Jurnal Info Sains : Informatika dan Sains
Volume 15, Number 02, 2025, DOI 10.54209/infosains.v15i02
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

Sentiment Analysis of ChatGPT Application Reviews Using the BERT Algorithm

Farhan Naufal Sukmana[1], **Aji Primajaya** [2], **Aries Suharso**[3]
Computer Science, Singaperbangsa Karawang University, Karawang, Indonesia

Article Info
Keywords:
Sentiment Analysis,
BERT,
IndoBERT,
KDD,
ChatGPT,
Google Play Store.

ABSTRACT
The rapid growth of generative artificial intelligence applications, particularly ChatGPT, has resulted in a significant increase in user reviews on the Google Play Store. These reviews serve as valuable sources for understanding user perceptions, experiences, and concerns. This study aims to analyze sentiment in Indonesian-language reviews of the ChatGPT application using the Bidirectional Encoder Representations from Transformers (BERT) algorithm combined with the Knowledge Discovery in Database (KDD) methodology. The dataset was collected using web scraping via the google-play-scraper library, producing 1,806 reviews after data cleaning and preprocessing. The dataset was divided into training and testing sets with an 80:20 ratio. IndoBERT was employed as the pre-trained model. Evaluation results show that the model successfully classified positive, negative, and neutral sentiments with an accuracy of 93%, precision of 87%, recall of 83%, and an F1-score of 85%. Although performance on the neutral class was lower due to dataset imbalance, the model demonstrated strong overall results. This study confirms that BERT is effective for sentiment analysis of application reviews and can serve as a reference for improving application service quality by understanding user opinions.

Corresponding Author:
Farhan Naufal Sukmana
Computer Science, Singaperbangsa Karawang University
Karawang, Indonesia
2110631170064@student.unsika.ac.id

INTRODUCTION

The rapid development of generative AI applications since the release of ChatGPT in 2022 was followed by the launch of an official mobile application in 2023, which then became one of the AI applications with the highest user engagement. In 2025, industry reports showed a large scale of adoption, ranging from millions of reviews to tens of millions of downloads, as well as dominance of use in a number of countries. This confirms the relevance of understanding user perceptions and experiences of the ChatGPT application through sentiment analysis (Alqadah et al., 2025).

In the Google Play ecosystem, reviews and ratings play an important role in determining application visibility, download decisions, and software maintenance strategies. Since 2023, Google has introduced an AI-based review summary feature, which confirms the strategic value of reviews as a source of research information (Vincent, 2023). However, a recent study found that the level of developer engagement in responding to reviews is still low, even

Jurnal Info Sains : Informatika dan Sains
Volume 15, Number 02, 2025, DOI 10.54209/infosains.v15i02
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

though these responses have a significant impact on app ratings and downloads (Xia et al., 2025). Additionally, Google Play's app curation policy for 2024–2025 further emphasizes the importance of monitoring user opinions (App Radar, 2023).

The urgency of this research is further strengthened by the fact that ChatGPT user reviews contain rich information about the quality of AI responses, ease of use, technical issues, and overall satisfaction levels. (Liu et al., 2024) shows that reviews on generative AI applications not only reflect user opinions but also capture important issues related to performance and user expectations of the technology. Thus, sentiment analysis of ChatGPT user reviews is important for identifying patterns of satisfaction, dominant complaints, and trends in public perception that are relevant to application development and academic literature related to generative AI technology.

At the scientific level, sentiment analysis has undergone significant developments. Deep learning techniques such as BERT, LSTM, BiLSTM, and CNN are considered more effective in handling the complexity of natural language than traditional methods (Yadav et al, 2024). A recent review shows that deep learning provides superior performance in polarity classification and opinion mining, while facing challenges such as class imbalance, language ambiguity, and domain variation (Basiri et al., 2025).

The selection of the Bidirectional Encoder Representations from Transformers (BERT) algorithm in this study is based on its advantage in understanding the context of sentences in two directions (bidirectional), allowing the model to capture the relationship between words more accurately. The transformer architecture in BERT, which is based on self-attention, allows the model to understand the nuances of informal, mixed, or concise language commonly found in user reviews of applications. Research (Kusuma et al., 2021) shows that BERT is capable of achieving up to 99% accuracy in analyzing the sentiment of Indonesian-language app reviews, while research (Lazuardi et al., 2023) shows a significant improvement in positive and negative sentiment classification using BERT compared to the baseline. The use of IndoBERT as a pre-trained Indonesian language model further improves text processing performance in the Indonesian language domain.

A number of studies have utilized app reviews on Google Play as a source of sentiment analysis data. For example, studies (Putra et al, 2024) and (Nurwidyantoro et al, 2025) explored thousands of reviews by comparing traditional classification models and deep learning. Meanwhile, large-scale studies on generative AI applications (such as ChatGPT) have analyzed hundreds of thousands of reviews to identify dominant topics, expectations, and major issues experienced by users (Liu et al., 2024; Yang et al., 2025). This confirms that application reviews are a rich empirical data source for evaluating user experience.

On the other hand, studies on generative AI applications such as ChatGPT are still limited, especially those that apply the Knowledge Discovery in Database (KDD) methodology end-to-end—from data collection, data preprocessing, data transformation, data mining, to evaluation. Therefore, this research is relevant by combining the KDD methodology and the BERT algorithm to produce an accurate sentiment analysis model and provide insights into the perceptions of ChatGPT application users on the Google Play Store platform.

Jurnal Info Sains : Informatika dan Sains
Volume 15, Number 02, 2025, DOI 10.54209/infosains.v15i02
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

Thus, this research not only has practical urgency for application developers to improve service quality, but also academic urgency in enriching studies related to sentiment analysis, NLP, and the development of generative AI applications in Indonesia.

METHODS

In this study, the object is the user review of ChatGPT users sentiment regarding application of ChatGPT on the Google Play Store. The data used in 2025 uses Indonesian language data. The algorithm used in this study is Bidirectional Encoder Representations from Transformer (BERT) with the Knowledge Discovery In Database (KDD) method which is a series of steps designed to find patterns or knowledge from datasets. The stages of this process include data selection, data preprocessing, data transformation, data mining, and evaluation (Rahayu et al., 2024).

Data Selection

In this first stage, descriptive statistics are conducted to obtain an overview of the review data regarding the sentiment of user on the ChatGPT application. Data is taken using a scraping technique by using the library google-play-scraper. Web scraping is a technique for automatically extracting data from web pages using programs or bots. This technique is commonly used in research to collect data from digital platforms, including the Google Play Store (Kumar et al., 2025). Search parameters are set by entering the application ID to com.openai.chatgpt to get user reviews, language and country to Indonesia, sort by most relevant to get the most relevant reviews, limit the scraping up to 2000 reviews and setting the time span of review during October 2025 in Indonesian. The data obtained will be classified into positive, negative and neutral labels by researchers manually and then validated and reviewed by an Indonesian language teacher.

Data Preprocessing

Data cleaning is performed during the preprocessing stage to remove and reduce data noise. This stage is conducted to ensure that the collected data is clean and easier to process in subsequent steps. The preprocessing process aims to improve data quality so that the results of the calculations and analysis become more optimal. The preprocessing stages are as follows.

1. Cleaning
   This stage aims to reduce noise in user review data. Duplicate data and irrelevant elements such as emoji, punctuation marks, and other special symbols are removed.
2. Case Folding
   All characters in the dataset are converted to lowercase to simplify the processing steps and maintain data consistency.
3. Normalization
   Non-standard or informal words are transformed into their standard forms according to linguistic rules.
4. Filtering
   Words that do not carry significant meaning, such as “di”, “ke”, “lah”, ”pun” and similar terms, are removed using a stopword list.

Jurnal Info Sains : Informatika dan Sains
Volume 15, Number 02, 2025, DOI 10.54209/infosains.v15i02
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

5. Stemming
   Words containing affixes are converted into their root or base forms to reduce word variations.

Data Transformation

Data Transformation In data analysis or modeling, transformation is used as a process of changing data from its original structure or format into a format that is more appropriate and informative for use. In this study, data transformation includes the following processes.

1. Token Embeddings
   Numerical vectors that represent each token, whether a word or subword, within the input text. These embeddings are derived from BERT's pre-trained vocabulary, which is designed to capture contextual and semantic information of words effectively.
2. Segment Embeddings
   Embedding vectors that function to differentiate between distinct segments of input text, such as in sentence-pair classification tasks. This mechanism enables the model to recognize and process multiple sentences within a single input representation.
3. Position Embeddings
   Vectors that encode the positional information of tokens in a sequence. Because Transformer-based models like BERT do not inherently model word order, position embeddings are essential to preserve the sequential structure of the input text.
4. Encoding
   A stage where token embeddings, segment embeddings, and position embeddings are integrated into a unified representation. This combined representation is then used as the primary numerical input for processing by the Transformer layers in the BERT model.

Data Mining

In the Knowledge Discovery in Database (KDD) data mining is the process of discovering meaningful patterns, relationships, and knowledge from large-scale data using techniques derived from statistics, machine learning, and artificial intelligence, data mining enables the extraction of valuable information from complex datasets to support predictive and descriptive analysis. In sentiment analysis research, data mining is commonly applied to classify textual data into predefined sentiment categories (Han et al. 2022). Classification is one of the most widely used data mining techniques for analyzing textual data, as it allows the grouping of data based on specific characteristics. In this study, data mining is utilized to classify user reviews of the ChatGPT application into positive, negative, and neutral sentiments, enabling a structured understanding of user perceptions (Rahayu et al. 2024).

The data mining stage involves splitting the labeled review dataset into two subsets, namely training data and testing data. The dataset is divided using an 80:20 ratio, with 80% allocated for model training and 20% for model evaluation. This proportion is selected based on the Pareto principle, which is commonly applied in machine learning modeling.

During the training phase, the BERT model utilizes the training data, where numerical representations obtained from the transformed data are processed through BERT's Transformer layers to capture contextual information in sentences bidirectionally.

Jurnal Info Sains : Informatika dan Sains
Volume 15, Number 02, 2025, DOI 10.54209/infosains.v15i02
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

Subsequently, the testing data is used to evaluate the trained model by predicting sentiment labels positive, neutral, or negative to assess the model's performance.

RESULTS AND DISCUSSION

Data Mining

The data mining process is carried out using Bidirectional Encoder Representation from Transformer (BERT). IndoBERT is used as a pre-trained model specifically for Indonesian. The split data operator is applied to divide the training data and testing data with a ratio of 80:20.

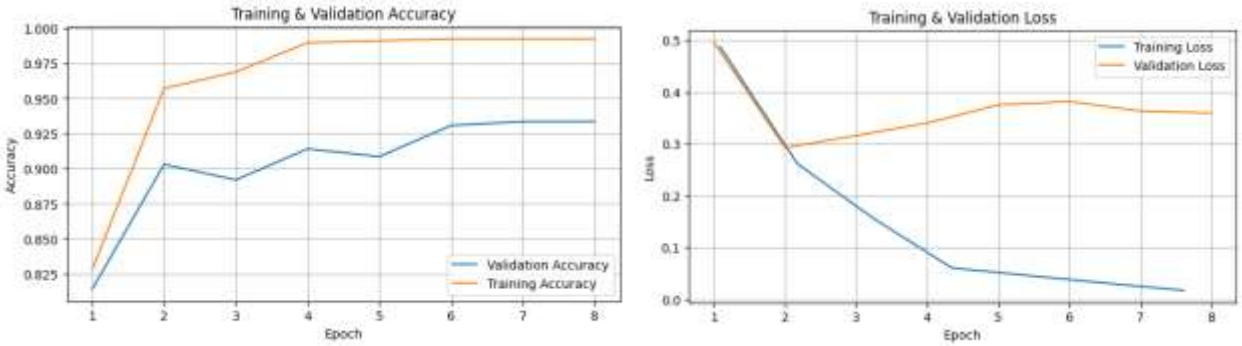

Figure 1. Training & Validation Accuracy dan Loss Graphic

After fine-tuning using BERT, it can be seen in Figure 1, the training accuracy graph has increased significantly up to 97%, while validation tends to start to stabilize at around 90%. Conversely, the loss graph in the training process decreases as the number of epochs increases. The loss in validation also decreases, but increases at the end of training

Evaluation

In this stage, an evaluation is carried out to measure the accuracy of the results of the BERT algorithm application that has been carried out previously. In this evaluation stage, the confusion matrix method is used to support the research objectives, with parameters including accuracy, precision, recall, and f1-score.

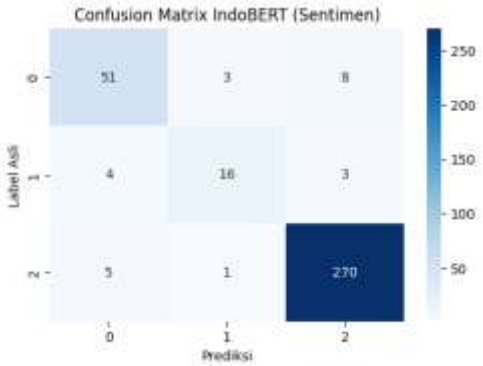

Figure 2. Confusion Matrix

Based on the confusion matrix results shown in Figure 2, it can be seen that the positive class (label 2) is classified very well, with 270 out of 276 data correctly classified. The neutral class (label 1) shows low performance, there are 7 data misclassifications (4 to negative and 3 to positive). The negative class (label 0) shows a decent performance, 51 of which are correctly classified, while the rest are incorrectly classified (3 to neutral and 8 to

Jurnal Info Sains : Informatika dan Sains
Volume 15, Number 02, 2025, DOI 10.54209/infosains.v15i02
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

positive). This indicates that the model has difficulty distinguishing neutral sentiment from other classes.

CLASSIFICATION REPORT:

	precision	recall	f1-score	support
0	0.85	0.82	0.84	62
1	0.80	0.70	0.74	23
2	0.96	0.98	0.97	276
accuracy			0.93	361
macro avg	0.87	0.83	0.85	361
weighted avg	0.93	0.93	0.93	361

Figure 3. Classification Report

As can be seen from Figure 3, from the three sentiment categories, the positive sentiment class shows good performance with a precision value of 0.96, recall 0.98, and f1-score 0.97. The neutral category shows poor values with a precision value of 0.80, recall 0.70, and f1-score 0.74. And the model performance for the negative category shows good values, with a precision of 0.85, recall 0.82, and f1-score 0.84 The macro average is the average value of precision, recall, and f1-score which show values of 0.87, 0.83, and 0.85 respectively. And the weighted average it weights each class contribution by its support value of precision, recall, and f1-score which show values of 0.93, 0.93, and 0.93 respectively. The final accuracy result of the model is 93%.

CONCLUSION

Based on the findings of this study, it can be concluded that sentiment analysis was successfully performed on user reviews of the ChatGPT application obtained from the Google Play Store. The research applied the Knowledge Discovery in Database (KDD) framework, which included dataset selection, data preprocessing, and transformation into numerical representations, followed by fine-tuning using the Bidirectional Encoder Representation from Transformers (BERT) model and evaluation through a confusion matrix. The performance evaluation results, using an 80:20 data split between training and testing data, indicate that the model achieved satisfactory performance, with an accuracy of 93%, precision of 87%, recall of 83%, and an F1-score of 85%. For future research, it is recommended to increase and balance the dataset to potentially enhance classification accuracy. Additionally, experimenting with different data split scenarios may help identify the most optimal model performance. Further studies may also explore the use of alternative algorithms or variations in hyperparameter settings to determine the most effective approach for sentiment analysis.

REFERENCE

Alqadah, F., Alenezi, M., & Alharbi, S. (2025). User perception and adoption of generative AI applications: A large-scale analysis of ChatGPT reviews. Journal of Artificial Intelligence Research. URL: https://www.jair.org

App Radar. (2023). Google Play Store policies and app curation guidelines 2024–2025. App Radar. URL: https://appradar.com/blog/google-play-store-policies

Basiri, M. E., Nemati, S., Abdar, M., & Acharya, U. R. (2025). A comprehensive survey on deep learning-based sentiment analysis. Expert Systems with Applications, 232, 120685. DOI URL: https://doi.org/10.1016/j.eswa.2023.120685

Han, J., Kamber, M., & Pei, J. (2022). Data mining: Concepts and techniques (4th ed.). Morgan Kaufmann. DOI URL: https://doi.org/10.1016/C2019-0-01632-7

Kumar, A., Singh, R., & Verma, P. (2025). Web scraping techniques for data collection from online platforms. International Journal of Data Science. URL: https://www.inderscience.com/jhome.php?jcode=ijds

Kusuma, R. A., Wibowo, A., & Nugroho, Y. (2021). Sentiment analysis of Indonesian application reviews using BERT. Journal of Information Systems, 17(2), 85–94. URL: https://journal.uii.ac.id/JIS

Lazuardi, R., Prasetyo, D., & Hidayat, A. (2023). Improving sentiment classification of Indonesian app reviews using BERT-based models. Journal of Big Data Analytics. URL: https://journal.springer.com/big-data-analytics

Liu, Y., Zhang, H., & Chen, Q. (2024). Understanding user expectations of generative AI through large-scale app review analysis. Computers in Human Behavior, 149, 107892. DOI URL: https://doi.org/10.1016/j.chb.2023.107892

Nurwidyantoro, A., Prabowo, R., & Sari, D. P. (2025). Comparative analysis of traditional and deep learning models for sentiment analysis on Google Play reviews. Journal of Software Engineering. URL: https://www.springer.com/journal/software-engineering

Putra, A. S., Rahman, F., & Wicaksono, A. (2024). Sentiment analysis of mobile application reviews using machine learning and deep learning approaches. Indonesian Journal of Computing. URL: https://journal.unnes.ac.id/sju/index.php/ijc

Rahayu, S., Fitriani, N., & Saputra, R. (2024). Implementation of Knowledge Discovery in Database (KDD) methodology for text classification. Journal of Information Science. URL: https://journals.sagepub.com/home/jis

Vincent, J. (2023). Google Play introduces AI-generated review summaries. The Verge. URL: https://www.theverge.com

Xia, M., Li, Z., & Wang, T. (2025). Developer responses to user reviews and their impact on mobile app success. Empirical Software Engineering. URL: https://link.springer.com/journal/10664

Yadav, A., Kumar, S., & Sharma, R. (2024). Deep learning models for sentiment analysis: A comparative study. International Journal of Intelligent Systems, 39(5), e22987. DOI URL: https://doi.org/10.1002/int.22987

Yang, L., Zhou, W., & Li, J. (2025). Topic and sentiment evolution in generative AI application reviews. IEEE Access, 13, 45678–45690. DOI URL: https://doi.org/10.1109/ACCESS.2025.3456789