JOIN (Jurnal Online Informatika)
p-ISSN: 2528-1682, e-ISSN: 2527-9165
Volume 10 Number 1 | June 2025: 132-142
DOI: 10.15575/join.v10i1.1472

Study of the Application of Text Augmentation with Paraphrasing to Overcome Imbalanced Data in Indonesian Text Classification

Mutiara Indryan Sari[1], Lya Hulliyyatus Suadaa[2]
[1,2]Computational Statistics, Politeknik Statistika STIS, Indonesia

Article Info

Article history:

Received September 20, 2024
Revised December 18, 2024
Accepted March 06, 2025
Published April 01, 2025

Keywords:

Imbalanced dataset
Paraphrase
Pre-trained model
Text augmentation
Text classification

ABSTRACT

Data imbalance in text classification often leads to poor recognition of minority classes, as classifiers tend to favor majority categories. This study addresses the data imbalance issue in Indonesian text classification by proposing a novel text augmentation approach using fine-tuned pre-trained models: IndoGPT2, IndoBART-v2, and mBART50. Unlike back-translation, which struggles with informal text, text augmentation using pre-trained models significantly improves the F1 score of minority labels, with fine-tuned mBART50 outperforming back translation and other models by balancing semantic preservation and lexical diversity. However, the approach faces limitations, including the risk of overfitting due to synthetic text's lack of natural variations, restricted generalizability from reliance on datasets such as ParaCotta, and the high computational costs associated with fine-tuning large models like mBART50. Future research should explore hybrid methods that integrate synthetic and real-world data to enhance text quality and diversity, as well as develop smaller, more efficient models to reduce computational demands. The findings underscore the potential of pre-trained models for text augmentation while emphasizing the importance of considering dataset characteristics, language style, and augmentation volume to achieve optimal results.

Corresponding Author:

Lya Hulliyyatus Suadaa,
Computational Statistics, Politeknik Statistika STIS
Jl. Otto Iskandardinata No. 64C, Jakarta Timur, DKI Jakarta, Indonesia 13330
Email: 222011536@stis.ac.id

1. INTRODUCTION

The rapid growth of technology and social media has led to an "always-on society" where constant digital interactions generate large volumes of data, commonly known as big data [1]. Text data, a significant component of big data, presents challenges in extracting relevant information due to its volume and diversity [2]. To address these challenges, NLP is used to computationally represent and analyze human language. One of the fundamental tasks in NLP is text classification, which involves labeling text into predefined categories [3]. However, text classification often faces the challenge of imbalanced data, where one or more classes have significantly fewer examples than others [4]. This imbalance leads to poor model performance, as classifiers tend to favor the majority class, resulting in inaccurate predictions for minority classes. Therefore, a high accuracy value does not indicate the overall performance of the classification model [5].

Imbalanced data is particularly problematic in text classification due to the high dimensionality and sparsity of text features [6]. Unlike structured data, text data requires models to capture semantic relationships, context, and linguistic nuances, which are difficult to learn when minority class examples are scarce. This issue is exacerbated in low-resource languages like Indonesian, where labeled datasets are often limited, and the diversity of linguistic expressions further complicates the learning process. For instance, informal language use (e.g., slang, abbreviation, reversed word), add layers of complexity to text classification tasks [7].

To address data imbalance, various methods have been proposed, including oversampling techniques like SMOTE (Synthetic Minority Oversampling Technique) [8]. The study conducted by Spelmen and Porkodi [8] have proven that SMOTE performs better than other algorithms in overcoming data imbalance. While SMOTE has proven effective for numerical data, its application to text data is limited. SMOTE generates synthetic samples by interpolating between feature vectors, treating words as numerical features without considering semantic relationships. This often results in nonsensical or grammatically incorrect sentences, particularly in morphologically rich languages like Indonesian [9]. Alternative approaches, such as text augmentation through back translation and paraphrasing, have shown promise [10]. However, back translation struggles with informal text, as it tends to normalize language, losing important context or stylistic elements. On the other hand, transformer-based paraphrasing techniques have demonstrated superior performance in maintaining semantic similarity and diversity, making them a more suitable choice for text augmentation [11].

Despite these advancements, few studies have focused on the unique challenges of imbalanced text data in low-resource languages like Indonesian. The scarcity of high-quality labeled datasets, combined with the linguistic complexity of Indonesian, underscores the need for tailored approaches to handle imbalanced text data effectively. This research aims to fill this gap by exploring the use of paraphrasing technique with pre-trained language models to generate high-quality synthetic data for Indonesian text classification. By doing so, this research aims to improve the performance of classification models on imbalanced datasets while preserving the linguistic richness and contextual nuances of Indonesian text.

2. METHOD

2.1 Data Source

The data sources used in this research are secondary and can be categorized into classification datasets and paraphrase datasets.

2.1.1 Classification Datasets

This research uses three classification datasets: clickbait, hate speech, and sentiment. These datasets were selected to represent diverse text classification tasks with varying linguistic styles (formal, informal, and mixed). The imbalance ratio in each dataset is 1:5. The clickbait dataset (CLICK-ID) contains Indonesian headlines collected from 12 local Indonesian news publishers [12]. The text tends to be formal, reflecting the language style of news headlines. This dataset consists of 5.297 non-clickbait samples and 1.060 clickbait samples. The hate speech dataset comprises user comment from Kompas.com, a popular Indonesian online news site [13]. The comments are predominantly informal or colloquial, reflecting real-world online interactions. In total, there are 263 data with the hate label and 1.307 data with the no hate label. Meanwhile, the sentence-level sentiment dataset includes comments and reviews in Indonesian obtained from various online platforms, namely Twitter, Zomato, TripAdvisor, Facebook, Instagram, and Qraved [14]. The language style in this dataset tends to be a mixture of formal and informal, representing diverse user-generated content. This dataset consists of 7.359 positively labeled comments and 1.472 negatively labeled comments.

2.1.2 Paraphrase Datasets

The paraphrase dataset is used to fine-tune the pre-trained models for the paraphrasing task. This research employs the ParaCotta dataset, which contains paraphrased text pairs synthesized using neural machine translation [15]. ParaCotta provides a large-scale collection of paraphrased text pairs (6 million pairs), ensuring diversity and coverage of various linguistic styles (formal and informal). The paraphrased pairs are semantically similar and lexically diverse, making them suitable for training models to generate high-quality synthetic data. Due to computational limitations, a subset of 200.000 randomly selected samples is used for fine-tuning.

2.2 Research Stages

In general, the research stages are shown in Figure 1. The research begins with a literature review to obtain the datasets to be used. Next, the datasets are augmented using back translation as baseline and paraphrasing with pre-trained language model. The augmented text is then evaluated using two approaches: automatic evaluation and human evaluation. The augmented datasets are classified using two classification model approaches. The results from the classification and assessment are interpreted to address the research objectives.

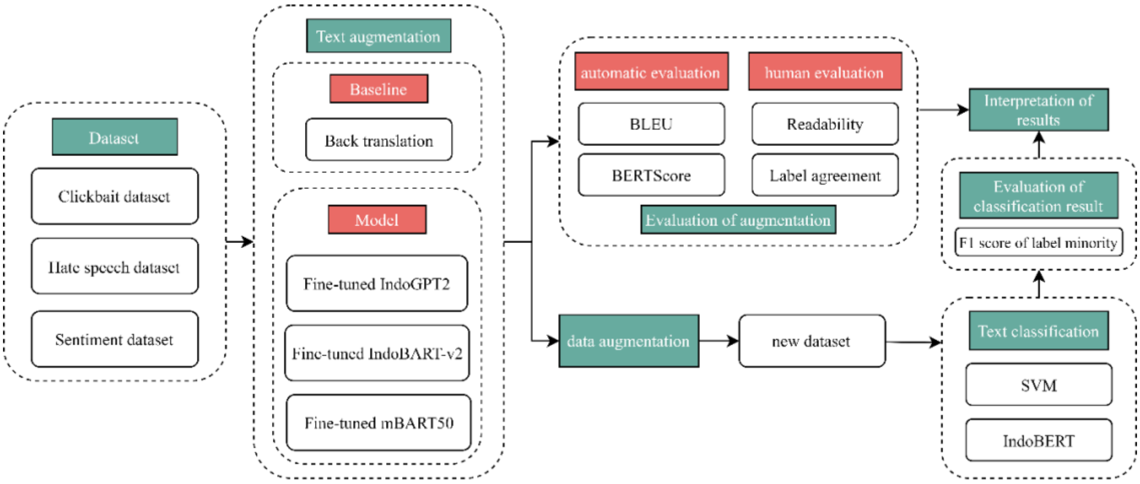

Figure 1. Research flow

2.3 Text Augmentation

Text augmentation is employed to generate new variations of the original text, increasing the amount and diversity of data. This research utilizes paraphrasing techniques using three pre-trained models: IndoGPT2, IndoBART-v2, and mBART50. These models are fine-tuned on the ParaCotta dataset to improve their performance in generating paraphrased text. Additionally, back translation is used as a baseline technique due to its proven effectiveness in text augmentation [9].

2.3.1 Back Translation

Back translation generates new text by translating the original text into an intermediary language and then translating it back into Indonesian [9]. This research uses the Google Translate API with four intermediary languages: English, Chinese, Malay, and Javanese, which is accessed through Python with the googletrans library. These languages were selected based on their linguistic proximity to Indonesian and their widespread use by Indonesian speakers, ensuring the validity and accuracy of the translations [9]. However, back translation may introduce biases, such as the normalization of informal text or the loss of contextual nuances, particularly in low-resource languages.

2.3.2 Pre-trained Models for Paraphrasing

Paraphrasing is a technique for restructuring text while preserving its semantic meaning [16]. This technique can be effectively implemented using language models, which are trained to generate fluent and accurate human language [17]. Among the available options, IndoGPT2, IndoBART-v2, and mBART50 stand out for their architectural strengths and suitability for the Indonesian language.\

IndoGPT2 is a decoder-only autoregressive model with 117 million parameters, pre-trained on Indonesian, Sundanese, and Javanese text corpora. Its unidirectional text generation process efficiently produces sequential text, where each token depends only on the previous ones [19]. IndoBART-v2, with 132 million parameters, adopts a bidirectional encoder and autoregressive decoder architecture. This structure enables the model to better capture the context from both directions, enhancing its text comprehension [19]. Similarly, mBART50, a multilingual Seq2Seq model with 610 million parameters, is pre-trained on 50 languages, including Indonesian. Its large-scale multilingual training makes it highly adaptable for diverse linguistic patterns [20], [21].

While these models offer advanced capabilities for paraphrasing, they may also introduce representational biases. Careful management of synthetic data distribution is essential, as excessive similarity or divergence from the original text can lead to overfitting or diminished model performance [22]. Maintaining this balance is crucial for ensuring robust and accurate paraphrasing outcomes.

2.4 Text Classification

Text classification is done using two approaches: machine learning with SVM and deep learning based on pre-trained Transformers with an IndoBERT model.

2.4.1 Support Vector Machine (SVM)

The primary goal of SVM is to find the optimal decision boundary or hyperplane by maximizing the margin, which is the distance between the decision boundary and the closest points from each class, known as the support vectors [23]. Before text classification, preprocessing is essential. Processing text involves converting text to lowercase (case folding), removing whitespace, stemming, and eliminating stopwords. Subsequently, weighting is applied using TF-IDF (term frequency-inverse document frequency), which accounts for the frequency of a word in a document and its importance across the entire set of documents [24]. Additionally, SVM involves hyperparameter tuning to select the optimal values for parameters C and gamma, while the choice of the linear kernel is fixed from the beginning.

2.4.2 IndoBERT

IndoBERT is a state-of-the-art model for the Indonesian language, built on the BERT architecture. This model has been trained on the Indo4B dataset, which comprises an Indonesian corpus sourced from various publicly available resources. In this research, the IndoBERT model is the "indobenchmark/indobert-base-p1" or IndoBERT$_{\text{BASE}}$ model, featuring 124.5 million parameters [25].

2.5 Classification Evaluation

The classification model evaluation uses 5-fold cross-validation, which is done manually for both SVM and IndoBERT. Synthetic text is added only to the training data, while validation and test data contain only original text. The dataset is split into 80% training, 10% validation, and 10% testing for each fold. F1 scores are calculated for each fold and averaged. The research emphasizes the F1 score, particularly for minority labels, to assess the impact of text augmentation on imbalanced data.

2.6 Evaluation of Text Augmentation

2.6.1 Automatic Evaluation

The automatic evaluation in this study uses BLEU and BERTScore. The BLEU score measures the lexical overlap between the synthetic text and the reference text, with scores ranging from 0 to 1. Meanwhile, the *Self-BLEU or* SBLEU score assesses the lexical diversity between synthetic texts. The lower the BLEU and SBLEU scores, the higher the diversity of the synthetic text [26].

On the other hand, BERTScore evaluates semantic similarity by comparing contextual embeddings of the generated and reference texts. The pre-trained transformer model employed is bert-base-multilingual-cased, which allows for scoring Indonesian text. A BERTScore close to 1 indicates more remarkable semantic similarity between the synthetic and original texts, reflecting a closer alignment in content and meaning [27]. In the context of text augmentation, a balance between high BERTScore (semantic preservation) and moderate BLEU score (lexical diversity) is ideal.

2.6.2 Human Evaluation

The human evaluation in this study examines two aspects: readability and label agreement. These aspects are critical for ensuring the quality and reliability of the synthetic text generated by the paraphrasing techniques. Three evaluators were involved in the assessment process, and the final results were determined by averaging their scores. Evaluators were selected based on their fluency in Indonesian and their familiarity with the specific text classification tasks. To ensure consistency, evaluators were provided with detailed guidelines, including criteria for readability and label agreement.

Readability evaluates whether the synthetic text resembles human-written text in terms of fluency, coherence, and grammatical accuracy [28]. This aspect is evaluated using Best-Worst Scaling (BWS) and analyzed with the MaxDiff approach [29]. Best-Worst Scaling (BWS) approaches are chosen because they efficiently capture subjective judgments, reducing evaluator bias by requiring evaluators to select only the best and worst options from a set. Additionally, BWS has been proven to be more reliable when used to annotate linguistically complex items [30]. The BWS score is calculated using the following formula (1).

$$BW_i = \left(\frac{A-B}{A+B}\right) \times 100\% \tag{1}$$

In equation (1), A represents the number of times method i is chosen as the better method, while B represents the number of times method i is chosen as the worst method.

Meanwhile, label agreement ensures that the augmentation process preserves the semantic and contextual integrity of the original text, preventing changes that may lead to incorrect labels in text classification tasks. The label agreement is measured through proportion analysis, where the percentage of synthetic texts that retain the original label is calculated. Proportion analysis provides a simple and quantifiable measure of label consistency, making it effective for evaluating the impact of text augmentation techniques.

3. RESULT AND DISCUSSION

3.1. Dataset Summary

Table 1 summarizes the characteristics of the classification datasets used to assess the impact of these characteristics on the classification and augmentation results. The study utilizes three classification datasets: hate speech, clickbait, and sentiment. Each of these datasets involves binary classification, as they contain two labels.

Table 1. Summary of Dataset Characteristics

Dataset	Label	Instance (µ)	Unique Words (µ)	Words per Instance (µ)
Clickbait	Binary	6.357	11.103	9
Hate Speech	Binary	1.570	6.973	19
Sentiment	Binary	8.831	14.247	37

Despite having fewer instances, the hate speech dataset is highly diverse, with 6,973 unique words and an average of 19 words per instance, surpassing the clickbait dataset. The sentiment dataset, however, has the most instances, unique words, and words per instance among all datasets.

3.2. Text Classification

Text classification is applied to each augmentation, from the original to the balanced dataset. The evaluation results, shown as a curve, plot the minority label F1 scores on the y-axis against the number of text augmentations on the x-axis. A red line represents the F1 score for the original dataset. The pink line corresponds to the paraphrasing technique using fine-tuned IndoGPT2, the yellow line to fine-tuned IndoBART-v2, and the green line to fine-tuned mBART50. Back translation, used as the baseline, is represented by a blue line.

3.2.1 SVM

Figure 2a, 2b, and 2c display the F1 scores for the minority labels in the SVM classification results across the three datasets. The curves in these figures demonstrate that text augmentation can enhance the classification model's performance. The green line consistently outperforms the others, indicating that the F1 score improves most when using the fine-tuned mBART50 model for augmentation.

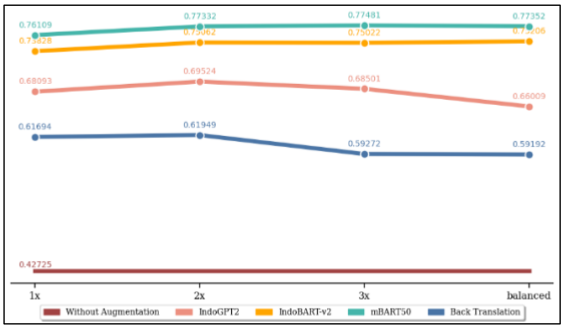

Figure 2a. F1 Score of Hate Label with SVM

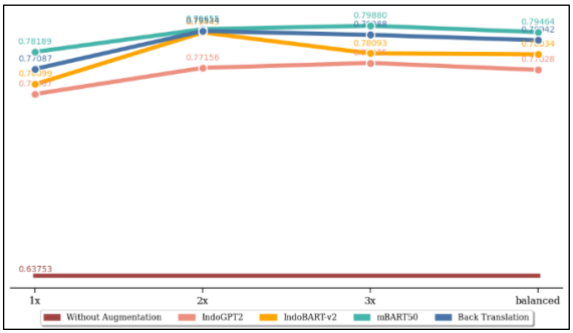

Figure 2b. F1 Score of Clickbait Label with SVM

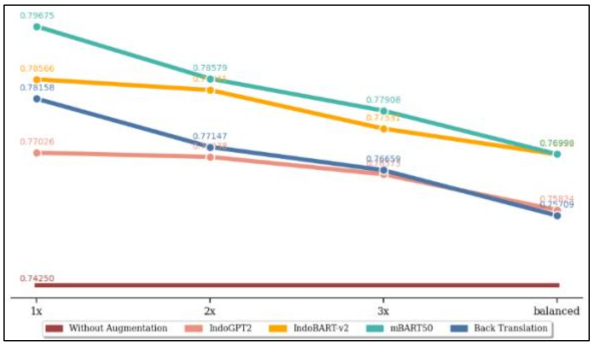

Figure 2c. F1 Score of Negative Label with SVM

On the clickbait dataset, the back translation technique still outperforms the fine-tuned IndoBART-v2 and IndoGPT2 models. However, on the hate speech and sentiment datasets, the F1 score for back translation is the lowest compared to the fine-tuned models. These contradictory results are due to the influence of language style in each dataset. Back translation is less effective on informal language datasets because it relies on Google Translate, which tends to normalize informal words, potentially altering the context or sentiment of the sentence [9]. As a result, back translation performs worse on the hate speech and sentiment datasets, which contain many informal words, than on the clickbait dataset.

When looking at the curve trend regarding the amount of augmentation performed, each dataset has its best augmentation scenario. The pattern of the best number of augmentation scenarios can be random. However, when the dataset is balanced, the F1 score tends to experience a downward trend. This trend shows that adding too much synthetic data can lead to overfitting, as excessive synthetic data input during training may degrade model performance.

3.2.2 IndoBERT

Figure 3a, 3b, and 3c display the F1 score curves for the minority label classification results using IndoBERT for each dataset. The classification results with IndoBERT match the classification results with SVM. However, the back translation technique outperforms the fine-tuned model on the clickbait dataset. The normalization process by Google Translate enhances the effectiveness of back translation on formal language datasets. The results also indicate that text augmentation is more effective in SVM than in IndoBERT; while SVM's F1 score can improve by up to 81%, IndoBERT's highest improvement reaches only 37%. This difference arises because IndoBERT is a pre-trained model trained with a large corpus, giving it a higher initial score than SVM.

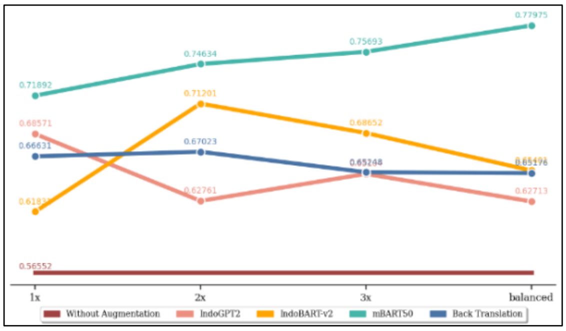

Figure 3a. F1 Score of Hate Label with IndoBERT

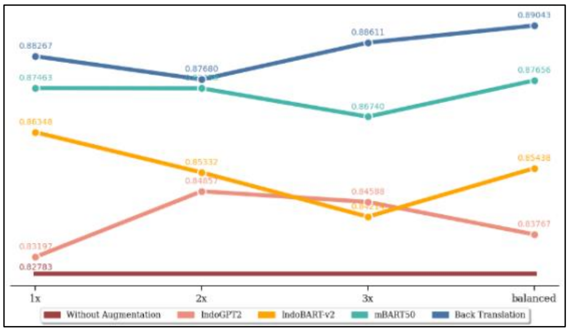

Figure 3b. F1 Score of Clickbait Label with IndoBERT

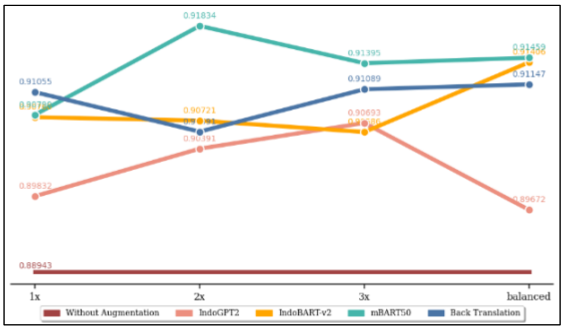

Figure 3c. F1 Score of Negative Label with IndoBERT

On the other hand, as a traditional machine learning model, SVM lacks the inherent ability to understand text context like IndoBERT. Therefore, augmenting an unbalanced dataset with synthetic text can diversify the data and aid the SVM model in understanding the context of the sentence. Furthermore, the results from both SVM and IndoBERT underscore the significant influence of text augmentation on the hate speech dataset. With the least number of instances, the addition of synthetic text to the original hate speech dataset can significantly enhance the model's knowledge.

3.3. *Evaluation of Text Augmentation*

Before being used for paraphrasing, the pre-trained model was fine-tuned using the ParaCotta dataset. This involved adjusting the model's parameters and training it on the ParaCotta dataset, a process that allows the model to learn the specific nuances of the paraphrasing task. The fine-tuned model was then evaluated with BLEU and BERTScore to assess its lexical diversity and semantic similarity aspects. The evaluation results of the fine-tuned model on the ParaCotta dataset can be seen in Table 2.

Table 2. Evaluation of Fine-tuned Model with ParaCotta

Model	BLEU	BERTScore
Fine-tuned IndoGPT2	0,31938	0,89423
Fine-tuned IndoBART-v2	0,32677	0,89645
Fine-tuned mBART50	0,33362	0,89764

Based on the evaluation results, the BLEU score of the three models is relatively low with a high BERTScore. This reassuringly shows that the generated texts have a balanced semantic similarity and lexical diversity, indicating the models' potential to solve paraphrasing tasks on different datasets. Furthermore, the augmented text on clickbait, hate speech, and sentiment datasets is evaluated based on four aspects: lexical diversity, semantic similarity, readability, and label agreement.

3.3.1 Lexical Diversity

This study calculates the BLEU score to measure the lexical diversity between the original and synthetic texts. Additionally, it calculates the self-BLEU (SBLEU) score to assess the lexical diversity among the synthetic texts. Table 3 presents the evaluation results for lexical diversity.

Table 3. Lexical Diversity Evaluation with BLEU

Dataset	Augmentation Technique	BLEU	SBLEU
Clickbait	Fine-tuned IndoGPT2	0,26757	0,71911
	Fine-tuned IndoBART-v2	0,28664	0,75523
	Fine-tuned mBART50	0,41293	0,89525
	Back Translation	0,36261	0,50789
Hate Speech	Fine-tuned IndoGPT2	0,13047	0,52411
	Fine-tuned IndoBART-v2	0,20369	0,54956
	Fine-tuned mBART50	0,29605	0,73855
	Back Translation	0,03281	0,32643
Sentiment	Fine-tuned IndoGPT2	0,20073	0,67328
	Fine-tuned IndoBART-v2	0,29305	0,71138
	Fine-tuned mBART50	0,33570	0,84301
	Back Translation	0,16219	0,45352

Based on the diversity evaluation results, the BLEU and SBLEU scores show the same pattern. The augmentation technique with back translation tends to have the lowest BLEU and SBLEU scores compared to other augmentation techniques. This indicates that the back translation technique can create more diverse synthetic texts. However, there is a different pattern of BLEU score on the clickbait dataset. In the clickbait dataset, the lexical diversity aspect is better when using the fine-tuned IndoGPT2 for augmentation. It is important to note that the hate speech and sentiment datasets consist of informal text, whereas the clickbait dataset contains only formal text. When informal text is augmented with back translation, the resulting synthetic text can vary significantly due to Google Translate's normalization process. This process can transform informal text into formal text, which can alter the context of the original sentence. For instance, the sentence "*banyak kali ngomong kau luhut kerja aja kau ngak becus*" in the hate speech dataset can change to "*saya sudah mengatakan berkali-kali bahwa anda hanya perlu bekerja dan saya tidak baik*" after back translation. Such changes in sentence context can have a detrimental effect on model training and lead to poor classification performance.

3.3.2 Semantic Similarity

This study uses the BERTScore metric to measure the semantic similarity between the synthetic and the original text. The results of the semantic similarity evaluation with BERTScore are shown in Table 4.

Table 4. Semantic Similarity Evaluation with BERTScore

Dataset	Augmentation Technique	BERTScore
Clickbait	Fine-tuned IndoGPT2	0,85433
	Fine-tuned IndoBART-v2	0,86588
	Fine-tuned mBART50	0,90881
	Back Translation	0,85583
Hate Speech	Fine-tuned IndoGPT2	0,79532
	Fine-tuned IndoBART-v2	0,82798
	Fine-tuned mBART50	0,87053
	Back Translation	0,74807
Sentiment	Fine-tuned IndoGPT2	0,81823
	Fine-tuned IndoBART-v2	0,85439
	Fine-tuned mBART50	0,88365
	Back Translation	0,83077

Evaluation with BERTScore shows that the paraphrasing technique with the fine-tuned mBART50 performs best across all three classification datasets. The high BERTScore value in the fine-tuned mBART50 indicates that the synthetic text has a similar meaning to the original text. Due to its training on a vast dataset, the fine-tuned mBART50 performs well in terms of semantic similarity and classification results. In addition, mBART50 architecture allows the model to better understand the context of sentences because BART has bidirectional and autoregressive components.

3.3.3 Readability

The readability aspect was evaluated using Best-Worst Scaling (BWS) and calculated with the MaxDiff approach. A score close to 100 indicates excellent readability of the synthetic text, while a score closes to -100 suggests poor readability. The readability evaluation results are derived from the average scores of three evaluators. Table 5 presents the results of the readability evaluation.

Table 5. Readability Evaluation

Dataset	Original	Fine-tuned IndoGPT2	Fine-tuned IndoBART-v2	Fine-tuned mBART50	Back Translation
Clickbait	68,33	-26,67	-15,00	-18,33	-8,33
Hate Speech	63,33	-56,67	-28,33	-11,67	33,33
Sentiment	61,67	-63,33	-51,67	0,00	53,33

The paraphrasing technique using the fine-tuned model generally produces negative scores, while the back translation technique shows negative scores only on the clickbait dataset. This pattern indicates that the synthetic text generated by back translation is more readable than the fine-tuned model. The better readability quality of back translation is due to Google Translate's ability to produce high-quality translations. Conversely, the synthetic text generated by the fine-tuned model is less readable, potentially due to the suboptimal quality of the training data used during fine-tuning. Furthermore, the readability evaluation results indicate that the original text did not achieve a perfect score of 100. This finding suggests that the synthetic text can deceive the evaluator regarding readability in certain text data.

3.3.4 Label Agreement

The label agreement evaluation score represents the percentage of synthetic text with labels that match the original text. According to the results in Table 6, each dataset achieves the highest score with different augmentation techniques. The clickbait dataset performs best in label agreement with the fine-tuned mBART50 model, while the sentiment dataset achieves the highest score with back translation. The hate speech dataset scores perfectly with the paraphrasing technique using the fine-tuned IndoBART-v2. Overall, the fine-tuned mBART50 model leads with the highest average score across the three datasets at 93.33%, followed by the fine-tuned IndoBART-v2 at 92.22%, back translation at 82.22%, and the fine-tuned IndoGPT2 with the lowest average score of 81.11%.

Table 6. Label Agreement Evaluation

Dataset	Fine-tuned IndoGPT2	Fine-tuned IndoBART-v2	Fine-tuned mBART50	Back Translation
Clickbait	90,00	86,67	96,67	63,33
Hate Speech	86,67	100,00	90,00	86,67
Sentiment	66,67	90,00	93,33	96,67

3.4. *Comparison of Text Augmentation Techniques*

The best augmentation technique for the three datasets is determined by comparing several evaluation aspects. Table 7 highlights the top-performing technique for each aspect.

Table 7. Comparison of the Best Augmentation Technique

Assesment Aspect	Best Technique		
	Clickbait	Hate Speech	Sentiment
Classification (SVM)	Fine-tuned mBART50	Fine-tuned mBART50	Fine-tuned mBART50
Classification (IndoBERT)	Back Translation	Fine-tuned mBART50	Fine-tuned mBART50
Lexical Diversity	Fine-tuned IndoGPT2	Back Translation	Back Translation
Semantic Similarity	Fine-tuned mBART50	Fine-tuned mBART50	Fine-tuned mBART50
Readability	Back Translation	Back Translation	Back Translation

Assesment Aspect	Best Technique		
	Clickbait	Hate Speech	Sentiment
Label Agreement	Fine-tuned mBART50	Fine-tuned IndoBART-v2	Back Translation

The fine-tuned mBART50 model emerges as the best augmentation method across the three datasets. However, when considering readability and lexical diversity, back translation stands out. Yet, caution is advised as low BLEU and BERTScores, such as those observed with back translation on the hate speech dataset, suggest that the increased linguistic diversity may lead to a loss of context, resulting in suboptimal paraphrasing. This, in turn, can negatively impact the performance of the classification model. Therefore, augmentation with the fine-tuned mBART50 is the most effective among the methods evaluated, outperforming back translation, fine-tuned IndoGPT2, and fine-tuned IndoBART-v2.

4. CONCLUSION

Based on the results and discussion, text augmentation with pre-trained language models, particularly the fine-tuned mBART50, IndoBART-v2, and IndoGPT2, effectively addresses data imbalance in Indonesian classification datasets. The study highlights that text augmentation can significantly enhance the F1 score of minority labels. However, its effectiveness depends on the dataset size, language style, and the amount of augmentation applied. Among these techniques, fine-tuned mBART50 is the most effective for paraphrasing, offering a good balance between all aspects. It also outperforms the back translation baseline, which has difficulty preserving the original meaning, especially on informal language datasets.

Despite its promising results, this study has several limitations. First, synthetic text, even when generated by advanced models, often lacks natural variations, so excessive use during training can lead to overfitting and degrade model performance. Second, the reliance on synthetic data, such as the ParaCotta dataset, may limit the generalizability of the findings, as the models may struggle with rare or unseen cases in real-world text. Third, the computational cost of fine-tuning and deploying large models like mBART50 may pose challenges for resource-constrained environments.

Future research should focus on addressing these limitations. One promising direction is the development of hybrid approaches that combine synthetic and real-world data to improve the diversity and quality of augmented text. Additionally, efforts should be made to develop smaller, more efficient models that can achieve comparable performance to large models like mBART50 while reducing computational costs. These advancements will enable more inclusive and equitable NLP systems that can handle the linguistic diversity of real-world data.

REFERENCES

[1] E. Olshannikova, T. Olsson, J. Huhtamäki, and H. Kärkkäinen, "Conceptualizing Big Social Data," *J. Big data*, vol. 4, no. 1, 2017, doi: 10.1186/s40537-017-0063-x.

[2] J. Kaur and J. R. Saini, "A Study of Text Classification Natural Language Processing Algorithms for Indian Languages," *Vnsgu J. Sci. Technol.*, vol. 4, no. 1, pp. 162–167, 2015.

[3] Y. Ko and J. Seo, "Automatic text categorization by unsupervised learning," pp. 453–459, 2000, doi: 10.3115/990820.990886.

[4] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Syst. Appl.*, vol. 73, pp. 220–239, 2017, doi: 10.1016/j.eswa.2016.12.035.

[5] N. A. Verdikha, T. B. Adji, and A. E. Permanasari, "Komparasi Metode Oversampling Untuk Klasifikasi Teks Ujaran Kebencian," *in Semin. Nas. Teknol. Inf. dan Multimed. 2018*, pp. 85–90, 2018.

[6] A. Sun, E. P. Lim, and Y. Liu, "On strategies for imbalanced text classification using SVM: A comparative study," *Decis. Support Syst.*, vol. 48, no. 1, pp. 191–201, 2009, doi: 10.1016/j.dss.2009.07.011.

[7] T. A. Le and D. Moeljadi, "Sentiment Analysis for Low Resource Languages: A Study on Informal Indonesian Tweets," pp. 123–131, 2016.

[8] V. S. Spelmen and R. Porkodi, "A Review on Handling Imbalanced Data," *in Proc. 2018 Int. Conf. Curr. Trends Towar. Converging Technol. ICCTCT 2018*, pp. 1–11, 2018, doi: 10.1109/ICCTCT.2018.8551020.

[9] I.A. Rahma and L. H. Suadaa "Penerapan Text Augmentation untuk Mengatasi Data yang Tidak Seimbang pada Klasifikasi Teks Berbahasa Indonesia," 2023.

[10] A. A. Tavor et al., "Do not have enough data? Deep learning to the rescue!" in 34th AAAI Conference on Artificial Intelligence, 2020, pp. 7383–7390. doi: 10.1609/aaai.v34i05.6233.

[11] D. R. Beddiar, M. S. Jahan, and M. Oussalah, "Data expansion using back translation and paraphrasing for hate speech detection," *Online Social Networks and Media*, vol. 24, 2021, doi: 10.1016/j.osnem.2021.100153.

[12] William and Y. Sari, "CLICK-ID: A novel *dataset* for Indonesian clickbait headlines," *Data in Brief*, vol. 32, p. 106231, 2020, doi: 10.1016/j.dib.2020.106231.

[13] A. D. Sanya and L. H. Suadaa, "Handling Imbalanced Dataset on Hate Speech Detection in Indonesian Online News

Comments," in *2022 10th ICoICT*, Bandung, Indonesia, 2022, pp. 380–385, doi: 10.1109/ICoICT55009.2022.9914883.

[14] A. Purwarianti and I. A. P. A. Crisdayanti, "Improving Bi-LSTM Performance for Indonesian Sentiment Analysis Using Paragraph Vector," in *2019 ICAICTA*, Yogyakarta, Indonesia, 2019, pp. 1-5, doi: 10.1109/ICAICTA.2019.8904199.

[15] A. F. Aji *et al.*, "ParaCotta: Synthetic Multilingual Paraphrase Corpora from the Most Diverse Translation Sample Pair," Proc. 35th Pacific Asia Conf. Lang. Inf. Comput. PACLIC 2021, pp. 666–675, 2021, doi: 10.48550/arXiv.2205.04651.

[16] A. Kumar, S. Bhattamishra, M. Bhandari, and P. Talukdar, "Submodular optimization-based diverse paraphrasing and its effectiveness in data augmentation," in *2019 Proceedings of NAACL-HLT*, Jun 2019, pp. 3609–3619, doi: 10.18653/v1/N19-1363.

[17] B. Li, Y. Hou, and W. Che, "Data augmentation approaches in natural language processing: A survey," *AI Open*, vol. 3, pp. 71–90, 2022, doi: 10.1016/j.aiopen.2022.03.001.

[18] J. Li, T. Tang, W. X. Zhao, and J. R. Wen, "Pretrained Language Models for Text Generation: A Survey," IJCAI Int. Jt. Conf. Artif. Intell., vol. 1, no. 1, pp. 4492–4499, 2021, doi: 10.24963/ijcai.2021/612.

[19] S. Cahyawijaya *et al.*, "IndoNLG: Benchmark and Resources for Evaluating Indonesian Natural Language Generation," In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, *Online* and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov 2021, pp. 8875–8898, doi: 10.48550/arXiv.2104.08200.

[20] Y. Liu *et al.*, "Multilingual denoising pre-training for neural machine translation," *Trans. Assoc. Comput. Linguist.*, vol. 8, pp. 726–742, 2020, doi: 10.1162/tacl_a_00343.

[21] Y. Tang *et al.*, "Multilingual Translation with Extensible Multilingual Pretraining and Finetuning," 2020, doi: 10.48550/arXiv.2008.00401.

[22] S. Y. Feng *et al.*, "A Survey of Data Augmentation Approaches for NLP," *Find. Assoc. Comput. Linguist. ACL-IJCNLP 2021*, pp. 968–988, 2021, doi: 10.18653/v1/2021.findings-acl.84.

[23] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," *Neurocomputing*, vol. 408, pp. 189–215, 2020, doi: 10.1016/j.neucom.2019.10.118.

[24] S. Qaiser and R. Ali, "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents," *Int. J. Comput. Appl.*, vol. 181, no. 1, pp. 25–29, 2018, doi: 10.5120/ijca2018917395.

[25] B. Wilie et al., "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," *Proc. 1st Conf. Asia-Pacific Chapter Assoc. Comput. Linguist. 10th Int. Jt. Conf. Nat. Lang. Process.*, pp. 843–857, 2020, doi: 10.48550/arXiv.2009.05387.

[26] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," *Proc. 40th Annu. Meet. Assoc. Comput. Linguist.*, pp. 311–318, 2002, doi: 10.3917/chev.030.0107.

[27] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating Text Generation with Bert," *8th Int. Conf. Learn*. Represent. ICLR 2020, pp. 1–43, 2020, doi: 10.48550/arXiv.1904.09675.

[28] A. Graefe, M. Haim, B. Haarmann, and H. B. Brosius, "Readers' perception of computer-generated news: Credibility, expertise, and readability," *Journalism*, vol. 19, no. 5, pp. 595–610, 2018, doi: 10.1177/1464884916641269.

[29] Orme, "MaxDiff Analysis: Simple Counting, Individual-Lelvel Logit, and HB" vol.98382, no. 360, 2009.

[30] S. Kiritchenko and S. M. Mohammad, "Best–Worst scaling more reliable than rating scales: A case study on sentiment intensity annotation," *ACL 2017 - 55th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap.*, vol. 2, pp. 465–470, 2017, doi: 10.18653/v1/P17-2074.