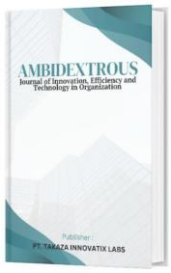
Ambidextrous: Journal of Innovation, Efficiency and Technology in Organization
https://journal.takaza.id/index.php/ambidextrous
Vol. 5, no. 2, 2026 , pp. 340-351
E-ISSN: 3031-7002
E-mail:ambidex@takaza.id



Comparison of IndoBERT and Naive Bayes Performance on Sentiment Analysis of Indonesian-Language E-Commerce Reviews

Tony Kurniawan[1*], Soffa Zahara [2], Yanuarini Nur Sukmaningtyas [3]
[1,2,3] Prodi Informatika, Universitas Islam Majapahit
Corresponding Author e-mail : tony.kurniawan036@gmail.com, offa.zahara@unim.ac.id, yanuarini.ft@unim.ac.id

Article History:
Received: 07-08-2026
Revised: 22-09-2026
Accepted: 22-09-2026

Keywords: Sentiment Analysis, IndoBERT, Naive Bayes, E-Commerce, Natural Language Processing

Abstract: *The growth of e-commerce platforms in Indonesia has generated a large number of user reviews containing important information about service quality and user satisfaction, yet manual analysis of these reviews remains inefficient, necessitating a Natural Language Processing (NLP)-based approach. This study aims to compare the performance of Naive Bayes and IndoBERT in sentiment classification of Indonesian-language e-commerce reviews from Tokopedia, Shopee, and Lazada applications. The dataset consists of 26,103 reviews obtained from the Google Play Store, which underwent sentiment labeling based on ratings, preprocessing with emoji retention, Term Frequency–Inverse Document Frequency (TF-IDF) weighting for Naive Bayes, and fine-tuning for IndoBERT. Model evaluation was conducted using a confusion matrix with accuracy, precision, recall, and F1-score metrics. The results show that Naive Bayes achieved an accuracy of 80.53%, while IndoBERT achieved 86.48% with a Macro F1-Score of 0.6515 and a Weighted F1-Score of 0.8723. Both models struggled with neutral sentiment classification due to data imbalance, with Naive Bayes unable to classify neutral reviews (F1-score: 0.00) and IndoBERT showing limited performance (F1-score: 0.17). These results demonstrate that IndoBERT is more effective in understanding the context of Indonesian text, particularly in handling informal language, negation, and semantic relationships, resulting in superior sentiment classification performance compared to the TF-IDF-based Naive Bayes approach. This study contributes a comprehensive comparative evaluation between traditional machine learning and Transformer-based models for Indonesian e-commerce sentiment analysis using a multi-platform dataset.*

How to Cite: First author., Second author., & Third author. (20xx). The title. Ambidextrous Journal of Innovation Efficiency and Technology in Organizations. 5(2). Pp 340-351 https://doi.org/10.61536/ambidextrous.v5i02.648

https://doi.org/10.61536/ambidextrous.v5i02.648

Introduction

The development of information technology has driven the growth of e-commerce platforms in Indonesia, such as Tokopedia, Shopee, and Lazada, which offer a variety of digital transaction services. The high number of users on these platforms has generated millions of reviews containing experiences, criticisms, suggestions, and satisfaction levels with the application services. User reviews are a crucial source of information for developers to evaluate service quality and understand user perceptions. However, the ever-increasing volume of reviews makes manual analysis inefficient, necessitating an automated approach based on Natural Language Processing (NLP) to process large amounts of data (Hermawan & Jowensen, 2023; Mao et al., 2024; Romadoni et al., 2020).

One widely used application of NLP is sentiment analysis, the process of classifying user opinions into positive, negative, or neutral categories, thus providing an overview of public perception of a product or service (Mao et al., 2024; Nurian et al., 2023). In Indonesian text classification, the Naive Bayes method remains widely used due to its simple training process, computational efficiency, and good performance in various sentiment analysis cases (Hermanto, 2021; Sanjaya et al., 2023). However, Naive Bayes uses a word frequency-based representation, making it less capable of understanding semantic relationships and sentence context, especially in text containing negation, ambiguity, abbreviations, and informal language commonly found in e-commerce reviews (Kesumawati & Thalib, 2023).

The development of Transformer architecture-based models has led to improved performance on various NLP tasks. One widely used model is Bidirectional Encoder Representations from Transformers (BERT), which is capable of generating contextual word representations through a self-attention mechanism (Devlin et al., 2020; Mohiuddin et al., 2023). For Indonesian, IndoBERT was developed using an Indonesian corpus, making it more effective in understanding the characteristics of the Indonesian language, including the use of informal language often found in app user reviews (Jayadianti et al., 2022; Wilie et al., 2020). Research published in JITET also shows that IndoBERT is capable of delivering good performance on Indonesian sentiment analysis tasks, making it a potential approach for text classification (Sumartha, 2025).

Although various studies have applied Naive Bayes and IndoBERT to sentiment analysis, most studies focus on only one classification method or use datasets from a single e-commerce platform (Kireyna Cindy Pradhisa, 2024; Maxmiliano et al., 2025; Putra & Ismarmiaty, 2025). Research comparing the two approaches on a combined dataset of Tokopedia, Shopee, and Lazada reviews while retaining emojis as part of the sentiment information is still relatively limited. However, emojis can represent users' emotional expressions, potentially providing additional information in the classification process.

Based on these gaps, this study compares the performance of the Naive Bayes and IndoBERT models in classifying the sentiment of Indonesian-language e-commerce reviews using a combined dataset of 26,103 user reviews of Tokopedia, Shopee, and Lazada applications grouped into three sentiment classes: positive, negative, and neutral. The pre-processing stage retains emojis as emotional representations, while model evaluation is conducted using a confusion matrix with accuracy, precision, recall, and F1-score metrics. This study contributes in the form of a comparative evaluation between the TF-IDF-based machine learning approach and the Transformer model (IndoBERT) in classifying the sentiment of Indonesian-language e-commerce reviews using a combined dataset from three platforms, so that it can be a reference in selecting a sentiment classification model for informal text data.

Literature review

Sentiment Analysis

Sentiment analysis is a field of Natural Language Processing (NLP) that aims to identify, extract, and classify opinions or emotions contained in text into sentiment categories, such as positive, negative, or neutral (Mao et al., 2024; Hermawan & Jowensen, 2023). In this study, sentiment analysis was applied

https://doi.org/10.61536/ambidextrous.v5i02.648

to classify user reviews of e-commerce applications to provide an overview of user perceptions of the quality of the services provided.

Natural Language Processing (NLP)

Natural Language Processing (NLP) is a branch of artificial intelligence that studies the interaction between computers and human language. NLP enables systems to understand, process, and analyze text data, enabling it to be utilized in various tasks, such as document classification, sentiment analysis, and entity recognition (Mao et al., 2024; Romadoni et al., 2020). In this study, NLP was used as the basis for preprocessing and sentiment classification of e-commerce user reviews.

Term Frequency–Inverse Document Frequency (TF-IDF) weighting

Term Frequency–Inverse Document Frequency (TF-IDF) is a word weighting method that measures the importance of a word based on its frequency of occurrence in documents and the entire document collection (Widianto et al., 2024). This weighting is capable of representing documents in the form of numeric vectors, thus allowing them to be used as input for classification algorithms. In this study, TF-IDF was used as a feature representation in the Naive Bayes model.

Naive Bayes

Naive Bayes is a classification algorithm based on Bayes' Theorem, which assumes that each feature is independent of all other features. This algorithm is widely used in text classification due to its simple training process, computational efficiency, and good performance on various sentiment analysis problems (Hermanto, 2021; Hidayati et al., 2022). In this study, Naive Bayes was used as a comparison model against the Transformer-based model.

Transformer, BERT, and IndoBERT

The Transformer architecture utilizes self-attention mechanisms to contextually capture relationships between words in a sentence, thereby improving the performance of various Natural Language Processing tasks. One implementation is Bidirectional Encoder Representations from Transformers (BERT), a pre-trained language model that understands word context bidirectionally. For Indonesian, IndoBERT was developed using an Indonesian corpus, making it more effective in understanding the characteristics of the Indonesian language, including informal language often found in app user reviews. In this study, IndoBERT was used through a fine-tuning process to perform sentiment classification (Hermanto, 2021; Hidayati et al., 2022).

Confusion Matrix and Evaluation Metrics

The Transformer architecture utilizes a self-attention mechanism to contextually capture relationships between words in a sentence, thereby improving the performance of various Natural Language Processing tasks. One implementation is Bidirectional Encoder Representations from Transformers (BERT), a pre-trained language model that understands word context bidirectionally (Devlin et al., 2020; Vaswani et al., 2020). For Indonesian, IndoBERT was developed using an Indonesian corpus, making it more effective in understanding the characteristics of the Indonesian language, including informal language often found in app user reviews (Baldwin, 2020; Wilie et al., 2020; Jayadianti et al., 2022). In this study, IndoBERT was used through a fine-tuning process to perform sentiment classification.

Research methodology

This study uses an experimental approach to compare the performance of the Naive Bayes and IndoBERT models in sentiment classification of Indonesian-language e-commerce reviews. The research stages include data collection, sentiment labeling, data preprocessing, TF-IDF weighting, training the Naive Bayes and IndoBERT models, and evaluation using a confusion matrix. The research flow is shown in Figure 1.

https://doi.org/10.61536/ambidextrous.v5i02.648

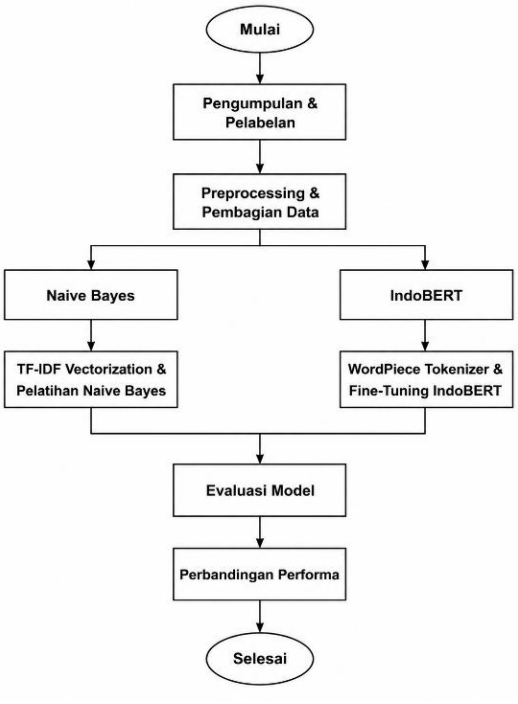


Figure 1. Research Flow.

Data collection

Data collection was conducted using web scraping techniques on user reviews of the Tokopedia, Shopee, and Lazada apps available on the Google Play Store. These three apps were selected because they are e-commerce platforms with a large user base in Indonesia, thus providing a sufficient volume of reviews for analysis. The data obtained consisted of user review texts along with supporting information, which were then compiled into a single research dataset.

In this study, 11,000 reviews were collected from each application, resulting in a total of 33,000 reviews as the initial dataset. The collected data included user names, review content, rating values, review time, and other supporting attributes. However, this study only utilized review text and rating values as the primary data. Review text served as input for the classification process, while rating values served as a reference for the sentiment labeling process. A summary of the data collection results is shown in Table 1.

Table 1. Number of Datasets from Data Collection Results.

Application	Number of Reviews
Tokopedia	11,000
Shopee	11,000
Lazada	11,000
Total	33,000

Sentiment Labeling

Sentiment labeling is performed to assign a sentiment category to each user review so that it can be used as a classification target (ground truth) in the supervised learning method. In this study, labeling is performed automatically based on rating values obtained from the Google Play Store. Although ratings are used as a reference in the labeling process, the primary data analyzed remains the review text, while ratings are only used to generate sentiment labels.

The labeling process was carried out by grouping each review into three sentiment categories: positive, neutral, and negative. Reviews with ratings of 4 and 5 were labeled positive, those with ratings

https://doi.org/10.61536/ambidextrous.v5i02.648

of 3 were labeled neutral, and those with ratings of 1 and 2 were labeled negative. This approach was chosen because ratings reflect users' general assessment of the application and have been widely used as a labeling reference in sentiment analysis research. The sentiment labeling criteria are shown in Table 2.

Table 2. Sentiment Labeling

Sentiment Label	Rating	Criteria
Positive	4-5	Reviews that indicate satisfaction with an application or service
Neutral	3	Reviews that are informative and do not show any clear sentimental tendencies
Negative	1-2	Reviews that indicate complaints, criticism, or dissatisfaction with an application or service

After the labeling process is complete, each review has a sentiment label, which is then used as ground truth in preprocessing, model training, and evaluation. Examples of sentiment labeling results are shown in Table 3.

Table 3. Example of Labeling Results

No	Rating	Review	Label
1	5	The app is very helpful and easy to use for shopping.	Positive
2	3	Delivery according to the estimate given by the application.	Neutral
3	1	The item received did not match the description and the return process was quite difficult.	Negative

Preprocessing

Preprocessing was performed to improve the quality of text data before it was used in the classification process. The steps applied included cleaning, case folding, tokenization, and stopword removal using the Naive Bayes model. In this study, emojis were retained because they were considered capable of representing users' emotional expressions and therefore potentially providing additional information in the sentiment classification process. Examples of the results of each preprocessing stage are shown in Table 4.

Table 4. Preprocessing Stage

Stages	Results
Original Text	Tokopedia is often having errors these days!!! 😭 Unable to make payment
Cleaning	Tokopedia often has errors these days 😭 Unable to make payment
Folding Case	Tokopedia often has errors these days 😭 unable to make payment
Tokenization (IndoBERT)	['tokopedia', 'times', 'this', 'often', 'error', ' 😭 ', 'no', 'can', 'do', 'payment']
Stopword Removal (Naive Bayes)	['tokopedia', 'often', 'error', ' 😭 ', 'payment']

Feature Representation and Dataset Sharing

After the preprocessing process is complete, the data used in the Naive Bayes model is represented numerically using the Term Frequency–Inverse Document Frequency (TF-IDF) method. This weighting is used to generate feature representations based on the frequency of word occurrences in each document, thus providing input for the classification process.

The dataset was then divided into training and test data using the Stratified Train-Test Split method with a 70%:30% ratio. This stratified approach was chosen to maintain the proportion of each sentiment class in both the training and test data, ensuring that the data distribution remains representative during the model training and evaluation process.

https://doi.org/10.61536/ambidextrous.v5i02.648

Model Training

The training phase was conducted using two classification models: Naive Bayes and IndoBERT. The Naive Bayes model uses the TF-IDF weighted features as input to build a classification model based on the posterior probabilities of each class.

Meanwhile, the IndoBERT model was trained through a fine-tuning process using a preprocessed dataset. Before training, each review was converted into a token using IndoBERT's built-in tokenizer. To mitigate the effects of class imbalance, the training process applied class weighting. The fine-tuned model was then used to perform sentiment classification on the test data.

Model Evaluation

The final stage of the research was to evaluate the performance of both models using a confusion matrix. Based on the confusion matrix, accuracy, precision, recall, and F1-score were calculated to measure each model's ability to classify e-commerce review sentiment. The evaluation results of both models were then compared to determine which model performed best on the research dataset.

Results and Discussion

Data Collection Results and Data Preprocessing

The data collection process successfully obtained 33,000 user reviews from the Tokopedia, Shopee, and Lazada apps on the Google Play Store. Each app contributed 11,000 reviews, resulting in the initial dataset used in the study. Next, the dataset underwent sentiment labeling based on rating values before being preprocessed to improve the quality of the text data.

The preprocessing stage is carried out through several processes, namely cleaning, case folding, tokenization, and stopword removal in the Naive Bayes model. The cleaning process aims to remove unnecessary characters, such as URLs, special characters, and duplicate data, while case folding converts all letters to lowercase to ensure a uniform text format. Next, tokenization breaks sentences into word tokens, while stopword removal is used to remove common words that do not have a significant contribution to the classification process in the Naive Bayes model. Unlike other characters, emojis are retained because they are considered capable of representing users' emotional expressions, thus potentially increasing the sentiment information contained in reviews.

After all preprocessing steps were completed, the dataset was reduced to 26,103 reviews, consisting of positive, negative, and neutral data. The reduction in data revealed that some reviews did not meet the research criteria, such as duplicate or invalid data, and were therefore removed during the preprocessing process. A summary of the dataset before and after preprocessing is presented in Table 1, while examples of the results of each preprocessing step are shown in Table 2.

A summary of the amount of data used in the study is shown in Table 5.

Table 5. Data Collection Results

Application	Number of Reviews
Tokopedia	11,000
Shopee	11,000
Lazada	11,000
Total	33,000

Table 6. Example of Data Preprocessing Results

Stages	Results
Original Text	I'm disappointed, the first time I installed it I got a 99% discount promo, but when I wanted to check out it couldn't be used. 😭
Folding Case	I'm disappointed, the first time I installed it I got a 99% discount promo, but when I wanted to check out it couldn't be used. 😭
Tokenization (IndoBERT)	['disappointed', 'yes', ',', 'first', 'time', 'install', 'get', 'promo', 'discount', 'price', '99%', 'ehh', 'just right', 'want', 'in', 'checkout', 'even', 'can't', 'use', '😭 ']

https://doi.org/10.61536/ambidextrous.v5i02.648

Stages	Results
Stopword Removal (Naive Bayes)	['disappointed', 'yeah', 'first', 'time', 'install', 'get', 'promo', 'discount', 'price', '99%', 'ehh', 'checkout', 'can't', 'use', ' 😭 ']

The data collection results show that user reviews encompass a wide range of opinions, ranging from positive experiences with the service, complaints about app usage issues, to neutral reviews. This diversity of review characteristics serves as the basis for evaluating the ability of the Naive Bayes and IndoBERT models to classify sentiment in Indonesian text data.

Sentiment Distribution and Word Cloud

Sentiment distribution was performed to determine the proportion of each sentiment class in the preprocessed dataset. Based on the labeling results, the dataset used in this study consists of three sentiment classes: positive, negative, and neutral. The distribution of data in each class is presented in Figure 2.

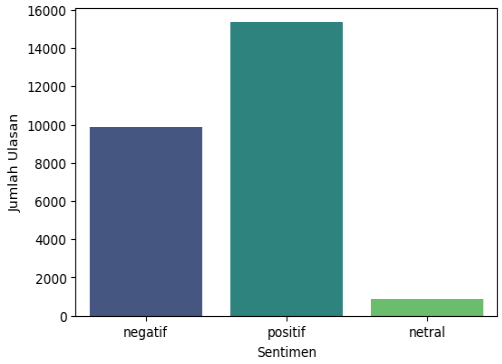


Figure 2. Sentiment Distribution

Based on Figure 2, the negative sentiment class has the largest amount of data compared to other classes, followed by the positive class, while the neutral class has the smallest amount of data. This condition indicates an unbalanced dataset distribution, especially for the neutral class, which has significantly less data. This imbalance has the potential to impact the model training process. Therefore, class weighting was applied in the IndoBERT fine-tuning process to reduce model bias toward classes with larger data sets.

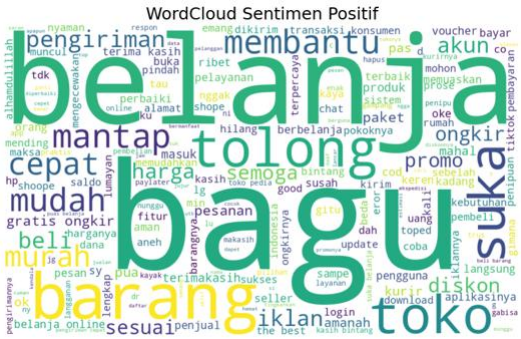


Figure 2. Word Cloud of Positive Sentiment Reviews

Figure 2 shows a word cloud visualization of reviews labeled with positive sentiment. The words "shopping," "good," "please," "goods," "great," "delivery," "fast," and "cheap" appear to have high occurrences. This indicates that the majority of positive reviews relate to the shopping experience, product quality, service, delivery speed, and various promotions offered by the e-commerce platform. Furthermore, informal terms such as "sukuh," "aja," and "gak" are still found, indicating that users tend to use everyday language when leaving reviews on the Google Play Store.

Naive Bayes classification results

https://doi.org/10.61536/ambidextrous.v5i02.648

A Naive Bayes model was trained using the Term Frequency–Inverse Document Frequency (TF-IDF) weighting feature from the preprocessed dataset. Model evaluation was performed using a classification report and confusion matrix to determine the model's ability to classify user review sentiment into three classes: negative, neutral, and positive.

Laporan Klasifikasi Naive Bayes:

	precision	recall	f1-score	support
negatif	0.73	0.83	0.78	2964
netral	0.00	0.00	0.00	263
positif	0.86	0.84	0.85	4604
accuracy			0.81	7831
macro avg	0.53	0.55	0.54	7831
weighted avg	0.78	0.81	0.79	7831

Figure 3. Classification Report Results of the Naive Bayes Model

Based on the classification report results, the Naive Bayes model achieved an accuracy of 80.53%, with a weighted average precision of 0.78, a recall of 0.81, and an F1-score of 0.79. These results indicate that the model performs quite well in classifying the overall sentiment of e-commerce reviews.

In the negative sentiment class, the model achieved a precision of 0.73, a recall of 0.83, and an F1-score of 0.78. The higher recall compared to precision indicates that the model is able to recognize most negative reviews, although there are still some false positive predictions. In the positive sentiment class, the model achieved a precision of 0.86, a recall of 0.84, and an F1-score of 0.85, indicating good classification ability for reviews with positive sentiment.

Meanwhile, the precision, recall, and F1-score for the neutral sentiment class were all 0.00. These results indicate that the model is unable to recognize the neutral sentiment class. This is due to the significantly smaller amount of neutral sentiment data compared to the positive and negative classes and the limitations of the TF-IDF-based feature representation in distinguishing the characteristics of reviews with similar vocabulary within both classes.

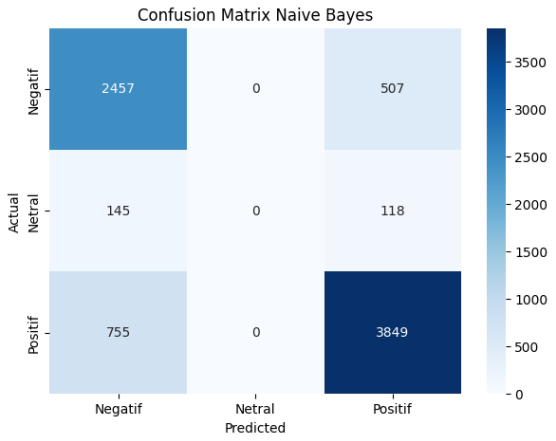


Figure 4. Confusion matrix

Based on Figure 4, the confusion matrix shows that the model successfully classified 2,457 negative sentiment data and 3,849 positive sentiment data correctly. In the negative class, 507 data were predicted as positive, while in the positive class, 755 data were predicted as negative.

Unlike the other two classes, the model was unable to correctly classify neutral sentiment data. A total of 145 neutral sentiment data were predicted as negative, and 118 as positive, resulting in no accurate predictions within the neutral class. These results indicate that the model is having difficulty recognizing the characteristics of neutral sentiment.

Overall, the evaluation results show that the Naive Bayes model performs quite well in both positive and negative sentiment classes. However, the model still has limitations in identifying the neutral

https://doi.org/10.61536/ambidextrous.v5i02.648

sentiment class due to the imbalance in data volume and the use of TF-IDF-based feature representation, which is unable to fully capture the context of a sentence.

IndoBERT Classification Results

The IndoBERT model was trained through a fine-tuning process using the preprocessed dataset. Model evaluation was performed using a confusion matrix and classification report to determine the model's ability to classify user review sentiment into three classes: negative, neutral, and positive. This is shown in Figure 5.

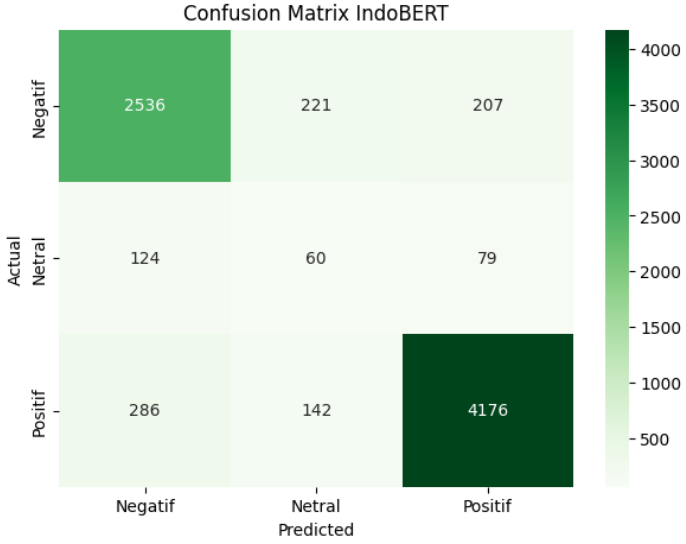


Figure 5. Confusion Matrix of IndoBERT Model

Based on Figure 5, the IndoBERT model was able to correctly classify most of the test data in both negative and positive sentiment classes. In the negative class, the model successfully classified 2,536 data points correctly, while 221 data points were predicted as neutral and 207 data points as positive. These results indicate the model's good ability to recognize reviews with negative sentiment.

In the neutral sentiment class, the model successfully classified 60 data points correctly. However, 124 data points were still predicted as negative and 79 as positive. This indicates that the neutral class remains the most difficult to identify because its review characteristics are similar to both the positive and negative classes and the data volume is relatively small compared to the other two classes.

Meanwhile, in the positive sentiment class, the model successfully classified4,176 data points were correctly predicted. Furthermore, 286 data points were predicted as negative and 142 as neutral. The high number of correct predictions indicates that IndoBERT is excellent at recognizing reviews with positive sentiment.

Overall, the confusion matrix results show that the IndoBERT model is able to classify negative and positive sentiment classes well, while the neutral class remains a challenge due to the imbalance in the amount of data and the similarity of characteristics with other sentiment classes.

```
--- EVALUASI AKHIR MODEL INDOBERT ---
Laporan Klasifikasi IndoBERT:
              precision    recall  f1-score   support

     negatif       0.86      0.86      0.86      2964
      netral       0.14      0.23      0.17       263
     positif       0.94      0.91      0.92      4604

    accuracy                           0.86      7831
   macro avg       0.65      0.66      0.65      7831
weighted avg       0.88      0.86      0.87      7831
```

Figure 6. Classification Report Results of the IndoBERT Model

Based on the classification report results, the IndoBERT model achieved an accuracy of 86.48%, with a weighted average precision of 0.88, a recall of 0.86, and an F1-score of 0.87. For the negative sentiment category, the model achieved a precision, recall, and F1-score of 0.86 each, indicating good classification capability for negative reviews.

https://doi.org/10.61536/ambidextrous.v5i02.648

The best performance was obtained in the positive sentiment class with a precision value of0.94, recall 0.91, and F1-score 0.92. Conversely, the neutral sentiment class still performed lower, with a precision of 0.14, a recall of 0.23, and an F1-score of 0.17. Nevertheless, these results indicate that IndoBERT is able to recognize some neutral reviews and performs better than the Naive Bayes model.

Overall, the evaluation results indicate that the IndoBERT model performed best in this study. Overall, the evaluation results indicate that the IndoBERT model performed well in classifying the sentiment of Indonesian-language e-commerce reviews.

Model Performance Comparison

A model performance comparison was conducted to evaluate the performance of Naive Bayes and IndoBERT in classifying user review sentiment for the Tokopedia, Shopee, and Lazada apps. The evaluation used Accuracy, Macro F1-Score, Weighted F1-Score, and F1-Score metrics for the neutral class. The comparison results for both models are presented in Figure 7.

```
======================================================
          TABEL PERBANDINGAN PERFORMA MODEL
======================================================
         Metrik Evaluasi Naive Bayes (Baseline) IndoBERT (Fine-Tuned)
       Accuracy Overall                 0.8053                0.8648
         Macro F1-Score                 0.5418                0.6515
      Weighted F1-Score                 0.7928                0.8723
F1-Score (Khusus Netral)                0.0000                0.1749
======================================================
```

Figure 7. Model Comparison

Based on Figure 6, the IndoBERT model performs better than Naive Bayes across all evaluation metrics. The accuracy value increased from 80.53% to 86.48%, while the Macro F1-Score increased from 0.5418 to 0.6515 and the Weighted F1-Score increased from 0.7928 to 0.8723. These improvements indicate that IndoBERT is capable of providing more consistent classification performance across all sentiment classes than Naive Bayes.

The performance difference was most pronounced in the neutral sentiment class. The Naive Bayes model achieved an F1-score of 0.0000, while IndoBERT achieved a score of 0.1749. Although the neutral class' performance was still lower than the positive and negative classes, these results indicate that IndoBERT has a better ability to recognize the characteristics of neutral reviews. The low performance in this class is influenced by the imbalance in data volume, where the amount of neutral sentiment data is much smaller than the positive and negative classes, making the model learning process more challenging.

IndoBERT's superiority stems from its ability to generate contextual word representations through self-attention. This capability allows IndoBERT to understand word relationships, sentence context, negation usage, and informal language commonly found in e-commerce reviews. In contrast, Naive Bayes uses a TF-IDF weighting-based representation that only considers word frequency, thus underutilizing contextual information. This results in Naive Bayes' performance declining when faced with reviews with implicit meanings or similar vocabulary across sentiment classes.

The results of this study are in line with the research on Sentiment analysis of Indonesian reviews using fine-tuning IndoBERT and R-CNN, which shows that the IndoBERT-based approach is able to provide better performance than conventional classification methods because it utilizes contextual language representation. This finding also supports the research on Comparison of Naive Bayes and BERT Algorithms for Sentiment Analysis of Shopee Product Reviews Based on Ratings and Product Attributes, which reports that the BERT-based model produces higher classification performance than Naive Bayes on sentiment analysis tasks.

Overall, the results show that IndoBERT is a more effective model for sentiment classification of Indonesian-language e-commerce reviews. With its better understanding of sentence context, IndoBERT performs better across all evaluation metrics, making it more suitable for classifying non-standard and diverse user review data.

In addition to demonstrating that IndoBERT performs better than Naive Bayes, this study also shows that using a combined dataset from three e-commerce platforms and retaining emojis in preprocessing

https://doi.org/10.61536/ambidextrous.v5i02.648

still yields good classification performance. These findings illustrate the effectiveness of the Transformer model in processing informal Indonesian-language user reviews.

Conclusion

a. This study compares the performance of the Naive Bayes and IndoBERT methods in classifying sentiment of Indonesian-language e-commerce reviews using 26,103 reviews of Tokopedia, Shopee, and Lazada apps obtained from the Google Play Store. The evaluation results show that the IndoBERT model provides better performance than Naive Bayes. The Naive Bayes model achieved an accuracy of 80.53%, while IndoBERT achieved 86.48%. In addition, IndoBERT produced a Macro F1-Score of 0.6515 and a Weighted F1-Score of 0.8723, higher than Naive Bayes. These results indicate that the contextual representation capability of IndoBERT is more effective in understanding the characteristics of Indonesian-language e-commerce reviews than the TF-IDF weighting-based approach.
b. However, both models still struggled to classify neutral sentiment due to the imbalance in data size and the similarity of review characteristics between the positive and negative classes. Therefore, future research could consider using a more balanced dataset and exploring techniques for handling data imbalance or developing other Transformer-based models to improve classification performance, particularly for the neutral sentiment class.

Thank-you note

The author would like to thank his supervisor for his guidance, input, and assistance during the research and preparation of this article. He also expresses his appreciation to the Informatics Study Program, Faculty of Engineering Technology, Majapahit Islamic University, for their support during the research.

Bibliography

A. Mao et, “Journal of King Saud University - Computer and 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.

A. Hermawan and I. Jowensen, “Implementation of Text-Mining for Sentiment Analysis on Twitter with Support Vector Machine Algorithm,” vol. 12, no. 1, pp. 129–137, 2023.

F. Romadoni, Y. Umaidah, and BN Sari, “Text Mining for Customer Sentiment Analysis of Electronic Money Services Using Support Vector Machine Algorithm,” vol. 09, pp. 247–253, 2020.

A. Nurian, BN Sari, US Karawang, and T. Timur, “Sentiment analysis of Google Play app user reviews using naive Bayes,” vol. 11, no. 3, pp. 829–835, 2023.

Agus Hermanto, “IMPLEMENTATION OF TEXT MINING USING NAIVE BAYES TO DETERMINE STUDENTS’ FINAL ASSIGNMENT CATEGORIES BASED ON THEIR ABSTRACT,” vol. 12, pp. 1–10, 2021.

Tania Puspa Rahayu Sanjaya, Ahmad Fauzi, and Anis Fitri Nur Masruriyah, “Sentiment analysis of reviews on e-commerce shopee using naive bayes algorithm and support vector machine,” INFOTECH J. Inform. Teknol., vol. 4, no. 1, pp. 16–26, 2023, doi: 10.37373/infotech.v4i1.422.

A. Kesumawati and AK Talib, "Hoax Classification with Term Frequency - Inverse Document Frequency Using Non-Linear SVM and Naïve Bayes," vol. 10, no. 3, 2023.

Devlin, M. C. Kenton, and L. Kristina, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” no. MLM, 2020.

K. Mohiuddin et al., “Retention Is All You Need,” Int. Conf. Inf. Knowl. Manag. Proc., no. Nips, pp. 4752–4758, 2023, doi: 10.1145/3583780.3615497.

B. Wilie et al., “IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding,” Proc. 1st Conf. Asia-Pacific Chapter Assoc. Comput. Linguist. 10th Int. Jt. Conf. Nat. Lang. Process. AACL-IJCNLP 2020, pp. 843–857, 2020, doi: 10.18653/v1/2020.aacl-main.85.

H. Jayadianti, W. Kaswidjanti, A.T. Utomo, S. Saifullah, F.A. Dwiyanto, and R. Drezewski, “Sentiment analysis of Indonesian reviews using fine-tuning IndoBERT and R-CNN,” Ilk. J. Ilm., vol. 14, no. 3, pp. 348–354, 2022, doi: 10.33096/ilkom.v14i3.1505.348-354.

https://doi.org/10.61536/ambidextrous.v5i02.648

DK Sumartha, “Public Opinion on the UKT Increase Policy in the Era of Government,” vol. 13, no. 3, 2025.

P. Maxmiliano, YF Riti, IY Nugroho, and CEC Juniarto, “Comparison of Naive Bayes and BERT Algorithms for Sentiment Analysis of Shopee Product Reviews Based on Ratings and Product Attributes (Color/Category),” J. Media Inform., vol. 6, no. 5, pp. 2552–2565, 2025.

RF Kireyna Cindy Pradhisa, “Sentiment Analysis of E-commerce User Reviews on Google Play Store Using the IndoBERT Method,” 2024.

A. Putra, A., Ismarmiaty, “Measuring the Level of Accuracy in E-Commerce Reviews Using the INDOBERT Method with Adam Optimizer,” J. Komputer, Inf. dan Teknol., vol. 5, no. 1, pp. 1–15, 2025.

M. Adi Widianto, Eka Pebriyanto, Fitriyanti, "Document Similarity using Term Frequency-Inverse Document Frequency Representation and Cosine Similarity," vol. 4, no. 2, pp. 149–153, 2024.

S. Hidayati, A. Nuraini, and F. Rahayu, “Designing a Posyandu nutrition education application for stunting prevention,” J. Abdika, vol. 6, no. 2, pp. 122–129, 2022, [Online]. Available: https://uniflor.ac.id/e-journal/index.php/abdika/article/view/3435

V. As. Np. Nu. Jj. Lg. Ak. Łp. I, “Attention Is All You Need,” no. Nips, 2020.

T. Baldwin, “IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP,” 2020.

https://doi.org/10.61536/ambidextrous.v5i02.648