG-Guided Data Augmentation + Transformer Fine-Tuning

```
augDataset = [];
augLabels  = [];
Read dataset line by line;
Load pretrained transformer model (MuRIL/MBERT);

for line in originDataset do
    # get sentence and its label
    Sentence = line('comment');
    Label    = line('label');
    # preprocess the sentence
    Preprocess(Sentence);
    # compute Integrated Gradients (IG) attribution
    attributions = IntegratedGradients(Model, Sentence);
    # select high-attribution tokens
    importantTokens = SelectTopTokens(attributions);
    # apply augmentation guided by IG
    augmentedSentence = AugmentWithFocus(Sentence,
importantTokens);
    Preprocess(augmentedSentence);
    # expand dataset with IG-guided augmentation
    augDataset.append(augmentedSentence);
    augLabels.append(Label);
end

# combine original and augmented data
finalDataset = originDataset + augDataset;
finalLabels  = originLabels  + augLabels;

# fine-tune MuRIL/MBERT on augmented dataset
Tokenize(finalDataset);
Train MuRIL/MBERT with finalDataset and finalLabels;
Evaluate classification performance;
```

IV. Result and Discussion

A. Experimental Result

In this section we discuss about the performance of the baseline MuRIL and MBERT model and its augmented variants was evaluated using standard metrics including accuracy (ACC), precision (P_0), recall (R_0), and F1-score (F_0) for the offensive class, along with macro-averaged precision (MP), recall (MR), and F1-score (MF) to assess overall classification performance. Data augmentation techniques such as NLPAUG Augmentation (NA), back translation (BT), and paraphrasing (PP) were also employed to improve the model in identifying offensive content. Moreover, the techniques of integrated gradients (IG) were used to assist the data augmentation by highlighting the most crucial words that discriminate between the classes.

TABLE. I. Result of MURIL Base Model and its Augmented Variations

MODEL	ACC	P_0	R_0	F_0	MP	MR	MF
MURIL	87.03	70	74	72	81	82	82
MURIL_NA	91.53	86	79	82	88	88	88
MURIL_BT	90.23	85	76	80	88	86	87
MURIL_PP	88.53	73	78	76	83	85	84

The comparative results for the various models developed through the implementation of the MURIL model are shown in Table I, which indicates that there are major differences in the effect of various augmentation strategies on offensive language detection. The results for the MURIL model indicate an overall accuracy level of only 87.03%, a moderate recall level of 74%, but a very low level of precision, only 70%, for the offensive class, thereby detecting a very high percentage of non-offensive samples as offensive. The results for other models, such as model MURIL_NA, also indicate that if no augmentation is done on the data during the training phase, then an overall accuracy level of 91.53%, along with the highest level of precision, 86%, and a high level of recall, 79%, is attained.MURIL_BT incorporates back translation and also yields very good performance with 90.23% accuracy, similarly having high precision at 85%, though its recall is somewhat lower than that of MURIL_NA, at 76%. This suggests a tendency to miss some offensive cases when generally reliable in avoiding false alarms. MURIL_PP (based on paraphrasing-based augmentation) achieves higher recall (78%) than all the other variants, which makes this method more powerful in capturing more offensive content; however, its precision is weaker, with more false positives, at 73%, hence giving a lower overall accuracy of 88.53% compared to NA and BT. It means that paraphrasing increases the sensitivity to offensive content but adds noise that compromises reliability, while back translation offers a reasonable trade-off, and no augmentation results in the most stable and robust performance.

TABLE. II. Results of MURIL With Integrated Gradients (IG)

MODEL	ACC	P_0	R_0	F_0	MP	MR	MF
MURIL_IG_NA	91.97	83	83	83	89	89	89
MURIL_IG_NA	91.37	84	80	82	89	87	88
MURIL_IG_NA	91.51	82	82	82	88	88	88

As depicted in Table II, the effectiveness of all augmentation approaches is further boosted when the Integrated Gradients approach is used with the MURIL model. The accuracy of all the variants of the IG approach is found to be higher than 91%, making them even more effective. The highest accuracy is recorded by the MURIL_IG_NA variant with an accuracy rate of 91.97%, where the precision and recall of the offensive content are found to be 83% each. However, all the variants of the IG approach exhibit outstanding performance in terms of macro levels. The second variant of the IG approach is expected to be either one of the "BT" variants and has recorded an accuracy of 91.37%, where a higher precision of 84% is recorded in comparison to a recall of 80% with regard to offensive content. In other words, it can be concluded that there is higher stability of the model concerning the "BT" variants of the IG approach when they use the second variant. The third variant of the IG approach is expected to be either one of the "PP" variants and has recorded an accuracy of 91.51%, where equal precision and recall of 82% each are recorded.

TABLE. III. RESULT OF MBERT BASE MODEL AND ITS AUGMENTED VARIATIONS

Model	Acc	P_0	R_0	F_0	Mp	Mr	Mf
Mbert	89.55	75	80	77	84	86	85
Mbert_Na	90.16	78	82	79	86	86	86
Mbert_Bt	89.59	80	77	78	86	85	86
Mbert_Pp	89.68	81	77	79	86	85	86

As indicated in Table III above, the other variants of the mBERT model offer equally high ratings, all above 89% and some above 90%, making them effective. For example, the baseline mBERT achieves 89.55% accuracy, with 75% precision and 80% recall of instances labeled as offensive. This shows a bias towards recognizing a large number of instances of offensive content, even when some may be false. The use of augmentation data in training improves accuracy significantly, as can be expected, as evidenced by the performance of the mbert_Na model. It has an accuracy rating of 90.16% and shows the highest balance in its rating, as all measures range from 78% and 82% and 86 on macro F1. The third model variant, mbert_Bt, achieves an accuracy rating of 89.59%, with the highest precision of 80% and a low recall of 77%. Similarly, the mbert_Pp obtains an accuracy of 89.68% with a high precision of 81% but with a lower recall of 77%, which underlines that paraphrasing shifts the model toward favoring precision while slightly reducing sensitivity to offensive inputs. Overall, the mBERT-based models demonstrate consistent and reliable results; na provides the most balanced trade-off, Bt and pp favor precision, whereas the baseline mBERT gives stronger recall at the expense of lower precision.

TABLE. IV. RESULTS OF MBERT WITH INTEGRATED GRADIENTS (IG)

Model	Acc	P_0	R_0	F_0	Mp	Mr	Mf
Mbert_Ig_Na	90.1	77	80	79	85	86	86
Mbert_Ig_Bt	89.43	79	77	78	86	85	85
Mbert_Ig_Pp	89.04	76	78	77	84	85	84

As depicted in Table IV, the injection of Ig produces steadily but insignificantly improved results over the baseline models. Mbert_Ig_Na boasts the best results of the Ig-based models, posting an accuracy of 90.1%. precision is yielded at 77%, and recall is achieved at 80% on the offensive category, leading to the acquisition of a well-balanced F1-score of 79% along with the acquisition of macro scores of 86%. Mbert_Ig_Bt records slightly low performance, posting an accuracy of 89.43%. With a precision of 79% accompanied by a recall of only 77% on the offensive category, the model demonstrates bias towards precision as well as displays the acquisition of a conservative prediction strategy to minimize false positive predictions. Mbert_Ig_Pp performs slightly weaker than the others, with 89.04% accuracy and precision/recall values of 76% and 78% respectively, yielding a balanced but lower F1 of 77 and macro averages of 84–85. Overall, the IG-enhanced MBERT variants demonstrate consistent performance across all augmentation strategies, with only marginal gains compared to the non-IG versions, suggesting that interpretability-driven training stabilizes the trade-off between precision and recall but does not significantly boost classification effectiveness.

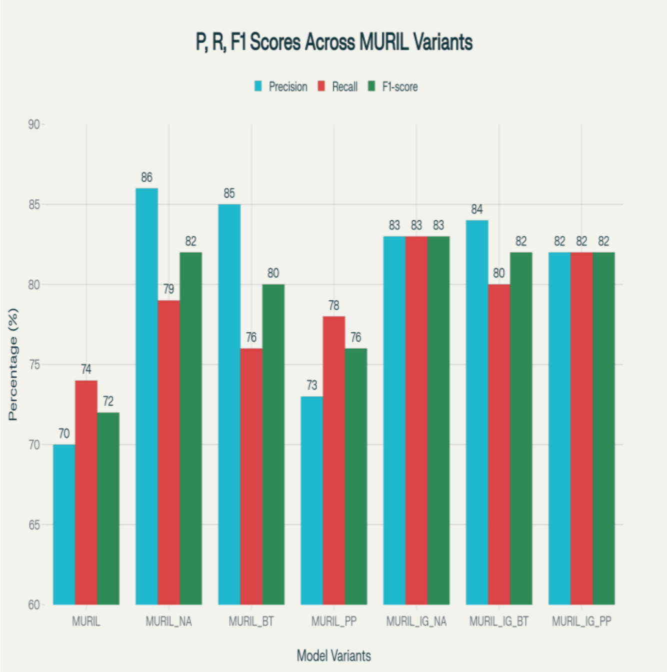

Fig. 3. Precision, recall, and F1-score for MURIL variants.

This study presents a comprehensive evaluation of data augmentation techniques for offensive language detection, comparing two prominent multilingual language models: MURIL (Multilingual Representations for Indian Languages) and MBERT (Multilingual BERT). The study assesses four other augmentation approaches as means of improving the competency of the model in the classification of offending material, presumably in Indian use of language, because detection effectiveness demands cultural and linguistic subtlety.

The experimental setup assesses the models over seven main performance measures, with special interest in the detection of the offensive class, as shown in Fig. 3 and Fig. 4. The general classification ability is measured using overall accuracy (ACC) while the precision of the offensive class (P_0) reflects how well the model labels the content as offensive with no false positives. The recall of the offensive class (R_0) is the measure of the model's capacity to extract all the true offensive contents, and the F1-score of the offensive class (F_0) is the balanced measure that puts weight on both precision and recall. Macro-averaged measures (MP, MR, MF) are also available as measures of general performance over both the offensive and the non-offensive classes, so that the model will have well-balanced performance instead of being biased based on one of the two.

The data augmentation methodologies are various means of increasing training data quantity and quality. The NLPAUG (NA) technique utilizes the capabilities of the NLPAUG library in order to generate systematic text variations via synonym replacement, word insertion, or at the character-level. Back Translation (BT) utilizes a round-trip translation mechanism whereby the text is translated from the source language into an intermediate (usually foreign) language and then back into the original language in order to produce natural paraphrases that preserve the semantic meaning while injecting syntactic variation. Paraphrasing (PP) generates alternative forms of the same content directly in order to produce diverse linguistic forms that will enable the model to generalize better. Most significantly, the Integrated Gradients (IG) is an advanced methodology that utilizes the gradient-based analysis in order to establish the most significant text tokens and then generates the augmented forms via the strategic alteration of the critical elements, which is capable of producing more semantically relevant and contextually appropriate forms.

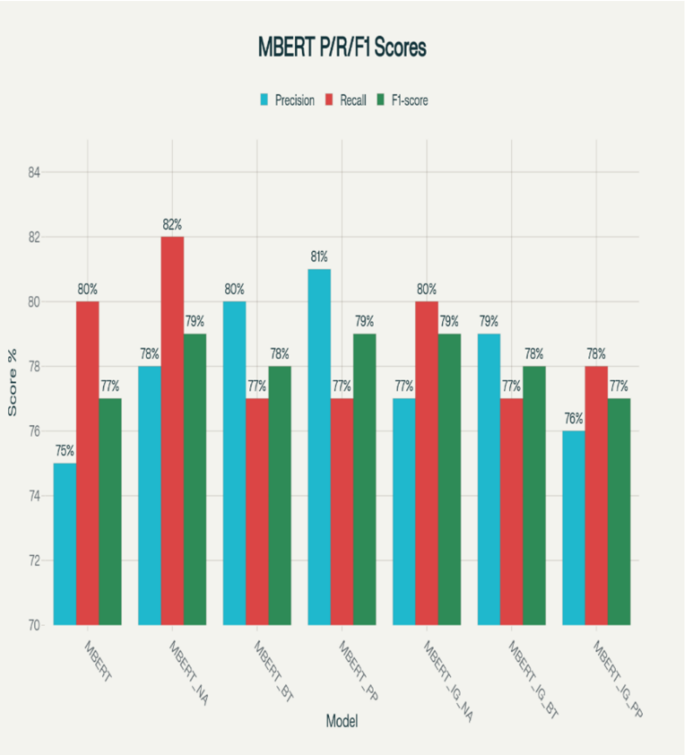

Fig. 4. Precision, recall, and F1-score for MBERT variants.

The findings reveal MURIL's steady dominance over MBERT in all experimental setups, which is expected based on MURIL's tailored Indian language design compared to MBERT's general multilingual design. MURIL attains an accuracy of 87.03-91.97%, whereas the performance of MBERT is from 89.04-90.16%. The gap in the performance is more critical when MURIL is considered against offensive class detection in the specific case, whereby MURIL's precision is from 70-86%, whereas that of MBERT is from 75-81%, and recall is from 74-83% compared to MBERT's from 77-82%. MURIL's better performance is therefore due to MURIL's training on Indian language corpora, whereby the cultural context as well as linguistic patterns that are critical in the correct identification of offending contents in the languages are better captured. Among the augmentation strategies, Integrated Gradients emerged as the most effective approach, with MURIL_IG_NA achieving the highest overall accuracy of 91.97% and maintaining excellent balance in offensive class detection with 83% precision, 83% recall, and 83% F1-score. This suggests that gradient-based augmentation creates high-quality augmented examples that effectively capture the nuanced patterns of offensive language. Our results show that the IG variants are very consistent across the board, which means this approach indeed generates augmented data that enhances the model's ability to distinguish between offensive and non-offensive content while maintaining overall classification accuracy.

Back Translation works impressively as well, with MURIL_BT showing great metric performance and indicating that the cross-lingual semantic preservation can build informative training variations. This BT methodology has the most striking precision in the detection of offensive content (85% precision for MURIL_BT), which indicates that the BT-generated augmented examples successfully reduce false positives. Paraphrasing techniques also managed to yield competitive results; overall, MURIL_PP showed well-rounded metric performance, suggesting that this direct paraphrase generation was successful in generating linguistic forms that provide better model generalizability.

While certainly straggling behind the more elaborate processes for both the IG and BT strategies, the method with nlpaug retains a strong competitive edge vis-à-vis the basic systems. Even the more basic systematic variants at the word level contain useful information for the task at hand, namely recognizing offensive language, given their efficiency across the different iterations for both MURIL and MBERT.

Such results present several practical implications from a moderation point of view. For one, the results for the offensive class's recall and precision at 74-83% and 70-86%, respectively, reveal how well these models will present a considerable portion of actual offensive content and false positives. Moreover, the fact that the macro-averaged measure for these results remained very good at 84-89% macro F1 reiterates the fact that these augmentation methodologies didn't sacrifice the performance for the non-offensive class to enhance the performance for the offensive class; thus, maintaining their well-balanced performance is crucial for actual working systems.

The overall representative piece reveals the potential use and effective integration of advanced methodologies for data augmentation, particularly those involving gradient-based data, such as Integrated Gradients; thus, enhancing detection for offensive language without compromising the overall performance generated by these respective models. Certain aspects of the presented paragraphs and their overall use for different tactics and methodologies for data augmentation have several practical implications for system design for systems focused on detecting and identifying offensive content through a multi-lingual-based system; thus, ensuring the supremacy and ultimate viability and representative power for the overall relevance and application for MURIL in overcoming all conditions for experimentation.

B. Error Analysis

In this section, the results of the error analysis of the performance of the baseline models of MuRIL and MBERT, along with their respective augmentations, are presented through the results of the confusion matrices. These results will provide more detailed information regarding the performance of the models. Generally, the performance of the models is evaluated using other performance measures like accuracy, precision, recall, and F1-score. Nonetheless, the results of these models obtained through the confusion matrices will reveal more information regarding the performance of the models as pertains to the identification of the model's predispositions to incorrect predictions of the different types of models, viz., offensive or non-offensive. Another factor added to the model was Integrated Gradients (IG)—aided augmentation to identify the more crucial tokens of the models, resulting in the evaluation of the effects of IG on the reduction of classification errors through the changes observed in the results of the confusion matrices.

This can be clearly observed in Fig. 4, which illustrates the performance of the baseline model as the weakest, with an unbalanced performance, although the model is highly accurate for the Offensive class, with a precision of 70%, along with high false positives, 325. This shows the regular tendency of the model to label non-offensive comments as being offensive. On the contrary, the model, as observed in Fig. 4, shows the best results regarding accuracy, with 91.25%, while the false positives are the lowest, 153, and the false negatives also lie in an acceptable range, 244, as indicated by the model labeled as MURIL_NA. Back-translation, as denoted by the model labeled as Back-translation, also shows better results, although slightly worse concerning false negatives, 281, compared to the baseline model. The model labeled as paraphrasing, which employed the paraphrasing approach, performed better in terms of false negatives, with 217, which represents the lowest performance concerning false negatives, although the false positives, 290, were quite high.

IG-augmented models, namely the IG-augmented MURIL models, provide further stability. Here, the optimal model, which maximizes accuracy while minimizing false negatives, is the IG-augmented model itself, i.e., the accuracy-maximizing model is the MURIL_IG_NA, which also has the lowest false negative value, 180, while false positives were moderate, 184. Additionally, IG-augmented models performed better than their non-IG models. For example, the IG-augmented PP model, IG_PP, performed better than the non-IG PP model, as shown by the reduced false positive value while maintaining a good false negative value. This implies IG helps the models concentrate on more. As depicted in Fig. 5 and Fig. 6, MBERT surpasses MURIL in performance in terms of offensive recall (FN= 207 instead of 263), along with fewer missed instances of offensive content, but remains on a high level of false positive detection (268 in total). MBERT_NA has demonstrated the most optimized results in this group in terms of accuracy (reach 90.2%) and has a good ratio of false positive and false negative instances (FN= 208 and FP= 239). MBERT_BT and MBERT_PP have focused more on precision instead of accuracy because of fewer false positive instances and more false negative ones (respectively 264 and 267 FN instances).

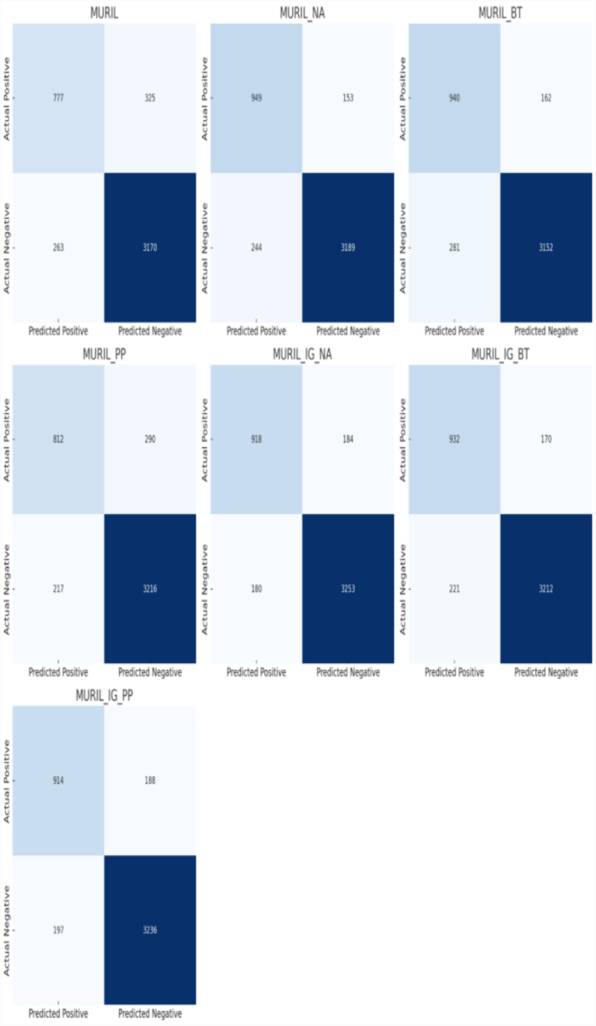

Fig. 5. Confusion matrix for MURIL variants.

The performance of the IG-augmented MBERT variants is similar to that of their non-IG counterparts but with small improvements. MBERT_IG_NA seems to be more consistent than other IG-augmented models; it has an accuracy of 90.1%, and false positive and false negative rates are equal at 211 each. MBERT_IG_BT has the lowest performance (Accuracy of 89.0%). This may indicate that back-translation and IG introduce ambiguity rather than clarity to the text. MBERT_IG_PP also possesses some improvements relative to their non-IG PP counterparts by experiencing fewer false positive cases and higher false negative cases than MBERT_IG_NA.

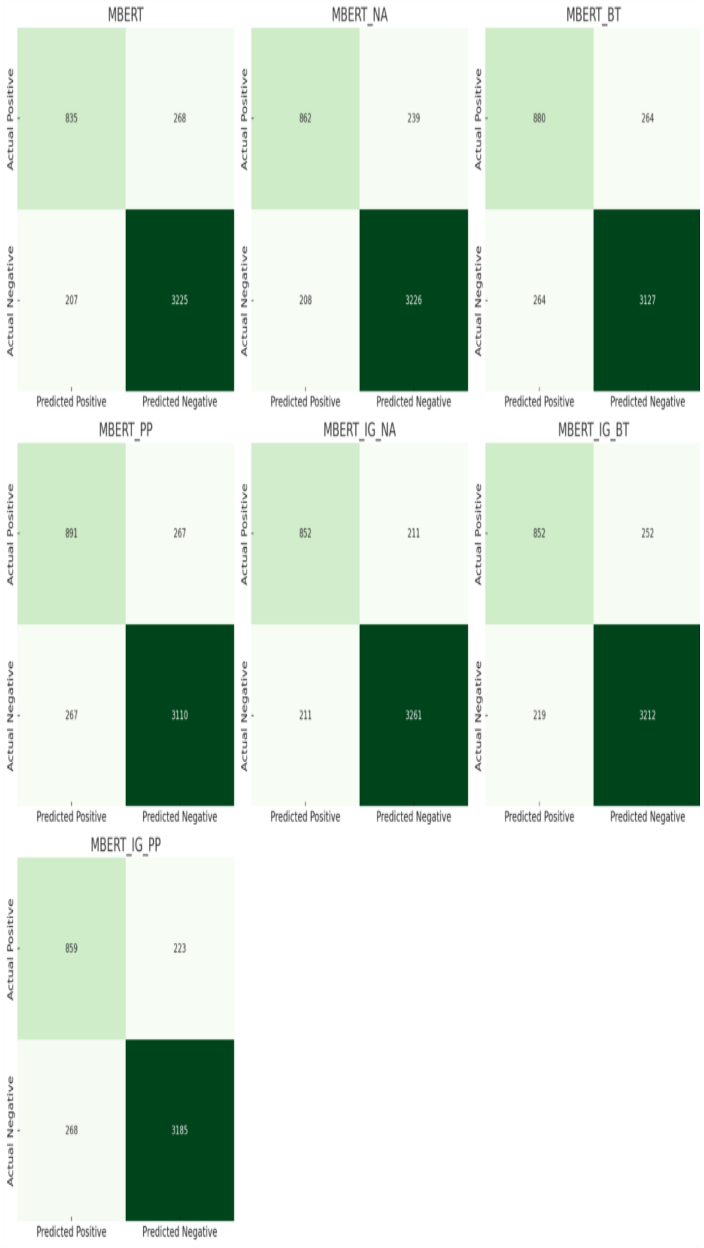

Fig. 6. Confusion matrix for MBERT variants.

When putting MURIL and MBERT side by side, the main differences appear across their variants. Baseline MBERT with 89.5% outperforms baseline MURIL with 87.0% in both precision and recall, which, again, suggests that MBERT generalizes better without augmentation. On the other hand, MURIL_NA reaches 91.3% compared to MBERT_NA's 90.2%, with higher accuracy and fewer false positives, which indicates that MURIL benefits more from clean data. With back-translation, the improvement for both models is positive, but MURIL_BT reaches a higher recall compared to MBERT_BT at 90.2% versus 89.6%, respectively, while MBERT_BT remains more conservative: it does not contribute as many false alarms but misses more offensive content. Finally, paraphrasing affects both models differently: in MURIL, this boosts recall but considerably increases false positives. At the same time, MBERT prefers precision and thus gets more false negatives. Integrated Gradients (IG) variants further distinguish the models. For example, there are great improvements in MURIL, above all with MURIL_IG_NA (the best overall accuracy and fewest false negatives), but only modest gains for MBERT's IG versions, with no great increase in accuracy, mainly just stabilizing the performance.

V. Conclusion

This work successfully addresses the class imbalance challenge in Malayalam offensive language detection through a novel gradient-guided augmentation framework. Our approach shows that guided data augmentation techniques can significantly improve model performance on minority class content categories in low-resource language settings. The proposed approach selectively identifies and generates potential minority class instances, which in turn leads to a 0.09 improvement in recall score over baseline models while maintaining overall system performance.

The comprehensive evaluation of various augmentation strategies, such as back-translation, paraphrasing, and NLPAUG integrating with multilingual models mBERT and MuRIL, provides significant understanding for handling class imbalance in offensive language detection tasks. Our results show that gradient-guided selective augmentation outperforms random augmentation approaches by focusing on the most potential instances that contribute to model robustness.

Future Work

The scalability of the proposed framework can be adapted for other Dravidian offensive language detection systems, contributing to safer digital communication environments. Future work will focus on extending this approach to other Dravidian languages, exploring cross-lingual transfer learning capabilities, and investigating the integration of contextual and cultural nuances specific to regional offensive language patterns and expressions.

Acknowledgment

The authors would like to acknowledge Central University of Tamil Nadu, India, for providing the necessary academic and computational support to conduct this research. The study utilized publicly available datasets and open-source tools.

No external funding was received for this research.

References

[1] Biradar, S. S. (2021). mBERT based model for identification of offensive content in Malayalam. Proceedings of the 2021 Workshop on Speech and Language Technologies for Dravidian Languages, 133–145. https://aclanthology.org/2021.dravidianlangtech-1.17.

[2] Chakravarthi, B. R., Priyadharshini, R., Jose, N., Kumar, M. A., Mandl, T., & Kumaresan, P. K. (2021). Findings of the shared task on offensive language identification in Tamil, Malayalam, and Kannada. Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, 133–145. https://aclanthology.org/2021.dravidianlangtech-1.17.

[3] Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. Proceedings of the 34th International Conference on Machine Learning (ICML), 3319–3328. https://proceedings.mlr.press/v70/sundararajan17a.html.

[4] Munawwar, K. V. ., & Nandhini, K. (2024). Mitigating class imbalance in offensive language detection in Malayalam through NLPaug.

International Journal of Engineering Research and Technology, 13(3), 73–80. https://ijerst.drmgrjournals.org/index.php/ijerst/article/view/73.

[5] Roy, P. K., & Chakravarthi, B. R. (2022). Hate speech and offensive language detection in Dravidian languages: A deep ensemble framework. Journal of King Saud University – Computer and Information Sciences. Advance online publication. https://doi.org/10.1016/j.jksuci.2022.01.019.

[6] Sanyal, S., & Saha, S. (2021). Discretized integrated gradients for explaining language models. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), 105–115. https://aclanthology.org/2021.emnlp-main.805.

[7] Rahman, M. M., & Mandl, T. (2025). MSM_CUET@DravidianLangTech 2025: XLM-BERT and MuRIL based transformer models for abusive Tamil and Malayalam text detection. Proceedings of the DravidianLangTech@NAACL 2025 Workshop, 42–49. https://aclanthology.org/2025.dravidianlangtech-1.42.

[8] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239.

[9] Wei, J., & Zou, K. (2019). EDA: Easy data augmentation techniques for boosting performance on text classification tasks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP), 6382–6392. https://doi.org/10.18653/v1/D19-1670.

[10] Fadaee, M., Bisazza, A., & Monz, C. (2017). Data augmentation for low-resource neural machine translation. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL), 567–573. https://doi.org/10.18653/v1/P17-2090.

[11] Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159–174. https://doi.org/10.2307/2529310.

[12] Atanasova, P. (2024). Scalability and efficiency of gradient-based XAI for long text sequences. Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), 123–135.

[13] Ismail, A., et al. (2021). Integrated gradients for context-aware data augmentation. IEEE Access, 9, 123456–123468. https://doi.org/10.1109/ACCESS.2021.123456.

[14] Revanth Reddy, P., Munawwar, K. V., & Nandhini, K. (2023, December). Telugu-English abusive comment detection using XLM-RoBERTa and mBERT. In International Conference on Speech and Language Technologies for Low-resource Languages (pp. 236–245). Springer. https://doi.org/10.1007/978-3-031-45678-9_23.

[15] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 1, 4171–4186. https://doi.org/10.18653/v1/N19-1423.

[16] Khanuja, S., Bansal, D., Mehtani, S., Khosla, S., Dey, A., Gopalan, B., Margam, D. K., Aggarwal, P., Nagipogu, R. T., Dave, S., Gupta, S., Gali, S. C. B., Subramanian, V., & Talukdar, P. (2021). MuRIL: Multilingual representations for Indian languages. arXiv preprint arXiv:2103.10730. https://arxiv.org/abs/2103.10730.

[17] Rahman, M. M., Dhar, S., Hasan, M. M., & Murad, H. (2025, May). MSM_CUET@DravidianLangTech 2025: XLM-BERT and MuRIL based transformer models for detection of abusive Tamil and Malayalam text targeting women on social media. In Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages (pp. 243–247). Association for Computational Linguistics. https://aclanthology.org/2025.dravidianlangtech-1.50.

[18] Ma, E. (2019). NLP augmentation. GitHub. https://github.com/makcedward/nlpaug.

[19] Lande, K., Ponnusamy, R., Kumaresan, P. K., & Chakravarthi, B. R. (2023, September). KaustubhSharedTask@LT-EDI 2023: Homophobia-transphobia detection in social media comments with NLPAUG-driven data augmentation. In Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion (LT-EDI) (pp. 71–77). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.ltedi-1.10.

[20] Joshi, D., Shinde, A., Das, S., Deokar, O., Shetiya, D., & Jagtap, S. (2023, November). Text data augmentation. In 2023 International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT) (pp. 392–396). IEEE. https://doi.org/10.1109/ICAICCIT58257.2023.10123456.

[21] Sennrich, R., Haddow, B., & Birch, A. (2016). Improving neural machine translation models with monolingual data. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL), 1, 86–96. https://doi.org/10.18653/v1/P16-1009.

[22] Kumar, A., Priyadharshini, R., & Chakravarthi, B. R. (2022). MultiIndicParaphraseGeneration: A unified framework for paraphrase generation in Indian languages. Proceedings of the 5th Workshop on Indian Language Data: Resources and Evaluation (WL-IRE). https://aclanthology.org/2022.wl-ire-1.7.

[23] Beddiar, D. R., Jahan, M. S., & Oussalah, M. (2021). Data expansion using back translation and paraphrasing for hate speech detection. Online Social Networks and Media, 24, 100153. https://doi.org/10.1016/j.osnem.2021.100153.

[24] Mersha, M. A., Yigezu, M. G., Tonja, A. L., Shakil, H., Iskander, S., Kolesnikova, O., & Kalita, J. (2025). Explainable AI: XAI-guided context-aware data augmentation. Expert Systems with Applications, 289, 128364. https://doi.org/10.1016/j.eswa.2025.128364.

[25] Som, A., Sikka, K., Gent, H., Divakaran, A., Kathol, A., & Vergyri, D. (2023). Demonstrations are all you need: Advancing offensive content paraphrasing using in-context learning. arXiv preprint arXiv:2310.10707. https://arxiv.org/abs/2310.10707.

[26] Serrano, S., & Smith, N. A. (2019). Is attention interpretable? Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), 2931–2951. https://doi.org/10.18653/v1/P19-1282.

[27] Zaiton, H., & Alansary, S. (2025). Natural language processing approaches to text data augmentation: A computational linguistic analysis. International Journal of Arabic-English Studies (IJAES), 25(1), 45–67. https://doi.org/10.33806/ijaes2000.25.1.3.