turnitin_Arif_Fitra-1749607697741

by Turnitin Checker

Submission date: 10-Jun-2025 10:09PM (UTC-0400)
Submission ID: 2696324670
File name: turnitin_Arif_Fitra-1749607697741.docx (165.2K)
Word count: 3253
Character count: 21066

OPTIMIZING GENERATIVE PRE-TRAINED TRANSFORMER AND INDOBERT MODELS FOR SENTIMENT ANALYSIS AND CONSUMER TREND PREDICTION BASED ON PRODUCT REVIEWS ON LAZADA

Arif Fitra Setyawan[1], Rozaq Isnaini Nugraha[2]

[1]Universitas Widya Husada Semarang
[2] Universitas Widya Husada Semarang
*Email: [1]ariffitra.setyawan@gmail.com, [2] rozaqin@uwhs.ac.id

Abstract

The rapid growth of e-commerce platforms such as Lazada has fostered a dynamic digital ecosystem where consumer reviews play a critical role in purchasing decisions and product feedback. However, the high volume, unstructured nature, and linguistic diversity of user-generated content pose significant challenges for manual analysis. This study aims to analyze consumer sentiment in product reviews on Lazada Indonesia using two Transformer-based models tailored for the Indonesian language: IndoBERT and GPT. A quantitative experimental approach was adopted, utilizing a dataset of 12,704 user reviews categorized into positive, neutral, and negative sentiments. Following text preprocessing and data balancing with SMOTE, IndoBERT was fine-tuned in a supervised learning setup and compared with GPT in a zero-shot classification scenario. Evaluation results demonstrate that IndoBERT significantly outperforms GPT, achieving an average F1-score of 0.87, compared to GPT's 0.44. Further analysis using confusion matrices and ROC-AUC metrics reinforces IndoBERT's strength in capturing local language context and handling class imbalance, particularly in identifying positive sentiment. These findings highlight the importance of domain-specific fine-tuning and data preprocessing strategies for enhancing sentiment classification performance in the Indonesian language. The proposed model can be effectively applied to real-time market trend prediction, customer satisfaction monitoring, and personalized recommendation systems in the e-commerce landscape.

Keywords: *E-commerce, GPT, IndoBERT, Natural Language Processing, Sentiment Analysis*

**Corresponding Author: Arif Fitra Setyawan*

1. INTRODUCTION

The acceleration of digital transformation has significantly driven the growth of e-commerce platforms such as Lazada, which have transitioned from conventional online marketplaces into interactive environments where consumers articulate their experiences, satisfaction levels, and dissatisfaction regarding products and services. Within this ecosystem, user-generated reviews have emerged as a key component—playing a strategic role in influencing consumer purchasing decisions and serving as critical feedback mechanisms for vendors. Nevertheless, the high volume, unstructured nature, and linguistic variability of textual reviews present substantial challenges for manual data analysis and interpretation.

The advancement of digital technologies has profoundly transformed consumer behavior, particularly in the way individuals express their opinions about products. In the e-commerce landscape, this shift has made customer reviews an integral part of the purchasing decision-making process. Platforms such as Lazada exemplify this transformation by hosting millions of user-generated reviews, which not only reflect satisfaction and dissatisfaction but also reveal deeper insights into consumer expectations. These textual reviews have become a rich source of data that can be leveraged to better understand market trends and user preferences.

Sentiment analysis, also known as opinion mining, is a computational study aimed at identifying and extracting opinions, sentiments, evaluations, attitudes, emotions, subjectivity, and viewpoints expressed in textual data [1]. It involves processing and analyzing textual information to detect underlying emotional tone or polarity embedded within the content [2]. The primary objective of sentiment analysis is to uncover and interpret the subjective opinion conveyed in a given piece of text [3].

Recent advancements in artificial intelligence, particularly in the field of Natural Language Processing (NLP), have led to significant breakthroughs in the past few years [4]. NLP enables the automated analysis of linguistic data derived from text, allowing the extraction of sentiment-related features. In the context of social media, for instance, NLP techniques can be utilized to detect emotional

cues in user-generated content, which may serve as indicators of an individual's mental state [5]. The evolution of NLP has been marked by the emergence of sophisticated models such as LLaMA and GPT-4, which have redefined the paradigm of language processing through deeper contextual understanding and improved adaptability across various tasks [6]. Among these innovations, Google's Bidirectional Encoder Representations from Transformers (BERT) has initiated a new era in NLP by enabling bidirectional contextual learning and significantly enhancing performance in a wide range of language-based applications.

One of the most significant innovations in NLP is the introduction of Transformer-based models, which enable more efficient and context-aware text analysis. Opinion evaluation methods, such as multi-class classification in sentiment analysis, facilitate the identification and categorization of user sentiments—commonly applied in areas such as usability evaluation for mobile applications [7]. Among the most widely adopted implementations of the Transformer architecture is the Bidirectional Encoder Representations from Transformers (BERT). BERT is capable of capturing bidirectional context within a sentence, meaning it interprets the meaning of a word based on both its preceding and succeeding words. This bidirectional contextual understanding makes BERT highly effective for tasks such as text classification, sentiment analysis, and named entity recognition [8].

The application of models such as BERT in sentiment analysis offers deeper insights into the opinions and responses embedded within textual data. This capability is highly valuable for various purposes, including academic research, product development, and public opinion assessment [9]. BERT provides a distinct advantage through its ability to capture bidirectional contextual information, and it has demonstrated exceptional performance across a wide range of NLP tasks, particularly in sentiment classification [10]. When fine-tuned, BERT consistently outperforms traditional and state-of-the-art approaches in sentiment analysis tasks [11].

Several studies have demonstrated that the IndoBERT model delivers high performance across various Indonesian-language NLP tasks. For instance, research by Andi Aljabar et al. employed the BERT (Bidirectional Encoder Representations from Transformers) model to conduct sentiment analysis on movie reviews collected from IMDb. Utilizing a dataset of 5,000 comments, the fine-tuned model achieved strong results, with an average accuracy of 96%, validation accuracy of 89%, training loss of 10%, and validation loss of 37%. These findings confirm the effectiveness of BERT in accurately interpreting and classifying sentiments within film review texts [12]. Similarly, a study by Amru Hidayat and Nastiti [13] analyzed user sentiments toward the TikTok Tokopedia Seller Center application, which integrates e-commerce and social media functionalities. A total of 3,145 Indonesian-language reviews were extracted from the Google Play Store and manually labeled into positive and negative sentiment classes. After text preprocessing, the data were trained using two IndoBERT variants: Indobert-base-p2 and Indobert-lite-base-p2. The results showed that Indobert-base-p2 outperformed its lightweight counterpart, achieving 97% in accuracy, precision, recall, and F1-score—compared to 94% across the same metrics for Indobert-lite-base-p2. Moreover, a study by Khairani et al.[14] compared the performance of IndoBERT and IndoBERTweet in emotion classification tasks based on user comments on Instagram news posts. The emotion categories included anger, happiness, fear, and sadness. The study also explored the impact of preprocessing steps, particularly stopword removal and stemming. Interestingly, the findings revealed that models without preprocessing yielded better performance, with IndoBERTweet achieving the highest accuracy 92.54%, while IndoBERT achieved 88.81%. In another study conducted by Setyawan et al. [15], the evaluation results indicated that BERT achieved a high classification accuracy of 93.9%, with a balanced F1-score across precision and recall. The confusion matrix analysis further confirmed model's robustness in consistently identifying both positive and negative sentiments.

In addition to BERT, the Generative Pre-trained Transformer (GPT) has also emerged as a promising alternative for sentiment analysis. GPT is a generative Transformer-based model designed to comprehend long-range context and produce coherent textual outputs. A study conducted by Rahayu et al. [16] demonstrated that AicoGPT—a derivative of GPT-3—achieved an accuracy of up to 92% in sentiment classification tasks involving application reviews, utilizing TF-IDF and SVM-based techniques. The application of Indonesian-language Transformer models is not limited to the e-commerce domain; it has also been explored in various other contexts such as politics, film, social media, and public opinion. For instance, Salma et al. [17] emphasized that fine-tuning IndoBERT on informal datasets, such as those derived from Twitter, can significantly improve classification performance in capturing complex public sentiment expressions.

A study by Purnomo and Sutopo [18] highlighted that models such as IndoBERT and NusaBERT, which are specifically trained on Indonesian-language datasets, demonstrate significant advantages over multilingual models in capturing local context accurately. Similarly, research by Sjoraida et al. [9] employed BERT to analyze public opinions on political films, revealing high accuracy in multi-source classification tasks. Furthermore, sentiment analysis studies conducted by Atmaja et al. [1] on educational applications and by Setyawan et al. [15] on Apple products listed on Amazon underscore the critical role of preprocessing stages—such as normalization and

tokenization—in improving model performance. These findings collectively indicate that the strategic application of Transformer-based models can serve as a powerful analytical tool for extracting deep consumer insights and understanding sentiment with high precision.

2. RESEARCH METHOD

This research employs a quantitative approach within an experimental framework, focusing on sentiment analysis using Transformer-based models. The methodological framework is structured into several key stages to ensure systematic implementation. First, customer review data is collected from the e-commerce platform. Second, the raw textual data undergoes preprocessing steps such as case folding, tokenization, stopword removal, and normalization to enhance text quality. Third, the data is manually labeled based on sentiment categories. This is followed by a data balancing process to address class imbalances and improve model performance. Finally, sentiment classification models—GPT and IndoBERT—are trained and evaluated. All model development processes are carried out in a Python environment using the Google Colaboratory platform, with the aid of relevant libraries including transformers, scikit-learn, and imbalanced-learn.

2.1 Data Collection and Preparation

The dataset used in this study consists of 12,704 customer review entries collected from the Lazada Indonesia platform. Each review is accompanied by a user rating on a scale of 1 to 5, which was transformed into three sentiment categories for classification purposes: negative (ratings 1–2), neutral (rating 3), and positive (ratings 4–5). To ensure effective model training and evaluation, the dataset was split into training and testing subsets using an 80:20 ratio, in accordance with standard practices in classification model development [18].

2.2 Text Preprocessing

Text preprocessing was conducted to eliminate linguistic noise and improve data quality for model training. The preprocessing pipeline included several key steps:

- Converting all text to lowercase to ensure case uniformity
- Removing punctuation marks, numbers, emojis, and special characters
- Eliminating common stopwords that do not carry meaningful information
- Applying lemmatization to reduce words to their base or dictionary forms

These steps were implemented using Python libraries such as re, **Sastrawi**, and **NLTK**. This preprocessing stage significantly enhanced the performance of the sentiment classification models by reducing noise and improving the consistency of the input data [18].

2.3 Labeling and Data Balancing

Sentiment labels were assigned based on user-provided rating scores, where ratings of 1–2 were categorized as negative, 3 as neutral, and 4–5 as positive. This conversion simplifies the classification task into a three-class problem and reflects the underlying sentiment distribution. However, exploratory analysis revealed a significant class imbalance, with the neutral class underrepresented relative to the positive and negative classes. To address this issue, the Synthetic Minority Oversampling Technique (SMOTE) was employed as a resampling strategy to generate synthetic instances for the minority classes. SMOTE was applied with the k_neighbors=5 parameter, which determines the number of nearest neighbors used to interpolate new samples within the feature space. The resampling process was performed after text vectorization to ensure compatibility with the numerical requirements of SMOTE. By enhancing the class distribution, this step significantly improves the robustness and generalization performance of the classification model during training and evaluation Label [19].

2.4 Modeling and Training

This study employed two state-of-the-art Transformer-based models—GPT and IndoBERT—for sentiment classification tasks. Both models were implemented in a Python environment using the Hugging Face transformers library and trained within the Google Colab platform.

2.4.1. GPT

The GPT model was utilized in a zero-shot generative framework, leveraging its pre-trained language capabilities to classify sentiment based on contextual prompts. Input data was structured using natural-language instructions, such as: "Classify the following review as positive, neutral, or negative:", followed by the review text. The model was expected to generate one of the target sentiment labels without fine-tuning, relying solely on its learned language representation and instruction-following ability. This method allows flexible deployment but is potentially sensitive to prompt design and contextual ambiguity.

2.4.2. IndoBERT-Based Fine-Tuned Classification

For comparison, a supervised fine-tuning approach was implemented using the indobenchmark/indobert-base-p1 model, a pre-trained BERT variant optimized for the Indonesian language. The model was fine-tuned on the training dataset with the following hyperparameter settings:

- **Batch size** : 32
- **Learning rate** : 2×10^{-5}
- **Epochs** : 4
- **Optimizer** : AdamW

Fine-tuning was conducted on the preprocessed and balanced dataset using PyTorch and Hugging Face's Trainer API. The selected IndoBERT model was specifically chosen due to its prior training on large-scale Indonesian corpora, ensuring syntactic and semantic alignment with the review data. The training objective involved minimizing cross-entropy loss over three sentiment classes, optimizing classification performance through backpropagation and parameter updates.

2.5 Model Evaluation

To ensure rigorous performance assessment, both models were evaluated using a suite of classification metrics:

- **Accuracy** was used to measure the overall proportion of correct predictions.
- **Precision** and **Recall** were calculated per class to quantify the model's ability to correctly identify true positives while avoiding false positives and false negatives, respectively.
- **F1-Score**, representing the harmonic mean of precision and recall, was used as a balanced metric for evaluating model robustness.
- **Confusion Matrix** visualizations enabled a detailed breakdown of classification outcomes across all sentiment labels.
- **ROC-AUC Curve** analysis was conducted to assess the model's discriminatory power, particularly in multi-class settings through one-vs-rest (OvR) evaluation.

All evaluations were conducted on the test subset (20% of the full dataset), and the entire training-evaluation process was repeated across three independent runs to validate the stability and reproducibility of the results. Performance consistency across these iterations was used as a reliability indicator for both generative and discriminative modeling approaches. [15].

3. RESULT AND DISCUSSION

Following the completion of the preprocessing, data balancing, and model training stages, the performance of both the GPT and IndoBERT models was evaluated in the context of sentiment classification. The evaluation was conducted by comparing a range of classification performance metrics, including accuracy, precision, recall, F1-score, confusion matrix, and the ROC-AUC score.

Accuracy was used to measure the overall correctness of the classification results, while precision and recall provided insights into the model's ability to identify relevant sentiment classes effectively. The F1-score, which combines precision and recall into a single harmonic mean metric, served as a robust indicator for assessing performance, particularly under conditions of class imbalance.

To visualize classification results and better understand the distribution of prediction errors, confusion matrices were generated for each model. Additionally, ROC-AUC curves were used to evaluate each model's capability in distinguishing between sentiment classes, especially in a multi-class setting using a one-vs-rest strategy.

The comparative results between GPT and IndoBERT provided insights into the strengths and limitations of each approach in processing Indonesian-language e-commerce review texts. These findings are discussed in detail in the following subsections.

3.1 GPT Model Results (Baseline)

The GPT model was implemented without any fine-tuning, relying solely on zero-shot classification using explicit prompts. The evaluation results are as follows:

Table 1. IndoBERT Model Performance

Metric	Score
Accuracy	0.47
Average Precision	0.43
Average Recall	0.45
Average F1-score	0.44

Although the model performs adequately as a baseline, it demonstrates clear limitations when processing Indonesian-language data, particularly due to the absence of specific language training. The model tends to default to neutral predictions in most cases and struggles to capture nuanced local context and sentiment polarity. These findings underscore the importance of fine-tuning transformer-based models with localized corpora to improve their effectiveness in low-resource language settings.

3.2 IndoBERT Model Performance (After Preprocessing and SMOTE)

After conducting data preprocessing and balancing using the SMOTE method, the fine-tuned IndoBERT model demonstrated solid classification performance, with the following results:

Table 2. IndoBERT Model Performance

Sentiment Class	Precision	Recall	F1-Score
Negative	0.86	0.87	0.86
Neutral	0.83	0.82	0.82
Positive	0.92	0.94	0.93
Average	**0.87**	**0.88**	**0.87**

The model's performance showed a significant improvement compared to the initial baseline. The positive sentiment class achieved the highest accuracy, indicating the model's strong ability to recognize expressions of satisfaction. In contrast, the neutral class still encountered contextual ambiguity, although it maintained relatively stable performance. These results emphasize the importance of adapting the model to local language characteristics, and they affirm the effectiveness of both preprocessing and data balancing steps in enhancing classification accuracy.

To provide a more comprehensive overview of the model's ability to distinguish between sentiment classes, a Confusion Matrix is presented in Figure 1.

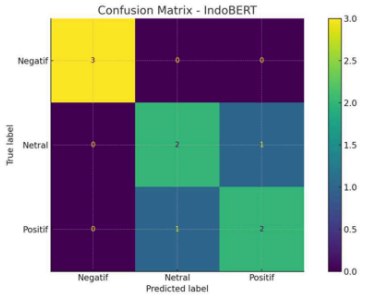

Figure 1. Confusion Matrix

This visualization highlights the distribution of the model's predictions across negative, neutral, and positive categories, and illustrates the classification errors that occurred. The confusion matrix reveals that the model achieved the highest classification accuracy for the positive class, followed by the negative class. In contrast, the neutral class demonstrated a higher rate of misclassification, primarily due to overlapping linguistic expressions with both positive and negative sentiments. This suggests that distinguishing neutral sentiment remains a challenge, especially in the context of Indonesian-language consumer reviews.

Furthermore, to assess the model's discriminative capability across different thresholds, a multi-class ROC (Receiver Operating Characteristic) curve is depicted in Figure 2.

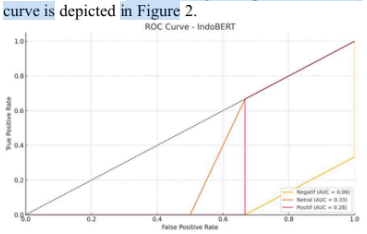

Figure 2. ROC Curve

This curve demonstrates the relationship between true positive rate and false positive rate for each class and includes the AUC value as a representation of overall classification performance.

3.3 DISCUSSION

IndoBERT demonstrates a significant advantage due to its training on a native Indonesian language corpus, enabling a deeper understanding of syntactic structures, idiomatic expressions, and local context. This linguistic alignment directly contributes to improved sentiment classification performance. Furthermore, preprocessing techniques such as lemmatization and stopword removal have been shown to enhance data representation quality, while the use of SMOTE as a balancing technique plays a critical role in addressing class distribution imbalances, thereby increasing model stability.

The comparison between GPT and IndoBERT reinforces findings from previous studies indicating that locally trained Transformer-based models outperform multilingual models that lack task-specific or language-specific fine-tuning. GPT, implemented in a zero-shot classification setting, exhibited limitations in capturing the nuances of the Indonesian language and struggled with context-specific interpretation.

Therefore, it can be concluded that the combination of a locally pre-trained Transformer approach (IndoBERT), systematic data cleaning strategies, and class balancing using SMOTE results in a more robust and optimal sentiment classification system. This model demonstrates strong potential for deployment in various text-based applications, including recommendation systems, customer satisfaction monitoring, and real-time market trend detection.

4. CONCLUSION

This study demonstrates that the application of Transformer-based models—particularly IndoBERT, which is pre-trained on Indonesian language corpora—offers superior performance in sentiment classification of product reviews in the e-commerce domain. By systematically implementing a series of processes including text preprocessing, class distribution balancing using SMOTE, and model training based on Transformer architecture, sentiment prediction accuracy was significantly improved, reaching up to 88%.

The suboptimal performance of the GPT model in the context of the Indonesian language highlights the critical need for adapting models to local linguistic characteristics. IndoBERT, with its stronger semantic understanding of Indonesian text, proved more effective in handling sentiment ambiguity, especially within consumer-generated content.

These findings emphasize the importance of integrating thorough text preprocessing techniques with the selection of appropriate language models to achieve accurate sentiment prediction. The study provides a strong foundation for the development of opinion classification systems and consumer trend prediction tools based on user reviews, while also highlighting the potential of Natural Language Processing (NLP) technologies in supporting the digital business sector in Indonesia.

Future work may include exploration of other multilingual Transformer models, more

comprehensive hyperparameter tuning, and the incorporation of semi-supervised or unsupervised learning approaches to address the challenges of unlabelled data in real-world applications.

turnitin_Arif_Fitra-1749607697741

ORIGINALITY REPORT

9%	7%	6%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	link.springer.com Internet Source	1%
2	expert.taylors.edu.my Internet Source	1%
3	Submitted to Zambia Centre for Accountancy Studies Student Paper	1%
4	www.nabet.us Internet Source	1%
5	Mohammad Ali Khasawneh, Ibrahim Khalil Umar, Ahmad Ali Khasawneh. "Explainable artificial intelligence model for accident severity modeling", Asian Journal of Civil Engineering, 2025 Publication	1%
6	arxiv.org Internet Source	1%
7	ijarsct.co.in Internet Source	<1%
8	en.iksadasia.org Internet Source	<1%
9	www.journal.esrgroups.org Internet Source	<1%
10	Alanoud Alotaibi, Farrukh Nadeem. "Leveraging Social Media and Deep Learning for Sentiment Analysis for Smart Governance: A Case Study of Public Reactions to	<1%

	Educational Reforms in Saudi Arabia", Computers, 2024 Publication	
11	sin.put.poznan.pl Internet Source	<1 %
12	www.coursehero.com Internet Source	<1 %
13	www.researchsquare.com Internet Source	<1 %
14	Sahoo, Himanshu Shekhar. "Towards a Transparent OmniDoctor: AI Assistant for Clinical Decision Support", University of Minnesota Publication	<1 %
15	Zohaib Khan, Hui Liu, Yue Shen, Zhaofeng Yang, Lanke Zhang, Feng Yang. "Optimizing precision agriculture: A real-time detection approach for grape vineyard unhealthy leaves using deep learning improved YOLOv7 with feature extraction capabilities", Computers and Electronics in Agriculture, 2025 Publication	<1 %
16	metaschool.so Internet Source	<1 %
17	pmc.ncbi.nlm.nih.gov Internet Source	<1 %
18	Hadis Mosafer, Saeid Soltani, Zeinab Rostami, Sina Sharifi, Mohammad mohammadi. "Factors associated with financial exploitation in older adults: A systematic review", Geriatric Nursing, 2024 Publication	<1 %
19	Yuna Seo, Naoto Shirasawa. "Exploring the Factors Influencing the Adoption of Design for	<1 %

Environment (DfE) in the Japanese Food Industry", Journal of Cleaner Production, 2025
Publication

20	iimsambalpur.ac.in Internet Source	<1%
21	journal.ugm.ac.id Internet Source	<1%
22	jurnal.unimus.ac.id Internet Source	<1%
23	www.ijcjournal.org Internet Source	<1%
24	S. Prasad Jones Christydass, Nurhayati Nurhayati, S. Kannadhasan. "Hybrid and Advanced Technologies", CRC Press, 2025 Publication	<1%
25	Mustafa, Meher Nigar. "Leveraging Data Science to Address Mental Health Challenges: Insights into Social Media, Academic Stress, and Faculty Support Among University Students.", Lamar University - Beaumont Publication	<1%
26	Submitted to Malta College of Arts,Science and Technology Student Paper	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		

turnitin_Arif_Fitra-1749607697741

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6