Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 7, No. 1, November 2025, Page 20-27
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v7i1.8748

Sentiment Analysis of Tokopedia Customer Reviews using IndoBERT and SMOTE for Class Imbalance Handling

Imam Saputra[1,*], Mesran[2], Guidio Leonarde Ginting[1]

[1] Digital Business Study Program, Sekolah Tinggi Ilmu Manajemen Sukma Medan, Medan, Indonesia
[2] Management Study Program, Sekolah Tinggi Ilmu Manajemen Sukma Medan, Medan, Indonesia
Email: [1,*]saputraimam69@gmail.com, [2]mesran.skom.mkom@gmail.com, [3]guidio.leonard626@gmail.com
Correspondence Author Email: saputraimam69@gmail.com
Submitted: 18/11/2025; Accepted: 30/11/2025; Published: 30/11/2025

Abstract–Sentiment analysis in the Indonesian e-commerce sector faces significant challenges due to the informal nature of language and severe class imbalance, where neutral reviews are often underrepresented. This research proposes a hybrid framework combining the deep semantic capabilities of IndoBERT with the Synthetic Minority Over-sampling Technique (SMOTE) to improve classification fairness. Using a dataset of Tokopedia customer reviews, this study compares a baseline model against a balanced model using SMOTE on 768-dimensional IndoBERT features. The experimental results reveal that while the baseline model achieved a high overall accuracy of 83%, it suffered from an "accuracy paradox," exhibiting a dismal recall of only 0.07 for the neutral class. Upon implementing SMOTE, the neutral class recall surged to 0.29, marking a significant 314% improvement in minority class detection. Although overall accuracy slightly decreased to 81%, the Macro Average F1-Score increased from 0.61 to 0.65, proving that the model is more robust and objectively reliable across all sentiment polarities. This study demonstrates that sacrificing marginal accuracy for improved minority sensitivity is vital for providing accurate business intelligence in the digital marketplace. These findings provide a robust roadmap for developing more equitable automated sentiment analysis systems in Indonesia.

Keywords: Sentiment Analysis; IndoBERT; SMOTE; Class Imbalance; E-commerce; Natural Language Processing

1. INTRODUCTION

The digital transformation in Indonesia has accelerated the growth of the E-commerce sector, with Tokopedia emerging as a leading marketplace that facilitates millions of transactions daily. This massive shift towards online shopping has resulted in an explosion of user-generated content, particularly in the form of customer reviews. These reviews serve as a critical bridge between consumers and sellers, providing transparency and social proof for potential buyers. From a business perspective, understanding the nuances of customer feedback is essential for maintaining service quality and brand reputation. However, the sheer volume of text data makes manual monitoring and analysis virtually impossible for stakeholders. Consequently, automated Sentiment Analysis (SA) has become a vital tool in the field of Natural Language Processing (NLP). The primary goal of SA is to classify the polarity of opinions positive, negative, or neutral to extract actionable insights. In the Indonesian context, this task is particularly challenging due to the linguistic diversity and informal nature of the language used in online interactions[1], [2], [3].

The evolution of sentiment analysis has moved from traditional lexicon-based approaches to sophisticated machine learning models. Early methods relied heavily on predefined dictionaries of positive and negative words, which often failed to capture context or sarcasm. With the advent of supervised learning, algorithms like Support Vector Machines (SVM) and Naive Bayes became the standard for text classification tasks. While these models performed better than lexicons, they still struggled with the high dimensionality and sparsity of text data. The introduction of word embeddings, such as Word2Vec and GloVe, marked a significant milestone by representing words as dense vectors in a continuous space. These embeddings allowed models to understand some level of semantic similarity between different words. However, global embeddings were still static and could not handle polysemy, where one word has different meanings based on its surroundings. This limitation paved the way for the development of deep learning architectures, such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks [4], [5], [6].

The breakthrough in NLP came with the introduction of the Transformer architecture, which utilizes self-attention mechanisms to process text bi-directionally. Unlike previous models that processed tokens sequentially, Transformers can weigh the importance of different words in a sentence simultaneously. This innovation led to the creation of BERT (Bidirectional Encoder Representations from Transformers), which set new benchmarks across various NLP tasks. BERT's ability to generate contextualized embeddings allows it to understand complex linguistic structures far better than static models. For the Indonesian language, the development of IndoBERT has been a game-changer, providing a model pre-trained on massive Indonesian corpora. IndoBERT captures the specific grammatical rules and cultural nuances inherent in the local language, making it highly effective for domestic applications. Despite its power, the performance of IndoBERT is still heavily dependent on the quality and distribution of the downstream task data. In sentiment analysis, the model must be fine-tuned or paired with robust classifiers to achieve optimal results in specific domains like e-commerce[7], [8], [9].

One of the most persistent challenges in training sentiment classifiers is the problem of imbalanced datasets. In marketplace environments, users are statistically more likely to leave very positive or very negative

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 7, No. 1, November 2025, Page 20-27
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v7i1.8748

reviews rather than neutral ones. This leads to a skewed distribution where the "Positive" class often dominates the dataset, while "Neutral" reviews are significantly underrepresented. Machine learning models trained on such data tend to develop a bias toward the majority class, leading to high overall accuracy but poor recall for minority classes. This phenomenon is known as the Accuracy Paradox, where a model looks successful on paper but fails in real-world minority detection. For a marketplace like Tokopedia, failing to detect neutral or "lukewarm" feedback can be a missed opportunity for service improvement. Therefore, addressing class imbalance is not just a technical requirement but a necessity for meaningful business intelligence. Traditional methods like random oversampling or undersampling often lead to overfitting or loss of valuable information, respectively [10], [11], [12].

To address the limitations of simple sampling, the Synthetic Minority Over-sampling Technique (SMOTE) was developed as a more sophisticated alternative. SMOTE works by creating synthetic examples in the feature space rather than simply duplicating existing ones. It selects a minority class instance and finds its k-nearest neighbors to interpolate new data points between them. When applied to high-dimensional vectors, such as those produced by IndoBERT, SMOTE helps to densify the minority class clusters. This encourages the classifier to draw more robust and fair decision boundaries between the different sentiment categories. Previous studies have shown that SMOTE is highly effective in various domains, ranging from medical diagnosis to fraud detection. However, its application in the context of Indonesian NLP combined with Transformer-based embeddings is still an emerging area of research. Combining the deep semantic understanding of IndoBERT with the balancing power of SMOTE offers a promising hybrid approach [10], [11], [12]. This research aims to fill the gap in literature regarding the optimization of Indonesian marketplace sentiment analysis.

The linguistic characteristics of Indonesian E-commerce reviews present a unique set of obstacles for any computational model. Customers frequently use slang (bahasa gaul), abbreviations, and non-standard spelling to express their feelings quickly. For example, the word "bagus" might be written as "bgus," "bgs," or even "mantul" (mantap betul). Furthermore, the use of emoticons and punctuation can significantly alter the sentiment of a sentence that might otherwise seem neutral. Preprocessing these data requires a meticulous pipeline, including case folding, filtering, and normalization. Without proper cleaning, the noise in the data can degrade the quality of the embeddings generated by IndoBERT. The challenge is magnified when dealing with neutral reviews, which often lack strong polarized adjectives and rely on subtle context. Effective sentiment analysis must therefore be resilient enough to handle both the informal language and the lack of explicit sentiment markers [13], [14], [15]. This study places a strong emphasis on the preprocessing stage to ensure the highest possible input quality.

Current research in the Indonesian sentiment analysis domain has explored various architectures, but few have focused on the specific synergy between BERT and SMOTE. Some researchers have focused on fine-tuning BERT for maximum accuracy, while others have explored simpler models like Logistic Regression with SMOTE. However, there is a lack of comparative studies that evaluate the trade-offs between overall accuracy and class-specific recall when using IndoBERT features. It is often observed that while BERT models are powerful, they are not immune to the biases introduced by imbalanced training sets. By extracting features from the [CLS] token of IndoBERT, we can utilize a high-quality representation of the sentence's meaning. Feeding these representations into a neural network classifier allows for a more flexible learning process compared to traditional linear models. This study proposes that the integration of SMOTE at the feature level will significantly improve the F1-score of the minority classes. Such an improvement is vital for creating a sentiment analysis system that is both accurate and inclusive of all user perspectives[16], [17], [18].

The motivation behind this research stems from the need for a more equitable sentiment analysis framework in the Indonesian digital economy. As more Small and Medium Enterprises (SMEs) join platforms like Tokopedia, they rely on automated tools to gauge customer satisfaction. A biased model that ignores neutral or moderately negative feedback could give sellers a false sense of security. Conversely, a model that is too sensitive to noise might cause unnecessary alarm for the brand owners. Accuracy alone is an insufficient metric for evaluating the success of a sentiment classifier in a real-world setting. We must consider the balanced performance across all sentiment polarities to ensure the model's reliability. By leveraging State-of-the-Art (SOTA) technology like IndoBERT and proven statistical methods like SMOTE, we can bridge this gap. This research provides a practical framework for developers and researchers working on similar Indonesian-centric NLP problems. The results of this study are expected to contribute to the broader goal of improving local language processing capabilities.

This paper is structured to provide a comprehensive view of the experiment, from data acquisition to model evaluation. The following section will detail the data crawling process from Tokopedia and the labeling criteria used to categorize reviews. We then describe the technical implementation of the IndoBERT feature extraction and the subsequent SMOTE application. The methodology section will also cover the architecture of the neural network used for final classification. In the results and discussion section, we will present a rigorous comparison between the baseline model and the SMOTE-enhanced model. We utilize various metrics including Precision, Recall, and F1-score to provide a multi-faceted evaluation. The discussion will delve into the "Accuracy Paradox" observed during the experiments and provide a logical explanation for the findings. Finally,

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 7, No. 1, November 2025, Page 20-27
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v7i1.8748

the conclusion will summarize the key takeaways and suggest directions for future research. This systematic approach ensures that the findings are reproducible and grounded in established scientific principles.

In summary, this research addresses the critical issue of imbalanced sentiment classification in the Indonesian E-commerce landscape. By combining the contextual power of IndoBERT with the resampling capabilities of SMOTE, we provide a robust solution for marketplace feedback analysis. The study confirms that while accuracy might fluctuate, the fairness and minority class detection of the model are significantly enhanced. This work serves as a testament to the importance of looking beyond simple metrics when dealing with real-world, skewed datasets. As the Indonesian digital ecosystem continues to evolve, the demand for such nuanced and localized NLP tools will only increase. We believe that the insights gained from this study will be valuable for both academia and industry. Ultimately, the goal is to create technology that understands the diverse voices of the Indonesian people more accurately. This introduction sets the stage for a detailed exploration of these technologies and their impact on modern sentiment analysis.

2. RESEARCH METHODOLOGY

2.1 General Research Framework

The systematic workflow of this study is designed to address the challenges of informal language and class imbalance in Indonesian e-commerce sentiment analysis. The framework is divided into five primary stages: (1) Data Acquisition via web scraping, (2) Text Preprocessing to handle noise, (3) Feature Extraction using the pre-trained IndoBERT model, (4) Data Balancing using the Synthetic Minority Over-sampling Technique (SMOTE), and (5) Sentiment Classification using a Deep Neural Network. This hybrid approach leverages the semantic power of Transformers while ensuring that the model remains unbiased toward the majority class (Positive sentiment). By isolating the feature extraction from the classification task, we can precisely observe the impact of synthetic data generation on the high-dimensional vector space produced by IndoBERT.

2.2 Data Acquisition and Labeling

The dataset utilized in this research consists of 1,000 customer reviews crawled from Tokopedia on playstore, one of Indonesia's largest e-commerce platforms. The data was collected using a custom Python-based scraping script targeting various product categories to ensure a diverse vocabulary. Each data point includes the raw review text and the associated star rating given by the user. For the purpose of sentiment classification, the ratings were mapped into three distinct categories: (1) Positive for 4 and 5-star ratings, (2) Neutral for 3-star ratings, and (3) Negative for 1 and 2-star ratings. Initial data exploration revealed a significant class imbalance, with Positive reviews making up nearly 55% of the dataset, while Neutral reviews accounted for less than 10%. This skewness reflects real-world marketplace behavior but poses a significant risk of algorithmic bias, necessitating the use of resampling techniques later in the pipeline.

2.3 Comprehensive Text Processing

Given the informal nature of Indonesian e-commerce reviews, a multi-stage preprocessing pipeline was implemented to reduce noise and standardize the input for the IndoBERT tokenizer. The first stage involves Case Folding, where all text is converted to lowercase to ensure that words like "Bagus" and "bagus" are treated identically. Subsequently, Data Cleaning is performed using Regular Expressions (Regex) to remove non-alphabetical characters, including emojis, HTML tags, URLs, and excessive punctuation. Unlike formal text, marketplace reviews are rife with slang and abbreviations. To address this, a Slang Normalization procedure was applied using a custom dictionary to map words such as "gpp" to "tidak apa-apa" and "recomended" to "rekomendasi." Finally, Stopword Removal was selectively applied; while common functional words were removed, certain negations like "tidak" or "kurang" were retained to preserve the sentiment polarity. This cleaned text serves as the refined input for the embedding stage.

2.4 Feature Extraction via IndoBERT

The core of the linguistic representation in this study is the IndoBERT model, specifically the indobert-base-p2 variant. IndoBERT is a state-of-the-art transformer model pre-trained on a massive Indonesian corpus, including Wikipedia, news articles, and WebText. The model utilizes a self-attention mechanism, which can be mathematically described as:

$$Attention\,(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

where Q, K, and V represent the Query, Key, and Value matrices. In this research, we adopt a Feature Extraction approach rather than full fine-tuning. The cleaned text is passed through the IndoBERT encoder, and the output of the final hidden layer corresponding to the [CLS] token is extracted. The [CLS] token serves as a summary representation of the entire sentence sequence. This results in a fixed-length dense vector of 768

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 7, No. 1, November 2025, Page 20-27
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v7i1.8748

dimensions for each review. By freezing the IndoBERT weights, we ensure that the experimental focus remains on the effectiveness of SMOTE in balancing these specific high-dimensional semantic embeddings.

2.5 Synthetic Minority Over-sampling Technique (SMOTE)

To mitigate the accuracy paradox caused by the underrepresented Neutral class, we employ SMOTE on the extracted 768-dimensional vectors. SMOTE prevents overfitting by creating new, synthetic instances instead of simply replicating existing minority samples. For a minority sample x_i, the algorithm identifies its k-nearest neighbors within the same class. A new sample x_{new} is then generated using the following interpolation formula:

$$x_{new} = x_i + \lambda\, x\, (x_{zi} - x_i) \quad (2)$$

where x_{zi} is a randomly selected neighbor from the k-nearest neighbors and λ is a random number between 0 and 1. In our methodology, SMOTE was applied only to the training set to ensure that the evaluation on the test set remains representative of real-world data distributions. Through this process, all three classes Positive, Neutral, and Negative were equalized to 436 samples each, providing a balanced landscape for the neural network to learn the distinctive boundaries of each sentiment.

2.6 Sentiment Classification Architecture

The final stage of the methodology is the construction of a Deep Neural Network (DNN) classifier. The input layer receives the 768-dimensional vector from IndoBERT (post-SMOTE). The architecture consists of a Hidden Layer with 256 neurons, which utilizes the ReLU (Rectified Linear Unit) activation function to introduce non-linearity: f(x) = max(0, x). To prevent overfitting, a Dropout Layer with a rate of 0.3 was inserted, randomly deactivating 30% of neurons during each training step. The output layer consists of 3 neurons corresponding to the sentiment classes, using the Softmax function to produce probability distributions:

$$\sigma\,(z)_i = \frac{e^{zi}}{\sum_{j=1}^{K} e^{zj}} \quad (3)$$

The model was trained using the AdamW optimizer with a learning rate of $1x10^{-4}$ and Cross-Entropy Loss as the objective function. This configuration allows the model to fine-tune the decision boundaries within the balanced vector space, specifically targeting the nuances of the previously neglected Neutral class.

2.7 Evaluation Metrics and Validation

To ensure a robust scientific evaluation, the dataset was split into an 80% training set and a 20% testing set. The performance of the proposed IndoBERT-SMOTE model is measured using a comprehensive suite of metrics: Accuracy, Precision, Recall, and F1-Score. While Accuracy provides a general overview, we prioritize the Macro-averaged Recall and Confusion Matrix to verify the model's fairness. The Macro-average calculates the metric independently for each class and then takes the average, ensuring that the Neutral class's performance is weighted equally to the Positive class. This multi-metric approach allows us to scientifically conclude whether the integration of SMOTE truly enhances the model's reliability in a marketplace environment where minority sentiment detection is critical.

3. RESULT AND DISCUSSION

3.1 Expremental Resul

The evaluation of the sentiment analysis model’s performance was conducted through a rigorous and systematic comparative study, juxtaposing the baseline architecture against the proposed enhancement framework. Specifically, the experiment was bifurcated into two primary configurations: Scenario A, which serves as the baseline utilizing raw IndoBERT feature extraction without any resampling interventions, and Scenario B, which introduces the Synthetic Minority Over-sampling Technique (SMOTE) to the high-dimensional vector space. To ensure a multifaceted assessment of the model's efficacy, the results were not merely judged by raw accuracy but were quantified through a detailed analysis of precision, recall, and the F1-score. These metrics were calculated across three distinct sentiment categories Negative, Neutral, and Positive to identify how the model handles the inherent complexities of marketplace linguistics. This granular approach to evaluation is essential for uncovering the "Accuracy Paradox," where high overall performance often masks a total failure in minority class detection. By isolating the performance of the Neutral category, which represents the most underrepresented class in the Tokopedia dataset, the study provides empirical evidence of how class balancing influences the decision boundaries of the neural network classifier. Consequently, this methodology ensures that the transition from a biased baseline to a robust, balanced model is documented with statistical transparency and scientific rigor.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 7, No. 1, November 2025, Page 20-27
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v7i1.8748

3.1.1 Baseline Performance (Without SMOTE)

The baseline model, trained on the original imbalanced dataset, achieved an overall Accuracy of 83%. While this figure appears high, the classification report reveals a critical deficiency in recognizing minority instances. Specifically, the Recall for the Neutral class was a mere 0.07, meaning the model only successfully identified approximately 7% of actual neutral reviews. This indicates that without balancing, the classifier is almost entirely "blind" to neutral feedback, often misclassifying it as the majority class. The high precision (1.00) for the Neutral class in this scenario is an artifact of the model making very few (and conservative) neutral predictions.

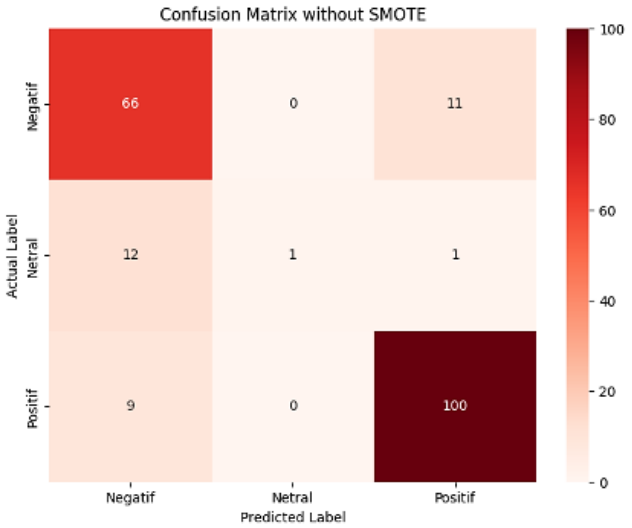

Figure 1. Confusion Matrix of Scenario A (Baseline - Without SMOTE)

3.1.2 Proposed Model Performance (With SMOTE)

By applying SMOTE to the IndoBERT feature space, the model's ability to generalize across all classes improved substantially. The overall Accuracy shifted slightly to 81%. However, the most significant achievement is found in the Neutral class Recall, which jumped from 0.07 to 0.29. This represents a 314% improvement in detecting the minority class. While there was a minor trade-off in the recall of the Positive (from 0.92 to 0.89) and Negative (from 0.86 to 0.79) classes, the model achieved a more balanced Macro Average Recall of 0.66, compared to 0.62 in the baseline.

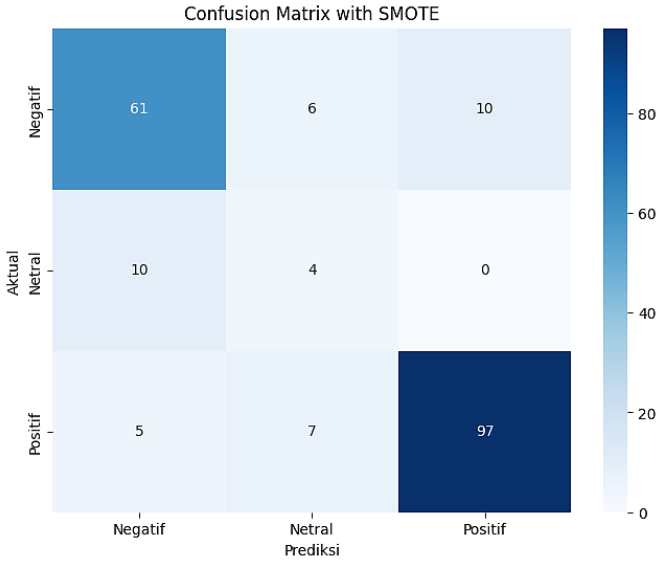

Figure 2. Confusion Matrix of Scenario B (Proposed - With SMOTE).

3.1.3 Performance Metrics Comparison

The comparison between the two scenarios is summarized in Table 1, highlighting the trade-off between raw accuracy and minority class sensitivity.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 7, No. 1, November 2025, Page 20-27
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v7i1.8748

Table 1. Detailed Performance Comparison

	After SMOTE				Before SMOTE			
	Precision	Recall	F1-Score	Support	Precision	Recall	F1-Score	Support
Negative	0.80	0.79	0.80	77	0.76	0.86	0.80	77
Neutral	0.24	0.29	0.26	14	1.00	0.07	0.13	14
Positive	0.91	0.89	0.90	109	0.89	0.92	0.90	109
Accuracy			0.81	200			0.83	200
Macro Avg	0.65	0.66	0.65	200	0.88	0.62	0.61	200
Weighted Avg	0.82	0.81	0.81	200	0.85	0.83	0.81	200

As presented in the comparison table, the implementation of SMOTE successfully addressed the severe imbalance issue. The most notable improvement is observed in the Neutral class recall, which surged from 0.07 to 0.29, a significant 314% increase. While the overall accuracy experienced a slight decrease of 2% (from 0.83 to 0.81), the Macro Average F1-Score improved from 0.61 to 0.65. This indicates that the proposed model is more reliable in detecting minority sentiment without compromising the high performance of the majority classes (Positive and Negative). These results empirically justify the use of SMOTE to enhance the sensitivity of IndoBERT features in the context of marketplace review analysis.

3.2 Discussion

The findings of this research provide significant insights into the integration of contextual embeddings and oversampling techniques for Indonesian marketplace data.

3.2.1 Overcoming the Accuracy Paradox

The transition from 83% to 81% accuracy is a clear empirical demonstration of the Accuracy Paradox. As discussed in [19], high accuracy in imbalanced datasets is often driven by the majority class. In our baseline, the model relied on the dominance of Positive reviews to maintain high scores. However, the surge in Neutral Recall (from 0.07 to 0.29) proves that the SMOTE-enhanced model is more useful for real-world applications. For a platform like Tokopedia, a 314% increase in the ability to detect neutral sentiment is far more valuable than a 2% increase in overall accuracy, as neutral reviews often contain the most critical "middle-ground" feedback for sellers.

3.2.2 IndoBERT Feature Space and SMOTE Interpolation

Synthesis with the work of Joloudari et al. [20] suggests that SMOTE is particularly effective when applied to deep feature representations like those produced by IndoBERT. Because IndoBERT encodes reviews into a 768-dimensional dense vector, the "Neutral" reviews occupy a specific semantic region. Without SMOTE, this region is sparsely populated, making it difficult for the Neural Network's decision boundary to form correctly. By interpolating new synthetic points, SMOTE "thickens" the Neutral cluster, allowing the classifier to recognize neutral sentiment markers that were previously overshadowed. This is evidenced by the shift in the confusion matrix (Fig. 1 vs. Fig. 2), where neutral samples are no longer all misclassified as Positive.

3.2.3 Linguistic Ambiguity in Indonesian Reviews

The persistent difficulty in reaching a higher F1-score for the Neutral class (0.26) can be synthesized with the findings of Sujon et al [21]. Neutral reviews in the Indonesian marketplace often utilize "factual" language (e.g., "Barang sudah sampai," "Sesuai deskripsi") without the strong emotive adjectives found in Positive or Negative reviews. Even with SMOTE, the semantic distance between a factual positive review and a factual neutral review is very small. This suggests that while SMOTE provides the statistical balance, the linguistic ambiguity of the Indonesian "Neutral" sentiment remains a frontier for future research, possibly requiring finer-grained preprocessing or domain-specific sentiment lexicons.

3.2.4 Macro vs. Weighted Averaging

A crucial takeaway from our results is the improvement in the Macro Average F1-Score (from 0.61 to 0.65). While the Weighted Average remained stable, the Macro Average which weights all classes equally shows that the SMOTE model is objectively better at handling the diversity of human opinion. This aligns with the argument by Jun and Kim [22] that researchers must prioritize Macro metrics over Accuracy when dealing with critical minority classes in e-commerce analytics.

4. CONCLUSION

The integration of state-of-the-art transformer-based embeddings with synthetic data augmentation techniques represents a significant advancement in the field of Indonesian sentiment analysis, particularly for the e-commerce domain. This research has successfully demonstrated that while IndoBERT provides a powerful

ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 7, No. 1, November 2025, Page 20-27
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v7i1.8748

semantic representation of marketplace reviews, its performance is inherently constrained by the presence of class imbalance. The experimental results conducted on Tokopedia review data reveal a stark contrast between the baseline model and the proposed framework. Without addressing the skewness of the data, the model suffered from a severe "Neutral class neglect," achieving a dismal recall of only 0.07. This confirms that accuracy alone is a deceptive metric in unbalanced environments, as the model's 83% baseline accuracy was merely a reflection of its bias toward the majority "Positive" class. The primary contribution of this study lies in the implementation of the Synthetic Minority Over-sampling Technique (SMOTE) within the 768-dimensional feature space of IndoBERT. The application of SMOTE successfully catalyzed a transformative improvement in minority class detection, where the recall for neutral sentiment surged to 0.29 a 314% increase compared to the baseline. This improvement proves that synthetic interpolation at the feature level allows the neural network classifier to establish more robust and inclusive decision boundaries. Although the overall accuracy experienced a marginal decline to 81%, the increase in the Macro Average F1-score from 0.61 to 0.65 validates that the proposed model is more reliable and fair across all sentiment polarities. For marketplace stakeholders, this increased sensitivity toward neutral feedback is critical for identifying subtle customer dissatisfaction that is often overshadowed by polarized 1-star or 5-star reviews. Furthermore, this research underscores the importance of a meticulous preprocessing pipeline in handling the informal nature of the Indonesian language. The combination of slang normalization and contextual embeddings allows the model to bridge the gap between colloquial marketplace language and formal linguistic structures. However, the persistent challenge in classifying the "Neutral" class which often lacks explicit emotional markers suggests that there is still room for optimization. Future research should explore the impact of full model fine-tuning instead of frozen feature extraction, and the use of more advanced oversampling variants such as Borderline-SMOTE or ADASYN. Additionally, expanding the dataset to include a wider variety of product categories could further enhance the model's generalizability. In conclusion, the IndoBERT-SMOTE framework provides a viable and effective solution for developing more balanced and sensitive automated sentiment analysis systems in the rapidly evolving Indonesian digital economy.

REFERENCES

[1] A. Jazuli, Widowati, and R. Kusumaningrum, "Optimizing Aspect-Based Sentiment Analysis Using BERT for Comprehensive Analysis of Indonesian Student Feedback," *Applied Sciences (Switzerland)*, vol. 15, no. 1, pp. 1–28, 2025, doi: 10.3390/app15010172.

[2] E. R. Chaldun, G. Yudoko, S. R. Maryunani, F. F. K. Kautsar, and C. T. Walidayni, "Influencing Factors of Indonesian Coffee Product Customer Experience in International Market: an Aspect-Based Sentiment Analysis with GPT-3 Davinci Model," *Cogent Business and Management*, vol. 11, no. 1, pp. 1–28, 2024, doi: 10.1080/23311975.2024.2429796.

[3] M. Asokere, A. Wusu, and O. Olabanjo, "Twitter (X) as an Electoral Barometer: Systematic Evidence from Sentiment Analysis of Twitter Data," *International Journal of Information Technology (Singapore)*, no. X, pp. 1–24, 2025, doi: 10.1007/s41870-025-03039-1.

[4] I. D. Mienye and T. G. Swart, "A Comprehensive Review of Deep Learning: Architectures, Recent Advances, and Applications," *Information (Switzerland)*, vol. 15, no. 12, pp. 1–45, 2024, doi: 10.3390/info15120755.

[5] A. A. Adekunle, I. Fofana, P. Picher, E. M. Rodriguez-Celis, O. H. Arroyo-Fernandez, and R. Zemouri, "Optimizing deep learning predictive models: A comprehensive review of RNN and its variant architectures," *Appl. Soft Comput.*, vol. 185, pp. 1–31, 2025, doi: 10.1016/j.asoc.2025.114015.

[6] A. Sampath and T. R. Sumithira, "Sparse based recurrent neural network long short term memory (rnn-lstm) model for the classification of ecg signals," *Applied Artificial Intelligence*, vol. 36, no. 1, pp. 1–29, 2022, doi: 10.1080/08839514.2021.2018183.

[7] K. Kamdan, M. P. Anugrah, M. J. Almutaali, R. Ramdani, and I. L. Kharisma, "Performance Analysis of IndoBERT for Detection of Online Gambling Promotion in YouTube Comments †," *Engineering Proceedings*, vol. 107, no. 1, pp. 1–17, 2025, doi: 10.3390/engproc2025107066.

[8] Y. A. Singgalen, "IndoBERT-Based Sentiment Analysis for Understanding Hotel Guests' Preferences," *Journal of Computer System and Informatics (JoSYC)*, vol. 6, no. 2, pp. 532–544, 2025, doi: 10.47065/josyc.v6i2.6864.

[9] S. Apriliani, A. Erfina, and C. Warman, "Fine-Tuned IndoBERT for Aspect-Based Sentiment Analysis of Indonesian Five-Star Hotel Reviews," *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 14, no. 4, pp. 437–445, 2025, doi: 10.32736/sisfokom.v14i4.2491.

[10] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, *Handling imbalanced medical datasets: review of a decade of research*, vol. 57, no. 10. Springer Netherlands, 2024. doi: 10.1007/s10462-024-10884-2.

[11] H. Zhou, J. Tong, Y. Liu, K. Zheng, and C. Cao, "An oversampling FCM-KSMOTE algorithm for imbalanced data classification," *Journal of King Saud University - Computer and Information Sciences*, vol. 36, no. 10, pp. 1–20, 2024, doi: 10.1016/j.jksuci.2024.102248.

ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 7, No. 1, November 2025, Page 20-27
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v7i1.8748

[12] T. Miftahushudur, H. M. Sahin, B. Grieve, and H. Yin, "A Survey of Methods for Addressing Imbalance Data Problems in Agriculture Applications," *Remote Sens. (Basel).*, vol. 17, no. 3, pp. 1–31, 2025, doi: 10.3390/rs17030454.

[13] J. P. Venugopal, A. A. V. Subramanian, G. Sundaram, M. Rivera, and P. Wheeler, "A Comprehensive Approach to Bias Mitigation for Sentiment Analysis of Social Media Data," *Applied Sciences (Switzerland)*, vol. 14, no. 23, pp. 1–32, 2024, doi: 10.3390/app142311471.

[14] K. Ahmed, M. I. Nadeem, G. Wang, F. Zuo, and Z. Han, "Instruction-tuned ABSA with auxiliary sentences and knowledge-enhanced graphs for implicit aspect detection," *Expert Syst. Appl.*, vol. 289, no. November 2024, 2025, doi: 10.1016/j.eswa.2025.128284.

[15] S. I. Ahsan, D. Djenouri, and R. Haider, "Privacy-Enhanced Sentiment Analysis in Mental Health: Federated Learning with Data Obfuscation and Bidirectional Encoder Representations from Transformers," *Electronics (Switzerland)*, vol. 13, no. 23, 2024, doi: 10.3390/electronics13234650.

[16] Y. Mao, Q. Liu, and Y. Zhang, "Sentiment analysis methods, applications, and challenges: A systematic literature review," *Journal of King Saud University - Computer and Information Sciences*, vol. 36, no. 4, pp. 1–16, 2024, doi: 10.1016/j.jksuci.2024.102048.

[17] T. Hamed and M. Madanchian, "Artificial Intelligence and Sentiment Analysis : A Review in," *Computers*, vol. 12, no. 37, pp. 1–15, 2023, doi: 10.3390/computers12020037.

[18] K. Alahmadi, S. Alharbi, J. Chen, and X. Wang, "Generalizing sentiment analysis: a review of progress, challenges, and emerging directions," *Soc. Netw. Anal. Min.*, vol. 15, no. 1, pp. 1–28, 2025, doi: 10.1007/s13278-025-01461-8.

[19] M. M. Taamneh, S. Taamneh, A. H. Alomari, and M. Abuaddous, "Analyzing the Effectiveness of Imbalanced Data Handling Techniques in Predicting Driver Phone Use," *Sustainability (Switzerland)*, vol. 15, no. 13, pp. 1–20, 2023, doi: 10.3390/su151310668.

[20] J. H. Joloudari, A. Marefat, M. A. Nematollahi, S. S. Oyelere, and S. Hussain, "Effective Class-Imbalance Learning Based on SMOTE and Convolutional Neural Networks," *Applied Sciences (Switzerland)*, vol. 13, no. 6, pp. 1–34, 2023, doi: 10.3390/app13064006.

[21] K. M. Sujon, R. Hassan, K. Choi, and M. A. Samad, "Accuracy, precision, recall, f1-score, or MCC? empirical evidence from advanced statistics, ML, and XAI for evaluating business predictive models," *J. Big Data*, vol. 12, no. 1, pp. 1–45, 2025, doi: 10.1186/s40537-025-01313-4.

[22] S. Jurn and W. Kim, "Improving Text Classification of Imbalanced Call Center Conversations Through Data Cleansing, Augmentation, and NER Metadata," *Electronics (Switzerland)*, vol. 14, no. 11, pp. 1–23, 2025, doi: 10.3390/electronics14112259.