Enhancing Hate Speech Detection in Low-Resource Code-Mixed Indonesian Tweets via GPT-Based Data Augmentation

Endang Wahyu Pamungkas
Department of Informatics Engineering, Universitas Muhammadiyah Surakarta, Surakarta, Indonesia | Social Informatics Research Center, Universitas Muhammadiyah Surakarta, Surakarta, Indonesia
endang.wahyu@ums.ac.id (corresponding author)

Dian Purworini
Department of Communication Science, Universitas Muhammadiyah Surakarta, Surakarta, Indonesia | Social Informatics Research Center, Universitas Muhammadiyah Surakarta, Surakarta, Indonesia
dian.purworini@ums.ac.id

Widi Widayat
Department of Informatics Engineering, Universitas Muhammadiyah Surakarta, Surakarta, Indonesia | Social Informatics Research Center, Universitas Muhammadiyah Surakarta, Surakarta, Indonesia
widi.widayat@ums.ac.id

Divi Galih Prasetyo Putri
Department of Electrical Engineering and Informatics, Vocational College, Universitas Gadjah Mada, Yogyakarta, Indonesia
divi.galih@ugm.ac.id

Ikhlasul Amal
Department of Artificial Intelligence, Universitas Gadjah Mada, Yogyakarta, Indonesia
ikhlasulamal@mail.ugm.ac.id

Received: 27 August 2025 | Revised: 29 September 2025 | Accepted: 6 October 2025

 DOI: https://doi.org/10.48084/etasr.14342

ABSTRACT

Automatic hate speech detection in low-resource, code-mixed languages, such as Indonesian social media environments, presents significant challenges due to the scarcity of annotated data and the linguistic variability introduced by code-mixing. However, due to the growing prevalence of hate speech on social media, there is a need for robust hate speech detection systems. This study investigates the effectiveness of data augmentation strategies, specifically Generative Pretrained Transformer (GPT)-based paraphrasing and aggressive text transformation, in enhancing the performance of hate speech detection models for Indonesian code-mixed tweets. To achieve that, we employed traditional machine learning models, Recurrent Neural Network (RNN)-based models, and transformer-based models to assess the impact of these augmentation strategies. Our findings reveal that GPT-generated data improve model performance, with transformer-based models, including Indonesian Bidirectional Encoder Representations from Transformers (IndoBERT) and the Cross-lingual Language Model Robustly Optimized BERT Pretraining approach (XLM-RoBERTa).

Keywords-hate speech detection; low-resource language; code-mixed language; data augmentation; large language model

I. INTRODUCTION

The development of social media platforms has revolutionized communication, enabling people to connect across borders and share their views and opinions easily [1]. Social media platforms have grown significantly, with X reporting 353.9 million active users and Facebook hosting 3 billion users who generate vast amounts of content daily [2]. Nevertheless, these platforms often facilitate the spread of hate speech, contributing to cyber-hate incidents and real-world violence [3]. Online hateful content can significantly harm society by fostering division, targeting vulnerable groups, and even disrupting technological developments, such as chatbots that learn from user interactions [4]. Moreover, detecting and mitigating hate speech is further complicated by the massive volume of online content, the subjective and context-dependent nature of hateful expressions, and the multilingual character of social media communication [5]. As a result, automated detection systems employing Natural Language Processing (NLP) and machine learning have become indispensable, although the evolving and context-dependent nature of hate speech makes the incorporation of such systems complex [6, 7].

In Indonesia, the proliferation of hate speech on social media is a growing concern [8-10]. According to the Criminal Investigation Agency of the Indonesian National Police, cybercrime cases related to hate speech increased from 143 in 2015 to 199 in 2016, reflecting a persistent and escalating issue. Indonesia's complex socio-cultural landscape, characterized by ethnic, religious, and linguistic diversity within a Muslim-majority population, further heightens tensions, with hate speech often directed at minority groups and contributing to real-world conflicts [11]. This complexity is compounded by Indonesia's rich linguistic diversity. Low-resource, code-mixed forms such as Bahasa mixed with Javanese or Sundanese introduce linguistic structures that differ significantly from the monolingual datasets typically used to develop NLP models, making automated hate speech detection even more challenging [12].

Prior research on Indonesian hate speech detection has made notable progress but still faces several limitations. Most studies have relied on monolingual datasets from platforms such as X [13], Facebook [14], and Instagram [15], applying traditional machine learning models including Naïve Bayes, Support Vector Machines (SVM), and Random Forests [16, 17], as well as deep learning models such as Convolutional Neural Networks (CNN) [18] and Long Short-Term Memory (LSTM) networks [19, 20]. Although these approaches yield competitive performance, they do not adequately address the code-mixed nature of real-world Indonesian social media, while slang, informal expressions, and linguistic variation add further complexity [8].

Moreover, research explicitly targeting code-mixed Indonesian data remains limited. For instance, authors in [21] investigated Javanese-Indonesian and Sundanese-Indonesian datasets using traditional and neural models, while another study found that multilingual models do not consistently outperform monolingual Indonesian ones in such settings [22].

Similar challenges have been reported in other multilingual environments, such as Hindi-English [23], Tamil-English [24], and Swahili-English [25], where language switching complicates syntactic and semantic representation. Additionally, the application of transformer-based models for Indonesian hate speech detection remains underexplored, despite their proven effectiveness in other NLP tasks. Furthermore, a persistent barrier is the scarcity of annotated datasets, as data collection and annotation are costly, time-consuming, and labor-intensive [26], especially in low-resource linguistic settings.

Data augmentation is one strategy to mitigate this issue, but traditional augmentation techniques often rely on heuristic or rule-based transformations that often fall short in terms of applicability and effectiveness [27, 28]. However, recent studies have demonstrated that Generative Pretrained Transformer (GPT) models can generate high-quality synthetic text for various NLP tasks. For instance, authors in [29] successfully leveraged GPT-based text augmentation in sentiment analysis, while other studies have shown that GPT-generated data enhances model performance in fake news detection [30] and can provide annotations comparable in quality to human-annotated ones. Despite these advances, the use of GPT-based augmentation for code-mixed Indonesian hate speech detection remains largely unexplored.

To address these limitations, this study leveraged GPT-based data augmentation to enhance hate speech detection in low-resource, code-mixed Indonesian languages, specifically Javanese–Indonesian (JV-ID) and Sundanese–Indonesian (SN-ID). We utilized GPT's generative capabilities to create synthetic training data using two augmentation strategies: paraphrasing, which produces semantically equivalent variations, and aggressive transformation, which generates more diverse textual forms while preserving the original level of abusiveness.

We evaluated the impact of these augmentation strategies across three model categories: traditional machine learning models, including Linear SVM, Decision Tree, Logistic Regression, and Random Forest; Recurrent Neural Network (RNN)-based models, including LSTM, Gated Recurrent Unit (GRU); and transformer-based models, including Cross-lingual Language Model (XLM), Cross-lingual Language Model Robustly Optimized BERT Pretraining approach (XLM-RoBERTa), Multilingual Bidirectional Encoder Representations from Transformers (BERT), Multilingual DistilBERT, and Indonesian BERT (IndoBERT). All models' performance was evaluated before and after the data augmentations to assess the impact of synthetic data. By combining GPT-based augmentation with state-of-the-art models, we aim to overcome the dual challenges of data scarcity and code-mixing complexity in Indonesian hate speech detection.

II. METHOD

A. Data Augmentation Approach

The dataset used in this study was sourced from [21], comprising tweets labeled as either hate speech or non-hate speech. To expand the dataset for hate speech detection in

code-mixed JV-ID and SN-ID languages, we applied a GPT-based data augmentation strategy consisting of two techniques: paraphrasing and aggressive transformation. Specifically, we employed OpenAI's GPT-4o model accessed through its Application Programming Interface (API) using default generation settings and using specialized prompts for each language:

```
Paraphrasing Prompt:
You are an Indonesian linguist. Help me to
augment the following text using
paraphrasing techniques into 5 new data.
The augmented data will be used as
training data for developing machine
learning models to detect hate speech in
Indonesian social media. Please produce
the augmented data by following the
criteria below:
(a) The input sentence is code-mixed data
[Javanese-Indonesian or Sundanese-
Indonesian], so make sure the output
sentence is also code-mixed data as the
original language.
(b) Maintain the abusive context of the
text, including abusive words that exist.
(c) You will be given information
regarding the label of the text, whether
it is hate speech or not. Make sure the
output data will have the same label as
the original data.
(d) The format of the output must be as
follows: 1. augmented text 2. augmented
text. 3. augmented text. 4. augmented
text. 5. augmented text. + text

Aggressive Transformation Prompt:
You are an Indonesian linguist. Help me to
augment the following text into 3 new
data. The augmented data will be used as
training data for developing machine
learning models to detect hate speech in
Indonesian social media. Please produce
the augmented data by following the
criteria below:
(a) The augmented sentence is a sentence
with a new theme but is still inspired by
the original sentence.
(b) The input sentence is code-mixed data
[Javanese-Indonesian or Sundanese-
Indonesian], so make sure the output
sentence is also code-mixed data as the
original language.
(c) Maintain the abusive context of the
text, including abusive words that exist.
(d) You will be given information
regarding the label of the text, whether
it is hate speech or not. Make sure the
output data will have the same label as
the original data.
(e) The format of the output must be as
follows: 1. augmented text 2. augmented
text. 3. augmented text. + text
```

1) Paraphrasing Strategy

The paraphrasing strategy aims to generate alternative expressions of the same content while preserving the original meaning. This method introduces minor variations in sentence structure and wording, ensuring contextual consistency with the original dataset. Although paraphrasing maintains high semantic fidelity, it may provide limited linguistic diversity, potentially limiting the model's ability to generalize from the training data.

2) Aggressive Transformation Strategy

The aggressive transformation strategy generates content that differs significantly from the original tweet while preserving the level of abusiveness. This method broadens thematic variation and linguistic diversity, potentially improving the model's generalization capabilities. However, there is a risk that as the generated content becomes more distinct from the original, the resulting text might fail to accurately reflect the intended abusiveness, potentially introducing inconsistencies into the dataset.

B. Data Augmentation Workflow

Figure 1 illustrates the workflow of the data augmentation process utilized in this study for expanding the training dataset.

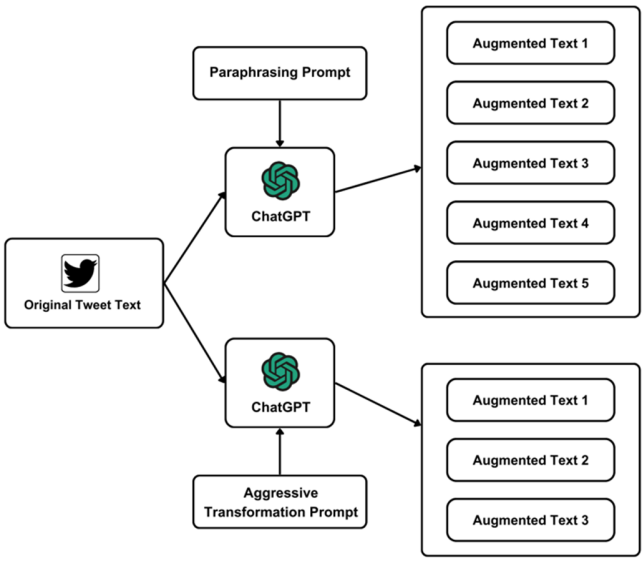

Fig. 1. Workflow of data augmentation strategy.

The process begins with an original tweet, which is independently processed using both augmentation strategies. The paraphrasing strategy generates five paraphrased versions of the original tweet, while the aggressive transformation strategy generates three transformed versions. Outputs from

both strategies are then integrated into the augmented training dataset.

C. Dataset Statistics Before and After Augmentation

Table I presents the dataset statistics before and after applying the two augmentation strategies. The original dataset [21] consisted of 890 JV-ID hate speech tweets and 2,635 non-hate speech tweets, and 504 SN-ID hate speech tweets and 1,557 non-hate speech tweets. After paraphrasing, the JV-ID dataset expanded to 21,150 samples (15,810 non-hate speech, 5,340 hate speech), and the SN-ID dataset to 12,366 samples (9,332 non-hate speech, 3,024 hate speech). On the other hand, after the aggressive transformation strategy, the JV-ID dataset expanded to 14,100 samples (10,540 non-hate speech, 3,560 hate speech), and the SN-ID dataset to 8,244 samples (6,228 non-hate speech, 2,016 hate speech).

TABLE I. TRAINING DATA STATISTICS BEFORE AND AFTER THE AUGMENTATION PROCESS

Language	Label	Original Data	Paraphrasing	Agg. Transf.
JV-ID	Hate Speech	890	5,340	3,560
	Non-hate Speech	2,635	15,810	10,540
SN-ID	Hate Speech	504	3,024	2,016
	Non-hate Speech	1,557	9,332	6,228

The dataset was partitioned into training and testing sets following the original split. Testing sets included 1,512 JV-ID samples and 725 SN-ID samples.

D. Models and Training Configuration

We employed three variants of machine learning models: traditional machine learning models, RNN-based models, and transformer-based models.

1) Traditional Machine Learning Models

Four traditional classifiers were implemented, including Linear SVM, Decision Tree, Logistic Regression, and Random Forest. These models used a Bag-of-Words (BoW) representation and were implemented in scikit-learn 1.6.1 with default parameters to ensure fair comparison and minimize confounding variables. Model configurations included:

- Linear SVM: penalty = l2, C = 1.0, max_iter = 1000.
- Logistic Regression: penalty = l2, solver = lbfgs, max_iter = 100.
- Decision Tree: criterion = gini, max_depth = None.
- Random Forest: n_estimators = 100, criterion = gini, max_depth = None

2) RNN-Based Models

Two RNN architectures were employed, including LSTM and Gated Recurrent Unit (GRU). Our model architecture begins with an embedding layer consisting of 128 dimensions, which is essential for learning representations from the input data since pre-trained models specific to JV-ID and SN-ID are unavailable. The embeddings generated from this layer were fed into either the LSTM or GRU layer, each comprising 64 units. The LSTM or GRU layer is followed by a dense layer with 16 units, utilizing the Rectified Linear Unit (ReLU) activation function to ensure non-linearity. The final layer is a dense layer with a sigmoid activation function, which outputs the probability of the input text being hate speech. To refine the model's performance, we experimented with various batch sizes (16, 32, 64) and epochs ranging from 1 to 5.

3) Transformer-Based Models

Additionally, five multilingual transformer-based models were employed, including XLM, XLM-RoBERTa, Multilingual BERT, Multilingual DistilBERT, and IndoBERT. Minimal hyperparameter tuning was applied to ensure comparability across configurations. All transformer models used the default configuration of available models in HuggingFace and employed AdamW optimizers with a learning rate of 1e-5. Also, an Early Stopping method was implemented to optimize the number of epochs and avoid overfitting.

4) Evaluation Strategy

Each model was trained on both the original and augmented training datasets to evaluate the impact of the additional augmented data. As the evaluation metric, macro F1-score was used as it assesses the performance of the model regardless of class imbalance.

The experiments were conducted using Google Colab, which provides an environment equipped with Python 3.10, a NVIDIA A100 Graphics Processing Unit (GPU), and TensorFlow 2.16 as the primary deep learning framework.

III. RESULTS AND DISCUSSION

Figure 2 and Figure 3 illustrate the macro F1-scores achieved by all the models employed across the JV-ID and SN-ID original and augmented datasets, while Table II summarizes the percentage macro F1-score improvement compared to the original dataset.

For the original dataset, IndoBERT obtained the best macro F1-score for the JV-ID dataset (0.743), while for the SN-ID dataset, it achieved 0.687. Traditional machine learning models such as Logistic Regression and Linear SVM also demonstrated competitive performance, with Logistic Regression achieving macro F1-scores of 0.717 (JV-ID) and 0.706 (SN-ID), while Linear SVM of 0.728 (JV-ID) and 0.705 (SN-ID).

Augmentation using the paraphrasing strategy improved performance across the models, with the only exceptions being the Linear SVM (-1.8%) and Decision Tree (-2.0%) in the JV-ID dataset. For the transformer-based models, IndoBERT achieved an increased macro F1-score of 0.754 (+1.5%) for JV-ID and 0.705 (+2.6%) for SN-ID, while XLM-RoBERTa and Multilingual BERT also benefited, with XLM-RoBERTa reaching 0.760 (+2.3%) for JV-ID and Multilingual BERT reaching 0.734 (+1.4%) for SN-ID. Additionally, the XLM achieved the highest improvement in both datasets of 27.9% (JV-ID) and 15% (SN-ID).

Moreover, the aggressive transformation strategy also led to performance enhancements, except for the traditional machine learning models in the JV-ID dataset, which achieved between

-2.9% and -4.4% lower macro F1-score values, and the LSTM and GRU models in the SN-ID dataset, which achieved -1.1% and -1.5% lower macro F1-score values, respectively. However, IndoBERT achieved a macro F1-score of 0.768 (+3.4%) for JV-ID and 0.738 (+7.4%) for SN-ID. XLM-RoBERTa also performed strongly, reaching 0.764 (+2.3%) for JV-ID and 0.755 (+2.4%) for SN-ID. The highest improvement was again seen on the XLM model, achieving improvements of 27.0% and 23.1% in the JV-ID and SN-ID, respectively.

Comparing the two augmentation strategies, aggressive transformation generally outperformed paraphrasing; however, aggressive transformation may introduce noise if generated samples diverge too far from the original context, which can negatively affect simpler models. Additionally, among the models, IndoBERT and XLM-RoBERTa consistently emerged as the most suitable models for hate speech detection in low-resource, code-mixed Indonesian languages, demonstrating robust performance across different augmentation strategies. The superior performance of IndoBERT and XLM-RoBERTa can be attributed to their ability to understand and process complex language structures, which is essential for handling code-mixed data. IndoBERT, being specifically trained on Indonesian language data, has a nuanced understanding of the linguistic patterns and nuances in the JV-ID code-mixed dataset, while XLM-RoBERTa, with its robust multilingual training, effectively captured the diverse expressions in the SN-ID dataset, making it well-suited for this task.

Overall, in most cases, the augmentation strategies positively impacted the performance of the models, highlighting the effectiveness of using GPT for data augmentation in low-resource settings. This suggests that more substantial changes in the training data can help models learn to generalize better across different contexts and expressions of hate speech. Nevertheless, some models showed slightly worse performance with certain augmentation strategies, suggesting that not all models benefit equally from synthetic data augmentation. For the traditional machine learning models, this worse performance with augmented data suggests that simpler models may struggle to generalize from highly diverse synthetic data, potentially due to overfitting to augmented patterns that differ from the test distribution.

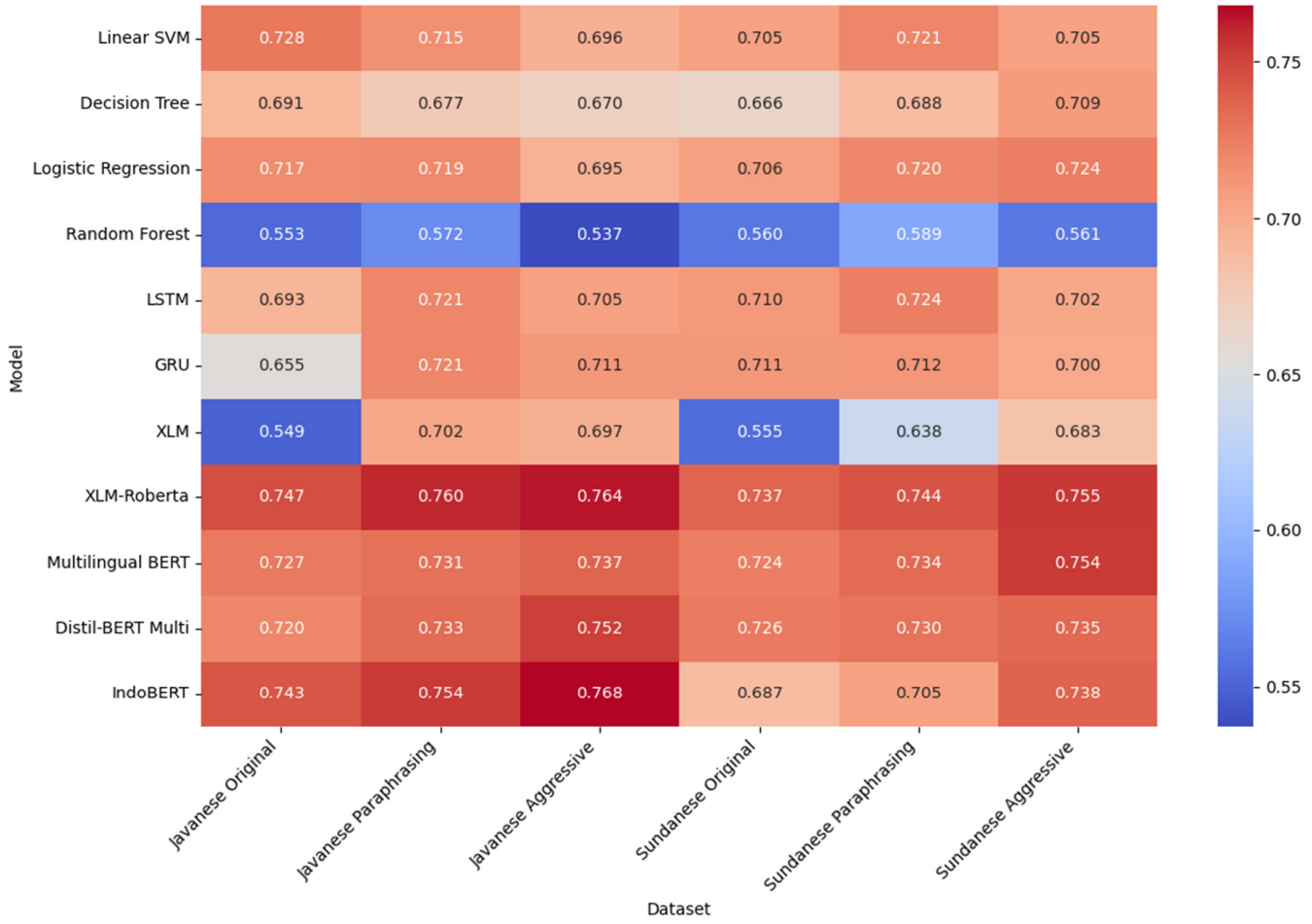

Fig. 2. Comparison of model performance across datasets and language settings.

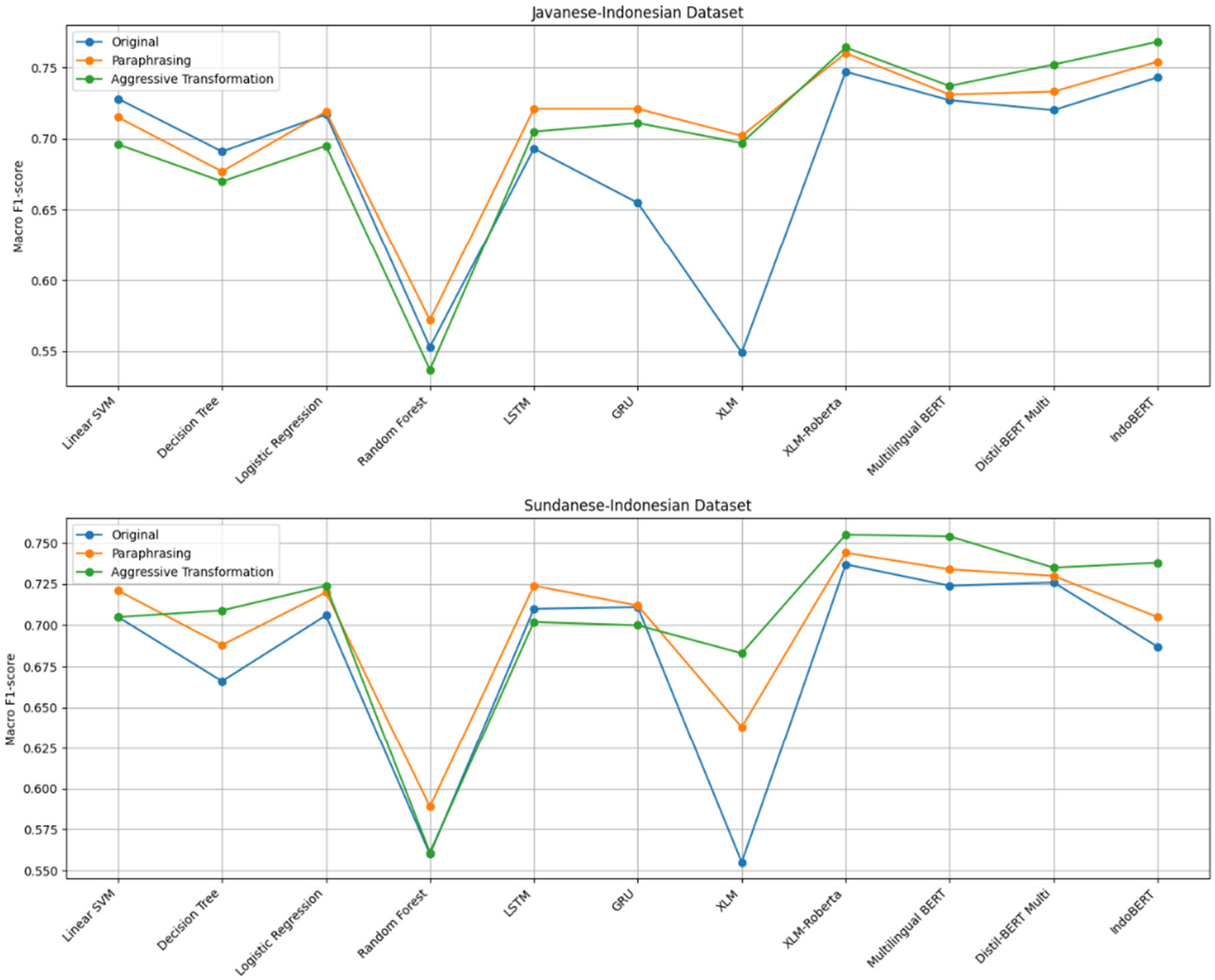

Fig. 3. Comparison of models' performance in three different datasets.

TABLE II. PERCENTAGE IMPROVEMENT IN MACRO F1-SCORE COMPARING ORIGINAL DATASET VERSUS AUGMENTED DATASETS

Model	JV-ID Paraphrasing (%)	JV-ID Agg. Transf. (%)	SN-ID Paraphrasing (%)	SN-ID Agg. Transf. (%)
Linear SVM	-1.8	-4.4	+2.3	0.0
Decision Tree	-2.0	-3.0	+3.3	+6.5
Logistic Regression	+0.3	-3.1	+2.0	+2.5
Random Forest	+3.4	-2.9	+5.2	+0.2
LSTM	+4.0	+1.7	+2.0	-1.1
GRU	+10.1	+8.5	+0.1	-1.5
XLM	+27.9	+27.0	+15.0	+23.1
XLM-RoBERTa	+1.7	+2.3	+0.9	+2.4
Multilingual BERT	+0.6	+1.4	+1.4	+4.1
DistilBERT Multi	+1.8	+4.4	+0.6	+1.2
IndoBERT	+1.5	+3.4	+2.6	+7.4

A. Comparison with Previous Studies

To further contextualize our findings, we compare them with similar studies that also employed GPT-based augmentation, as presented in Table III. For instance, ValizadehAslani et al. [31] demonstrated +1.5% improvement on the SST-2 sentiment analysis dataset by applying a two-stage fine-tuning strategy with GPT-augmented data, where the F1-score increased from 0.8812 to 0.8957 using a BERT baseline. Similarly, Woźniak and Kocoń [29] reported a +5% gain on the PerSenT dataset, with F1-score rising from 0.38 to 0.43 when using RoBERTa-base, and a +1% gain on the MultiEmo dataset, where RoBERTa-base improved from 0.87 to 0.88. Our best-performing models show comparable improvements in the domain of hate speech detection for code-mixed Indonesian datasets, with IndoBERT achieving +3.4% improvement on Javanese-Indonesian using aggressive transformation, while XLM-RoBERTa achieved +2.4% improvement on Sundanese-Indonesian also using aggressive transformation. These findings reinforce the generalizability of GPT-based augmentation, showing its potential to enhance model performance across both monolingual and code-mixed natural language processing tasks.

TABLE III. COMPARISON OF GPT-BASED AUGMENTATION IMPROVEMENTS ACROSS DIFFERENT STUDIES

Task/Dataset	Baseline Score	GPT-Augmented Score	Improvement
Sentiment Analysis (SST-2) [31]	0.8812	0.8957	+1.5%
Sentiment Analysis (PerSenT) [29]	0.38	0.43	+5%
Sentiment Analysis (MultiEmo) [29]	0.87	0.88	+1%
Hate Speech Detection (JV-ID) (This work)	0.743	0.768	+2.5%
Hate Speech Detection (SN-ID) (This work)	0.737	0.755	+1.8%

IV. CONCLUSION

In this study, we explored the effectiveness of data augmentation strategies for improving the performance of hate speech detection models in low-resource, code-mixed languages, specifically Javanese-Indonesian and Sundanese-Indonesian. We employed both paraphrasing and aggressive transformation strategies using GPT to augment the training datasets and evaluated the impact on various machine learning models. Our findings demonstrate that GPT is an effective tool for augmenting data in low-resource settings. By generating meaningful and contextually relevant paraphrases and transformations, we successfully increased the diversity and volume of the training datasets. The experimental results showed significant improvements in macro F1-scores across various models when trained on augmented datasets. The paraphrasing strategy improved the performance of 9 out of 11 models for the Javanese-Indonesian dataset and all 11 models for the Sundanese-Indonesian dataset. The aggressive transformation strategy further enhanced performance, with 8 out of 11 models improving for the Javanese-Indonesian dataset and 10 out of 11 models for the Sundanese-Indonesian dataset. Transformer-based models, particularly IndoBERT and XLM-RoBERTa, consistently outperformed other models in both language settings. IndoBERT achieved the highest macro F1-score of 0.768 for the Javanese-Indonesian dataset, while XLM-RoBERTa attained the highest macro F1-score of 0.755 for the Sundanese-Indonesian dataset when using aggressively transformed data.

This study underscores the importance of data augmentation in enhancing the performance of hate speech detection models, particularly in low-resource, code-mixed language settings. Future work could focus on several key areas. Enhancing augmentation techniques, including context-aware transformations and the incorporation of additional linguistic features, could provide more diverse and representative training data. Integrating multimodal data, such as images and videos, along with text, could improve the detection of hate speech on social media platforms where users often communicate using multiple modalities. Leveraging cross-lingual transfer learning approaches to adapt models trained on high-resource languages to low-resource languages could further enhance performance and generalizability. These advancements will contribute to the development of more robust and scalable hate speech detection systems.

ACKNOWLEDGMENT

This work has been funded by the Indonesian Ministry of Research and Higher Education under Grant Number 108/E5/PG.02.00.PL/2024 with title "Pendeteksian Ujaran Kebencian untuk Bahasa Code-Mixed pada Media Sosial Berbahasa Indonesia".

Augmented dataset and associated resources are available upon reasonable request from the authors.

REFERENCES

[1] A. Hande, K. Puranik, R. Priyadharshini, S. Thavareesan, and B. R. Chakravarthi, "Evaluating Pretrained Transformer-based Models for COVID-19 Fake News Detection," in *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India, Apr. 2021, pp. 766–772, https://doi.org/10.1109/ICCMC51019.2021.9418446.

[2] M. S. Jahan and M. Oussalah, "A systematic review of hate speech automatic detection using natural language processing," *Neurocomputing*, vol. 546, Aug. 2023, Art. no. 126232, https://doi.org/10.1016/j.neucom.2023.126232.

[3] N. S. Mullah and W. M. N. W. Zainon, "Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review," *IEEE Access*, vol. 9, pp. 88364–88376, 2021, https://doi.org/10.1109/ACCESS.2021.3089515.

[4] G. Kovács, P. Alonso, and R. Saini, "Challenges of Hate Speech Detection in Social Media: Data Scarcity, and Leveraging External Resources," *SN Computer Science*, vol. 2, no. 2, Apr. 2021, Art. no. 95, https://doi.org/10.1007/s42979-021-00457-3.

[5] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, and V. Patti, "Resources and benchmark corpora for hate speech detection: a systematic review," *Language Resources and Evaluation*, vol. 55, no. 2, pp. 477–523, June 2021, https://doi.org/10.1007/s10579-020-09502-8.

[6] A. Rawat, S. Kumar, and S. S. Samant, "Hate speech detection in social media: Techniques, recent trends, and future challenges," *WIREs Computational Statistics*, vol. 16, no. 2, Mar. 2024, Art. no. e1648, https://doi.org/10.1002/wics.1648.

[7] Z. Mansur, N. Omar, and S. Tiun, "Twitter Hate Speech Detection: A Systematic Review of Methods, Taxonomy Analysis, Challenges, and Opportunities," *IEEE Access*, vol. 11, pp. 16226–16249, 2023, https://doi.org/10.1109/ACCESS.2023.3239375.

[8] E. W. Pamungkas, D. G. P. Putri, and A. Fatmawati, "Hate Speech Detection in Bahasa Indonesia: Challenges and Opportunities," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 6, 2023, https://doi.org/10.14569/IJACSA.2023.01406125.

[9] E. Fauziati, S. D. Amalia, H. A. Zahra, S. E. Ningrum, Y. Sidiq, and A. Budiono, "Hate Speech Typology of Selected Controversial Figures on Social Media: A Discourse-Analytic Perspective," *Wseas Transactions on Information Science and Applications*, vol. 22, pp. 552–564, July 2025, https://doi.org/10.37394/23209.2025.22.46.

[10] M. Ridenhour, A. Bagavathi, E. Raisi, and S. Krishnan, "Detecting Online Hate Speech: Approaches Using Weak Supervision and Network Embedding Models," in *Social, Cultural, and Behavioral Modeling*, vol. 12268, R. Thomson, H. Bisgin, C. Dancy, A. Hyder, and M. Hussain, Eds. Cham: Springer International Publishing, 2020, pp. 202–212.

[11] M. Lim, "Freedom to hate: social media, algorithmic enclaves, and the rise of tribal nationalism in Indonesia," *Critical Asian Studies*, vol. 49, no. 3, pp. 411–427, July 2017, https://doi.org/10.1080/14672715.2017.1341188.

[12] S. G. Cahyani, A. B. Wahyudi, Markhamah, and A. Sabardila, "Code Mixing on News Accounts Catch Me Up! on Twitter in News Text Learning," in *Proceedings of the International Conference on Learning*

and Advanced Education (ICOLAE 2022), Paris, France, 2023, vol. 757, pp. 2024–2039, https://doi.org/10.2991/978-2-38476-086-2_162.

[13] M. O. Ibrohim and I. Budi, "Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter," in *Proceedings of the Third Workshop on Abusive Language Online*, Florence, Italy, 2019, pp. 46–57, https://doi.org/10.18653/v1/W19-3506.

[14] N. Aulia and I. Budi, "Hate Speech Detection on Indonesian Long Text Documents Using Machine Learning Approach," in *Proceedings of the 2019 5th International Conference on Computing and Artificial Intelligence*, Bali Indonesia, Apr. 2019, pp. 164–169, https://doi.org/10.1145/3330482.3330491.

[15] I. G. M. Putra and D. Nurjanah, "Hate Speech Detection In Indonesian Language Instagram," in *2020 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, Depok, Indonesia, Oct. 2020, pp. 413–420, https://doi.org/10.1109/ICACSIS51025.2020.9263084.

[16] I. Alfina, R. Mulia, M. I. Fanany, and Y. Ekanata, "Hate speech detection in the Indonesian language: A dataset and preliminary study," in *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, Bali, Oct. 2017, pp. 233–238, https://doi.org/10.1109/ICACSIS.2017.8355039.

[17] A. D. Sanya and L. H. Suadaa, "Handling Imbalanced Dataset on Hate Speech Detection in Indonesian Online News Comments," in *2022 10th International Conference on Information and Communication Technology (ICoICT)*, Bandung, Indonesia, Aug. 2022, pp. 380–385, https://doi.org/10.1109/ICoICT55009.2022.9914883.

[18] B. P. Putra, B. Irawan, C. Setianingsih, A. Rahmadani, F. Imanda, and I. Z. Fawwas, "Hate Speech Detection using Convolutional Neural Network Algorithm Based on Image," in *2021 International Seminar on Machine Learning, Optimization, and Data Science (ISMODE)*, Jakarta, Indonesia, Jan. 2022, pp. 207–212, https://doi.org/10.1109/ISMODE53584.2022.9742810.

[19] H. Imaduddin, L. A. Kusumaningtias, and F. Y. A'la, "Application of LSTM and GloVe Word Embedding for Hate Speech Detection in Indonesian Twitter Data," *Ingénierie des systèmes d information*, vol. 28, no. 4, pp. 1107–1112, Aug. 2023, https://doi.org/10.18280/isi.280430.

[20] A. T. Azar, H. M. Noori, A. R. Mahlous, A. Al-Khayyat, and I. K. Ibraheem, "Quasi-Reflection Learning Arithmetic Firefly Search Optimization with Deep Learning-based Cyberbullying Detection on Social Networking," *Engineering, Technology & Applied Science Research*, vol. 14, no. 5, pp. 17162–17169, Oct. 2024, https://doi.org/10.48084/etasr.8314.

[21] E. W. Pamungkas, A. Fatmawati, and F. D. Salam, "Hate Speech Detection on Indonesian Social Media: A Preliminary Study on Code-Mixed Language Issue," in *Proceedings of the 2022 6th International Conference on Natural Language Processing and Information Retrieval*, Bangkok Thailand, Dec. 2022, pp. 104–109, https://doi.org/10.1145/3582768.3582771.

[22] E. W. Pamungkas, A. Fatmawati, Y. S. Nugroho, D. Gunawan, and E. Sudarmilah, "Hate Speech Detection in Code-Mixed Indonesian Social Media: Exploiting Multilingual Languages Resources," in *2022 Seventh International Conference on Informatics and Computing (ICIC)*, Denpasar, Bali, Indonesia, Dec. 2022, pp. 1–5, https://doi.org/10.1109/ICIC56845.2022.10006940.

[23] A. Bohra, D. Vijay, V. Singh, S. S. Akhtar, and M. Shrivastava, "A Dataset of Hindi-English Code-Mixed Social Media Text for Hate Speech Detection," in *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, New Orleans, Louisiana, USA, 2018, pp. 36–41, https://doi.org/10.18653/v1/W18-1105.

[24] B. R. Chakravarthi, A. K. M, J. P. McCrae, B. Premjith, K. P. Sorman, and T. Mandl, "Overview of the track on HASOC-Offensive Language Identification-DravidianCodeMix," in *FIRE 2020: Forum for Information Retrieval Evaluation*, Hyderabad, India, Dec. 2020.

[25] E. Ombui, L. Muchemi, and P. Wagacha, "Hate Speech Detection in Code-switched Text Messages," in *2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, Ankara, Turkey, Oct. 2019, pp. 1–6, https://doi.org/10.1109/ISMSIT.2019.8932845.

[26] Z. Tan *et al.*, "Large Language Models for Data Annotation and Synthesis: A Survey," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA, 2024, pp. 930–957, https://doi.org/10.18653/v1/2024.emnlp-main.54.

[27] J. Wei and K. Zou, "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, 2019, pp. 6381–6387, https://doi.org/10.18653/v1/D19-1670.

[28] H. Q. Abonizio, E. C. Paraiso, and S. Barbon, "Toward Text Data Augmentation for Sentiment Analysis," *IEEE Transactions on Artificial Intelligence*, vol. 3, no. 5, pp. 657–668, Oct. 2022, https://doi.org/10.1109/TAI.2021.3114390.

[29] S. Woźniak and J. Kocoń, "From Big to Small Without Losing It All: Text Augmentation with ChatGPT for Efficient Sentiment Analysis," in *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*, Shanghai, China, Dec. 2023, pp. 799–808, https://doi.org/10.1109/ICDMW60847.2023.00108.

[30] Q. Zhang, S. Shi, K. Zhang, Z. Lu, T. Zhang, and X. Xie, "Data Augmentation for Fake News Using ChatGPT," in *2023 International Conference on Intelligent Management and Software Engineering (IMSE)*, Rome, Italy, Sept. 2023, pp. 12–17, https://doi.org/10.1109/IMSE61332.2023.00009.

[31] T. ValizadehAslani *et al.*, "Two-stage fine-tuning with ChatGPT data augmentation for learning class-imbalanced data," *Neurocomputing*, vol. 592, Aug. 2024, Art. no. 127801, https://doi.org/10.1016/j.neucom.2024.127801.