**Optimizing Hate Speech Detection in Malayalam-English Code-Mixed Text: Handling Women's Abuse by Synthetic Data Augmentation**

Dhanya L K[[1]](#footnote-1)\*, Kannan Balakrishnan2

1,2\* Cochin University of Science and Technology, Kochi, Kerala, India.

1Mar Baselios College of Engineering and Technology, Trivandrum, Kerala, India

dhanya.lk@cusat.ac.in; [bkannan@cusat.ac.in](mailto:bkannan@cusat.ac.in)

**Abstract**

The escalation of online hate speech, especially directed at women, has generated considerable apprehension in digital environments. Women often encounter mistreatment through disparaging remarks, body shaming, and explicit material, rendering the automated identification of such detrimental speech an essential endeavour. The difficulty escalates in low-resource languages such as Malayalam, particularly when code-mixing with English occurs. This study examines the efficacy of synthetic data augmentation methods—Machine Translation (MT), Masked Language Modeling (MLM), and Few-Shot Learning (FSL)—in improving hate speech detection for Malayalam-English code-mixed text. We utilize a multilingual BERT (mBERT) classifier, combining empirical data with synthetic data produced by masked language Modeling (MLM) and few-shot learning (FSL). Our results demonstrate that this method markedly enhances classification performance, attaining an F1-score of 86.42%. Through LIME-based explainability analysis, we demonstrate that contextual meaning is pivotal in the model's decision-making process. Furthermore, our comparative analysis of false positive and false negative rates between authentic and synthetic datasets underscores the capacity of MLM and FSL to foster a more equitable and impartial classification system. This study is the inaugural investigation into synthetic data generation for the detection of hate speech in code-mixed languages, concurrently addressing fairness issues, especially regarding online abuse directed at women.

**Keywords:** Synthetic data generation**,** Malayalam-English code-mixed data, Hate speech detection on women abuse comments, Fairness in NLP, Masked language modelling (MLM), few-shot learning, mBERT, LIME.

**1.Introduction**

Hate speech detection is becoming a crucial task in the realm of social media, where individuals articulate their opinions through various social media platforms. They are unconcerned to the substance of the opinion. It may contain unacceptable content. Therefore, the oversight of content on social media is the responsibility of every technology professional. This content may propagate hate against society based on sex, race, caste, politics, etc. We refer to such content as hate speech. Detecting hate speech is a prominent research area in natural language processing, where researchers employ machine learning and deep learning algorithms to identify hate speech content across various social media platforms. Hate speech is regarded as a subset of aggressive and abusive language; nonetheless, all these forms possess an offensive characteristic [1].

The language is not a factor when expressing ideas or opinions. It may be unilingual, biliteral, or multilingual. It is a common observation that websites frequently feature multilingual expressions or writing, which presents an additional obstacle. From where does code mixing originate? Individuals who are members of multilingual communities frequently communicate in multiple regional languages. This type of text is referred to as code-mixed data and is represented using a variety of languages [2]. Hate speech detection is already a difficult task; however, it is even more difficult to detect hate speech in code mixed data due to the unreliability of the annotated dataset, morphological complexity, and syntax variety.

Research on code-mixed data in Indian languages, particularly Dravidian languages, is notably scarce. Our research examines code-mixed data of Malayalam and English, where Malayalam is a Dravidian language spoken by the Malayali people in the Indian state of Kerala and the union territories of Lakshadweep and Puducherry (Mahé district). It is one of the 22 officially recognized languages of India. In 2013, Malayalam was officially recognized as a "Classical Language of India [3]." When individuals incorporate both Malayalam and English within a sentence while composing comments on social media platforms such as Facebook and YouTube, it is referred to as code-mixed text content.

As a low-resource language, Malayalam is receiving limited research attention in the field of natural language processing, particularly concerning code-mixed data. In our prior research, we gathered code-mixed comments from social media platforms, particularly YouTube and Facebook, which were subsequently pre-processed and labelled by various annotators for hate speech detection tasks. The manual collection of data and annotation has become a tedious task for the research work. We are motivated to engage in synthetic data generation. Synthetic data generation denotes the artificial production of data that simulates real-world data, created through computational methods rather than sourced from actual environments [4]. This methodology is especially beneficial in low-resource environments, such as for the detection of code-mixed hate speech in Malayalam-English, where the acquisition and annotation of authentic data is costly or time-consuming.

This study employed three synthetic data generation methodologies: 1) Translation-based augmentation involved utilizing the Hinglish dataset (Hindi-English) [5] from Yadav, A. K. et al. (2023) and executing machine translation via Google Translate after segregating Hindi and English words, which will be elaborated upon in Section 3. 2) Lexical substitution utilizing the XLM Roberta model [6], wherein entities have been replaced. 3) Contextual Sentence Generation utilizing the mT5 model [7], wherein we formulate prompts to enable the model to produce hateful posts within the specified context upon receiving several examples of hate speech. We performed several experiments to examine the efficacy of the proposed data augmentation methods in Malayalam-English (Manglish) code-mixed data. This article's primary contributions are as follows:

1. This is the first study to investigate the potential of synthetic data generation methods—Machine Translation (MT), Masked Language Modeling (MLM), and Few-Shot Learning (FSL)—to enhance the detection of hate speech in code-mixed text.
2. We have emphasized the necessity of further research into the development of synthetic hate speech data and, more broadly, the inclusion of additional low-resource languages in big language models for future application.
3. Our study of explainability underscores the significance of AI methods that are explainable in identifying critical linguistic indicators of hate speech. This enables the enhancement of detection models, transparency, and accountability.
4. In order to assess the influence of real versus synthetic data on model biases, we implement a fairness analysis. Our results demonstrate that the False Positive Rate (FPR) and False Negative Rate (FNR) are reduced by the integration of MLM and FSL-based synthetic data, thereby ensuring a more equitable classification across various data distributions.

This paper is organized as follows: Section 2 provides a synthesis of the literature. Section 3 delineates the dataset utilized in this study. Section 4 delineates our proposed methodology. Section 5 provides detailed experiments and an in-depth examination of the results, followed by concluding observations in Section 6.

**2.Related work**

In our literature review, we initiated the development of datasets for English hate speech identification by analysing common sources and methodologies for dataset creation. The predominant labelled datasets for abusive language identification have been created using data from Twitter (now X), primarily because of the extended availability of its data collection APIs relative to other platforms [8]. Nonetheless, alternative platforms have been examined, albeit to a lesser extent. Facebook has been utilized [9,10], while Instagram has been employed in more recent studies [11,12], highlighting the importance of multimedia platforms in the identification of abusive language. Some other studies employed comments from YouTube [13,14]. Data manual extraction, scraping, annotation, and preprocessing on all platforms are consistently challenging, time-consuming, and costly tasks [1]. Following the analysis of English dataset generation, we focused on the creation of code-mixed datasets, addressing the linguistic complexities inherent in multilingual and multicultural contexts, particularly on social media. These databases are crucial for understanding hate speech in multilingual communities, especially in regions where code-mixed communication prevails. However, aside from English, our emphasis on other languages pertains to multilingual usage, specifically code mixing, which is our primary focus area Hate speech detection efforts are notably more prevalent in Hinglish (Hindi-English) [16,17], where the curation of hate speech collected from social media environments is a laborious task. If our attention shifted to Dravidian languages, Tamil-English works have been utilized [18,19].

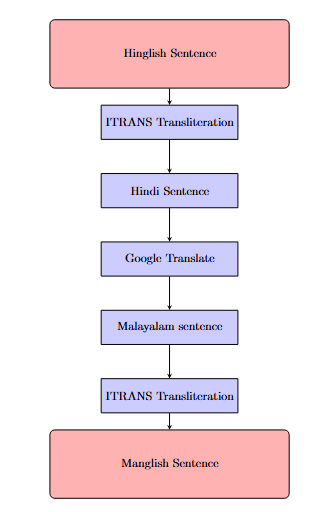
We expanded our literature review to include synthetic data generation for hate speech identification. This emerging field employs generative models to surmount the limitations of traditional data sources, enabling the creation of diverse and high-quality datasets to enhance model performance, particularly in resource-constrained settings. Recent advancements in image generation, text generation, and speech synthesis have facilitated the emergence of a research domain where in model outputs can be utilized to retrain subsequent models. This mitigates annotation expenses, preserves data confidentiality, and addresses issues of data imbalance and scarcity [15].One of the significant works published in 2020 was by Rui Cao et al., who proposed HateGAN, a deep generative reinforcement learning model that addresses the problem of class imbalance by augmenting the dataset with hateful tweets. They conduct extensive experiments to augment two commonly used hate speech detection datasets with tweets produced by HateGAN. Their experimental findings demonstrate that HateGAN improves the detection effectiveness of the hate speech category, regardless of the classifiers and datasets utilized in the detection task [19]. Thomas Hartvigsen et.al [20] endeavored to enhance current toxic content datasets by utilizing GPT-3, a text generation model, to produce extensive data on toxic and benign statements aimed at minority identity groups. The authors discovered that the machine-generated dataset was of superior quality, and toxicity detection models trained on it markedly surpassed those trained on pre-existing human-curated toxicity datasets. Like these works, we seek to create synthetic data to raise the machine learning classifier's hate speech detecting accuracy. But we position our work especially to increase hate speech detection accuracy for constrained data settings. especially mixed data using Malayalam-English codes

**3.Methodology**

Here we have covered our strategy for creating synthetic data for hate speech detection in Malayalam-English code-mixed language. Here are the generating methods:(1) Machine Translation (2) Masked language modelling (3) few shot learning using mT5

**3.1 Machine Translation(MT)**

Machine translation is an efficient technique for producing synthetic data, especially for the detection of hate speech in code-mixed languages. The scarcity of labelled hate speech data often complicates the training of robust models. Utilizing machine translation, we can produce varied synthetic data that reflects linguistic diversity while maintaining the semantic and contextual integrity of hate speech expressions. This study employs machine translation to transform Hinglish (Hindi-English code-mixed text) into Manglish (Malayalam-English code-mixed text), generating synthetic samples for model training. The procedure commences with the transliteration of Hinglish into Hindi utilizing the ITRANS system [21], transforming Romanized Hindi text into the Devanagari script. Subsequently, Google Translate is employed to convert the Hindi text into Malayalam[22], ensuring the preservation of the fundamental meaning. The Malayalam script is ultimately transliterated into Manglish (Romanized Malayalam) via ITRANS, resulting in a synthetically generated code-mixed sentence. This pipeline enhances hate speech datasets by incorporating new linguistic variations, thereby augmenting the robustness of detection models. We intend to mitigate data scarcity and enhance generalization across various code-mixed language patterns by incorporating machine translation into the data generation process. The entire process is illustrated in Figure 1, and a sample example is provided in Figure 2

****

**ഈ സിനിമ വളരെ നല്ലതായിരുന്നു, I really enjoyed it**

**यह मूवी बहुत अच्छी है, I really enjoyed it**

**ee cinema valare nallathayirunnu, I really enjoyed it**

**yeh movie bohot acchi hai, I really enjoyed it**

Figure :Machine translation approach for synthetic data generation Figure 2: Sample translation from Hinglish to Manglish

One disadvantage of this method is that during human evaluation of sentence efficiency, we observed that intra-sentential code-mixing sentences were ineffective, whereas inter-sentential code-mixing sentences performed better. Intra-sentential code-mixing occurs when multiple languages are integrated within a single sentence. This frequently entails the incorporation of words, phrases, or clauses from one language into another. Conversely, inter-sentential code-mixing transpires when a speaker alternates languages between sentences or clauses.[23]. In figure 2 inter sentential code-mixing sentence is given as an example.

**3.2 Masked Language Modelling(MLM)**

Our second approach is Masked language modelling(MLM).This is a method employed in natural language processing (NLP). It enables computers to comprehend and produce human language [24]. Envision reading a sentence with certain words obscured, necessitating the inference of those words. Masked language Modeling performs this function. It conceals certain words within a sentence and subsequently endeavours to predict the omitted words. This approach instructs models to comprehend context and significance. We have implemented this method using the language model XLM-RoBERTa. The detailed methodology is given in figure 3.

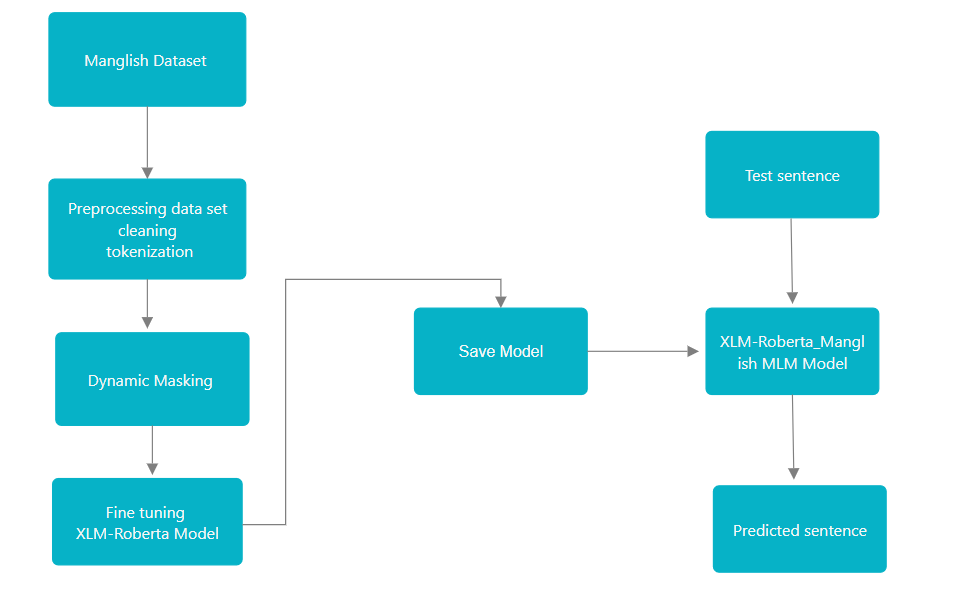


Figure 3: Methodology for MLM for synthetic data generation

Masked language models are essential as they can analyze the contextual meaning of words more efficiently than conventional models. By utilizing both the left and right context surrounding a masked word, masked language models comprehend language in a manner that closely resembles human understanding. We have experimented with an MLM model for synthetic data generation. We have dynamically obscured specific words in a sentence and employed XLM-RoBERTa to predict the masked word in a Manglish sentence. As a multilingual model, our expectations for the output were elevated. However, the actual output produced by the pretrained XLM-RoBERTa was subpar and chaotic.

We have considered fine-tuning XLM-Roberta utilizing the existing Manglish dataset employed in our prior research [3,25]. Prior to finetuning, we eliminated extraneous noise from the text data, tokenized it, and subsequently employed it for finetuning, utilizing dynamic masking [24]. In dynamic masking, randomly introduce [MASK] tokens within the text to facilitate MLM training. This facilitates the model's acquisition of contextual word substitutions. The finetuned model has been saved as ***XLM-Roberta\_Manglish MLM Model***.

**3.2.1 XLM-RoBERTa**

The Facebook AI team launched XLM-RoBERTa (**Robustly Optimized BERT Pretraining Approach)** in November 2019. This is purely a Transformer-based language model efficient in processing text from 100 distinct languages. The primary enhancement that XLM-Roberta provides compared to the original is a markedly greater volume of training data [26]. The Transformer fundamentally relies on a self-attention mechanism that enables the model to assess the significance of various words in a sentence in relation to one another. This facilitates MLMs in comprehending context with nuance. Transformers process entire sentences concurrently, rather than sequentially analysing one word at a time as older models did, enabling them to simultaneously capture relationships among all words. Through self-attention, the model can concentrate on pertinent words, irrespective of their location within the sentence. The pretrained XLM-Roberta MLM model is fine-tuned on a smaller, task-specific Manglish (Malayalam-English code-mixed) dataset. Pre-training establishes the foundation, while fine-tuning enhances specialization. After tokenization, we randomly obscure a portion of the tokens (usually 15% of the tokens in a sentence) prior to inputting them into the model. The model's function is to forecast these concealed tokens. Example of proposed MLM model is given in Table 1 where masked word of input sentence is getting 4 different outputs.

**3.3 Few shot Learning(FSL)**

The necessity for substantial training data in deep neural networks to attain favourable outcomes renders these models inefficient in scenarios where training data is scarce. Hate speech and offensive language are prevalent phenomena on social media that disregard linguistic barriers; consequently, the scarcity of adequately labelled data in certain languages, particularly low-resource ones, renders automatic detection algorithms impractical [27].

|  |
| --- |
| **Manglish Sentence**: Avan oru **[MASK]** jaathi aanu, avanodu sammathikkanda. |
| **English Meaning**: He belongs to **[MASK]** caste, don't agree with him |
| MLM output: Avan oru **veruppikkunna** jaathi aanu, avanodu sammathikkanda.  Avan oru **nashicha** jaathi aanu, avanodu sammathikkanda  Avan oru **mandan** jaathi aanu, avanodu sammathikkanda.  Avan oru **cheriya** jaathi aanu, avanodu sammathikkanda. |

Table 1: Sample output of Proposed MLM Model\*

This is where Few-shot learning (FSL) comes in. It's done in a slightly different way. This lets a model learn how to do new things with just a few examples. In situations where it's hard to get a lot of labelled data, this method comes in very handy. FSL uses what you already know and quickly and easily applies it to new tasks. Few-shot learning lets large language models (LLMs) create fake data without needing a lot of fine-tuning. This makes it an efficient and cost-effective way to add to data. The quality of the hate speech data that is made depends a lot on how well the prompts are designed [28]. This is where well-structured examples tell the model what to do based on the context. But validation is still an important step to make sure that the generated text fits the target context and shows hate speech patterns correctly. It is possible to make hate speech detection models much more effective by adding validated synthetic data to the training dataset. This leads to better generalization and robustness in real-world applications.

First, a small group of real hate speech examples in Manglish (Malayalam-English code-mixed) are gathered to prepare the ground for few-shot learning. After that, carefully thought-out prompts are made using these examples in a structured way to help the language model (LLM) make fake hate speech. A few examples (usually 3–5) are used in few-shot learning to help the language model learn the style, structure, and context of hate speech in Manglish. The model, mT5, makes up new fake hate speech based on the patterns it has learned after it has been trained. The outputs are then post-processed, which means they are looked over and filtered to get rid of text that isn't relevant or isn't good enough. This is done by manually verifying, toxicity filtering, and de-duplicating the text. Lastly, the validated synthetic hate speech dataset is put together and saved so that it can be used to train and improve models that find hate speech. The detailed methodology is depicted in figure 4.

**3.3.1 mT5(Multilingual T5)**

The mC4 dataset has natural text in 101 languages and is used to train the mT5 model. The model can understand shared representations and common patterns across different languages thanks to this in-depth training [30]. There are several good reasons why mT5 (Multilingual T5) is a good choice for making fake data for Manglish (Malayalam-English code-mixed) hate speech detection:

1. mT5 is already trained on several languages, such as English and Malayalam, which makes it a good choice for working with code-mixed text like Manglish.
2. mT5 uses the text-to-text paradigm of the T5 model, which makes it very adaptable to tasks like creating text, translating it, and sorting it into groups. This makes it easy to change for making fake data with well-thought-out prompts.
3. mT5 can make relevant fake hate speech samples with just a few examples (few-shot learning) because it has been trained on different types of text.\

Figure 4: Few shot learning methodology to create synthetic data

**3.3.2 Prompting**It involves making and improving the instructions that are given to a Large Language Model (LLM) so that it can produce accurate and useful results [31]. It is very important to carefully plan the prompts so that the model can find hate speech in code-mixed Malayalam and English data. In few-shot learning, prompts are provided with 3 to 4 labelled examples, as illustrated in Table 2. Here Every sentence adheres to the abusive patterns outlined in the prompt. It adversely affects women, perpetuating patterns of hate speech. The syntax and style are preserved in Manglish to maintain authenticity.

**4.Experiments and Results**

In all experiments, we optimized the cased multilingual BERT model [32]. We employed mBERT's subword tokenizer to tokenize the pre-processed input post and encode it into 650 dimensions utilizing mBERT embeddings. The encoding layer is succeeded by a dropout layer with a probability of 0.1, followed by a linear output layer that projects the 650-dimensional embedding into a two-dimensional vector.

|  |  |
| --- | --- |
| **Sample Prompt** | **Output** |
| # Few-shot learning prompt  prompt = """Here are some examples of hate speech targeting women in Manglish:  1. Aval oru buddhi illa, njan arinjirunnu.  2. Pennungaline okke vishwasikkan kazhiyilla.  3. Aval oru pandaram, arayille?  4. Ee pennu onnum ariyilla, thirinju nikkatte.  5. Ninte mathiri pennungal onnum venda namukku.  Now, generate another hate speech comment in Manglish:""" | Generated Hate Speech Sentences:  1. Aval oru mandan pennu, avalude karyangal onnum nannayilla.  2. Ee pennu nannayi samsarikkilla, avalde buddhi thazhacheduthu.  3. Pennungal onnum work cheyyan ariyilla, avarkku mathrame abhiprayam undu.  4. Ithra neram moola nirthunna pennu aale pidikkilla.  5. Avalde mathiri pennungal undenkil society nashtapedum. |

Table 2: Sample output of mT5 after few shot learning\*

The model is trained using the Cross-Entropy loss function and the Adam optimizer. To ensure our experimental conditions reflect real-world scenarios, our test dataset comprises solely original posts selected from the restricted data context. This dataset consists of Malayalam–English code-mixed posts sourced from YouTube, specifically concerning women's abuse. Hyperparameters and their value are given in table 3. Detailed distribution dataset and their sources for each experiment is given in table 4. Following are the different experimental setups done for this research work. Multilingual BERT (mBERT) is the optimal selection for Malayalam-English code-mixed data owing to its proficiency in managing multilingual text efficiently. In contrast to standard BERT-large, which is exclusively trained on English, mBERT[32] is trained on 104 languages, including Malayalam, enabling it to capture cross-lingual representations and comprehend Malayalam-specific linguistic structures more effectively. This renders it especially appropriate for code-mixed contexts, wherein both Malayalam and English lexicon, alongside transliterations and grammatical variations, coexist within the same text. Furthermore, mBERT's tokenizer accommodates Malayalam subwords, minimizing excessive token fragmentation that may impair performance in monolingual models such as BERT-large. A significant advantage of mBERT is its capacity to utilize transfer learning from its multilingual pretraining corpus, enhancing performance despite the availability of limited Malayalam-English training data. Conversely, standard BERT does not possess this multilingual pretraining, rendering it less effective for low-resource languages such as Malayalam.

|  |  |
| --- | --- |
| **Model** | mBERT-large-cased |
| **Max Sequence Length** | 256 |
| **Batch Size** | 8 |
| **Learning Rate** | 2e-5 |
| **Weight Decay** | 0.01 |
| **Dropout** | 0.1 |
| **Optimizer** | AdamW |
| **Epochs** | 15-20 |
| **Early Stopping** | patience = 5 |

Table 3: Hyper parameters and their values of the model

**4.1 Experiment on Baseline Model**

In Experiment 1, we utilized authentic data collected from 48 YouTube videos sourced from prominent Malayalam channels. Due to confidentiality concerns, we are unable to disclose channel names here. Due to their popularity, these videos garnered a substantial number of comments and reactions. We have concentrated on a channel predominantly led by women, featuring content such as cooking, beauty vlogs, and dieting vlogs. Utilizing the YouTube Data API, we collected comprehensive information on each video, encompassing its ID, title, comments, replies to comments, likes, date, and time. We collected approximately 5,500 hate comments related to these videos. The predominant comments were submitted from January 2024 to January 2025. The extracted comments and replies from each video were stored in distinct CSV files in chronological order. Ultimately, we amalgamated the dataset into a singular CSV file comprising solely the “comments” column. We must focus solely on Manglish comments. The presence of 5500 comments is attributed to the fact that many channels have removed negative comments, specifically those related to bullying, from their videos, which presents a challenge. From a total of 5500, we selected only 3000 for dataset balancing and gathered an additional 3000 non-hateful comments concerning women. In this study, two annotators proficient in both Malayalam and English manually annotated the dataset to identify hate speech. Cohen’s Kappa[34] was employed to evaluate inter-annotator agreement for hate speech annotations across two collections of 10,000 code-mixed texts, resulting in a Kappa score of 0.94, signifying high-quality annotations. In the study, hate speech is designated as "HATE," whereas non-hateful messages are designated as "NHATE."

Code-mixed data is sourced from various origins and frequently comprises substantial noise, such as punctuation, whitespace, numerals, special characters, excessive spaces, emojis, and stop words. This unrefined code-mixed data poses difficulties for the analytical process. Consequently, data preprocessing is essential to convert raw data into a more refined format. Data preprocessing is crucial for ensuring proper formatting, which facilitates more effective outcomes when employing this processed data in diverse models [33]. In the field of hate speech identification research, there has been comparatively less focus on data "cleaning" than in other natural language processing tasks. This is probably attributable to the intricate characteristics of hate speech language, which necessitates a more profound level of analysis than conventional text. Certain users ingeniously bypass platform speech limitations by replacing letters with symbols to discreetly communicate otherwise forbidden messages. Researchers frequently utilize standard techniques such as transforming text to lowercase, tokenizing tweets, and eliminating URLs. Emojis, while potentially beneficial for improving NLP task performance, are often omitted [33].

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Experiments** | **Training dataset** | | | | | **Test dataset (Real data)** |
| **Real data** | | **Synthetic data (Method)** | | **Total** |
|  | **hateful** | **nonhateful** | **hateful** | **nonhateful** |  |  |
| Experiment1 | 3000 | 3000 | - | | 6000 | 1200 |
| Experiment2 | 1500 | 1500 | 1400(MT) | 1492(MT) | 5892 | 1200 |
| Experiment 3 | 1500 | 1500 | 1450(MLM) | 1410(MLM) | 5860 | 1200 |
| Experiment 4 | 1500 | 1500 | 1450(FSL) | 1407(FSL) | 5857 | 1200 |
| Experiment 5 | 1500 | 1500 | 1000(MT)+1000(MLM)+1000(FSL)**[Equal selection of hateful and non-hateful posts]** | | 6000 | 1200 |

Table 4: Distribution of data for different experiments

We collected an additional 1,200 comments, both hateful and non-hateful, for training purposes, as this is standard practice for all experiments.

The mBERT model was trained on an authentic dataset and assessed for its capacity to detect hate speech. The model attained an accuracy of 81.08%, a precision of 80.03%, a recall of 83.08%, and an F1-score of 81.53%. The classifier accurately identified 501 instances of hate speech (TP) but erroneously classified 125 non-hate speech cases as hate speech (FP), suggesting a propensity for false positives. Furthermore, 102 instances of hate speech (FN) went undetected, indicating that certain subtle patterns of hate speech were not adequately identified. The comparatively high recall indicates the model's efficacy in detecting hate speech, yet the moderate incidence of false positives underscores the necessity for additional refinement. Prospective advancements, including synthetic data augmentation and enhanced fine-tuning methodologies, could mitigate misclassifications and elevate overall detection efficacy.

**4.2. Experiment on Machine Translation-based Synthetic Data**

In Experiment 2, we combined 3,000 posts from authentic data with 2,892 posts from synthetic data generated through the machine translation method discussed in Section 3.1. The comprehensive distribution of datasets and sources is presented in Table 4. The decline in performance noted when integrating machine-translated (MT) synthetic data into training indicates that the quality and distribution of generated samples are vital to model efficacy. The reduction in True Positives (TP) from 501 to 422 and the escalation in False Positives (FP) from 125 to 180 suggest that the model is misclassifying a greater number of instances, probably owing to discrepancies between actual and synthetic hate speech patterns. Machine translation can produce unnatural sentence structures, grammatical inaccuracies, or shifts in context that diverge from authentic hate speech, resulting in classification ambiguity. The rise in False Negatives (FN) from 102 to 148 indicates that the model is increasingly unable to accurately identify instances of genuine hate speech, thereby diminishing its recall and overall efficacy. A significant factor may be overfitting to synthetic patterns, wherein the model acquires repetitive or less diverse representations from machine translation-generated data instead of generalizable linguistic cues. This is evidenced by the decline in precision (80.03% → 70.1%), F1-score (81.53% → 72.01%), and overall accuracy (81.08% → 72.67%). The creation of synthetic data through machine translation of Hinglish sentences, combined with authentic data sourced from social media, may result in increased confusion and diminished effectiveness. The writing style, specifically the sentence structure in Manglish and Hinglish, differs, which has also contributed to the issue. The lack of success of this experiment necessitates a third experiment that integrates real data with synthetic data generated through the MLM method.

**4.3 Experiment on MLM-based Synthetic Data**

In Experiment 3, we combined 3,000 posts from authentic data with 2,860 posts from synthetic data generated through the masked language modelling method discussed in Section 3.2. The comprehensive distribution of datasets and sources is presented in Table 4. The results indicate that incorporating MLM-generated synthetic data enhanced the model's performance relative to both Machine Translation (MT)-based synthetic data and the real dataset baseline. The accuracy rose to 84.42%, and the F1-score attained 84.48, representing a substantial enhancement compared to the MT-based synthetic dataset (72.67% accuracy, 72.01 F1-score).

Synthetic data augmentation based on Masked Language Modeling (MLM) surpassed that of Machine Translation (MT) owing to its capacity to produce more natural and contextually relevant data. MLM produces synthetic samples by predicting masked words within a given linguistic context, thereby maintaining the original grammatical structure and fluency of code-mixed Malayalam-English text. Conversely, MT-based data frequently results in translation inaccuracies, awkward phrasing, or misinterpretations of colloquial language, thereby diminishing training efficacy. The dataset utilized for Experiment 1 is specifically employed for MLM-based synthetic data generation using XLM-Roberta. Our proposed "**XLM-Roberta\_Manglish MLM Model**" generates more diverse and contextual sentences from provided real data.

A significant factor underlying this success is the diminished incidence of false positives (FP = 98) and false negatives (FN = 89) achieved through MLM-based augmentation. A reduced false positive count indicates that fewer non-hate speech instances were erroneously categorized as hate speech, thereby enhancing precision. Likewise, the decrease in false negatives guarantees that genuine hate speech is not disregarded. XLM-RoBERTa, employed in the MLM methodology, enhances the model's multilingual proficiency by utilizing a more extensive and varied pretraining corpus than mBERT, thereby augmenting its capacity to manage intricate cross-lingual and code-mixed situations.

**4.4 Experiment on Few-shot Learning-based Synthetic Data**

In Experiment 4, we combined 3,000 posts from authentic data with 2,857 posts from synthetic data generated through the few shot learning using mT5 model discussed in Section 3.3. The comprehensive distribution of datasets and sources is presented in Table 4. The Few-Shot Learning (FSL) approach utilizing synthetic data markedly enhanced the model's efficacy in detecting hate speech in Malayalam-English (Manglish). In contrast to Machine Translation (MT) and Masked Language Modeling (MLM)-based augmentation, FSL-generated samples yielded a more contextually precise and varied array of hate speech variations. This is demonstrated by the enhancement in precision (85.2%) and recall (86.62%), resulting in an overall F1-score of 85.9%. The capacity of mT5 to produce novel data from limited examples likely enhanced generalization, effectively capturing nuanced linguistic variations and slang prevalent in online hate speech. Moreover, the False Positives (FP) and False Negatives (FN) were further diminished (90 and 80, respectively), signifying a more balanced classifier with reduced misclassifications. In comparison to MLM-based augmentation, FSL attained a marginally superior accuracy (85.83%) owing to its capacity to generate synthetically diverse yet authentic hate speech patterns. This indicates that FSL-generated synthetic data is a superior augmentation method compared to MT and MLM for detecting code-mixed hate speech. Future research may concentrate on refining the FSL model to augment diversity and reduce biases, potentially employing contrastive learning techniques to enhance semantic representation in hate speech classification.

**4.5 Experiment on Combination of Synthetic Data**

In Experiment 5, we amalgamated 3,000 posts from genuine data with 3,000 posts from synthetic data produced via three distinct methods: 1,000 from Machine Translation (MT), 1,000 from Masked Language Modeling (MLM), and 1,000 from Few-Shot Learning (FSL). Table 4 presents the extensive distribution of datasets and sources. Experiment 5 yielded a modest improvement in accuracy and F1 score compared to prior experiments.

The integration of Machine Translation (MT), Masked Language Modeling (MLM), and Few-Shot Learning (FSL) for synthetic data generation has markedly enhanced the efficacy of mBERT in detecting hate speech within Malayalam-English code-mixed data. The model attained an F1-score of 86.42%, surpassing all prior experiments. The precision (85.65%) and recall (87.2%) demonstrate a well-calibrated classifier that adeptly distinguishes between hateful and non-hateful content. The decrease in false negatives (77) indicates that the model has improved its ability to identify hate speech instances while maintaining accuracy. The incorporation of various data augmentation techniques has enhanced generalization and robustness. The enhanced efficacy of this method is due to the variety of synthetic data. MT introduces novel sentence structures, MLM guarantees linguistic fluency, and FSL produces contextually pertinent variations. Utilizing various augmentation strategies enables the model to acquire more nuanced linguistic representations, thereby diminishing bias and improving classification efficacy. The findings suggest that integrating various synthetic data generation methods is an effective approach for enhancing hate speech detection, especially in underrepresented multilingual contexts. Subsequent investigation may entail refining synthetic data selection and testing contrastive learning methodologies to improve feature representation.

**4.6 Results**

The performance evaluation of the mBERT model for detecting hate speech in data with mixed Malayalam and English codes is shown in this section. Real data and a variety of synthetic data augmentation methods, such as Machine Translation (MT), Masked Language Modeling (MLM), and Few-Shot Learning (FSL), were used in the experiments. Assessing the effects of synthetic data on model performance in terms of precision, recall, accuracy, and F1-score was the aim of these experiments. Table 5 depicts performance analysis of mBERT on various combination of datasets.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Dataset** | **TP** | **FP** | **TN** | **FN** | **Precision** | **Recall** | **Accuracy** | **F1-score** |
| **mBERT** | **Real Dataset** | **501** | **125** | **472** | **102** | **80.03** | **83.08** | **81.08** | **81.53** |
| **mBERT** | **Real + Synthetic (MT)** | **422** | **180** | **450** | **148** | **70.1** | **74.04** | **72.67** | **72.01** |
| **mBERT** | **Real + Synthetic (MLM)** | **509** | **98** | **504** | **89** | **83.86** | **85.12** | **84.42** | **84.48** |
| **mBERT** | **Real + Synthetic (FSL)** | **518** | **90** | **512** | **80** | **85.2** | **86.62** | **85.83** | **85.9** |
| **mBERT** | **Real + Synthetic (MT+MLM+FSL)** | **525** | **88** | **520** | **77** | **85.65** | **87.2** | **86.15** | **86.42** |

Table 5: Performance of mBERT on different datasets

A comparison of Precision, Recall, Accuracy, and F1-score for various dataset configurations used to train the mBERT model is shown in Figure 5. All metrics showed balanced performance for the baseline model, which was trained solely on real data. However, the precision and recall of the model trained using synthetic data produced by Machine Translation (MT) decreased, suggesting that this augmentation technique introduced some inconsistencies. On the other hand, adding artificial data produced by Few-Shot Learning (FSL) and Masked Language Modeling (MLM) greatly enhanced every performance metric, with the FSL-based model receiving the highest ratings. With the highest F1-score, recall, and accuracy, the combined synthetic dataset (MT + MLM + FSL) showed the best overall performance, proving that a variety of synthetic data sources improves model generalization and robustness.

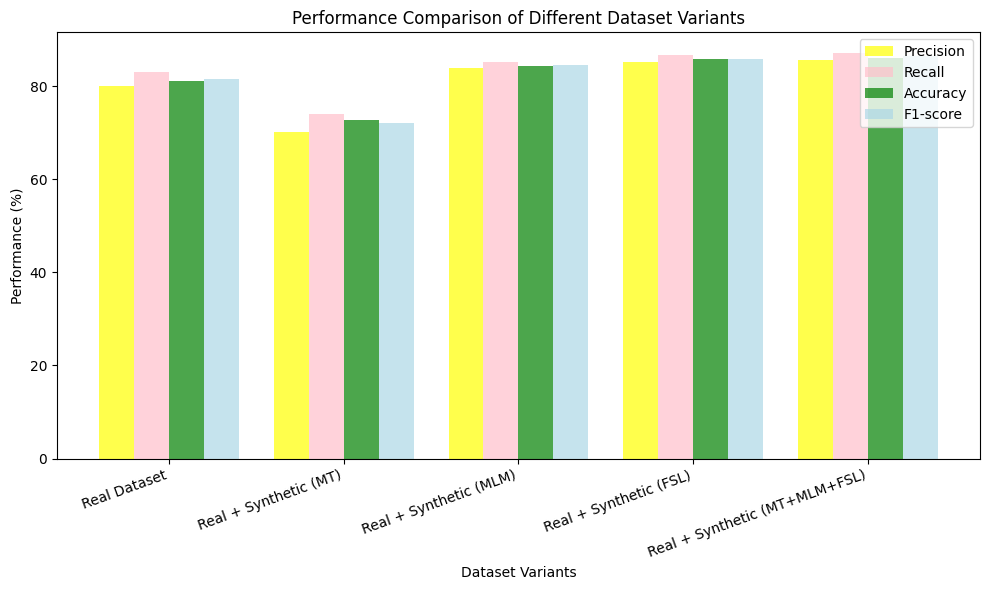


Figure 5: Performance Analysis of mBERT on different datasets

The F1-score is an essential metric for assessing hate speech detection models as it equilibrates precision and recall, accounting for both false positives and false negatives. In hate speech detection, precision quantifies the proportion of predicted hateful comments that are genuinely hateful, whereas recall assesses the extent to which the model accurately identifies actual hateful comments. An elevated F1-score guarantees that the model is both precise and thorough in identifying harmful content. The results clearly indicate that synthetic data significantly enhances model generalization. Although distinct synthetic methodologies may yield performance discrepancies, the integration of MT, MLM, and FSL synthetic data results in a more resilient model with diminished errors. This indicates that meticulously crafted synthetic datasets can significantly improve model performance, reduce biases, and enhance hate speech classification in Malayalam-English code-mixed data.

The F1-score line plot (Figure 6) illustrates how different synthetic data augmentation techniques impact the model’s classification performance. The baseline model, trained solely on real data, achieves an F1-score of 81.53%. However, when Machine Translation (MT) synthetic data is introduced, the F1-score drops to 72.01%, suggesting that the synthetic data generated using MT might introduce noise or unnatural linguistic patterns that reduce classification effectiveness.

Conversely, the Masked Language Modeling (MLM) and Few-Shot Learning (FSL) techniques improve the F1-score to 84.48% and 85.9%, respectively. This indicates that synthetic data generated using contextual learning and few-shot learning approaches contributes positively to hate speech detection. The best performance (86.42%) is observed when all three methods (MT, MLM, and FSL) are combined, highlighting the advantage of using a diverse set of synthetic data generation techniques.

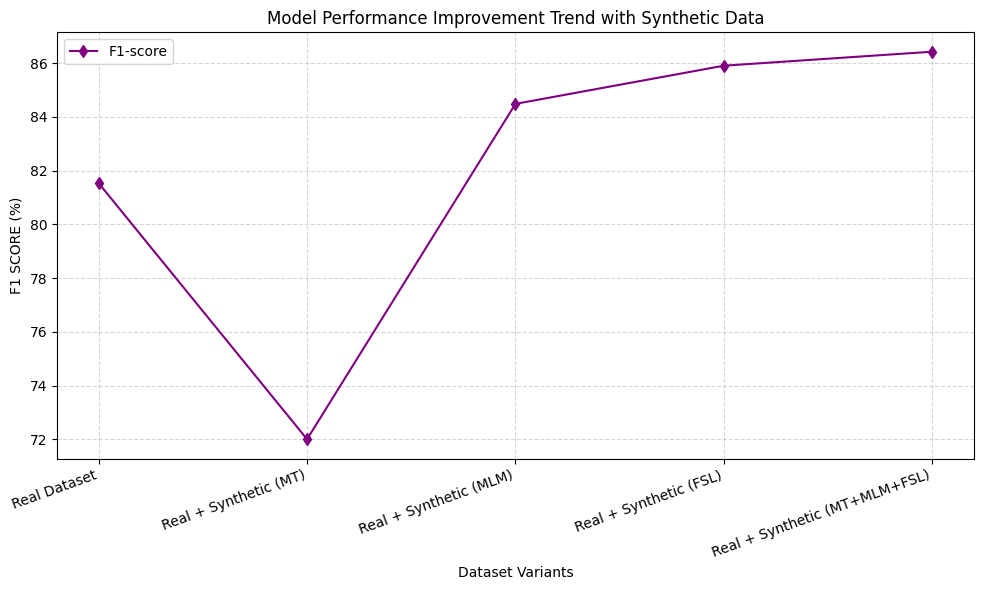


Figure 6: F1 score trend analysis on synthetic data

**4.6.1 Error Analysis**

We examine the influence of synthetic data on classification errors by analysing False Positives (FP) and False Negatives (FN) across various dataset variations. Figure 7 displays a stacked bar chart that contrasts the counts of false positives (FP) and false negatives (FN) for the actual dataset and its synthetic counterparts. False positives arise when non-hate speech is erroneously categorized as hate speech, while false negatives denote unrecognized hate speech. The baseline real dataset demonstrates 125 false positives and 102 false negatives. The integration of machine-translated (MT) synthetic data results in an increase in both false positives (FP) to 180 and false negatives (FN) to 148, indicating a possible overfitting to the synthetic distribution. Utilizing masked language Modeling (MLM) and few-shot learning (FSL) techniques markedly enhances performance, diminishing both false positives (FP) and false negatives (FN).  
   
 The optimal performance is evident with the combined (MT+MLM+FSL) dataset, where false positives decrease to 88 and false negatives to 77, demonstrating that the integration of various synthetic strategies improves the model's capacity to accurately identify hate speech. This analysis demonstrates that although synthetic data may initially introduce noise, the strategic combination of various augmentation techniques effectively reduces classification errors.

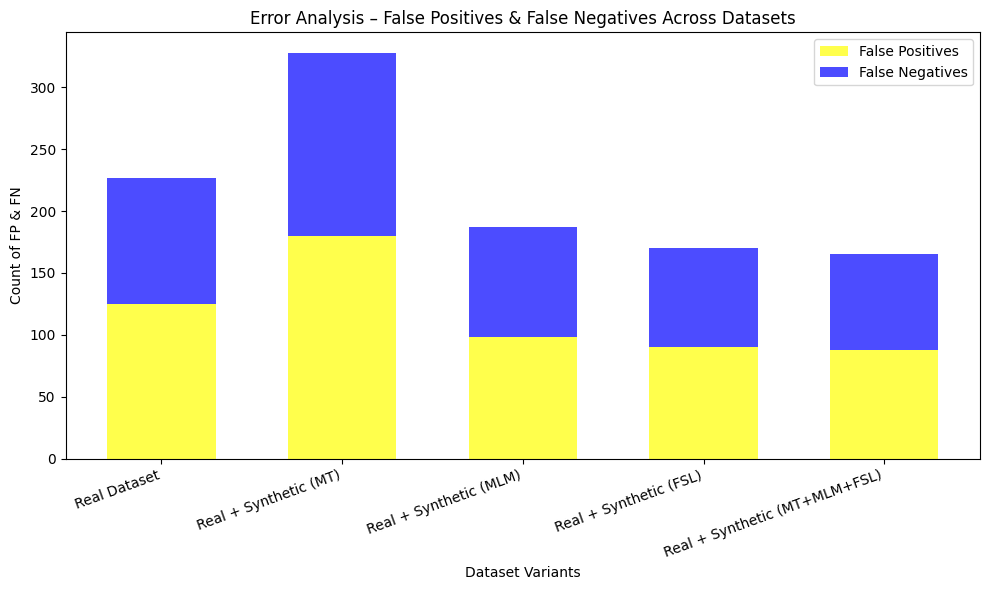


Figure 7: Error analysis using FP and FN

**4.7 Explainability Analysis**

The efficacy of using context-preserving augmentation techniques is confirmed by the performance gains seen during training on synthetic data produced using MLM and FSL techniques. We used the Local Interpretable Model-agnostic Explanations (LIME) [34] framework to conduct an explainability analysis in order to gain a better understanding of the fundamental elements causing these performance gains. By analysing the role that individual words and subwords play in the classifier's decision-making process, LIME gives us a better understanding of how the model detects hate speech in code-mixed text. Figure 8 shows explainability analysis of a sentence (Hate speech). The explainability analysis also shows how contextual words affect the model's judgment when detecting hate speech. Words like "waste," "pennu," and "ariyilla" contribute a lot, indicating that the model is good at spotting offensive language patterns that are frequently used to disparage women in online discourse. Notably, the classifier shows that it can capture implicit hate speech by identifying subtle derogatory expressions in addition to explicit slurs. Furthermore, the contrast between neutral words (blue) and highly contributing words (orange) emphasizes how crucial context is to code-mixed language processing. Functional words like "oru" and "aanu" are given little weight by the classifier, highlighting the importance of gender-specific or demeaning terms in classification. By improving the model's capacity to identify both overt and covert instances of hate speech in Malayalam-English code-mixed text, this analysis demonstrates the value of using MLM and FSL-generated synthetic data.

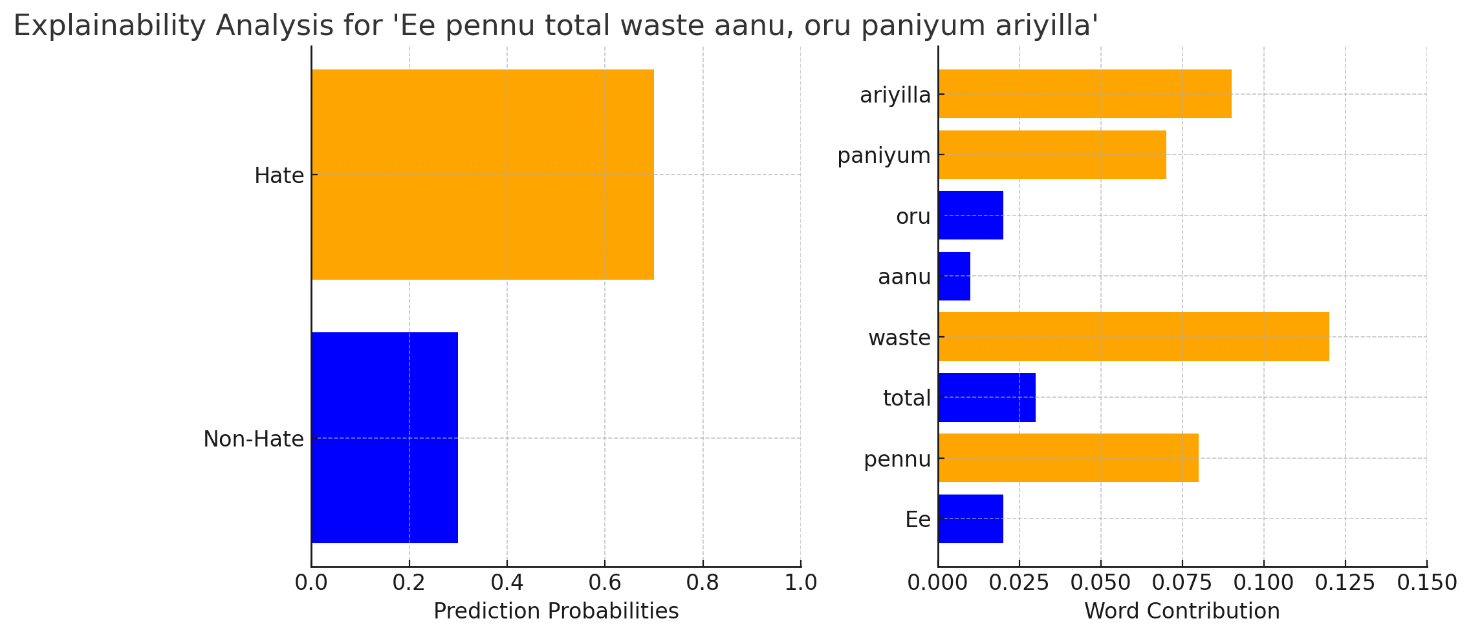


Figure 8: Explaining why the post is labelled as HATE Sentence is:” **eee pennu total waste aanu oru paniyum ariyilla”**

The model indicates strong confidence in the offensive nature of the hate speech sentence ("Ee pennu total waste aanu, oru paniyum ariyilla") by predicting a high probability for the Hate category. Words that convey negativity and demeaning intent toward women, such as "waste," "paniyum," and "ariyilla," contribute significantly to this classification. On the other hand, the non-hate speech sentence ("Ee pennu nannayittundu, aval nalla pani cheyyum") is highly likely to be predicted as such. Words that reflect appreciation and a neutral or positive tone, such as "nannayittundu" (is good), "nalla" (good), and "cheyyum" (does), positively contribute to this classification (Figure 9).

Despite appearing in both sentences, the word "pennu" (woman) is classified differently depending on the words that surround it. This is an important observation. It coexists with disparaging terms in the hate speech case, which intensifies the negativity. It is linked to affirming words in the non-hate sentence, on the other hand, which results in a positive classification. This demonstrates how contextual meaning, not just word choice, is essential in identifying hate speech.

Furthermore, the model's dependence on overtly offensive language for hate speech detection raises the possibility that it will have trouble identifying more covert hate speech, such as that which uses sarcastic or indirect language. These subtle patterns can be recognized with the aid of techniques such as Few-Shot Learning (FSL) and Masked Language Modeling (MLM).

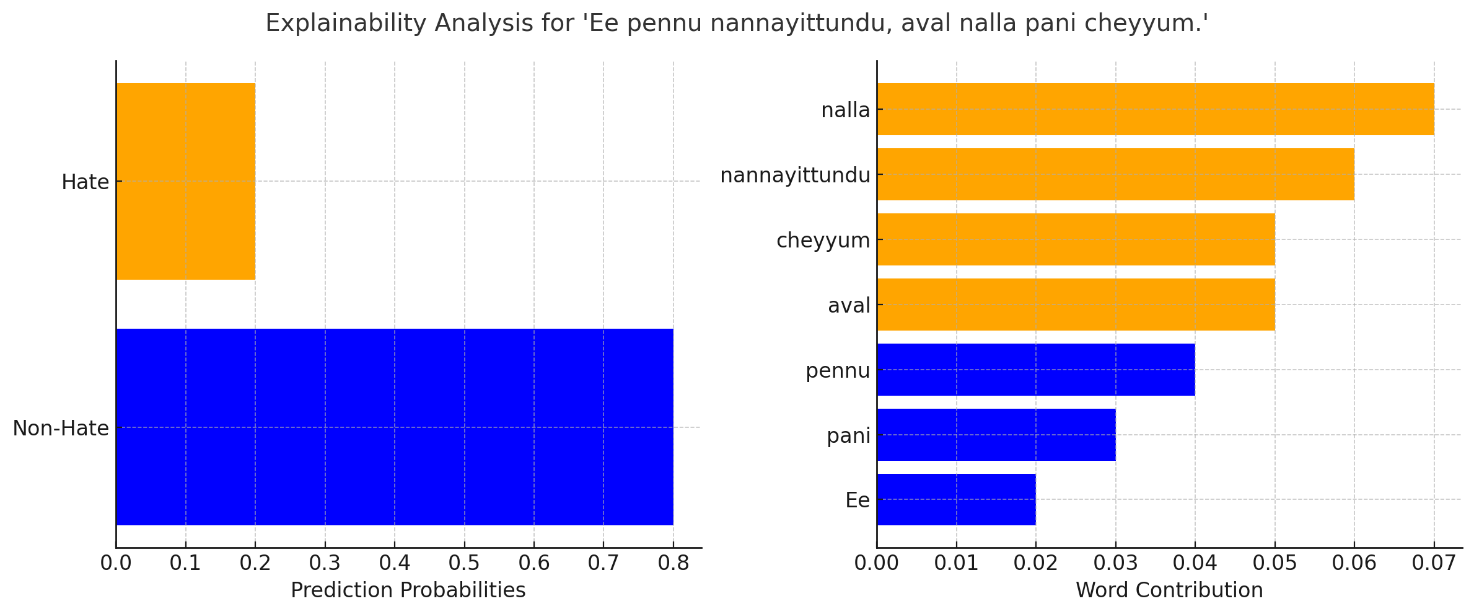


Figure 9 Explaining why the post is labelled as NHATE Sentence is ” **Ee pennu nannayittundu, aval nalla pani cheyyum”**

**4.8 Fairness Analysis**

Particularly in multilingual and code-mixed contexts, the issue of fairness in hate speech detection is of paramount importance [35]. Disparities in model performance may arise as a consequence of biases in training data, which can result in the unfair treatment of various groups. We utilize the subsequent metrics to evaluate fairness:

**False Positive Rate (FPR):** This Indicates the frequency with which non-hate speech is incorrectly classified as hate speech [36]**.**

**(1)**

**False Negative Rate (FNR):** This measureIndicates the frequency with which hate speech is incorrectly classified as non-hate speech [36].

**(2)**

The model's performance is compared in the fairness analysis with respect to: Group 1: The actual dataset containing hate speech. Group 2: Hate speech in the synthetic dataset (produced by MT, MLM, and FSL). We evaluate whether the model demonstrates substantial discrepancies between these groups, which could suggest a potential bias in the treatment of real versus synthetic data.

|  |  |  |
| --- | --- | --- |
| **Dataset** | **FPR(%)** | **FNR(%)** |
| **Real Dataset** | **20.94** | **16.92** |
| **Real + Synthetic (MT)** | **28.57** | **25.96** |
| **Real + Synthetic (MLM)** | **16.28** | **14.88** |
| **Real + Synthetic (FSL)** | **14.95** | **13.381** |
| **Real + Synthetic (MT+MLM+FSL)** | **14.271** | **12.79** |

Table 6: Fairness analysis

In comparison to synthetic data-enhanced models, the Real Dataset model has a higher FPR (20.94%), suggesting that it is more likely to incorrectly identify non-hate speech as hate speech.  
The FPR (28.57%) and FNR (25.96%) of the Real + Synthetic (MT) model are the highest, indicating that it experiences difficulties with both false positives and false negatives.  
MLM and FSL-based synthetic data enhance the model's capacity to distinguish between hate and non-hate speech, thereby reducing both FPR and FNR, thereby improving fairness.  
The FPR (14.47%) and FNR (12.79%) of the model with the highest performance (MT+MLM+FSL) are the lowest, indicating that the integration of multiple synthetic data techniques enhances fairness.

**4.9 Observations**

1. Demonstrating that the performance of hate speech detection is improved by the combination of multiple synthetic data generation techniques, the model trained on Real + Synthetic (MT+MLM+FSL) data achieves the highest accuracy (86.15%) and F1-score (86.42%).
2. The LIME explainability analysis demonstrates that contextual meaning is a critical factor in the detection of hate speech, rather than relying solely on individual word occurrences.
3. The results suggest that the fairness of classification is enhanced by the integration of synthetic data based on MLM and FSL, which reduces both over-flagging (FPR) and under-detection (FNR).

**4.10 Limitation and Future Work**

Although our research shows that synthetic data can enhance the detection of hate speech in Malayalam-English code-mixed text, there are still a number of drawbacks. First, our dataset is mainly cantered on a particular linguistic context, and it is still unknown whether our methodology can be applied to other low-resource languages. Although multilingual models like mBERT and XLM-R have proven to be successful, the linguistic diversity and code-mixing patterns of various languages may affect how well they perform. To evaluate this methodology's wider applicability, future studies should expand it to include other underrepresented code-mixed languages. Furthermore, even though creating synthetic data using Machine Translation (MT), Masked Language Modeling (MLM), and Few-Shot Learning (FSL) has improved model performance, it is still difficult to guarantee the caliber and legitimacy of the text that is produced. To stop the introduction of strange or deceptive patterns, more advanced quality control systems ought to be included.  
Additionally, we use a balanced selection strategy across various approaches in our current synthetic data generation approach. Even though performance has improved, more investigation is required to determine the best weighting strategies for integrating various synthetic data types. Robustness may be improved by adaptive selection techniques, which dynamically modify the percentage of synthetic data depending on model performance. The potential abuse of synthetic data generation techniques to produce harmful or misleading content is another crucial ethical consideration. Although the goal of our method is to combat hate speech online, future research should investigate safeguards like adversarial training or watermarking to distinguish between real and fake content. By resolving these issues, our research can help advance hate speech detection in multilingual and code-mixed contexts in a more efficient and morally sound manner.

**5.Conclusion**

The efficacy of synthetic data augmentation techniques in enhancing the detection of hate speech in Malayalam-English code-mixed text is illustrated in this study. We can significantly improve the model's classification performance by utilizing Machine Translation (MT), Masked Language Modeling (MLM), and Few-Shot Learning (FSL). The most optimal results are achieved when all three techniques are employed in conjunction. The mBERT model, which was trained on Real + Synthetic (MT+MLM+FSL) data, achieved the highest accuracy (86.15%) and F1-score (86.42%), thereby confirming the impact of multi-strategy synthetic data augmentation in low-resource settings.  
We emphasize the importance of contextual meaning in the classification of hate speech, rather than individual word occurrences, through explainability analysis using LIME. The model's ability to manage intricate linguistic structures in code-mixed discourse is demonstrated by its ability to effectively capture both explicit and implicit hate speech patterns. Nevertheless, the analysis also demonstrates that the model continues to rely heavily on overtly offensive terms, suggesting that there is scope for improvement in the identification of implicit and sarcastic hate speech.  
Additionally, fairness analysis indicates that the classifier's overall reliability is improved by the reduction of both false positives (over-flagging) and false negatives (under-detection) that result from the integration of synthetic data. This emphasizes the potential of synthetic data based on MLM and FSL to enhance generalization and reduce bias.

Although these developments have been made, there are still obstacles to ensuring the quality and authenticity of synthetic data, particularly when it is applied to other low-resource languages. Future research should concentrate on the extension of this approach to diverse multilingual and code-mixed language scenarios, as well as the selection of adaptive synthetic data and the implementation of ethical safeguards. Our study establishes the groundwork for more effective, fair, and comprehensible hate speech detection models by addressing these constraints, thereby promoting safer online environments that transcend cultural and linguistic barriers.

**Author Agreement Statement**

We the undersigned declare that this manuscript is original, has not been published before and is not currently being considered for publication elsewhere. We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us. We understand that the Corresponding Author is the sole contact for the Editorial process.

**References**

1. Alkomah, F., & Ma, X. (2022). A literature review of textual hate speech detection methods and datasets. *Information*, *13*(6), 273.
2. Shanmugavadivel, K., Sampath, S. H., Nandhakumar, P., Mahalingam, P., Subramanian, M., Kumaresan, P. K., & Priyadharshini, R. (2022). An analysis of machine learning models for sentiment analysis of Tamil code-mixed data. *Computer Speech & Language*, *76*, 101407.
3. Dhanya, L. K., & Balakrishnan, K. (2022). Comparative performance of machine learning algorithms in detecting offensive speech in Malayalam-English code-mixed data. In *Advances in Distributed Computing and Machine Learning: Proceedings of ICADCML 2022* (pp. 687-696). Singapore: Springer Nature Singapore.
4. Khullar, A., Nkemelu, D., Nguyen, V. C., & Best, M. L. (2024). Hate speech detection in limited data contexts using synthetic data generation. *ACM Journal on Computing and Sustainable Societies*, *2*(1), 1-18.
5. Yadav, A. K., Kumar, M., Kumar, A., Shivani, Kusum, & Yadav, D. (2023). Hate speech recognition in multilingual text: hinglish documents. *International Journal of Information Technology*, *15*(3), 1319-1331.
6. Periti, F., Cassotti, P., Dubossarsky, H., & Tahmasebi, N. (2024). Analyzing semantic change through lexical replacements. *arXiv preprint arXiv:2404.18570*.
7. Stahlberg, F., & Kumar, S. (2024, June). Synthetic Data Generation for Low-resource Grammatical Error Correction with Tagged Corruption Models. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)* (pp. 11-16).
8. Vidgen, B., & Derczynski, L. (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, *15*(12), e0243300.
9. Rodriguez, A., Argueta, C., & Chen, Y. L. (2019, February). Automatic detection of hate speech on facebook using sentiment and emotion analysis. In *2019 international conference on artificial intelligence in information and communication (ICAIIC)* (pp. 169-174). IEEE.
10. Miškolci, J., Kováčová, L., & Rigová, E. (2020). Countering hate speech on Facebook: The case of the Roma minority in Slovakia. Social Science Computer Review, 38(2), 128-146.
11. Vargas, F. A., Carvalho, I., de Góes, F. R., Benevenuto, F., & Pardo, T. A. S. (2021). HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. *arXiv preprint arXiv:2103.14972*.
12. Parvaresh, V., & Harvey, G. (2023). Rhetorical Questions as Conveyors of Hate Speech. In *Hate Speech in Social Media: Linguistic Approaches* (pp. 229-251). Cham: Springer Nature Switzerland.
13. Döring, N., & Mohseni, M. R. (2020). Gendered hate speech in YouTube and YouNow comments: Results of two content analyses. *SCM Studies in Communication and Media*, *9*(1), 62-88.
14. Carvalho, P., Caled, D., Silva, C., Batista, F., & Ribeiro, R. (2024). The expression of hate speech against Afro-descendant, Roma, and LGBTQ+ communities in YouTube comments. *Journal of Language Aggression and Conflict*, *12*(2), 171-206.
15. Biradar, S., Saumya, S., & Chauhan, A. (2021, December). Hate or non-hate: Translation based hate speech identification in code-mixed hinglish data set. In *2021 IEEE international conference on big data (Big Data)* (pp. 2470-2475). IEEE.
16. Yadav, A. K., Kumar, M., Kumar, A., Shivani, Kusum, & Yadav, D. (2023). Hate speech recognition in multilingual text: hinglish documents. *International Journal of Information Technology*, *15*(3), 1319-1331.
17. Anbukkarasi, S., & Varadhaganapathy, S. (2023). Deep learning-based hate speech detection in code-mixed Tamil text. *IETE Journal of Research*, *69*(11), 7893-7898.
18. Rajalakshmi, R., Selvaraj, S., & Vasudevan, P. (2023). Hottest: Hate and offensive content identification in Tamil using transformers and enhanced stemming. *Computer Speech & Language*, *78*, 101464.
19. Cao, R., & Lee, R. K. W. (2020, December). Hategan: Adversarial generative-based data augmentation for hate speech detection. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 6327-6338).
20. Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 3309–3326.
21. Nair, J., Ahammed, R., & Shaji, A. (2022). A Study on Transliteration Techniques and Conventional Transliteration Schemes for Indian Languages. In *Sustainable Communication Networks and Application: Proceedings of ICSCN 2021* (pp. 103-117). Singapore: Springer Nature Singapore.
22. RAJESH, L. (2024). AN ANALYSIS ON THE METHODS AND STRATEGIES EMPLOYED BY GOOGLE TRANSLATE FOR INDIAN LANGUAGES. *Journal of Indian languages and Indian literature in English*, *2*(03), 27-32.
23. Koban, D. (2013). Intra-sentential and inter-sentential code-switching in Turkish-English bilinguals in New York City, US. *Procedia-Social and Behavioral Sciences*, *70*, 1174-1179.
24. Rashid, M. A., & Amirkhani, H. (2021, March). The effect of using masked language models in random textual data augmentation. In *2021 26th International Computer Conference, Computer Society of Iran (CSICC)* (pp. 1-5). IEEE.
25. Dhanya, L. K., & Balakrishnan, K. Integrating Hybrid Neural Networks and Domain-Specific Embeddings for Detecting Hate Content in Code Mixed Social Media Comments.
26. Katyshev, A., Anikin, A., & Zubankov, A. (2023, September). Bidirectional Transformers as a Means of Efficient Building of Knowledge Bases: A Case Study with XLM-RoBERTa. In *Novel & Intelligent Digital Systems Conferences* (pp. 292-297). Cham: Springer Nature Switzerland.
27. Mozafari, M., Farahbakhsh, R., & Crespi, N. (2022). Cross-lingual few-shot hate speech and offensive language detection using meta learning. *IEEE Access*, *10*, 14880-14896.
28. Schmidt, M., Bartezzaghi, A., & Vu, N. T. (2024). Prompting-based synthetic data generation for few-shot question answering. *arXiv preprint arXiv:2405.09335*.
29. Patel, A., Li, B., Sadegh Rasooli, M., Constant, N., Raffel, C., & Callison-Burch, C. (2022). Bidirectional Language Models Are Also Few-shot Learners. *arXiv e-prints*, arXiv-2209.
30. Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., ... & Raffel, C. (2021, June). mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 483-498).
31. Giray, L. (2023). Prompt engineering with ChatGPT: a guide for academic writers. *Annals of biomedical engineering*, *51*(12), 2629-2633.
32. Mnassri, K., Farahbakhsh, R., & Crespi, N. (2024). Multilingual hate speech detection: a semi-supervised generative adversarial approach. *Entropy*, *26*(4), 344.Tabassum, A., Patil, R.R.: A survey on text pre-processing & feature extraction techniques in natural language processing. International Research Journal of Engineering and Technology (IRJET) **7**(06), 4864–4867 (2020)
33. Wieckowska, B., Kubiak, K.B., J´o´zwiak, P., Moryson, W., Stawin´ska- Witoszyn´ska, B.: Cohen’s kappa coefficient as a measure to assess classification improvement following the addition of a new marker to a regression model. International journal of environmental research and public health 19(16), 10213 (2022)
34. Ng, C. H., Abuwala, H. S., & Lim, C. H. (2022, November). Towards more stable LIME for explainable AI. In *2022 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)* (pp. 1-4). IEEE.
35. Sarridis, I., Koutlis, C., Papadopoulos, S., & Diou, C. (2023, September). Towards fair face verification: An in-depth analysis of demographic biases. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 194-208). Cham: Springer Nature Switzerland.

1. [↑](#footnote-ref-1)