VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8327

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

IMPACT OF TEXT AUGMENTATION ON INDOBERT PERFORMANCE FOR HOSPITAL REVIEW SENTIMENT ANALYSIS

Yoga Anugrah Pratama. SY[1]; Ali Ibrahim[1*]

Magister of Computer Science[1]
Sriwijaya University, Palembang, Indonesia[1]
https://unsri.ac.id[1]
yogaanugrahpsy@gmail.com, aliibrahim@unsri.ac.id[*]

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract— *Sentiment analysis of hospital patient reviews plays a critical role in evaluating healthcare service quality. However, limited labeled data and class imbalance often affect classification performance and reduce minority-class detection. This study empirically investigates the impact of text augmentation techniques on improving IndoBERT performance for sentiment classification of patient reviews at Dr. Mohammad Hoesin Palembang General Hospital. A total of 1,464 reviews were collected, preprocessed, and weakly labeled using a VADER-based approach, resulting in 1,168 positive and 296 negative instances. To address class imbalance, augmentation was applied exclusively to the training set using back-translation and Easy Data Augmentation (EDA), including synonym replacement, random insertion, random swap, and random deletion. IndoBERT was fine-tuned under consistent hyperparameter settings and evaluated using accuracy, precision, recall, F1-macro, and AUC. The baseline model achieved an F1-macro of 70.2%, indicating limited minority-class sensitivity. After augmentation, the random swap technique achieved the highest observed performance within the experimental setup, reaching 96.8% accuracy, 95.3% F1-macro, and 98.9% AUC. These results suggest improved performance within the experimental setting, particularly in minority-class detection. However, it should be noted that the labels were generated through a translation-based weak labeling approach, which may introduce noise and affect the accuracy of the labels. Therefore, the findings should be interpreted within the scope of this experimental setting.*

Keywords*: Easy Data Augmentation, Hospital Reviews, Imbalanced Dataset, IndoBERT, Sentiment Analysis.*

Intisari—*Analisis sentimen terhadap ulasan pasien rumah sakit memiliki peran penting dalam mengevaluasi kualitas layanan kesehatan. Namun, keterbatasan data berlabel dan ketidakseimbangan kelas sering memengaruhi kinerja klasifikasi serta menurunkan kemampuan dalam kelas minoritas. Penelitian ini bertujuan menganalisis secara empiris dampak teknik augmentasi teks terhadap peningkatan kinerja IndoBERT dalam klasifikasi sentimen ulasan pasien di RSUP Dr. Mohammad Hoesin Palembang. Sebanyak 1.464 ulasan dikumpulkan, dipraproses, dan diberi label secara lemah menggunakan pendekatan berbasis VADER, menghasilkan 1.168 ulasan positif dan 296 ulasan negatif. Untuk mengatasi ketidakseimbangan kelas, augmentasi diterapkan secara eksklusif pada data latih menggunakan teknik back-translation dan Easy Data Augmentation (EDA), yang meliputi synonym replacement, random insertion, random swap, dan random deletion. Model IndoBERT kemudian di-fine-tune dengan konfigurasi hiperparameter yang konsisten dan dievaluasi menggunakan metrik accuracy, precision, recall, F1-macro, dan AUC. Model dasar memperoleh nilai F1-macro sebesar 70,2%, yang menunjukkan keterbatasan dalam mendeteksi kelas negatif. Setelah penerapan augmentasi, teknik random swap menunjukkan kinerja tertinggi yang diamati dalam eksperimen ini, dengan akurasi sebesar 96,8%, F1-macro 95,3%, dan AUC 98,9%. Hasil penelitian ini mengindikasikan adanya peningkatan kinerja dalam kondisi eksperimen yang dilakukan, khususnya pada kemampuan deteksi kelas minoritas. Namun, perlu diperhatikan bahwa label yang digunakan dihasilkan melalui pendekatan pelabelan lemah berbasis terjemahan, yang berpotensi mengandung noise dan memengaruhi akurasi label. Oleh karena*

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8327

itu, temuan penelitian ini perlu diinterpretasikan dengan mempertimbangkan keterbatasan dataset dan kondisi eksperimen yang digunakan.

***Kata Kunci**: Augmentasi Data Sederhana, Analisis Sentimen, IndoBERT, Ketidakseimbangan Dataset, Ulasan Rumah Sakit.*

INTRODUCTION

Healthcare institutions are increasingly evaluated through patient feedback, which serves as an important indicator of service quality and institutional performance. With the rapid advancement of digital technology, patients actively share their experiences through various online platforms such as health applications, social media, and public review websites [1], [2]. These online reviews provide a continuously updated and publicly accessible source of information for assessing healthcare services [1], [3], [4]. However, patient feedback is typically expressed in unstructured free text form, making manual analysis inefficient, subjective, and difficult to replicate [5], [6]. The growing volume of textual feedback further complicates systematic evaluation and highlights the need for automated analytical approaches. Consequently, computational techniques such as sentiment analysis are increasingly employed to transform qualitative patient opinions into structured, measurable, and actionable insights [5], [6], [7].

Sentiment analysis, a key task in Natural Language Processing (NLP), aims to determine the polarity of opinions expressed in textual data [7], [8], [9]. In healthcare contexts, sentiment analysis has been widely applied to examine patient satisfaction and service perceptions. Prior studies have developed sentiment-based evaluation frameworks for hospital reviews in various countries [1], [2], [3], [4]. Similar analytical approaches have also been employed in physician review platforms to identify clinical attributes influencing patient perceptions [10]. Furthermore, transformer-based architectures such as BERT have demonstrated superior performance compared to traditional machine learning methods in healthcare-related sentiment classification tasks [11], [12] due to their contextual representation capability.

In Indonesia, the development of sentiment analysis research has increasingly leveraged transformer-based language models tailored to the Indonesian language. IndoBERT, introduced within the IndoLEM benchmark initiative, provides a pre-trained model specifically designed for Indonesian NLP tasks [13]. Subsequent studies have reported strong performance of IndoBERT across various classification scenarios, including healthcare-related sentiment analysis [5], [8]. This aligns with broader findings indicating that transformer architectures dominate contemporary sentiment analysis research because of their ability to capture contextual and semantic nuances effectively [9]. Such contextual modeling is particularly important for Indonesian patient reviews, which frequently contain informal expressions, abbreviations, colloquial terms, and mixed linguistic styles [5], [9].

Despite these advancements, several challenges remain. First, most existing studies rely on manually labeled datasets, whereas in real hospital environments, including Dr. Mohammad Hoesin Palembang General Hospital, patient reviews are typically unlabeled. As a result, large-scale manual annotation becomes resource intensive and impractical. Second, limited dataset size and class imbalance, where positive reviews substantially outnumber negative ones, can introduce model bias and reduce sensitivity toward minority-class complaints [14], [15].

Since negative reviews frequently contain critical feedback regarding service quality, this imbalance poses a significant limitation for reliable healthcare evaluation systems. Third, although text augmentation techniques such as Easy Data Augmentation and back-translation have been reported to enhance classification performance under low-resource conditions [14], [15], [16], controlled empirical comparisons that isolate augmentation effects on IndoBERT performance within the Indonesian healthcare domain remain limited. Furthermore, the effectiveness of augmentation may vary depending on dataset characteristics, labeling quality, and domain-specific language patterns. Therefore, evidence obtained from a single dataset should be interpreted cautiously and may not necessarily generalize to other healthcare review settings.

Recent empirical surveys emphasize the importance of systematically evaluating augmentation strategies in transformer-based models under limited and imbalanced data conditions [17]. However, many prior studies apply augmentation without explicitly controlling training configurations, which makes it difficult to attribute performance improvements solely to augmentation techniques. To address these gaps, this study develops a sentiment analysis pipeline integrating VADER-based weak labeling and IndoBERT fine-tuning under explicitly imbalanced conditions. By maintaining identical hyperparameter settings across experimental scenarios, this research isolates augmentation as the primary experimental variable and enables clearer attribution of performance differences.

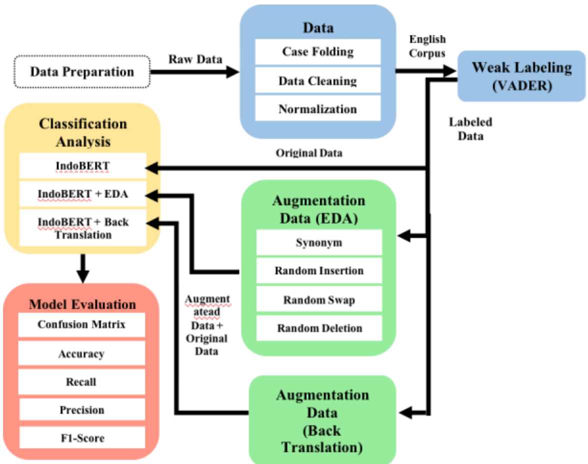
The objectives of this study are threefold. First, to construct a structured corpus of Indonesian hospital patient reviews using a weak labeling approach. Second, to fine-tune IndoBERT for binary sentiment classification under imbalanced conditions. Third, to empirically evaluate the comparative effectiveness of multiple text augmentation strategies in improving classification performance. The findings of this study are intended to provide empirical evidence within the specific dataset and experimental setting examined rather than to establish broad conclusions regarding robustness or generalizability across healthcare domains.

This study contributes to the field in three main aspects: (1) integrating a translation-based weak-labeling approach with IndoBERT for unlabeled Indonesian healthcare data; (2) providing a controlled empirical comparison of multiple token-level and back-translation augmentation strategies; and (3) comparing augmentation effectiveness against alternative class-imbalance handling techniques (e.g., class weighting and oversampling). By explicitly isolating augmentation effects, this study contributes methodological clarity to augmentation research while offering practical guidance for developing robust and data-driven patient feedback monitoring systems in healthcare institutions. This study hypothesizes that text augmentation techniques may improve classification performance under imbalanced conditions and that different augmentation techniques may exhibit varying levels of effectiveness.

MATERIALS AND METHODS

This study employed a quantitative experimental design to analyze the effect of text augmentation techniques on IndoBERT performance in classifying patient review sentiment. The overall research procedure consisted of data collection and preprocessing, weak sentiment labeling using a VADER Sentiment Analyzer, class imbalance handling through text augmentation, dataset splitting into training, validation, and testing sets, IndoBERT fine-tuning, and model performance evaluation using standard classification metrics.

During preprocessing, the text was normalized into standard Indonesian form using IndoNLP to improve consistency and reduce noise. Detailed preprocessing, translation, and weak labeling procedures are described in the Weak Labeling and Class Distribution subsection. The methodological design was informed by prior studies in healthcare sentiment analysis that employed Transformer-based architectures such as BERT and IndoBERT [5], [11], [12], lexicon-based weak labeling approaches including VADER [18], and text augmentation strategies such as Easy Data Augmentation and back-translation [14], [15], [16]. However, unlike previous studies that applied classification models directly to labeled datasets, this research integrates weak labeling and augmentation within a controlled experimental pipeline to isolate and evaluate their specific impact on model performance.



Source: (Research Results, 2026)
Figure 1. Research Method Flow

Data Collection and Selection

The dataset was obtained from the RSMH patient satisfaction survey database covering the period from December 2022 to October 2025. All entries containing textual reviews were extracted for analysis. Data cleaning involved removing empty responses, duplicate records, symbolic texts, and filtering out short texts containing fewer than four tokens to reduce noise and improve overall data quality [4], [9], [19]. The cleaned dataset was subsequently used for labeling and modeling processes.

Data Preprocessing

Preprocessing was conducted to enhance corpus consistency and improve model readiness [5], [7]. The procedure included case folding, removal of URLs, HTML tags, mentions, hashtags,

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8327

and non-alphabetic characters, slang normalization using the IndoNLP dictionary, reduction of repeated characters, and mapping of non-standard terms to standardized forms. Stopword removal and manual tokenization were intentionally omitted because IndoBERT utilizes an internal WordPiece-based tokenizer. Previous studies have reported that removing function words may negatively affect the performance of BERT-based architectures [20], [21], [22].

Weak Labeling and Class Distribution

Sentiment labeling was performed using VADER, a lexicon-based sentiment analysis tool that produces compound polarity scores ranging from −1 to +1. Lexicon-based approaches such as VADER are commonly used for automatic labeling in low-resource and domain-specific settings due to their efficiency and minimal annotation requirements [18]. VADER was specifically selected because it is computationally lightweight, publicly available, and widely adopted for weak label generation in low-resource sentiment analysis. In healthcare sentiment analysis, VADER has also been evaluated alongside deep learning models and remains effective for short informal texts [12]. The dataset consists of user-generated patient reviews, which often contain informal language, abbreviations, and non-standard expressions.

To enable weak labeling using VADER, the normalized Indonesian reviews were translated into English using the Helsinki-NLP translation model. During preprocessing, the text was normalized into standard Indonesian form by applying slang normalization and basic text cleaning using IndoNLP. This step aims to reduce informal variations, including abbreviations, slang expressions, and inconsistent writing styles commonly found in patient reviews. Since VADER is designed for English, all Indonesian-language reviews were translated into English using the Helsinki-NLP/opus-mt-id-en model, a transformer-based neural machine translation system from the OPUS-MT framework. The normalization process helps mitigate issues related to informal language, including slang and negation patterns, before translation. However, it is acknowledged that the translation process may introduce noise or subtle semantic shifts. Therefore, the resulting sentiment labels are treated as weak labels rather than ground truth annotations. This approach is commonly adopted in low-resource scenarios where manual annotation is limited.

Class mapping followed the standard VADER threshold configuration, where reviews with compound scores greater than or equal to 0.05 were labeled as positive, scores less than or equal to −0.05 were labeled as negative, and scores between these thresholds were considered neutral. For the main experiment, only positive and negative reviews were retained to form a binary classification setting. Binary sentiment classification is commonly adopted in limited and imbalanced data scenarios to improve model stability and interpretability [19]. To assess the validity of the weak labels, a stratified random sampling approach was applied to select 120 reviews from the dataset. This subset was manually annotated purely as a preliminary validation step to verify VADER's viability as an automatic labeler, rather than to serve as a ground-truth test set for the final model. This corresponds to approximately 8.2% of the dataset and was used to obtain a representative subset while maintaining feasibility for manual annotation, consistent with common practices in NLP validation studies. The sampling preserved the original class distribution, consisting of 96 positive and 24 negative reviews.

The selected reviews were manually annotated by three independent annotators without access to the VADER outputs to avoid bias. To establish a single ground-truth reference, the individual labels from the three annotators were aggregated using a majority voting approach to produce a 'Final Human Consensus' label. Subsequently, the agreement between the VADER-generated labels and the Final Human Consensus labels was evaluated using Cohen's Kappa. Because this procedure evaluates exactly two sets of ratings (VADER vs. Human Consensus), Cohen's Kappa is an appropriate statistic for measuring their agreement while accounting for chance agreement. Cohen's Kappa is commonly applied to assess the reliability and consistency of annotated datasets [19]. To further assess the reliability of the labeling process, the weak labels generated through translation were validated using a manually annotated subset of the dataset, as reported in the Results section. The resulting dataset exhibited a substantial imbalance between positive and negative classes, which motivated the application of data augmentation techniques to improve class balance and model performance.

Data Augmentation

To address class imbalance and limited data availability, augmentation was applied exclusively to the training dataset [14], [16]. Two augmentation approaches were implemented, namely Easy Data Augmentation and back-translation. The Easy Data Augmentation method consists of four token-level operations including synonym replacement, random insertion, random swap, and random deletion. Back-translation was performed using an Indonesian–English–Indonesian translation scheme to generate paraphrased variations while preserving semantic meaning. For synonym-based augmentation, lexical substitutions were performed using multiple resources, including a manually curated dictionary, the multilingual WordNet from WordNet, and a dataset-driven synonym dictionary constructed using similarity-based clustering. To maintain data quality, augmented samples were filtered based on minimum token length and duplicate removal. Back-translation was conducted using an Indonesian–English–Indonesian scheme with neural machine translation models from the Helsinki-NLP OPUS-MT framework to generate semantically equivalent variations.

The augmentation process was designed not only to increase data diversity but also to address class imbalance by generating additional samples for the minority class. The augmented samples were used to construct a balanced training dataset, resulting in an equal distribution between positive and negative classes. The validation and testing sets remained unchanged and consisted only of original data to prevent data leakage and ensure unbiased evaluation. In this study, the baseline model refers to IndoBERT trained on the original dataset without augmentation, while augmented datasets were used to evaluate the impact of text augmentation techniques. This experimental design follows a controlled setup in which the model architecture and training configuration are kept consistent, allowing performance differences to be attributed specifically to augmentation strategies.

This augmentation-based approach can be interpreted as a form of data-level imbalance handling, similar to oversampling, but with the advantage of introducing linguistic variation rather than simple duplication. The primary experiment focused on augmentation-based approaches. To provide a broader comparative evaluation, additional imbalance-handling strategies were subsequently included and evaluated using the same data partitions and evaluation procedures.

Class Imbalance Handling Strategies

In addition to text augmentation, several class imbalance handling strategies were evaluated to provide a broader comparison framework. Three approaches were considered: class weighting, random oversampling, and a TF-IDF + Support Vector Machine (SVM) baseline classifier. For class weighting, inverse-frequency class weights were incorporated into the cross-entropy loss function during IndoBERT fine-tuning, assigning greater importance to minority-class samples. For random oversampling, minority-class instances in the training set were duplicated until both sentiment classes contained equal numbers of samples. As a traditional machine-learning baseline, a TF-IDF + SVM classifier was implemented. Texts were represented using TF-IDF unigram and bigram features with a maximum vocabulary size of 10,000 terms and a minimum document frequency of 2. Classification was performed using a linear Support Vector Machine (SVM) with a regularization parameter of C = 1.0 and balanced class weighting to mitigate class imbalance. All methods were evaluated using the same training, validation, and testing partitions as the augmentation experiments to ensure a fair comparison.

Fine-tuning IndoBERT

The dataset was divided into training, validation, and testing sets using a stratified 70:15:15 ratio to maintain the original class distribution. The training set was used for model training and augmentation procedures, while the validation set was utilized for monitoring training performance and preventing overfitting. The testing set was reserved exclusively for final performance evaluation to ensure unbiased assessment. The testing set was completely isolated from all training, augmentation, hyperparameter tuning, and validation procedures and was used only once for the final performance evaluation. For the sentiment classification task, IndoBERT was selected as the base model because it has been pre-trained on large-scale Indonesian corpora and has demonstrated strong performance across various NLP tasks [5], [8].

Text representations were extracted from the final hidden state of the CLS token and passed to a fully connected layer with a softmax activation function for binary classification. The model was initialized using the pre-trained checkpoint “indobenchmark/indobert-base-p2”. Input sequences were tokenized using a maximum sequence length of 128 tokens. Sequences longer than this limit were truncated, while shorter sequences were padded to maintain a consistent input size. Training was conducted using the AdamW optimizer with a linear learning rate scheduler and warm-up strategy. A warm-up ratio of 0.1 was applied during training to stabilize optimization in the early training stages. To ensure reproducibility, a fixed random seed of 42 was applied across all experiments. Early stopping based on validation loss was implemented to prevent overfitting.

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8327

Hyperparameter optimization was performed through controlled experiments by varying the number of training epochs (5, 6, and 7), learning rates (2e-5, 3e-5, and 5e-5), and batch sizes (16 and 32). The optimal hyperparameters were selected using a grid-search approach based on validation F1-score, while keeping all other settings constant, to ensure balanced performance under class-imbalanced conditions. The selected optimal hyperparameter configuration was then applied consistently across all IndoBERT-based experiments, including the baseline, augmentation, class-weighting, and random oversampling scenarios. This design ensured that observed performance differences could be attributed primarily to the evaluated imbalance-handling strategy rather than variations in training settings.

Model Evaluation

Model performance was evaluated using a confusion matrix, accuracy, precision, recall, F1-macro, and Area Under the Curve (AUC). The evaluation was conducted on a separate test set that was not used during training or validation to ensure an unbiased assessment of model performance. The F1-macro score was used as the primary evaluation metric because it provides a balanced assessment across classes under class-imbalanced conditions [7], [19]. In addition, AUC was included to evaluate the model's discriminative ability. The analysis particularly emphasized the model's ability to correctly identify negative reviews, which are fewer in number but contain critical information for healthcare service quality improvement.

RESULTS AND DISCUSSION

Data Preparation

The preprocessing and weak labeling procedures described in the Materials and Methods section produced a final dataset of 1,464 labeled reviews, consisting of 1,168 positive and 296 negative instances.

Table 1. Distribution of Sentiment Labels

Stage	Total
Total survey entries	24,359
Entries containing review text	7,456
Cleaned reviews (short texts containing fewer than four tokens removed)	1,783
Positive (*weak labeling*)	1,168
Negative (*weak labeling*)	296
Neutral (*weak labeling*)	319

Source: (Research Results, 2026)

As shown in Table 1, positive reviews accounted for 79.8% of the dataset, while negative reviews accounted for 20.2%, indicating substantial class imbalance. This imbalance may bias the model toward the majority class and reduce sensitivity to minority-class complaints. Therefore, data augmentation techniques were applied to improve minority-class representation and support more balanced sentiment classification.

Preprocessing Results

Following the preprocessing procedures described in Section 2.3, the review texts were cleaned and normalized to ensure consistency prior to labeling and modeling. The process successfully removed noisy characters, standardized informal expressions, and corrected non-standard word forms commonly found in patient reviews. Table 2 presents examples of text transformation from the original form to the normalized version and its corresponding English translation used for weak labeling. The examples illustrate how preprocessing helps standardize informal expressions and improve the consistency of the translated output prior to sentiment labeling.

Table 2. Examples of Preprocessing Results

No	Original Text	Normalized Text	English Translation
1	WC lantai 2 kurang bersih	wc lantai 2 kurang bersih	2nd floor wc is not clean
2	Utk pendaftaran sulit secara online,TDK bisa sampai selesai.	untuk pendaftaran sulit secara online tidak bisa sampai selesai	For the hard registration online, it can't be completed.
3	Tuk ruang laboratoriumny baik	untuk ruang laboratoriumny baik	For the lab room. Good.
4	Pelayanan graha eksekutif sdh sangat baik	pelayanan graha eksekutif sudah sangat baik	Executive service has been excellent.

Source: (Research Results, 2026)

Sentiment Label Distribution

Examples of sentiment labels generated through the weak labeling procedure are presented in Table 3. It should be noted that the sentiment labels are derived from translated text and may be affected by minor translation distortions. However, this limitation is mitigated through preprocessing and supported by manual validation.

Table 3. Sample of Sentiment Labeling Results

No	Translated Text	Compound Score	Final Label
1	2nd floor wc is not clean	-0.3089	Negative
2	For the hard registration online, it can't be completed.	-0.1027	Negative
3	For the lab room. Good.	0.4404	Positive
4	Executive service has been excellent.	0.5719	Positive

Source: (Research Results, 2026)

To evaluate the reliability of the automatically generated labels, a manual validation experiment was conducted using a stratified random sample of 120 reviews. The comparison between the VADER-generated labels and the Final Human Consensus annotations showed that 102 out of 120 samples were consistent, resulting in an agreement accuracy of 85%. Furthermore, the agreement between the VADER labels and the human consensus was evaluated using Cohen's Kappa, yielding a score of 0.53, a value generally interpreted in statistical practice as indicating a moderate level of agreement. Cohen's Kappa is a widely used metric for assessing the reliability and consistency of annotated datasets [19]. While the validation results support the practical use of weak labeling for large-scale dataset construction, the moderate agreement level suggests the presence of labeling uncertainty. Consequently, some classification outcomes may be influenced by label noise introduced during the weak labeling process. A summary of the manual validation results is presented in Table 4.

Table 4. Manual Validation Results of Weak Labels

Metric	Value
Total samples	120
Agreement	102
Disagreement	18
Accuracy	85%
Cohen's Kappa	0.53

Source: (Research Results, 2026)

Representative examples of the comparison between VADER-generated labels and human annotations are presented in Table 5.

Table 5. Sample of Manual Annotation Comparison

Original Text	VADER	Annotator 1	Annotator 2	Annotator 3	Final
pelayanan sudah cukup baik	**Positive**	Positive	Positive	Positive	**Positive**
pengeras suara waktu pamanggilan pasien kurang maksimal	**Negative**	Positive	Negative	Negative	**Negative**
tolong kebersihan kamar mandi diperbaiki	**Positive**	Positive	Negative	Negative	**Negative**

Source: (Research Results, 2026)

Figure 2 illustrates the word cloud visualization of positive and negative reviews after preprocessing. In negative reviews, frequently occurring terms such as "tidak", "menunggu", "lama", and "parkir" indicate common issues related to waiting time, service delays, and facility limitations. In contrast, positive reviews are dominated by words such as "terima kasih", "baik", "ramah", and "cepat", reflecting patient satisfaction with service quality. Although the word cloud is primarily descriptive, it provides contextual insight into dominant terms within each sentiment class and supports the interpretation of classification results. In addition, it serves as a preliminary qualitative check to ensure that the extracted patterns are consistent with expected sentiment characteristics. These observations are aligned with the model's ability to distinguish between positive and negative sentiment, particularly in capturing key polarity-indicating terms.

Source: (Research Results, 2026)

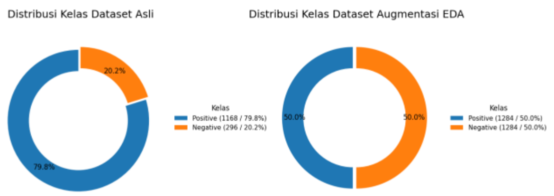
Figure 2. Word cloud of positive and negative reviews highlighting dominant terms associated with patient satisfaction and service-related complaints.

Figure 2 is included as an exploratory visualization to illustrate the vocabulary characteristics of positive and negative patient reviews. While it provides descriptive insight into sentiment-related terms, the primary conclusions of this study are derived from the quantitative evaluation results presented in the subsequent sections.

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)
VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8327

Data Augmentation

To address class imbalance in the base dataset (1,168 positive reviews and 296 negative reviews), text augmentation techniques were applied to the training set. Augmentation was primarily focused on the negative class until a balanced training distribution was achieved, resulting in 2,568 training examples consisting of 1,284 positive reviews and 1,284 negative reviews. The distribution before and after augmentation is shown in Figure 3.



Source: (Research Results, 2026)

Figure 3. Distribution of sentiment classes before and after augmentation, illustrating the reduction of class imbalance and improved minority-class representation.

The augmentation techniques used include Easy Data Augmentation (EDA) operations such as synonym replacement, random insertion, random swap, random deletion, and back-translation, as well as several hybrid combinations of these methods. Table 6 presents examples of sentence transformations generated using different augmentation techniques. It can be observed that the surface structure of the sentences changes across methods, while the core semantic meaning of the complaint is generally preserved. In some cases, particularly with token-level operations such as synonym replacement and random swap, the augmented sentences may appear less fluent or syntactically distorted. For example, synonym replacement may introduce contextually less appropriate words, while random swap may alter the natural word order of the sentence. Despite these variations, the overall sentiment expressed in the sentence remains consistent. Therefore, the original sentiment label can be retained for the augmented text. These variations are acceptable in the context of data augmentation, as they help increase data diversity and improve model robustness, which has been widely reported in previous studies [16], [17]. To mitigate excessive noise, basic filtering strategies such as minimum token length constraints and duplicate removal were applied.

Table 6. Examples of augmentation results based on the original sentence “untuk orang tua sulit untuk naik tangga”

No	Method	Augmented Text
1	*Synonym Replacement*	untuk orang tua sukar untuk naik tangga
2	*Random Insertion*	untuk orang tua banyaknya sulit untuk naik tangga
3	*Random Swap*	untuk orang tua sulit tangga naik untuk
4	*Random Deletion*	untuk tua sulit untuk naik tangga
5	*Back-Translation*	Untuk seorang pria tua sulit untuk menaiki tangga
6	SR + BT	Bagi orang tua, sulit untuk menaiki tangga
7	RI + BT	Bagi orang tua, sulit untuk menaiki tangga orang.
8	RS + BT	Untuk mendaki sulit untuk orang tua tangga
9	RD + BT	Bagi orang tua, tangga mendakinya sulit

Source: (Research Results, 2026)

The examples in Table 6 illustrate that augmentation techniques introduce different levels of lexical and structural variation. While certain transformations, particularly Synonym Replacement, can occasionally alter the contextual meaning, the core sentiment-bearing cues (e.g., *susah* / difficult) are generally preserved. In sentiment classification, these token-level perturbations may act as a form of regularization by encouraging the model to focus on sentiment-bearing cues rather than specific lexical items, thereby supporting the retention of the original sentiment labels despite minor semantic drift.

Fine-tuning Configuration

Before conducting the augmentation experiment, hyperparameter tuning was performed to determine the most suitable tuning configuration. Two model variants were evaluated: IndoBERT-base-p2 and IndoBERT-base-uncased, using a combination of epochs (5–7), learning rates (2e-5, 3e-5, and 5e-5), and batch sizes (16 and 32). The evaluation results are summarized in Table 7. Overall, IndoBERT-base-p2 consistently achieved higher accuracy and F1 scores than IndoBERT-base-uncased in most configurations. The best performance on the validation set was obtained using IndoBERT-base-p2 with 5 epochs, a learning rate of 5e-5, and a batch size of 16. This configuration was then adopted as the standard setting for all IndoBERT-based experimental scenarios, including the baseline, augmentation, class-weighting, and random oversampling experiments. Only the best-performing hyperparameter configuration identified for each IndoBERT variant is presented in Table 7 for brevity.

Table 7. Selected Hyperparameter Configuration for IndoBERT Models

Model	Epoch	Learning Rate	Batch Size	Accuracy	F1 Score
IndoBERT-base-p2	5	5e-5	16	0.8818	0.8797
IndoBERT-base-uncased	7	5e-5	16	0.8682	0.8519

Source: (Research Results, 2026)

For the IndoBERT-based experiments (baseline, augmentation, class weighting, and random oversampling), the same optimal hyperparameter configuration obtained from the tuning process was applied, consisting of 5 epochs, a learning rate of 5e-5, and a batch size of 16. Maintaining identical training settings ensured that performance differences could be attributed primarily to the evaluated imbalance-handling strategy rather than variations in model configuration. The TF-IDF + SVM baseline used TF-IDF unigram and bigram features with a linear Support Vector Machine (SVM) classifier and balanced class weighting.

Performance Comparison

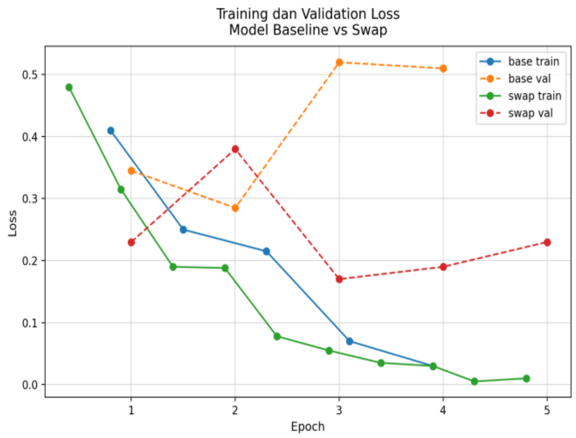
Overall, IndoBERT performance showed improvement under the experimental setup after the application of text augmentation. The baseline model achieved an accuracy of 0.8500 and an F1-macro of 0.7022, indicating limited balance in handling positive and negative classes. Among the evaluated augmentation strategies, the random swap technique achieved the highest overall performance, with an accuracy of 0.9682, F1-macro of 0.9530, precision-macro of 0.9371, recall-macro of 0.9717, and AUC of 0.9890, as summarized in Table 8. Random insertion and full EDA also demonstrated strong results, with F1-macro values above 0.92. In contrast, augmentation methods involving back-translation (BT and hybrid variants) consistently yielded lower performance compared to token-level EDA operations.

Table 8. Performance Comparison of Augmentation Techniques

Technique	Accuracy	F1-macro	Precision-macro	Recall-macro	AUC
Random Swap	**0.9682**	**0.9530**	**0.9371**	**0.9717**	**0.9890**
insertion	0.9636	0.9467	0.9284	0.9689	0.9902
eda	0.9500	0.9273	0.9075	0.9521	0.9860
synonym	0.9364	0.9067	0.8901	0.9270	0.9724
deletion	0.9273	0.8901	0.8840	0.8965	0.9736
eda_bt	0.8955	0.8352	0.8444	0.8270	0.9600
insertion_bt	0.8864	0.8296	0.8222	0.8378	0.9462
bt	0.8818	0.8014	0.8371	0.7771	0.9185
synonym_bt	0.8773	0.7917	0.8313	0.7660	0.9360
deletion_bt	0.8773	0.7830	0.8447	0.7495	0.9545
swap_bt	0.8636	0.7614	0.8128	0.7327	0.9335
base	0.8500	0.7022	0.8275	0.6663	0.9297

Source: (Research Results, 2026)

As shown in Figure 4, the baseline model exhibits a clear divergence between training and validation loss after the second epoch. While training loss continues to decrease steadily, validation loss increases sharply starting from epoch three, indicating overfitting. The widening gap between training and validation curves suggests that the model increasingly memorizes training data rather than generalizing to unseen data. In contrast, the random swap configuration shows more stable training dynamics. Both training and validation losses decrease more consistently, and the gap between them remains relatively small throughout the epochs. Although minor fluctuations are observed, the validation loss does not exhibit the sharp increase seen in the baseline model. This behavior suggests that augmentation via random swap may help improve model stability and reduce overfitting within the experimental setup.

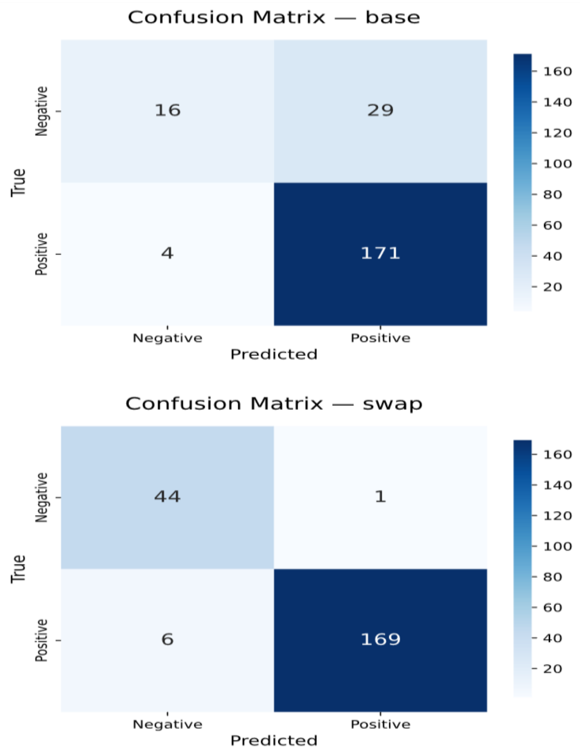


Source: (Research Results, 2026)
Figure 4. Training and Validation Loss Comparison.

The confusion matrix comparison in Figure 5 highlights improved minority-class detection after augmentation. In the baseline model, only 16 negative reviews were correctly classified, while 29 were misclassified as positive. After applying random swap augmentation, correctly classified negative reviews increased to 44, with only 1 misclassification. Meanwhile, positive class performance remained high (169 correct predictions and 6 misclassifications). These results show a substantial improvement in sensitivity toward negative reviews while maintaining strong performance for the positive class.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8327



Source: (Research Results, 2026)

Figure 5. Confusion Matrix Comparison Between the Baseline Model and the Random Swap Model.

While Figure 5 demonstrates the effect of random swap augmentation on minority-class detection, additional experiments were conducted using class weighting, random oversampling, and a TF-IDF + SVM baseline classifier to provide a broader evaluation beyond augmentation-based methods. All methods presented in Table 9 were evaluated using identical training, validation, and testing partitions, as well as the same evaluation protocol, to ensure a fair and consistent comparison.

Table 9. Comparative Performance of Class Imbalance Handling Strategies

Technique	Accuracy	F1-macro	Precision-macro	Recall-macro	AUC
Random Swap	**0.9682**	**0.9530**	**0.9371**	**0.9717**	**0.9890**
Class Weighting	0.8773	0.8031	0.8181	0.7908	0.9350
TF-IDF + SVM	0.8591	0.7779	0.7856	0.7711	0.9317
Random Oversampling	0.8591	0.7560	0.8009	0.7298	0.9314
Baseline	0.8500	0.7022	0.8275	0.6663	0.9297

Source: (Research Results, 2026)

As shown in Table 9, class weighting improved the F1-macro score from 0.7022 to 0.8031, indicating that adjusting the loss function partially mitigated the impact of class imbalance. Random oversampling produced only a modest improvement, achieving an F1-macro score of 0.7560. The TF-IDF + SVM baseline achieved an F1-macro score of 0.7779, demonstrating competitive performance despite its simpler architecture. Among all evaluated approaches, random swap augmentation achieved the highest performance, with an F1-macro score of 0.9530 and an AUC of 0.9890. Compared with class weighting, oversampling, and the TF-IDF + SVM baseline, random swap augmentation consistently produced superior overall classification performance and minority-class detection within the experimental setting of this study. Since all comparison methods were evaluated using identical training, validation, and testing partitions as well as the same evaluation protocol, the observed performance differences are more likely attributable to the evaluated imbalance-handling strategies rather than differences in experimental settings.

Discussion

The experimental results suggest that text augmentation is associated with improved transformer-based sentiment classification performance under imbalanced data conditions within the experimental setup of this study. The baseline model showed a tendency toward the majority class, particularly in detecting negative reviews. After augmentation was applied, especially through token-level strategies such as random swap and random insertion, class balance improved while maintaining overall performance. This improvement is particularly evident in minority-class detection, as reflected in the increase in F1-macro scores.

The comparison with alternative imbalance-handling methods presented in Table 9 further strengthens the interpretation of the augmentation results. Although class weighting and random oversampling improved performance relative to the baseline model, the magnitude of improvement remained substantially lower than that achieved through random swap augmentation. The TF-IDF + SVM baseline produced competitive results but was still outperformed by the augmentation-based IndoBERT configuration. These findings suggest that text augmentation may provide benefits beyond conventional imbalance-handling techniques by introducing greater lexical variation into the training data rather than merely adjusting class distributions.

The relatively strong performance of the Random Swap strategy suggests that lightweight token-level perturbations may introduce greater lexical variation while preserving sentiment-bearing keywords. Unlike synonym replacement, which can occasionally introduce semantic distortion or out-of-context nouns, random swap simply rearranges the existing sequence. This is less likely to introduce substantial semantic alterations and helps preserve the core sentiment-bearing cues and most of the original vocabulary remain intact, which likely explains its superior performance across the evaluated strategies (as seen in Table 8 and Table 9). Indonesian patient reviews frequently contain concise emotional expressions and informal language. These findings suggest that maintaining core polarity cues appears to be more beneficial than applying more complex transformations. The relatively lower performance of back-translation-based variants may be due to structural transformations that affect contextual sentiment signals captured by IndoBERT. This observation is consistent with recent findings that lightweight token-level augmentation strategies are associated with improved model performance in transformer-based text classification models [17].

These results are consistent with prior findings that simple augmentation methods can effectively enhance deep learning performance in low-resource scenarios. However, this study extends previous research by providing controlled empirical evidence within the Indonesian healthcare context, where linguistic variability and class imbalance pose unique challenges. Nevertheless, the results should be interpreted within the scope of this experimental setting, which involves weak labeling and a limited set of augmentation strategies. Methodologically, the use of consistent data partitions and hyperparameter configurations across all experimental scenarios strengthens internal validity and supports the interpretation that the observed performance differences are associated with the evaluated imbalance-handling strategies rather than differences in training settings.

From a practical perspective, improved detection of negative reviews may support healthcare service evaluation. Minority-class complaints often contain critical information regarding waiting time, facility conditions, and service responsiveness. By enhancing recall for negative sentiment, the proposed approach may contribute to more reliable patient feedback monitoring systems. Despite these findings, several limitations should be noted. The sentiment labels are generated using a translation-based weak labeling approach, which may introduce noise and affect the accuracy of the labels. The manual validation produced a Cohen's Kappa value of 0.53, indicating moderate agreement between weak labels and human annotations. Although this result supports the practical use of weak labeling for large-scale corpus construction, some degree of label noise is likely present. Consequently, the reported classification performance should be interpreted with consideration of potential labeling inaccuracies introduced by the weak-labeling process. Because the model evaluation relies on these weak labels rather than confirmed human ground-truth, the reported metrics primarily reflect the model's ability to approximate the noisy VADER distribution. Therefore, the strength of these findings lies in demonstrating the relative performance gains achieved by augmentation techniques over the baseline, rather than establishing absolute real-world classification benchmarks.

In addition, the experimental comparison remains limited to the methods evaluated in this study. Although class weighting, random oversampling, and a TF-IDF + SVM baseline were included, other imbalance-handling and classification approaches were not examined. Furthermore, the study was conducted using a single hospital review dataset and relied on weak labels generated through a translation-based procedure. Consequently, the reported performance improvements may be influenced by dataset-specific characteristics and labeling noise. The findings should therefore be interpreted within the scope of the evaluated methods and experimental conditions rather than as evidence of broad generalizability.

Future work should validate these findings using manually annotated datasets, external healthcare review datasets, and additional imbalance-handling approaches. Overall, the results suggest that token-level augmentation techniques may improve IndoBERT performance under imbalanced Indonesian patient review data within the defined experimental conditions. However, broader validation is required before general conclusions regarding robustness or applicability can be established. These findings provide methodological insights into augmentation research and may offer practical value for healthcare quality management.

CONCLUSION

This study evaluated the impact of text augmentation on IndoBERT-based sentiment classification of hospital patient reviews under limited and imbalanced data conditions. Within the evaluated experimental setting, augmentation-based approaches were generally associated with improved classification performance compared

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8327

with the baseline model, particularly in detecting minority-class negative reviews. The findings suggest that class imbalance reduced the baseline model's ability to detect minority-class negative reviews and that augmentation-based approaches improved this capability within the evaluated experimental setting. Among the evaluated strategies, random swap demonstrated relatively consistent improvements in classification balance and sensitivity toward negative reviews compared with back-translation-based approaches. The controlled experimental design suggests that the observed performance differences are associated with the evaluated augmentation strategies rather than variations in training configuration.

Additional comparison experiments demonstrated that random swap augmentation outperformed class weighting, random oversampling, and a TF-IDF + SVM baseline within the evaluated experimental setting. These results further support the potential effectiveness of augmentation-based approaches compared with the alternative imbalance-handling methods evaluated in this study. However, this study relies on weakly labeled data generated through a translation-based approach, which may introduce labeling noise. Because the test set was also labeled using the same automated procedure, the reported improvements should be interpreted as relative gains over the baseline within the evaluated experimental setting rather than as evidence of absolute human-level performance. Future work should validate these findings using manually annotated and external healthcare review datasets, compare additional imbalance-handling strategies, and incorporate statistical significance testing to further assess the robustness and generalizability of the proposed approach.

REFERENCE

[1] T. Zitek, J. Bui, C. Day, S. Ecoff, and B. Patel, “A cross-sectional analysis of Yelp and Google reviews of hospitals in the United States,” JACEP Open, vol. 4, no. 2, p. e12913, 2023, doi: 10.1002/emp2.12913.

[2] A. R. Pandey, M. Seify, U. Okonta, and A. Hosseinian-Far, “Advanced Sentiment Analysis for Managing and Improving Patient Experience: Application for General Practitioner (GP) Classification in Northamptonshire,” *Int. J. Environ. Res. Public Health*, vol. 20, no. 12, Jun. 2023, doi: 10.3390/ijerph20126119.

[3] M. P. Tse, I. Dhalla, and D. Nayyar, “Google star ratings of Canadian hospitals: a nationwide cross-sectional analysis,” BMJ Open Qual., vol. 13, no. 3, Jul. 2024, doi: 10.1136/bmjoq-2023-002713.

[4] I. Villanueva-Miranda, Y. Xie, and G. Xiao, “Sentiment analysis in public health: a systematic review of the current state, challenges, and future directions,” Front. Public Heal., vol. 13, p. 1609749, 2025, doi: 10.3389/fpubh.2025.1609749.

[5] H. Imaduddin, F. Y. A’la, and Y. S. Nugroho, “Sentiment Analysis in Indonesian Healthcare Applications using IndoBERT Approach,” Int. J. Adv. Comput. Sci. Appl., vol. 14, no. 8, pp. 113–117, 2023, doi: 10.14569/IJACSA.2023.0140813.

[6] O. S. Alkhnbashi, R. Mohammad, and M. Hammoudeh, “Aspect-Based Sentiment Analysis of Patient Feedback Using Large Language Models,” *Big Data Cogn. Comput.*, vol. 8, no. 12, Dec. 2024, doi: 10.3390/bdcc8120167.

[7] Y. Mao, Q. Liu, and Y. Zhang, “Sentiment analysis methods, applications, and challenges: A systematic literature review,” J. King Saud Univ. - Comput. Inf. Sci., vol. 36, no. 4, p. 102048, 2024, doi: 10.1016/j.jksuci.2024.102048.

[8] N. C. Mei, S. Tiun, and G. Sastria, “Multi-label aspect-sentiment classification on Indonesian cosmetic product reviews with IndoBERT model,” International Journal of Advanced Computer Science and Applications, vol. 15, no. 11, 2024, doi: 10.14569/IJACSA.2024.0151168.

[9] J. R. Jim, M. A. R. Talukder, P. Malakar, M. M. Kabir, K. Nur, and M. F. Mridha, “Recent advancements and challenges of NLP-based sentiment analysis: A state-of-the-art review,” *Nat. Lang. Process. J.*, vol. 6, p. 100059, Mar. 2024, doi: 10.1016/j.nlp.2024.100059.

[10] S. H. Park, C. P. Cheng, N. J. Buehler, T. Sanford, and W. Torrey, “A sentiment analysis on online psychiatrist reviews to identify clinical attributes of psychiatrists that shape the therapeutic alliance,” Front. Psychiatry, vol. 14, p. 1174154, 2023, doi: 10.3389/fpsyt.2023.1174154.

[11] L. Gandy, L. Ivanitskaya, L. Bacon, and R. Bizri-Baryak, “Public health discussions on social media: evaluating automated sentiment analysis methods,” JMIR Formative Research, vol. 9, Art. no. e57395, 2025, doi: 10.2196/57395.

[12] I. Aggarwal, S. Joseph, N. Jaganathan, A. Patel, V. Kumar, and M. Devarapalli, “Sentiment

Analysis in Healthcare: A Comparison of VADER, BERT, and Flair NLP Models on Patient Reviews of Pain Management Physicians," Cureus, vol. 17, no. 7, 2025, doi: 10.7759/cureus.88902.

[13] 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 *Proc. 28th Int. Conf. Comput. Linguist. (COLING)*, Barcelona, Spain, 2020, pp. 757–770, doi: 10.18653/v1/2020.coling-main.66.

[14] R. Kimera, D. N. Heo, D. N. Rim, and H. Choi, "Data Augmentation With Back translation for Low Resource languages: A case of English and Luganda," in NLPIR 2024 - 2024 8th International Conference on Natural Language Processing and Information Retrieval, 2025, pp. 142–148. doi: 10.1145/3711542.3711594.

[15] Natasya and A. S. Girsang, "Modified EDA and Backtranslation Augmentation in Deep Learning Models for Indonesian Aspect-Based Sentiment Analysis," Emerg. Sci. J., vol. 7, no. 1, pp. 256–272, 2023, doi: 10.28991/ESJ-2023-07-01-018.

[16] J. Wei and K. Zou, "EDA: Easy data augmentation techniques for boosting performance on text classification tasks," in *Proc. 2019 Conf. Empirical Methods Natural Lang. Process. and 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, Hong Kong, China, 2019, pp. 6382–6388, doi: 10.18653/v1/D19-1670.

[17] J. Chen, D. Tam, C. Raffel, M. Bansal, and D. Yang, "An Empirical Survey of Data Augmentation for Limited Data Learning in NLP," Trans. Assoc. Comput. Linguist., vol. 11, pp. 191–211, 2023, doi: 10.1162/tacl_a_00542.

[18] S. Biswas, K. Young, and J. Griffith, "A Comparison of Automatic Labelling Approaches for Sentiment Analysis," in *Proc. 14th Int. Joint Conf. on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2022) - KDIR*, Valletta, Malta, 2022, pp. 195–202, doi: 10.5220/0011265900003269.

[19] J. Opitz, "A Closer Look at Classification Evaluation Metrics and a Critical Reflection of Common Evaluation Practice," *Trans. Assoc. Comput. Linguist.*, vol. 12, pp. 820–836, 2024, doi: 10.1162/tacl_a_00675.

[20] M. Siino, I. Tinnirello, and M. La Cascia, "Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers," *Inf. Syst.*, vol. 121, Mar. 2024, doi: 10.1016/j.is.2023.102342.

[21] M. A. Palomino and F. Aider, "Evaluating the Effectiveness of Text Pre-Processing in Sentiment Analysis," *Appl. Sci.*, vol. 12, no. 17, Sep. 2022, doi: 10.3390/app12178765.

[22] S. Jahić and J. Vičič, "Impact of Negation and AnA-Words on Overall Sentiment Value of the Text Written in the Bosnian Language," *Appl. Sci.*, vol. 13, no. 13, Jul. 2023, doi: 10.3390/app13137760.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri