International Journal of Advances in Data and Information Systems
Vol. 7, No. 2, August 2026, pp. 716~737
ISSN: 2721-3056, DOI: 10.25008/ijadis.v7i2.1620

Evaluating GPT-4o-based Data Augmentation for Imbalanced Multiclass Sentiment Classification of GoPay Reviews Using IndoBERT-LoRA

Muhammad Yusran[1], Farit Mochamad Afendi[2], Anwar Fitrianto[3]
[1,2,3]Master Program in Statistics and Data Science, School of Data Science, Mathematics and Informatics, IPB University

Article Info

Article history:
Received May 18, 2026
Revised June 14, 2026
Accepted June 20, 2026

Keywords:
GPT-4o
Data augmentation
Sentiment classification
Class imbalance
IndoBERT-LoRA

ABSTRACT

This study evaluated GPT-4o-based data augmentation for imbalanced multiclass sentiment classification of GoPay user reviews using IndoBERT-LoRA. The main problem addressed in this study was the limited representation of minority sentiment classes, particularly the neutral class, which could reduce the model's ability to recognize all sentiment categories proportionally. The dataset consisted of 16,955 Google Play Store reviews that were manually labeled into positive, neutral, and negative classes. Two augmentation strategies were compared, namely prompt-based augmentation and fine-tuning augmentation. The generated synthetic data were evaluated using novelty, diversity, duplication, and manual validation of sampled reviews before being incorporated into the training data. The IndoBERT-LoRA model was trained under four scenarios: baseline, class weighting, prompt-based augmentation, and fine-tuning augmentation. The results showed that fine-tuning produced better lexical-level quality indicators, as indicated by more stable novelty and diversity scores and a lower duplication rate. Both augmentation strategies improved macro recall and macro F1-score compared with the baseline and class weighting scenario. The largest improvement occurred in the neutral class, where recall increased from 0.5633 to 0.7801 with prompt-based augmentation and to 0.8133 with fine-tuning augmentation. These findings indicate that GPT-4o-based augmentation improved minority-class recognition, although the improvement involved a trade-off between recall, precision, and implementation cost.

This is an open access article under the CC BY-SA license.

Corresponding Author:
Muhammad Yusran,
Master Program in Statistics and Data Science, School of Data Science, Mathematics and Informatics
IPB University
Jl. Raya Dramaga, Kampus IPB Dramaga, Bogor, 16680, West Java, Indonesia.
Email: muhammadyusran@apps.ipb.ac.id

1. INTRODUCTION

The rapid development of information and communication technology has driven the growth of digital financial services in Indonesia, including the use of digital wallets such as GoPay. This platform is not only used for various daily transactions but also generates digital traces in the form of user reviews on application distribution platforms such as the Google Play Store. These reviews represent user experiences in the form of opinions, appreciation, criticism, and complaints, making them a valuable source of information for understanding public perceptions of service quality. However, user review data naturally tend to have an imbalanced sentiment distribution, in which positive sentiment usually dominates [1], while other sentiments appear in much smaller proportions. This class imbalance condition poses an important challenge because it may cause the classification model to become biased toward the majority class, thereby failing to represent the overall sentiment distribution proportionally.

Sentiment analysis is one of the approaches in Natural Language Processing (NLP) used to identify, extract, and classify user opinions into specific sentiment categories. [2], [3]. However, most studies still focus on binary classification, which separates positive and negative sentiments, while the neutral class is often ignored. In fact, the neutral class plays an important role in representing opinions that are ambiguous, tentative, or informative without strong emotional polarity. Multiclass sentiment classification presents greater complexity due to differences in linguistic characteristics across classes and the difficulty of computationally capturing subjective human expressions [4]. This condition is further exacerbated by class imbalance, which causes models to become biased toward the majority class and to overlook minority classes [5].

Various approaches have been developed to address data imbalance, including oversampling techniques such as SMOTE. However, this approach is less suitable for text data because it generates synthetic data through numerical interpolation in the feature space without considering the syntactic structure or semantic meaning of language [6]. In the context of NLP, such an approach may degrade the quality of language representation [7]. Therefore, text-based data augmentation has become a more relevant alternative. Classical methods such as back translation and paraphrasing have been shown to improve classification performance, with increases in accuracy and F1-score of 4-10% [8]. Nevertheless, these approaches still have limitations in generating rich linguistic variation, maintaining contextual relevance, and optimally representing the diversity of language expressions in minority classes.

The development of Large Language Models (LLMs), particularly Generative Pre-trained Transformer (GPT), has opened new opportunities for text data augmentation through the generation of synthetic data that are more contextual, coherent, and semantically relevant. Several studies have shown that GPT-based approaches, especially prompt-based strategies such as paraphrasing and few-shot generation, can improve classification performance. For example, GPT-3.5-based augmentation improved accuracy by 4% [9], while few-shot prompting achieved an improvement of up to 8.2% [10]. GPT-3-based augmentation has also been reported to increase accuracy by up to 18% in short text classification [11]. In addition to prompt-based approaches, fine-tuning strategies for generative models have also begun to show promising potential, with GPT-3 fine-tuning improving augmentation-based classification accuracy by an average of 12.76% [12]. Nevertheless, LLM-based augmentation studies are still dominated by prompt-based approaches, whereas the exploration of fine-tuning approaches remains relatively limited and has not been widely examined comparatively. Furthermore, systematic evaluation of the quality of generated synthetic data, particularly in terms of novelty, diversity, and duplication, is still rarely conducted. With the emergence of GPT-4o, which has more advanced semantic understanding and text generation capabilities [13], there is an opportunity to generate higher-quality synthetic data; however, its use in the context of Indonesian sentiment data augmentation remains underexplored.

In addition to data quality, sentiment classification performance is also strongly influenced by the model architecture used. Transformer-based models such as BERT have been shown to be effective in understanding language context bidirectionally [14], and have been adapted to Indonesian through IndoBERT [15]. Previous research showed that IndoBERT achieved an overall accuracy of 92.52%, but still had difficulty classifying sentiments in minority classes [16]. Although transformer-based models can achieve high performance, their fine-tuning process generally requires substantial computational resources. To address this limitation, parameter-efficient fine-tuning approaches such as Low-Rank Adaptation (LoRA) have been introduced to reduce the number of trainable parameters without significantly compromising model performance [17]. This makes IndoBERT-LoRA an efficient yet robust classification framework for evaluating the impact of data augmentation on multiclass sentiment classification performance.

Based on this background, a research gap remains in examining the effectiveness of LLM-based data augmentation, particularly GPT-4o, for addressing class imbalance in Indonesian multiclass sentiment classification. Specifically, studies that comprehensively compare prompt-based and fine-tuning augmentation remain limited, both in terms of synthetic data quality and its impact on classification model performance. Therefore, this study evaluated the effectiveness of GPT-4o-based data augmentation through these two approaches in sentiment classification of GoPay user reviews using IndoBERT-LoRA. The quality of synthetic data was evaluated using

novelty, diversity, duplication, and manual validation of sampled synthetic reviews, while model performance was assessed using multiclass classification metrics with an emphasis on the model's ability to recognize minority classes. Thus, this study provides empirical insight into the relationship between the quality of LLM-generated synthetic data and the improvement of sentiment classification performance under imbalanced data conditions.

2. RESEARCH METHOD

This study used a comparative experimental approach to evaluate the effectiveness of GPT-4o-based data augmentation in addressing class imbalance in multiclass sentiment classification of GoPay user reviews. The research workflow was carried out systematically, starting from data collection, sentiment annotation, corpus characteristic exploration, text preprocessing, data augmentation using two different approaches, synthetic data quality evaluation, classification model training using IndoBERT-LoRA, and model performance evaluation. The overall research procedure is presented in Figure 1.

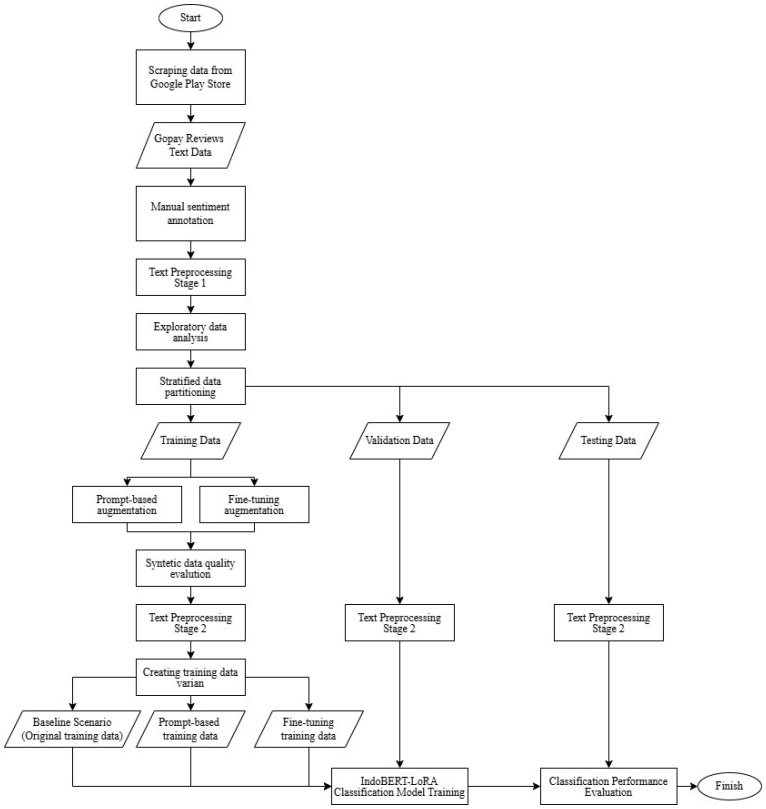


Figure 1. Flowchart of data analysis procedure

2.1. Dataset Collection and Annotation

The data used in this study consisted of GoPay application user reviews collected from the Google Play Store. Data collection was conducted using a web scraping technique with the assistance of the google-play-scraper library in the Python programming language. The reviews were collected from those published between January 1, 2025, and May 24, 2025, resulting in a total of 16,955 reviews as the main corpus of this study.

After data collection, all reviews were manually annotated by the researcher into three sentiment categories: positive, neutral, and negative. Manual annotation was used as the reference labeling approach because human judgment is important for capturing sentence context, semantic meaning, informal expressions, and sentiment nuances in Indonesian user reviews. This approach is consistent with [18], who treated manual annotation as a reference approach because human annotators are generally better able to capture contextual meaning and emotional cues in review texts.

To reduce subjectivity during the annotation process, this study applied a structured sentiment annotation guideline. Reviews were labeled as positive when they contained expressions of satisfaction, appreciation, ease of use, usefulness, or favorable evaluation of GoPay services. Reviews were labeled as negative when they contained complaints, dissatisfaction, transaction failures, balance-related problems, top-up issues, application errors, or customer service-related problems. Meanwhile, reviews were labeled as neutral when they were descriptive, informative, tentative, expressed expectations, described initial user experiences, or did not show clear positive or negative polarity. This criterion was particularly important because neutral sentiment tends to be more ambiguous than positive and negative sentiment in multiclass sentiment annotation.

Labeling consistency was maintained by applying the same annotation guideline throughout the labeling process. In addition, the researcher conducted a second-pass review of potentially ambiguous labels, particularly those assigned to the neutral class, and checked reviews with similar expressions or comparable sentiment tendencies to ensure consistent label assignment. This procedure served as an internal quality-control mechanism in the manual annotation process. The distribution of the annotated data is presented in Figure 2.

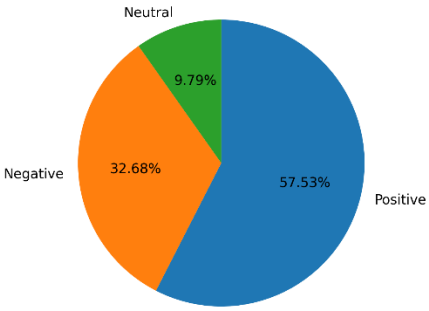


Figure 2. Sentiment distribution

Figure 2 shows that the positive sentiment class dominated the dataset with a proportion of 57.53%, followed by the negative sentiment class at 32.68%, while the neutral sentiment class had the smallest proportion, at 9.79%. This distribution indicates a clear class imbalance, particularly for the neutral and negative classes as minority classes. This condition is consistent with the general characteristics of user opinion data, in which class distributions are often uneven [19]. Such imbalance may cause classification models to become biased toward the majority class, thereby limiting their ability to recognize minority classes [20].

To further strengthen the reliability of the sentiment labels, an additional validation procedure was conducted on a sampled subset of the annotated data. A total of 300 reviews were selected using a stratified sampling approach, consisting of 100 positive, 100 neutral, and 100 negative reviews. These samples were independently reviewed by a second annotator using the same structured annotation guideline. The agreement between the original annotation and the second annotator was evaluated using percentage agreement and Cohen's Kappa. The validation results showed that all 300 sampled reviews received the same labels from both annotators, resulting in a percentage agreement of 100% and a Cohen's Kappa value of 1.000. This result indicates perfect agreement between annotators and provides additional support for the reliability of the sentiment labels, particularly for the neutral class.

2.2. Text Preprocessing

Text preprocessing is a fundamental stage in Natural Language Processing (NLP) because it transforms raw, unstructured data into a cleaner, more standardized format that is ready for downstream tasks such as classification, sentiment analysis, summarization, and intent detection. This stage is particularly important in Indonesian text processing because informal Indonesian often contains complex morphology, slang, abbreviations, non-standard spelling variations, and informal expressions [21]. In this study, preprocessing was conducted in two stages: the first-stage preprocessing and the second-stage preprocessing. This division was applied because each stage served a different purpose. The first-stage preprocessing was used to clean the data before

exploration and augmentation, while the second-stage preprocessing was used to prepare the data according to the input requirements of the IndoBERT-LoRA model.

The first-stage preprocessing was applied to all scraped review data before data exploration and augmentation. This stage aimed to produce clean and standardized text while preserving the general characteristics of user language, so that the original linguistic patterns of each sentiment class could still be observed and used as the basis for designing the augmentation strategy. The first-stage preprocessing included:

1. Case folding, which converted all text into lowercase letters [22].
2. Text cleaning, which removed noise from the data, such as usernames, URLs, hashtags, emoticons, numbers, punctuation marks, non-alphabetic characters, and excessive whitespace [22].

The second-stage preprocessing was conducted after the data augmentation process had been completed. This stage was applied to all data used in modeling, including the original training data, synthetic augmented data, validation data, and test data. Its purpose was to ensure that all data had a consistent text format before being processed by the IndoBERT-LoRA model. The second-stage preprocessing included:

1. Slang normalization, which standardized non-standard words, abbreviations, or informal spelling variations into more consistent forms. This step was performed to reduce differences among word forms with the same meaning, for example, converting “gk”, “ga”, or “nggak” into “tidak” [21].
2. Redundant character removal, which normalized words containing excessive repeated characters, especially those commonly used to express emotion or emphasis in informal text. For example, the form “baguuuusss” was corrected to “bagus” [21].

In this study, stopword removal and stemming were not applied. This decision was based on the consideration that, in transformer-based models, these processes may reduce model performance by disrupting contextual coherence [23]. Examples of text transformation at each preprocessing stage are presented in Table 1.

Table 1. Example of text preprocessing stages

Preprocessing stage	Operation	Example 1	Example 2
Original review	Original review	qrisnya tidak bagus...sering gagal...di scan ke apk laen bisa..di GoPay tidak bisa	canggih siiip deh😘😘🥰🥰
Stage 1	Case folding	qrisnya tidak bagus...sering gagal...di scan ke apk laen bisa..di gopay tidak bisa	canggih siiip deh😘😘🥰🥰
Stage 1	Text cleaning	qrisnya tidak bagus sering gagal di scan ke apk laen bisa di gopay tidak bisa	canggih siiip deh
Stage 2	Slang normalization	qrisnya tidak bagus sering gagal di scan ke aplikasi lain bisa di gopay tidak bisa	canggih siiip deh
Stage 2	Redundant character removal	qrisnya tidak bagus sering gagal di scan ke aplikasi lain bisa di gopay tidak bisa	canggih sip deh

2.3. Data Partitioning

The data were divided into three subsets: training, validation, and test sets, with a proportion of 60:20:20. The partitioning was performed using stratified splitting to maintain the sentiment class proportions in each subset. This process was carried out after the first-stage preprocessing and before data augmentation to prevent data leakage. The distribution of data in each subset is presented in Table 2.

Table 2. Dataset distribution after stratified splitting

Sentiment Class	Training Set	Validation Set	Test Set	Total
Positive	5852	1951	1951	9754
Neutral	996	332	332	1660
Negative	3325	1108	1108	5541
Total	10173	3391	3391	16955

2.4. GPT-4o-based Data Augmentation

Data augmentation was conducted to improve the representation of minority classes in the training data. Data augmentation aims to enrich the variation of training data without requiring

additional manual data collection, thereby improving the model's generalization ability [24], [25]. An augmentation approach based on Large Language Models (LLMs) was selected because generative models such as GPT can generate synthetic texts that are more coherent, contextual, and semantically relevant [26], compared with classical augmentation techniques, which often produce low-quality or noisy data [27].

In this study, augmentation was performed using GPT-4o version gpt-4o-2024-08-06. GPT-4o was selected because of its advanced contextual understanding and text generation capabilities, as well as its ability to process language across various domains [13]. The augmentation focused on the neutral and negative classes, which had fewer observations than the positive class. Augmentation was applied only to the training data to ensure that information from the validation and test data was not involved in the synthetic data generation process, thereby maintaining the integrity of model evaluation. Two augmentation approaches were evaluated in this study: prompt-based augmentation and fine-tuning augmentation.

1. Prompt-based augmentation was conducted using a few-shot prompting strategy. In this approach, the model was provided with several original reviews from the training data as references to generate new synthetic reviews corresponding to the target sentiment class. Few-shot prompting was used because it can provide the model with explicit context regarding the expected linguistic patterns, sentence structures, and semantic characteristics. The synthetic data generation process was performed iteratively. In each iteration, several original reviews were randomly selected from the training data according to the target class and included in the prompt as few-shot examples. The use of different examples in each iteration aimed to increase the variation of generated texts and reduce the possibility of producing outputs that were too similar to the original data. The prompts were designed specifically for each sentiment class. For the neutral class, the model was directed to generate reviews that were informative, descriptive, tentative, and did not contain strong emotional expressions. Meanwhile, for the negative class, the model was directed to generate reviews containing criticism, complaints, or problems relevant to the use of GoPay services. The generation parameter was set with a temperature value of 1.0 to increase linguistic variation, while the output length was limited to 30–220 tokens to keep the generated reviews realistic and consistent with the characteristics of user reviews.
2. Fine-tuning augmentation was conducted by adapting the GPT-4o generative model to the GoPay review domain. Unlike prompt-based augmentation, which depends on review examples in each prompt, the fine-tuning approach enabled the model to learn linguistic patterns, sentence structures, and sentiment characteristics from the training data in a more targeted manner. The fine-tuning data were prepared in JSON Lines (JSONL) format as prompt–completion pairs. The prompt section contained instructions or context that guided the model to generate reviews with a specific sentiment, while the completion section contained original reviews from the training data that matched the instruction. This format was used so that the model could learn the relationship between the generation instruction and the expected review form. The fine-tuning process was performed using GPT-4o version gpt-4o-2024-08-06 with a learning rate multiplier of 0.1, 4 epochs, and a batch size of 8. After fine-tuning was completed, the fine-tuned model was used to generate new synthetic reviews for the neutral and negative classes. During the generation stage, the temperature and token length settings were kept the same as those used in the prompt-based approach.

To ensure a fair comparison between prompt-based augmentation and fine-tuning augmentation, both approaches used the same base GPT-4o model, namely gpt-4o-2024-08-06, targeted the same minority classes, namely neutral and negative, and applied the same generation settings during the synthetic data generation stage, including temperature and output length limits. In addition, augmentation was applied only to the training data, while the validation and test sets were not used in the augmentation process. The number of generated synthetic reviews was also kept relatively comparable between the two approaches. With this setting, the comparison between prompt-based and fine-tuning augmentation focused on the difference in generation strategy rather than differences in the base model, target classes, or evaluation data. The number of synthetic reviews generated by each approach is presented in Table 3.

Table 3. Number of synthetic reviews generated

Sentiment Class	Prompt-Based	Fine-Tuning
Neutral	4855	4852
Negative	2527	2527

2.5. Synthetic Data Quality Evaluation

Synthetic data quality evaluation was conducted before the augmented data were incorporated into the training data. In this study, synthetic data quality was primarily evaluated using three lexical-level aspects: novelty, diversity, and duplication. These metrics were complemented by manual validation of sampled synthetic reviews to assess sentiment-label consistency. Novelty and diversity were used to assess the quality of the augmented data [12], while duplication was used to detect redundancy in synthetic texts that could affect the model learning process [28].

1. Novelty refers to the degree of uniqueness of the generated text compared with the training corpus. The novelty score was measured using the following equation:

$$Novelty(S_i) = 1 - \max\{\varphi(S_i, C_j)\}_{j=1}^{|C|} \quad (1)$$

where S_i denotes an augmented text, C denotes the set of texts in the training corpus, and $\varphi(S_i, C_j)$ denotes the Jaccard similarity function between S_i and C_j. A novelty score close to 1 indicates that the synthetic text has a high degree of novelty, whereas a score close to 0 indicates that the synthetic text is highly similar to the original data.

2. Diversity was used to measure the degree of variation among synthetic texts. This metric is important to ensure that the augmented data not only increase the number of samples but also expand the variation of linguistic expressions in the target class. The diversity score was measured using the following equation:

$$Diversity(S_i) = 1 - \max\{\varphi(S_i, S_j)\}_{j=1}^{j=|S|, j\neq i} \quad (2)$$

where S_i and S_j denote synthetic texts, $\varphi(S_i, S_j)$ denotes the Jaccard similarity function between S_i and S_j. A diversity score close to 0 indicates that the text is similar to other augmented texts, whereas a score close to 1 indicates that the text is different from the others.

3. Duplication was used to detect the presence of identical or highly similar texts within the synthetic data. Duplication checking was conducted to ensure that the generation results were not dominated by identical or repetitive texts.

After the quality evaluation process, synthetic data that passed the filtering stage were combined with the original training data to form the prompt-based and fine-tuning scenarios. Synthetic data identified as duplicates were removed so that the final training data composition increased not only in quantity but also maintained informational variation.

2.6. Sentiment Classification Using IndoBERT-LoRA

The classification model used in this study was IndoBERT with the Low-Rank Adaptation (LoRA) approach. IndoBERT is an adaptation of BERT for the Indonesian language, trained on a large Indonesian corpus, enabling it to better capture Indonesian linguistic characteristics and context [15]. Architecturally, IndoBERT adopts the BERT structure based on the Transformer Encoder. This model uses a bidirectional self-attention mechanism to generate contextual representations of each token by simultaneously considering the left and right contexts [14]. The transformer architecture underlying BERT was introduced in [29].

At the initial stage, each preprocessed review text was tokenized using the IndoBERT tokenizer. The special token [CLS] was added at the beginning of the text as the representation of the entire input, while the [SEP] token was used as an end marker or segment separator. Each token was then converted into an embedding representation. In BERT, the input representation is the sum of token embedding, segment embedding, and positional embedding, formulated as follows:

$$E_{input} = E_{token} + E_{segment} + E_{position} \quad (3)$$

where E_{input} denotes the input representation passed to the encoder, E_{token} denotes the token representation, $E_{segment}$ denotes the segment representation, and $E_{position}$ denotes the positional representation of tokens in the sequence.

The input representation was then processed through the Transformer Encoder layers. The main component of the encoder is the self-attention mechanism, which allows the model to attend to relationships among tokens within a text sequence. In general, the attention mechanism is formulated as follows:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (4)$$

where Q, K, and V denote query, key, and value, respectively, while d_k denotes the key dimension. Through this mechanism, IndoBERT can construct contextual representations that are relevant for sentiment classification tasks.

To improve the efficiency of BERT fine-tuning, this study used Low-Rank Adaptation (LoRA). LoRA is a parameter-efficient fine-tuning approach that freezes the main parameters of the pretrained model and trains only small adaptation matrices. The basic principle of LoRA is to represent the weight updates of the model through low-rank matrix decomposition, formulated as follows [17]:

$$W = W_0 + \Delta W = W_0 + BA \quad (5)$$

where $W_0 \in \mathbb{R}^{d\times k}$ is the weight matrix from the pretrained model, $\Delta W \in \mathbb{R}^{d\times k}$ is the weight update during fine-tuning, while $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$ with $r \ll \min(d, k)$ are two small trainable matrices. In the LoRA scenario, only the small matrices A and B are updated during fine-tuning, while the large weight matrix W_0 remains unchanged or frozen.

For the sentiment classification task, the contextual representation of the [CLS] token from the final IndoBERT layer was used as the representation of the entire review. This representation was then passed to a classification layer with a softmax activation function to produce prediction probabilities for three sentiment classes: positive, neutral, and negative. This process is formulated as follows:

$$\hat{y} = softmax\left(h_{[CLS]}W_c + b_c\right) \quad (6)$$

where $\hat{y} \in \mathbb{R}^C$ is the prediction probability vector for each class, $h_{[CLS]} \in \mathbb{R}^d$ is the [CLS] token representation, $W_c \in \mathbb{R}^{d\times C}$ is the classification weight matrix, $b_c \in \mathbb{R}^C$ is the bias, and C is the number of sentiment classes. The model was trained using the categorical cross-entropy loss function, formulated as follows:

$$L_{CCE} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \quad (7)$$

where y_i denotes the actual label for class i, while $\hat{y}_i$ denotes the model predicted probability for class i.

The IndoBERT-LoRA model was trained under four experimental scenarios. The first scenario was the baseline, in which the model was trained using the original training data. The second scenario was class weighting, which was added as a non-LLM imbalance-handling baseline. In this scenario, the model was still trained using only the original training data, but the loss function was weighted according to the class distribution in the training set so that misclassification of minority classes received a larger penalty [18]. The third and fourth scenarios were prompt-based augmentation and fine-tuning-based augmentation, respectively, in which the model was trained using the original training data combined with GPT-4o-generated synthetic reviews. The model configuration used in this study is presented in Table 4.

Table 4. Configuration of IndoBERT-LoRA model

Component	Configuration
Base Model	IndoBERT pretrained
Fine-tuning Strategy	Low-Rank Adaptation (LoRA)
LoRA Rank	4
Dropout Rate	0.1
Optimizer	AdamW
Learning Rate	2e-5

Number of Epochs	5
Loss Function	Categorical Cross-Entropy
Output Layer	Dense layer with softmax activation

2.7. Classification Performance Evaluation

Classification performance evaluation was conducted to determine how well the model correctly classified the data. Model performance was evaluated on the same test set for all experimental scenarios using a confusion matrix and several multiclass classification metrics, namely accuracy, precision, recall, and F1-score. A confusion matrix is a classification table that describes model performance by presenting the number of correct and incorrect predictions for each class [30]. In multiclass classification, the confusion matrix is a square matrix of size $k \times k$, where k is the number of classes, and contains the frequency of each predicted class compared with the actual label [31]. Table 5 shows an example of a confusion matrix for three classes.

Table 5. Confusion matrix for multiclass classification

Actual	Predicted		
	Class I	Class II	Class III
Class I	a_{11}	a_{12}	a_{13}
Class II	a_{21}	a_{22}	a_{23}
Class III	a_{31}	a_{32}	a_{33}

Based on the confusion matrix, accuracy, precision, recall, and F1-score are formulated as follows [31], [32]:

$$Accuracy = \frac{\sum_{i=1}^{k} a_{ii}}{\sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij}} \tag{8}$$

$$Precision_i = \frac{a_{ii}}{\sum_{j=1}^{k} a_{ji}} \tag{9}$$

$$Recall_i = \frac{a_{ii}}{\sum_{j=1}^{k} a_{ij}} \tag{10}$$

$$F1_i = 2 \times \frac{Precision_i \times Recall_i}{Precision_i + Recall_i} \tag{11}$$

Because the data had an imbalanced class distribution, the main evaluation focused on macro-average metrics. Macro-average calculates the average metric value across classes without considering the sample proportion of each class, making it more suitable for evaluating model performance under class imbalance conditions [33]. The macro-average for a metric M is formulated as:

$$Macro\text{-}M = \frac{1}{k}\sum_{i=1}^{k} M_i \tag{12}$$

where k denotes the number of sentiment classes. In addition to aggregate metrics, evaluation was also conducted at the class level to examine the impact of augmentation on the model's ability to recognize classes with low representation, especially the neutral class.

3. RESULTS AND DISCUSSION

3.1. Characteristics of Original GoPay Review Data

Exploration of the original data was conducted to understand the linguistic characteristics of GoPay user reviews before the augmentation process. This analysis used the text produced from the first-stage preprocessing, namely text that had been cleaned from noise while still preserving the natural language patterns of users, allowing the characteristics of each sentiment class to be observed.

As previously shown, the GoPay review data had an imbalanced class distribution. The positive class dominated the dataset, while the neutral and negative classes had smaller proportions and were categorized as minority classes. A wordcloud was then used to identify the dominant words in each sentiment class, as presented in Figure 3.



(a) (b) (c)

Figure 3. Wordcloud of original GoPay reviews by sentiment class: (a) positive, (b) neutral, and (c) negative.

Based on Figure 3(a), the positive sentiment class was dominated by words such as “sangat”, “bagus”, “mantap”, “membantu”, “mudah”, and “banget”. This pattern indicates that positive reviews were generally expressed through explicit appreciative expressions. These words represent user satisfaction with the ease of use, benefits, and service quality of GoPay. In addition, the occurrence of intensifying words such as “sangat” and “banget” suggests that the positive class had relatively clear lexical markers.

For the neutral sentiment class, Figure 3(b) shows the dominance of words such as “ok”, “oke”, “semoga”, “baru”, “dulu”, and “coba”. These words indicate that neutral reviews tended to be brief, tentative, and did not yet express strong emotional evaluation. Words such as “baru” and “coba” reflect users’ initial experiences, while “semoga” indicates expectations toward the service without a clear sentiment tendency. This characteristic makes the neutral class more difficult for the model to learn because it lacks explicit sentiment markers.

Meanwhile, Figure 3(c) shows that the negative class was dominated by words such as “tidak”, “bisa”, “saldo”, “masuk”, “top up”, and “customer service”. The dominance of the word “tidak” indicates that negation patterns were a major characteristic of negative sentiment expressions. The occurrence of words related to balance, transactions, top up, and customer service shows that negative reviews mainly contained complaints about the core functions of the application and problem resolution experienced by users.

Overall, the wordcloud exploration shows that each class had different linguistic characteristics. The positive class had explicit appreciative expressions, the negative class had strong complaint and negation patterns, while the neutral class tended to be more ambiguous, brief, and less emotionally expressive. These findings served as the basis for applying data augmentation to the neutral and negative classes, since the generated synthetic data needed to increase the representation of lower-proportion classes while preserving the semantic characteristics of the original data.

3.2. Characteristics of GPT-4o Synthetic Reviews

After the characteristics of the original data had been identified, data augmentation was applied to the neutral and negative classes using two approaches, namely prompt-based augmentation and fine-tuning augmentation. The synthetic data were then analyzed to determine whether the texts generated by GPT-4o still preserved the semantic characteristics of their original classes. This aspect is important because synthetic data used for model training should not only increase the number of observations, but also remain relevant to the language patterns and context of GoPay user reviews.

In the prompt-based approach, synthetic data were generated through instructions and examples of original reviews provided in the prompt. The wordclouds of prompt-based synthetic reviews for the neutral and negative classes are presented in Figure 4.



(a) (b)

Figure 4. Wordcloud of prompt-based synthetic reviews: (a) neutral and (b) negative

Based on Figure 4(a), the prompt-based synthetic data in the neutral class still exhibited informative and tentative language characteristics. Words such as “coba”, “fitur”, “cukup”, and “bisa” appeared dominantly. This pattern indicates that the neutral synthetic reviews still reflected comments that did not yet express strong emotional evaluation. These characteristics are consistent with the original neutral data, which frequently contained expressions of initial experience, expectations, or temporary assessments of the service.

In the negative class, as shown in Figure 4(b), words such as “saldo”, “tidak”, “sudah”, “padahal”, and “customer service” appeared dominantly. The presence of these words indicates that the prompt-based synthetic data still preserved the context of user complaints related to transactions, balance issues, and customer service. Negation patterns also remained an important feature of the negative class, indicating that the generated synthetic data were still relevant to the semantic characteristics of the original data.

Furthermore, the wordclouds of synthetic data generated through fine-tuning augmentation are presented in Figure 5.



(a) (b)

Figure 5. Wordcloud of fine-tuning synthetic reviews: (a) neutral and (b) negative.

Based on Figure 5(a), the fine-tuning synthetic data in the neutral class showed richer and more contextual vocabulary variation. Words such as “fitur”, “transaksi”, “pembayaran”, and “mudah” appeared prominently. This indicates that the fine-tuned model was able to generate neutral reviews that were more descriptive of the GoPay usage context, particularly those related to application features and transaction activities. Compared with the prompt-based approach, the texts generated through fine-tuning appeared more domain-specific and were not limited to short expressions.

In the negative class, as shown in Figure 5(b), the fine-tuning synthetic data still preserved complaint patterns relevant to the original data. Words such as “saldo”, “masuk”, “padahal”, “gak”, and “top up” indicate that the model was able to capture the language style used by users when expressing complaints. The use of informal language variations such as “gak” also suggests that the

synthetic data produced through fine-tuning were closer to the everyday language of users than the prompt-based outputs, which tended to be more structured.

Overall, both augmentation approaches appeared to generate synthetic data that preserved the main semantic characteristics of the target classes. Prompt-based augmentation tended to produce simpler texts that were closely aligned with the basic patterns provided in the prompt. In contrast, fine-tuning augmentation produced richer vocabulary and contextual variation because the model had been adapted to the GoPay review domain. These findings suggest that fine-tuning has the potential to generate more domain-specific and varied synthetic data, although this observation should be interpreted together with quantitative lexical-level evaluation and manual validation of sampled synthetic reviews.

3.3. Synthetic Data Quality Evaluation

Synthetic data quality evaluation was conducted to ensure that the augmented data not only increased the number of observations in classes with low representation but also had characteristics suitable for use in the model training process. In this study, synthetic data quality was evaluated using three aspects: novelty, diversity, and duplication.

1. Novelty

Novelty was used to examine the extent to which the generated synthetic texts differed from the original data. A higher novelty score indicates that the synthetic data have a higher degree of newness, whereas a lower score indicates that the synthetic texts tend to be similar to the training data. The comparison of novelty scores between prompt-based and fine-tuning augmentation is presented in Figure 6.

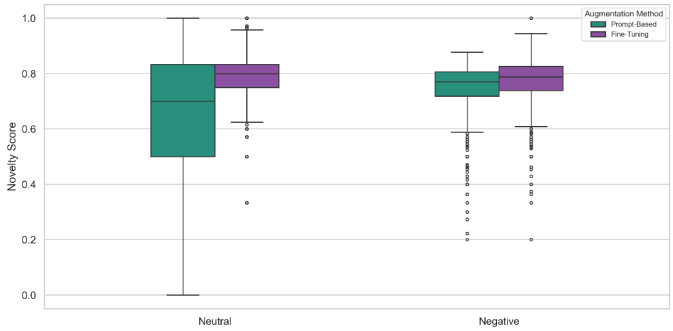


Figure 6. Comparison of novelty scores between prompt-based and fine-tuning augmentation

Based on Figure 6, the fine-tuning approach showed a more stable novelty pattern than the prompt-based approach, particularly in the neutral class. In the neutral class, fine-tuning produced a higher median novelty score, approximately 0.80, while prompt-based augmentation showed a lower score and a wider distribution. This indicates that the synthetic data generated through prompt-based augmentation for the neutral class varied more in their level of similarity to the original data, with some texts tending to remain close to the patterns of the training data.

In the negative class, both methods showed relatively comparable novelty scores, with median values ranging from approximately 0.77 to 0.79. Nevertheless, prompt-based augmentation still showed a wider score distribution than fine-tuning. This pattern suggests that fine-tuning was more consistent in generating novel synthetic texts, whereas prompt-based augmentation depended more heavily on the examples provided in the prompt, causing some outputs to have higher similarity to the original data.

Overall, the novelty results indicate that fine-tuning showed more stable lexical novelty than prompt-based augmentation. This may have occurred because the fine-tuned model had learned the domain patterns of GoPay reviews in a more targeted manner, enabling it to generate new text variations without relying too heavily on examples explicitly provided in each prompt.

2. Diversity

Diversity was used to assess the degree of variation among the generated synthetic texts. A higher diversity score indicates that the synthetic data contain a wider range of expressions,

whereas a lower diversity score indicates that the synthetic texts tend to be repetitive or share highly similar patterns. The comparison of diversity scores between the two augmentation methods is presented in Figure 7.

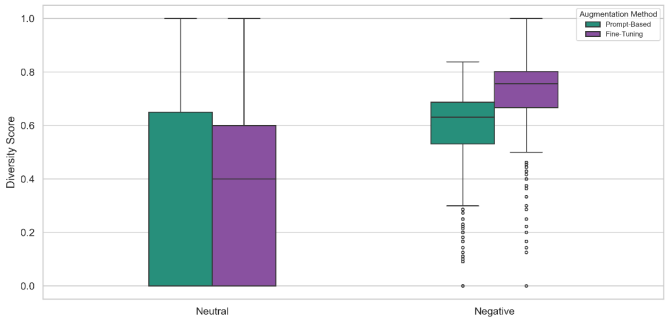


Figure 7. Comparison of diversity scores between prompt-based and fine-tuning augmentation

Based on Figure 7, fine-tuning showed higher diversity scores than prompt-based augmentation, especially in the negative class. In the negative class, the median diversity score of fine-tuning was around 0.76, while prompt-based augmentation was around 0.63. This indicates that fine-tuning was able to generate richer language variation in describing user complaints, particularly those related to transactions, balance, top up, and customer service.

In the neutral class, the diversity scores of both methods were relatively lower than those of the negative class. This indicates that the neutral class posed a specific challenge in the synthetic data generation process. The linguistic characteristics of neutral reviews, which tend to be brief, informative, and lack strong emotional expressions, made the model more likely to generate similar sentence patterns. Nevertheless, fine-tuning still showed better diversity than prompt-based augmentation, indicating that it was more effective in enriching language variation in the neutral class.

Overall, the diversity results show that fine-tuning achieved higher lexical diversity scores than prompt-based augmentation. This difference was particularly evident in the negative class, which linguistically provides a wider space for expression because it contains various forms of complaints and service-related problems. Meanwhile, the neutral class remained more difficult to diversify because its language characteristics were more limited and less expressive.

3. Duplication

In addition to novelty and diversity, evaluation was also conducted on the duplication rate of synthetic data. Duplication was used to determine whether the augmented data were dominated by identical or highly similar texts. A high duplication rate indicates that the synthetic data have not fully enriched informational variation because some generated texts tend to be repetitive. The duplication evaluation results for each augmentation method are presented in Table 6.

Table 6. Duplication rate of GPT-4o synthetic reviews

Augmentation Method	Sentiment Class	Unique Texts	Duplicate Texts	Unique Duplicate Texts	Pure Duplications	Duplicate Percentage
Prompt-based	Neutral	2764	2405	314	2091	49.54%
Prompt-based	Negative	2512	26	11	15	1.03%
Fine-tuning	Neutral	3508	1649	305	1344	33.99%
Fine-tuning	Negative	2526	2	1	1	0.08%

Based on Table 6, the most visible difference in duplication rate occurred in the neutral class. Prompt-based augmentation produced a duplication percentage of 49.54%, indicating that nearly half of the synthetic data in the neutral class consisted of repeated or highly similar texts. In contrast, fine-tuning reduced the duplication rate to 33.99%. Although this value was still relatively high, the result shows that fine-tuning was more capable of reducing repetition than prompt-based augmentation.

In the negative class, the duplication rate was much lower for both methods. Prompt-based augmentation produced a duplication rate of 1.03%, while fine-tuning produced only 0.08%. This indicates that the model was better able to generate unique variations of negative texts than neutral texts. The more expressive and problem-oriented nature of the negative class, along with its wider variety of complaint contexts, made negative synthetic data more diverse and less repetitive.

Overall, the evaluation of novelty, diversity, and duplication indicates that fine-tuning produced better lexical-level quality indicators than prompt-based augmentation. Specifically, fine-tuning generated synthetic reviews with higher novelty, greater diversity, and lower duplication rates. However, these metrics should be interpreted as surface-level or lexical-level indicators because they mainly capture word variation, text similarity, and redundancy, rather than semantic validity or sentiment-label correctness. The neutral class remained the main challenge in the augmentation process because its language characteristics tended to be brief, tentative, and repetitive. These findings indicate that the effectiveness of data augmentation depends not only on the generative strategy used but also on the linguistic characteristics of the target sentiment class. Therefore, a manual validation of sampled synthetic reviews was conducted to complement the lexical-level evaluation and to assess whether the generated texts were consistent with the intended sentiment labels.

4. Manual Validation of Synthetic Reviews

Manual validation of synthetic reviews was conducted to complement the evaluation of novelty, diversity, and duplication. This additional evaluation was necessary because these three metrics mainly represent lexical-level quality indicators and do not fully assess whether the synthetic texts are semantically consistent with the intended sentiment labels. Therefore, manual validation was performed to examine the sentiment correctness of the generated synthetic reviews. A total of 100 synthetic reviews were randomly selected from each combination of augmentation method and target class. Each sampled review was manually examined to determine whether the expressed sentiment was consistent with the intended target label. In this validation process, a review was categorized as correct when its expressed sentiment matched the target class. The manual validation results are presented in Table 7.

Table 7 Manual validation of synthetic reviews

Augmentation Method	Sentiment Class	Sample Size	Sentiment Correct	Correctness Rate
Prompt-based	Neutral	100	100	100%
Prompt-based	Negative	100	100	100%
Fine-tuning	Neutral	100	100	100%
Fine-tuning	Negative	100	100	100%

All manually validated synthetic review samples were consistent with their intended sentiment labels. This finding indicates that the synthetic data generated by both augmentation approaches not only showed lexical variation but also maintained sentiment consistency with the target classes in the evaluated samples. For the neutral class, this result is particularly important because neutral sentiment tends to be more ambiguous than negative sentiment. Nevertheless, this validation result should still be interpreted within the scope of the evaluated sample, and a broader human evaluation of the entire synthetic dataset may be considered in future studies.

3.4. Training Data Distribution after Augmentation Filtering

After the synthetic data quality evaluation was conducted, the augmented data that passed the filtering process were combined with the original training data to form new training scenarios. In this subsection, three training data composition scenarios are presented: baseline, prompt-based augmentation, and fine-tuning augmentation. The baseline scenario used the original training data without adding synthetic data, while the other two scenarios used the original training data combined with GPT-4o-generated synthetic data. The class weighting scenario used the same original training data as the baseline, but applied class weights during model training. The training data distribution for each scenario is presented in Table 8.

Table 8. Training data distribution after augmentation filtering

Sentiment Class	Baseline	Prompt-Based	Fine-Tuning
Positive	5852 (57.53%)	5852 (37.88%)	5852 (36.11%)
Neutral	996 (9.79%)	3760 (24.34%)	4504 (27.79%)
Negative	3325 (32.68%)	5837 (37.78%)	5851 (36.10%)
Total	10173 (100%)	15449 (100%)	16207 (100%)

Based on Table 8, the initial distribution in the baseline scenario showed a clear class imbalance. The positive class accounted for more than half of the training data, while the neutral class represented only 9.79%. This condition indicates that a model trained on the baseline data may learn more patterns from the positive class than from the neutral class.

After adding synthetic data, the class distribution in the prompt-based and fine-tuning scenarios became more proportional. In the prompt-based scenario, the number of neutral samples increased from 996 to 3760, while the number of negative samples increased from 3325 to 5837. The proportion of the neutral class increased to 24.34%, while the negative and positive classes were relatively balanced, at 37.78% and 37.88%, respectively. This indicates that prompt-based augmentation improved the representation of classes with low proportions, although the number of neutral samples retained after filtering was still lower than that of the other classes.

In the fine-tuning scenario, the training data distribution became slightly more balanced than in the prompt-based scenario. The number of neutral samples increased to 4504, or 27.79%, while the number of negative samples increased to 5851, or 36.10%. The proportions of the positive and negative classes were almost equal, while the neutral class showed a greater increase than in the prompt-based scenario. This difference indicates that more synthetic data generated through fine-tuning were retained after the quality filtering process, mainly due to the lower duplication rate compared with prompt-based augmentation.

Overall, both augmentation approaches successfully shifted the training data distribution from an imbalanced condition toward a more proportional distribution. However, fine-tuning produced a slightly better training data composition because more synthetic data passed the filtering process, especially in the neutral class. This result is consistent with the lexical-level quality evaluation, which showed that fine-tuning produced more retained synthetic samples after filtering, particularly for the neutral class.

3.5. Classification Performance of IndoBERT-LoRA

Classification performance evaluation was conducted to examine the effect of GPT-4o-based data augmentation on the ability of the IndoBERT-LoRA model to classify the sentiment of GoPay user reviews. The model was evaluated on the same test set across all experimental scenarios to ensure a fair performance comparison. In this study, model performance was evaluated across four experimental scenarios: baseline, class weighting, prompt-based augmentation, and fine-tuning-based augmentation. The class weighting scenario was included as a non-LLM imbalance-handling baseline, allowing GPT-4o-based augmentation to be compared not only with the original baseline but also with a simpler imbalance-handling strategy that does not rely on synthetic text generation. Evaluation was conducted at the overall level using accuracy, macro precision, macro recall, and macro F1-score, and at the class level using precision, recall, and F1-score for each sentiment class. The overall evaluation results are presented in Table 9.

Table 9. Overall performance of IndoBERT-LoRA across experimental scenarios

Scenario	Accuracy	Macro Precision	Macro Recall	Macro F1-score
Baseline	0.9378	0.9138	0.8389	0.8664
Class weighting	0.9316	0.8723	0.8633	0.8676
Prompt-based	0.9345	0.8713	0.8933	0.8811
Fine-tuning	0.9328	0.8648	0.9020	0.8803

Based on Table 9, the IndoBERT-LoRA model in the baseline scenario achieved the highest accuracy of 0.9378. However, because the dataset had an imbalanced class distribution, accuracy alone was not sufficient to describe the model's ability to recognize all sentiment classes

proportionally. Therefore, the interpretation should focus on macro-average metrics, particularly macro recall and macro F1-score.

The class weighting scenario was added as a non-LLM imbalance-handling baseline. Compared with the baseline, class weighting slightly decreased accuracy from 0.9378 to 0.9316, but increased macro recall from 0.8389 to 0.8633 and macro F1-score from 0.8664 to 0.8676. This result indicates that weighting the loss function based on class distribution helped the model pay more attention to underrepresented classes, although the improvement was still limited compared with GPT-4o-based augmentation.

After GPT-4o-based augmentation was applied, a more substantial improvement was observed in macro recall. Macro recall increased to 0.8933 in the prompt-based scenario and 0.9020 in the fine-tuning scenario. This improvement indicates that adding synthetic reviews enhanced the model's ability to capture patterns from all classes, especially those that were previously underrepresented. However, the increase in macro recall was followed by a decrease in macro precision, from 0.9138 in the baseline scenario to 0.8713 in the prompt-based scenario and 0.8648 in the fine-tuning scenario. This pattern suggests that augmentation made the model more sensitive to difficult classes, but also increased the risk of misclassification across classes.

In terms of macro F1-score, class weighting produced only a slight improvement over the baseline, whereas both GPT-4o-based augmentation scenarios showed clearer improvements. Prompt-based augmentation achieved the highest macro F1-score of 0.8811, while fine-tuning produced a nearly similar value of 0.8803. Although fine-tuning produced the highest macro recall, prompt-based augmentation achieved a slightly higher macro F1-score, indicating a more balanced trade-off between precision and recall. These findings suggest that class weighting is a useful non-LLM baseline, but GPT-4o-based augmentation provides additional benefits by introducing new linguistic variations into the minority classes.

Furthermore, class-level evaluation was conducted to examine the specific impact of augmentation on each sentiment class. The model performance for the positive class is presented in Table 10.

Table 10. Positive-class performance of IndoBERT-LoRA across experimental scenarios

Scenario	Precision	Recall	F1-score
Baseline	0.9393	0.9831	0.9607
Class weighting	0.9517	0.9590	0.9553
Prompt-based	0.9648	0.9549	0.9598
Fine-tuning	0.9736	0.9441	0.9586

Based on Table 10, the model performance for the positive class was very high across all scenarios. In the baseline scenario, the positive class achieved a precision of 0.9393, a recall of 0.9831, and an F1-score of 0.9607. The very high recall indicates that most positive reviews were correctly recognized by the model. This result is reasonable because the positive class had the largest number of samples, allowing the model to learn stronger positive language patterns during training.

After imbalance handling was applied, the precision of the positive class increased in all non-baseline scenarios. Class weighting increased positive-class precision to 0.9517, while prompt-based and fine-tuning augmentation further increased it to 0.9648 and 0.9736, respectively. However, positive-class recall decreased from 0.9831 in the baseline scenario to 0.9590 in class weighting, 0.9549 in prompt-based augmentation, and 0.9441 in fine-tuning augmentation. This pattern indicates that imbalance handling reduced the model's tendency to over-recognize the majority positive class and made the model more selective in assigning positive labels.

Despite this decrease in recall, the F1-score of the positive class remained very high across all scenarios. The baseline achieved an F1-score of 0.9607, followed by prompt-based augmentation with 0.9598, fine-tuning with 0.9586, and class weighting with 0.9553. These relatively small differences indicate that neither class weighting nor GPT-4o-based augmentation substantially disrupted the model's ability to recognize the majority class. The positive class remained well learned because it had strong representation and clear lexical markers, such as appreciative expressions and words associated with satisfaction.

The model performance for the neutral class is presented in Table 11.

Table 11. Neutral-class performance of IndoBERT-LoRA across experimental scenarios

Scenario	Precision	Recall	F1-score
Baseline	0.8500	0.5633	0.6775
Class weighting	0.7048	0.6687	0.6862
Prompt-based	0.6780	0.7801	0.7255
Fine-tuning	0.6522	0.8133	0.7239

Table 11 shows that the most substantial performance change occurred in the neutral class. In the baseline scenario, the neutral class achieved a precision of 0.8500 but a recall of only 0.5633. This condition indicates that when the model predicted a review as neutral, the prediction was relatively accurate. However, the low recall indicates that many neutral reviews were not correctly recognized and tended to be classified into other classes. This confirms that the neutral class was the most difficult class for the model to learn under the initial imbalanced data condition.

The class weighting scenario improved neutral-class recall from 0.5633 to 0.6687, showing that a non-LLM imbalance-handling strategy could help the model recognize more neutral reviews without adding synthetic data. However, this improvement was accompanied by a decrease in precision from 0.8500 to 0.7048. This indicates that class weighting increased the model's sensitivity to the neutral class, but also made the model more likely to assign neutral labels to some reviews from other classes.

GPT-4o-based augmentation produced a stronger increase in neutral-class recall. Prompt-based augmentation increased recall to 0.7801, while fine-tuning achieved the highest recall of 0.8133. These results indicate that synthetic reviews helped expand the representation of the neutral class and enabled the model to capture more neutral language patterns that were previously underrepresented. However, the increase in recall was also followed by a further decrease in precision, which dropped to 0.6780 in prompt-based augmentation and 0.6522 in fine-tuning. Therefore, the improvement in neutral-class recall should be interpreted as an increase in model sensitivity toward neutral reviews, rather than as an absolute improvement in the precision of neutral sentiment understanding.

In terms of F1-score, all imbalance-handling strategies improved neutral-class performance compared with the baseline. Class weighting increased the neutral F1-score from 0.6775 to 0.6862, while prompt-based and fine-tuning augmentation achieved higher F1-scores of 0.7255 and 0.7239, respectively. Although fine-tuning produced the highest neutral recall, prompt-based augmentation produced a slightly higher neutral F1-score because it retained better precision. This shows that prompt-based augmentation provided a more balanced precision-recall trade-off for the neutral class, whereas fine-tuning was more aggressive in increasing sensitivity.

These results reinforce the finding that the main difficulty in multiclass sentiment classification is not determined solely by the number of samples, but also by the linguistic characteristics of each class. The neutral class tends to be brief, tentative, and lacks explicit sentiment markers, making it more difficult to distinguish from positive and negative classes. Therefore, the increase in neutral-class recall demonstrates the benefit of imbalance handling, particularly GPT-4o-based augmentation, but the accompanying decrease in precision also highlights the need to interpret this improvement carefully.

The model performance for the negative class is presented in Table 12.

Table 12. Negative-class performance of IndoBERT-LoRA across experimental scenarios

Scenario	Precision	Recall	F1-score
Baseline	0.9522	0.9702	0.9611
Class weighting	0.9604	0.9621	0.9612
Prompt-based	0.9712	0.9449	0.9579
Fine-tuning	0.9687	0.9486	0.9585

Based on Table 12, the model performance for the negative class was also very high across all scenarios. In the baseline scenario, the negative class achieved a precision of 0.9522, a recall of 0.9702, and an F1-score of 0.9611. These results indicate that the model was already able to

recognize the negative class very well, although the number of negative samples was lower than that of the positive class. This suggests that classification difficulty is not solely determined by the number of samples, but also by the linguistic characteristics of the class.

The class weighting scenario produced a precision of 0.9604, a recall of 0.9621, and an F1-score of 0.9612 for the negative class. This result was nearly identical to the baseline and slightly higher in terms of F1-score, indicating that class weighting maintained strong performance for the negative class while improving attention to minority classes. Compared with class weighting, GPT-4o-based augmentation produced higher precision for the negative class, with 0.9712 in the prompt-based scenario and 0.9687 in the fine-tuning scenario. However, negative-class recall decreased to 0.9449 and 0.9486, respectively.

Despite these changes, the F1-score of the negative class remained very high across all scenarios. This indicates that neither class weighting nor GPT-4o-based augmentation produced a major disruption in negative-class classification. Negative reviews generally contained negation, complaints, and vocabulary related to transaction problems, balance, top up, and customer service. These characteristics made the negative class easier to distinguish from other classes than the neutral class.

Overall, the class-level performance results show that the greatest impact of imbalance handling occurred in the neutral class. The class weighting baseline improved neutral-class recall and slightly increased macro F1-score compared with the baseline, indicating that a non-LLM imbalance-handling strategy could help the model recognize more minority-class samples. However, the improvement achieved by class weighting was smaller than that achieved by GPT-4o-based augmentation, particularly for neutral-class recall and macro F1-score. This suggests that weighting the loss function alone helped address class imbalance to some extent, but it did not provide additional linguistic variation for the minority classes.

Compared with class weighting, prompt-based and fine-tuning augmentation provided stronger improvements in neutral-class recall. Fine-tuning achieved the highest macro recall and neutral-class recall, indicating greater sensitivity toward underrepresented neutral reviews. Meanwhile, prompt-based augmentation achieved the highest macro F1-score and neutral-class F1-score, suggesting a more balanced trade-off between precision and recall. Nevertheless, the increase in neutral-class recall across imbalance-handling scenarios was consistently accompanied by a decrease in neutral-class precision. Therefore, the improvement should be interpreted as increased sensitivity to the neutral class rather than a complete improvement in neutral sentiment understanding.

These findings indicate that GPT-4o-based data augmentation can provide additional benefits beyond class weighting by enriching the linguistic representation of underrepresented sentiment classes. However, the results also highlight an important trade-off between sensitivity and prediction precision, especially for the neutral class. Thus, prompt-based augmentation may be more suitable when balanced performance is prioritized, whereas fine-tuning augmentation may be more suitable when the main objective is to maximize recall for difficult minority classes.

3.6. Practical Cost and Complexity Analysis

In addition to classification performance, implementation cost and complexity are important practical considerations in GPT-based data augmentation. This analysis is necessary because prompt-based augmentation and fine-tuning augmentation have different workflows. Prompt-based augmentation only requires prompt design and iterative generation, whereas fine-tuning augmentation requires additional steps, including preparing training data in JSONL format, fine-tuning the model, and generating synthetic data using the fine-tuned model. The estimated token usage and API cost are presented in Table 13.

Table 13 Token usage and API cost of GPT-4o-based augmentation

Component	Prompt-based neutral	Prompt-based negative	Fine-tuning neutral	Fine-tuning negative
Training tokens	-	-	121528	928796
Input tokens	1434234	1016925	821625	371972
Output tokens	47725	103614	62449	93814
Training cost	-	-	$3.04	$23.22
Input cost	$3.59	$2.54	$3.08	$1.39

Output cost	$0.48	$1.04	$0.94	$1.41
Total cost	$4.06	$3.58	$7.06	$26.02

Prompt-based augmentation had an estimated total cost of $7.64, consisting of $4.06 for the neutral class and $3.58 for the negative class. In contrast, fine-tuning augmentation had an estimated total cost of $33.08, consisting of $7.06 for the neutral class and $26.02 for the negative class. Thus, fine-tuning augmentation was approximately 4.33 times more expensive than prompt-based augmentation. This cost difference was mainly caused by the fine-tuning training cost. In prompt-based augmentation, all costs came from the synthetic text generation process. In fine-tuning augmentation, however, the total cost consisted of two components: model training cost and generation cost using the fine-tuned model. The highest cost was observed in the fine-tuning augmentation for the negative class because it used a much larger number of training tokens than the neutral class. In terms of implementation complexity, prompt-based augmentation had a simpler workflow because it did not require model adaptation. In contrast, fine-tuning augmentation required additional steps, including JSONL file preparation, fine-tuning job execution, and synthetic data generation using the fine-tuned model. Therefore, the choice of augmentation strategy should consider the trade-off among classification performance, synthetic data quality indicators, cost, and implementation complexity. Fine-tuning augmentation may be more suitable when the main objective is to improve recall and obtain better lexical-level quality indicators, whereas prompt-based augmentation is more efficient when the balance between performance improvement and cost is prioritized.

4. CONCLUSION

This study showed that GPT-4o-based data augmentation can improve the recognition of underrepresented sentiment classes in multiclass sentiment classification of GoPay reviews using IndoBERT-LoRA. Based on the synthetic data quality evaluation, fine-tuning augmentation showed better lexical-level quality indicators than prompt-based augmentation, as reflected by more stable novelty and diversity scores and a lower duplication rate, especially in the neutral class. However, these indicators mainly capture lexical variation, text similarity, and redundancy; therefore, they should not be interpreted as a complete measure of semantic quality. To complement these lexical-level metrics, manual validation of sampled synthetic reviews was conducted, and all evaluated samples were consistent with their intended sentiment labels.

In terms of classification performance, both GPT-4o-based augmentation approaches improved the ability of the IndoBERT-LoRA model to recognize underrepresented sentiment classes compared with the baseline and class weighting scenarios. The most evident improvement occurred in the neutral class, where the recall of IndoBERT-LoRA increased from 0.5633 in the baseline scenario to 0.7801 with prompt-based augmentation and 0.8133 with fine-tuning augmentation. However, this increase in neutral-class recall was followed by a decrease in neutral-class precision, indicating a trade-off between model sensitivity and prediction accuracy. Therefore, the improvement should be interpreted as increased sensitivity of the IndoBERT-LoRA model toward neutral reviews rather than a complete improvement in neutral sentiment understanding.

Prompt-based augmentation provided slightly more balanced IndoBERT-LoRA performance based on macro F1-score and neutral-class F1-score, while fine-tuning augmentation achieved the highest macro recall and neutral-class recall. From a practical perspective, fine-tuning augmentation also required higher estimated API cost and implementation complexity than prompt-based augmentation. Therefore, fine-tuning augmentation may be more suitable when the main objective is to maximize the recall of IndoBERT-LoRA for difficult minority classes, whereas prompt-based augmentation may be more efficient when the balance between IndoBERT-LoRA performance improvement, cost, and implementation simplicity is prioritized.

This study has several limitations. First, although an additional validation procedure involving a second annotator was conducted on 300 sampled reviews, the full dataset was still annotated by a single primary annotator. Therefore, the inter-annotator agreement result should be interpreted as supporting evidence for label reliability on the sampled subset rather than as a

complete reliability assessment of the entire dataset. Future studies should involve multiple annotators in the full annotation process to further strengthen label reliability, particularly for the neutral class. Second, the manual validation of synthetic reviews was limited to sampled data; therefore, broader human evaluation involving more samples and multiple evaluators is recommended. Third, the prompt-based approach used class-specific few-shot prompts and did not involve extensive prompt-template tuning or systematic comparison of multiple prompt styles. Future studies may explore different prompt formulations, prompt optimization strategies, or prompt ensembles. Fourth, this study focused on IndoBERT-LoRA as the downstream classifier because it provides an efficient and relevant architecture for Indonesian sentiment classification. Future research may further examine whether the benefits of GPT-based augmentation remain consistent across other downstream models and application domains.

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest related to this work.

DATA AVAILABILITY

The data supporting the findings of this study are available from the corresponding author upon reasonable request. The raw review data are not publicly shared to maintain the ethical use of user-generated content.

AUTHOR CONTRIBUTIONS

Muhammad Yusran: Conceptualization; Data Curation; Data Labeling; Methodology; Software; Formal Analysis; Visualization; Writing–Original Draft. **Farit Mochamad Afendi**: Conceptualization; Methodology; Supervision; Validation; Writing–Review & Editing. **Anwar Fitrianto**: Conceptualization, Methodology; Supervision; Validation; Writing–Review & Editing.

REFERENCES

[1] V. Oktaviani, B. Warsito, H. Yasin, R. Santoso, and Suparti, "Sentiment analysis of e-commerce application in Traveloka data review on Google Play site using Naïve Bayes classifier and association method," *J. Phys. Conf. Ser.*, vol. 1943, no. 1, p. 012147, Jul. 2021, doi: 10.1088/1742-6596/1943/1/012147.

[2] Y. Kang, Z. Cai, C.-W. Tan, Q. Huang, and H. Liu, "Natural language processing (NLP) in Management Research: A literature review," *Journal of Management Analytics*, vol. 7, no. 2, pp. 139–172, 2020, doi: 10.1080/23270012.2020.1756939.

[3] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artif. Intell. Rev.*, vol. 55, pp. 5731–5780, 2022, doi: 10.1007/s10462-022-10144-1.

[4] M. Bouazizi and T. Ohtsuki, "Multi-class Sentiment Analysis on Twitter: Classification Performance and Challenges," *Big Data Mining and Analytics*, vol. 2, no. 3, pp. 181–194, 2019, doi: 10.26599/BDMA.2019.9020002.

[5] A. Mukherjee, S. Mukhopadhyay, P. K. Panigrahi, and S. Goswami, "Utilization of Oversampling For Multiclass Sentiment Analysis on Amazon Review Dataset," in *2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST)*, Morioka, Japan: IEEE, Oct. 2019, pp. 1–6. doi: 10.1109/ICAwST.2019.8923260.

[6] I. A. Rahma and L. H. Suadaa, "Automated Indonesian Text Augmentation with Web-Based Application Using Flask Framework," *Proceedings of The International Conference on Data Science and Official Statistics*, vol. 2023, no. 1, pp. 96–108, Dec. 2023, doi: 10.34123/icdsos.v2023i1.324.

[7] J. Jang, Y. Kim, K. Choi, and S. Suh, "Sequential Targeting: A Continual Learning Approach For Data Imbalance in Text Classification," *Expert Syst. Appl.*, vol. 179, p. 115067, Oct. 2021, doi: 10.1016/j.eswa.2021.115067.

[8] D. R. Beddiar, M. S. Jahan, and M. Oussalah, "Data Expansion Using Back Translation and Paraphrasing For Hate Speech Detection," *Online Soc. Netw. Media*, vol. 24, p. 100153, Jul. 2021, doi: 10.1016/j.osnem.2021.100153.

[9] 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: IEEE, Dec. 2023, pp. 799–808. doi: 10.1109/ICDMW60847.2023.00108.

[10] H. Dai *et al.*, "AugGPT: Leveraging ChatGPT for Text Data Augmentation," *IEEE Trans. Big Data*, vol. 11, no. 3, pp. 907–918, Jun. 2025, doi: 10.1109/TBDATA.2025.3536934.

[11] S. V. Balkus and D. Yan, "Improving short text classification with augmented data using GPT-3," *Nat. Lang. Eng.*, vol. 30, no. 5, pp. 943–972, 2024, doi: 10.1017/S1351324923000438.

[12] C. Suhaeni and H.-S. Yong, "Mitigating Class Imbalance in Sentiment Analysis through GPT-3-Generated Synthetic Sentences," *Applied Sciences*, vol. 13, no. 17, p. 9766, Aug. 2023, doi: 10.3390/app13179766.

[13] OpenAI, "GPT-4o System Card," https://openai.com/index/gpt-4o-system-card/.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[15] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," in *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain: International Committee on Computational Linguistics, Dec. 2020, pp. 757–770. doi: 10.18653/v1/2020.coling-main.66.

[16] Y. A. Singgalen, "Performance Analysis of IndoBERT for Sentiment Classification in Indonesian Hotel Review Data," *Journal of Information System Research (JOSH)*, vol. 6, no. 2, pp. 978–988, 2025.

[17] E. Hu *et al.*, "Lora: Low-Rank Adaptation of Large Language Models," *ICLR*, vol. 1, no. 2, pp. 3–22, 2022.

[18] R. Anadra, H. Wijayanto, and K. Sadik, "Sentiment Analysis of Tokopedia Customer Reviews Using BiLSTM and IndoBERT with Comparative Analysis of Preprocessing and Labeling Methods," *International Journal of Advances in Data and Information Systems*, vol. 6, no. 3, pp. 773–788, Dec. 2025, doi: 10.59395/ijadis.v6i3.1458.

[19] W. Chen, K. Yang, Z. Yu, Y. Shi, and C. L. P. Chen, "A Survey on Imbalanced Learning: Latest Research, Applications and Future Directions," *Artif. Intell. Rev.*, vol. 57, no. 6, pp. 137–187, 2024, doi: 10.1007/s10462-024-10759-6.

[20] P. Thölke *et al.*, "Class Imbalance Should Not Throw You Off Balance: Choosing The Right Classifiers and Performance Metrics For Brain Decoding with Imbalanced Data," *Neuroimage*, vol. 277, p. 120253, 2023, doi: 10.1016/j.neuroimage.2023.120253.

[21] C. Suhaeni, S. A. Kamila, F. Fahira, M. Yusran, and G. Alfa Dito, "Exploring a Large Language Model on the ChatGPT Platform for Indonesian Text Preprocessing Tasks," *Indonesian Journal of Statistics and Its Applications*, vol. 9, no. 1, pp. 100–116, Jun. 2025, doi: 10.29244/ijsa.v9i1p100-116.

[22] M. Yusran, Siswanto, and A. Islamiyati, "Comparison of Multinomial Naïve Bayes and Bernoulli Naïve Bayes on Sentiment Analysis of Kurikulum Merdeka with Query Expansion Ranking," *Sistemasi: Jurnal Sistem Informasi*, vol. 13, no. 1, pp. 96–106, 2024.

[23] U. Hasanah, A. Mohamad Soleh, C. Suhaeni, and A. Fitrianto, "EVALUATING ROBERTA AND GPT-BASED MODELS FOR SDG MULTICLASS TEXT CLASSIFICATION ACROSS DIFFERENT DOCUMENT LENGTHS," *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, vol. 20, no. 3, pp. 2645–2664, Apr. 2026, doi: 10.30598/barekengvol20iss3pp2645-2664.

[24] S. Y. Feng *et al.*, "A Survey of Data Augmentation Approaches for NLP," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Association for Computational Linguistics, 2021, pp. 968–988. doi: 10.18653/v1/2021.findings-acl.84.

[25] 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.

[26] H. Zhao, H. Chen, T. A. Ruggles, Y. Feng, D. Singh, and H.-J. Yoon, "Improving Text Classification with Large Language Model-Based Data Augmentation," *Electronics (Basel).*, vol. 13, no. 13, pp. 2535–2548, 2024, doi: 10.3390/electronics13132535.

[27] B. Ding *et al.*, "Data Augmentation using LLMs: Data Perspectives, Learning Paradigms and Challenges," in *Findings of the Association for Computational Linguistics ACL 2024*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2024, pp. 1679–1705. doi: 10.18653/v1/2024.findings-acl.97.

[28] X. Li and J. Li, "Generative Deduplication For Social Media Data Selection," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2024, pp. 5765–5776. doi: 10.18653/v1/2024.findings-emnlp.330.

[29] A. Vaswani *et al.*, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[30] N. A. Farhana, F. M. Afendi, A. Fitrianto, and S. H. Wijaya, "Classification modeling of support vector machine (SVM) and random forest in predicting pharmacodynamics interactions," *J. Phys. Conf. Ser.*, vol. 1863, no. 1, p. 012067, Mar. 2021, doi: 10.1088/1742-6596/1863/1/012067.

[31] S. Farhadpour, T. A. Warner, and A. E. Maxwell, "Selecting and Interpreting Multiclass Loss and Accuracy Assessment Metrics for Classifications with Class Imbalance: Guidance and Best Practices," *Remote Sens. (Basel).*, vol. 16, no. 3, p. 533, Jan. 2024, doi: 10.3390/rs16030533.

[32] K. Takahashi, K. Yamamoto, A. Kuchiba, and T. Koyama, "Confidence interval for micro-averaged F1 and macro-averaged F1 scores," *Applied Intelligence*, vol. 52, no. 5, pp. 4961–4972, Mar. 2022, doi: 10.1007/s10489-021-02635-5.

[33] A. E. Blanchard *et al.*, "A Keyword-Enhanced Approach to Handle Class Imbalance in Clinical Text Classification," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 6, pp. 2796–2803, Jun. 2022, doi: 10.1109/JBHI.2022.3141976.