Intechno Journal :
Information Technology Journal
Vol. 8 No. 1 July 2026
Homepage: https://jurnal.amikom.ac.id/index.php/intechno
INTECHNO Journal
INFORMATION TECHNOLOGY JOURNAL
e-ISSN: 2655-1438
p-ISSN: 2655-1632

# Sentiment Analysis and Emotion Detection of ChatGPT Users on Google Play Store Using the Naïve Bayes Algorithm

**Bagus Setya[1], Merlin Diana[2], Ikhwan[3]**
[1,2,3]*Information System, Universitas Pamulang, Serang, Indonesia*

**Article Info**

***Article history:***

Received: 30 June 2026
Revised: 17 July 2026
Accepted: 21 July 2026
Published: 31 July 2026

***Keywords:***

Sentiment Analysis
Emotion Detection
Naive Bayes
ChatGPT
Machine Learning

**ABSTRACT**

With the rapid adoption of artificial intelligence applications such as ChatGPT, understanding both users' sentiments and emotional expressions has become increasingly important for evaluating user experience and improving service quality. This study aims to analyze user sentiment and identify emotional tendencies expressed in Google Play Store reviews of the ChatGPT application. Methods: A total of 10,000 user reviews were collected through web scraping from the Google Play Store. The collected data underwent several preprocessing stages, including data cleaning, case folding, tokenization, normalization, stopword removal, and stemming. Sentiment analysis was performed using the Multinomial Naïve Bayes algorithm to classify reviews into positive, neutral, and negative categories. Emotion detection was subsequently conducted by interpreting dominant lexical patterns and frequently occurring terms within each sentiment category through Word Cloud visualization, enabling the identification of users' emotional tendencies rather than direct emotion classification. Model performance was evaluated using accuracy, precision, recall, and F1-score. Results: The sentiment distribution showed that 84.7% of the reviews were classified as positive, 3.9% as neutral, and 11.4% as negative. The best-performing model, using an 80:20 train–test split, achieved an accuracy of 86.8%, precision of 83.8%, recall of 86.8%, and an F1-score of 82.0%. The emotion detection results revealed that positive reviews were dominated by words reflecting satisfaction, usefulness, and productivity, whereas negative reviews mainly expressed frustration related to subscription costs, system errors, and application performance. Novelty: Unlike previous studies that primarily focused on sentiment classification, this research combines machine learning-based sentiment analysis with lexical interpretation of emotional tendencies derived from user-generated reviews. This integrated approach provides a more comprehensive understanding of users' perceptions and emotional responses toward ChatGPT, offering practical insights for improving AI-based application services.



***Corresponding Author:***

Bagus Setya
Email: dosen03357@unpam.ac.id

## 1. INTRODUCTION

The development of Artificial Intelligence (AI) in Indonesia has gained significant momentum in recent years, particularly in the sectors of education, healthcare, agriculture, and public services.

However, its implementation continues to face several challenges, including uneven digital infrastructure, a shortage of skilled human resources capable of developing and utilizing AI technologies, and regulatory frameworks that are not yet fully prepared to address issues related to ethics, privacy, and data security. For example, in the education sector, literature studies have identified that AI research in Indonesia is still predominantly focused on applications in education, economics, and healthcare, while research related to social, technological, ethical, and public aspects remains relatively underdeveloped and unevenly distributed.[1]

The level of public sentiment toward this application indicates a trend that warrants further analysis to determine whether user perceptions are predominantly positive or negative. The rapid advancement of computer-based information technology has significantly influenced transformations across various aspects of human life. One of the most important innovations emerging from this progress is Artificial Intelligence (AI), which represents a major achievement in the evolution of modern technology.[2] Artificial Intelligence (AI) enables computer systems to perform a wide range of tasks that were previously achievable only by humans, making it one of the most widely adopted technologies across various sectors today. Its presence has provided significant convenience by helping people meet diverse needs while improving the efficiency of daily activities.[3]

One example of this technology is a chatbot, a conversational system powered by Artificial Intelligence (AI). In general, a chatbot is a computer program designed and developed to interact with users through text- or voice-based inputs. This technology integrates AI with Natural Language Processing (NLP), enabling it to understand the context of conversations and generate intelligent, accurate, and contextually relevant responses to users' questions or commands.[4]

ChatGPT is an Artificial Intelligence (AI) technology developed by OpenAI that is trained to emulate human conversation through Natural Language Processing (NLP) techniques. In practice, it is capable of generating text in a variety of styles, ranging from academic and technical writing to formal communication, while maintaining a high level of coherence and contextual relevance. This capability is made possible through the use of deep learning algorithms, which enable the system to understand context, interpret user inputs, and generate responses that align with the instructions or prompts provided by users.[5]

The remarkable capabilities of ChatGPT have sparked considerable debate among academics and researchers. One of the primary concerns is that excessive reliance on this technology may reduce individuals' motivation to independently produce scholarly and scientific work. This concern stems from ChatGPT's ability to generate high-quality text that closely resembles academic and scientific writing. Given this phenomenon, it is important to conduct an analysis of public sentiment toward the use of ChatGPT as a representation of advancements in Artificial Intelligence (AI) technology. Sentiment analysis is a process used to identify and evaluate the attitudes, emotions, opinions, and judgments expressed by individuals toward a particular person, organization, product, service, or activity. The primary objective of this analysis is to determine whether the expressed perception reflects a positive, negative, or neutral sentiment.[6]

The Google Play Store provides application descriptions, user reviews, and ratings that offer insights into the strengths and weaknesses of available applications. In 2017, Hartmann-Boyce and colleagues conducted an evaluation of various applications on the Google Play Store to analyze user preferences and feedback, particularly for applications designed for weight loss and weight management. Their findings indicated that user ratings and reviews have a significant impact on an application's success in the digital marketplace. These findings are also relevant to AI-based applications such as ChatGPT, as well as to the development of other applications available on the Google Play Store. Furthermore, the study highlighted that discrepancies between user ratings and written reviews may reflect differences in users' perceptions of an application's quality, functionality, and overall performance. [7]

The Naïve Bayes algorithm is a probabilistic classification method based on Bayes' Theorem, which assumes that all features or variables are independent of one another. Due to its simplicity and computational efficiency, this algorithm is widely applied in various data analysis tasks, including text classification, spam detection, and sentiment analysis. In the context of sentiment analysis, Naïve Bayes operates by calculating the probability that a given text belongs to a particular sentiment category—such as positive, negative, or neutral—based on the frequency and distribution of specific words within the text.[8] Despite its simplicity, the Naïve Bayes algorithm is well known for its relatively high

classification accuracy and fast computational performance. These advantages make it a popular baseline method for the development of Artificial Intelligence (AI) applications and Natural Language Processing (NLP) systems, particularly in tasks involving text mining and sentiment analysis.[9]

Although recent Natural Language Processing (NLP) models, such as Long Short-Term Memory (LSTM), Bidirectional Encoder Representations from Transformers (BERT), and IndoBERT, have demonstrated superior performance in various text classification tasks, these approaches generally require substantially larger training datasets, higher computational resources, and longer training times. In contrast, the Naïve Bayes algorithm was selected in this study because it provides an efficient and computationally lightweight solution for sentiment classification, particularly when applied to medium-sized textual datasets with limited computational resources. Moreover, Naïve Bayes has consistently shown competitive performance in sentiment analysis tasks due to its simplicity, fast training process, and robustness in handling high-dimensional sparse text features generated through TF-IDF weighting. Therefore, this algorithm was considered appropriate for establishing a reliable baseline model for analyzing user reviews of the ChatGPT application while maintaining computational efficiency. Future studies may compare the performance of Naïve Bayes with more advanced transformer-based models, such as BERT or IndoBERT, to investigate potential improvements in classification accuracy and emotion recognition.

## 2. METHOD

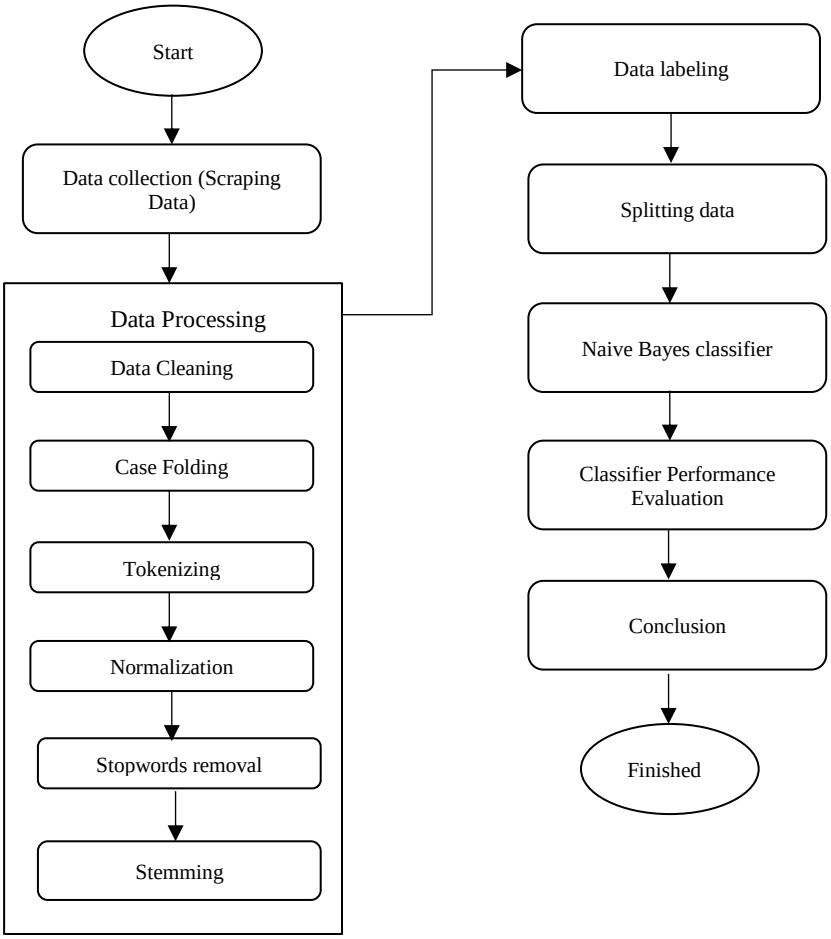



Figure 1. Reseach Flow

As illustrated in Figure 1, this study aims to analyze and classify users' emotions and sentiments toward the ChatGPT application based on reviews collected from the Google Play Store, using the Naïve Bayes algorithm as the classification method. The dataset consists of user-generated review texts that undergo several text preprocessing stages, including tokenization, stopword removal, and stemming, before being used for model

training and testing. The findings of this study are expected to provide a quantitative and objective overview of users' perceptions and emotional responses toward ChatGPT, enabling a more systematic and measurable evaluation of user feedback.

A. Data Collection (Data Scraping)

The data used in this study were collected from user reviews of the ChatGPT application available on the Google Play Store through a web scraping technique. The collected data included user review texts, star ratings, review dates, and other publicly available user information. The data extraction process was carried out using Google Colab and Python libraries such as Google Play Scraper and BeautifulSoup, which facilitated the automated retrieval of review data from the Google Play Store.[10]

B. Data Cleaning

After the data were successfully collected, a data cleaning process was performed to improve the overall quality and reliability of the dataset. This stage involved removing duplicate records, eliminating unnecessary symbols and special characters, handling missing values, and filtering out noise that could potentially interfere with the analysis process. The primary objective of this step was to ensure that the dataset was clean, consistent, and ready for subsequent processing and analysis.[11]

C. Text Preprocessing

The cleaned text data were subsequently processed through several Natural Language Processing (NLP) stages to prepare them for machine learning analysis. These stages included case folding, which converts all text into lowercase letters; tokenization, which breaks sentences into individual words or tokens; stopword removal, which eliminates common words that carry little semantic value; word normalization, which transforms non-standard or informal words into their standard forms; and stemming, which reduces words to their root forms. These preprocessing steps were conducted to produce a more structured and consistent dataset, thereby improving the effectiveness and accuracy of the machine learning algorithms used in the analysis.[12]

D. Labeling

Data labeling in this study was performed using a lexicon-based approach, a sentiment analysis method that relies on a sentiment lexicon containing words assigned with predefined sentiment scores or weights. These sentiment weights serve as references for determining and calculating the sentiment orientation of each word within the text. The labeling process utilized InSet Lexicon, an Indonesian-language sentiment dictionary specifically developed to support sentiment analysis of Indonesian texts. By employing this lexicon, each word in the dataset could be identified and classified according to its associated sentiment value, enabling the automatic assignment of sentiment labels to the collected reviews.[13]

E. Data Splitting

The collected review data were subsequently divided into two subsets: training data and testing data. The training data were used to build and train the classification model based on the Naïve Bayes algorithm, while the testing data were employed to evaluate the model's performance and classification accuracy. Through this evaluation process, the effectiveness of the algorithm in classifying previously unseen data could be objectively assessed, providing a reliable measure of its predictive capability.[14]

F. Use of Naïve Bayes

The classification model was subsequently developed using the Naïve Bayes algorithm. This algorithm learns word patterns from the training dataset and calculates the probability of word occurrences to determine the most appropriate sentiment category for each review. Based on the classification results, user reviews were grouped into three sentiment categories: positive, negative, and neutral. In addition to sentiment classification, the results were also utilized to provide insights into users' emotional responses and overall perceptions of the ChatGPT application.[15]

## 3. RESULTS AND DISCUSSIONS

### A. Data Collection Results

The research data were collected through a web scraping process from the ChatGPT application page on the Google Play Store. This process successfully retrieved a large number of user reviews, which were subsequently used as the research dataset. The collected data included review texts, star ratings, review dates, and other relevant supporting information. Following the data collection stage, a data cleaning process was conducted to remove duplicate records, special characters, hyperlinks, and irrelevant information. This step ensured that the resulting dataset was clean, consistent, and suitable for further analysis and model development

### B. Data Pre-processing Results

The preprocessing stage was carried out to improve the quality of the textual data before it was used for the classification process. This stage included several Natural Language Processing (NLP) techniques, namely case folding, tokenization, stopword removal, word normalization, and stemming. The outcome of these preprocessing steps was a more structured and standardized collection of words, which helped enhance the performance of the Naïve Bayes algorithm in identifying sentiment patterns within user reviews and improving the overall accuracy of the classification model.

### C. Web Scraping

The data were collected using a web scraping technique through Google Colaboratory by utilizing the google-play-scraper library to extract user reviews from the Google Play Store. A total of 10,000 reviews were successfully retrieved and used as the dataset for this study. The results of the raw data extraction process can be seen in Figure 2. Subsequently, the collected data were organized into four primary attributes—userName, at, score, and content—which served as the basis for the data filtering and preprocessing stages. This organization facilitated a more systematic and efficient analysis process in the subsequent stages of the research.

| userName | content | score | at |
|---|---|---|---|
| _MZ IBET_ | sangat naguss | 5 | 02/04/2026 10:38 |
| 06. Khoirul Anam | keren dan sangat membantu sekali | 5 | 03/04/2026 04:47 |
| 18 iputugedehendradarmawan | di suruh hd in foto malah ceramah astaga hampir banting hp | 1 | 02/04/2026 22:37 |
| 1D_021_nur hidayah | bagus | 5 | 03/04/2026 06:08 |
| 26_Rahmi Sahira Septiani | keren wak, bisa jadi "temen" ngobrol | 5 | 03/04/2026 21:29 |
| Aandri Sugianto | bagus | 5 | 02/04/2026 23:25 |
| Abdul Gofur | Bagus unut belajar bersama sama | 5 | 01/04/2026 18:32 |
| Abdul Muizz | bagus banget buat ini itu | 5 | 04/04/2026 21:11 |
| Abdul Rasyid Rifani | bisa diandalkan | 5 | 01/04/2026 18:27 |

Figure 2. Web Scraping Results

Since this study only required the content and score attributes, a data filtering process was performed to remove unnecessary columns, retaining only these two variables for further analysis. The data contained in the content and score columns were subsequently utilized to analyze user reviews and classify them into positive and negative sentiment categories, as illustrated in Figure 3. This filtering step helped streamline the dataset and ensured that only relevant information was used in the sentiment analysis process.

| | content | score |
|---|---|---|
| 0 | terbaik.. | 5 |
| 1 | saya suka sesuai dengan saya minta | 3 |
| 2 | benar benar sangat membantu.. | 5 |
| 3 | asisten chat yang paling sesuai dengan keingin... | 5 |
| 4 | sangat bagus | 5 |

Figure 3. Filter web scraping data results

### D. Data Cleaning

The dataset used in this study consisted of two attributes: content and score. The content column was stored as an *object* data type, which is commonly used to represent textual data or a combination of different data formats, while the score column was stored as an int64 data type, indicating that it contained integer numerical

values. Furthermore, no missing values were identified in either column of the DataFrame, ensuring the completeness of the dataset.

The next stage involved text mining or data cleaning, which aimed to remove unnecessary elements from the textual data. This process was conducted to eliminate noise, such as special characters, hyperlinks, and other irrelevant information that could negatively affect the analysis results. The outcome of this cleaning process is illustrated in Figure 4.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 2 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   content  10000 non-null  object
 1   score    10000 non-null  int64
dtypes: int64(1), object(1)
memory usage: 156.4+ KB
```

Figure 4. Data information

After the data cleaning process was completed, the dataset became more structured and suitable for further analysis, as irrelevant and non-essential elements had been removed. This preprocessing step improved the overall quality and consistency of the textual data, enabling more effective feature extraction and sentiment classification. The results of the cleaning process are presented in Figure 5.

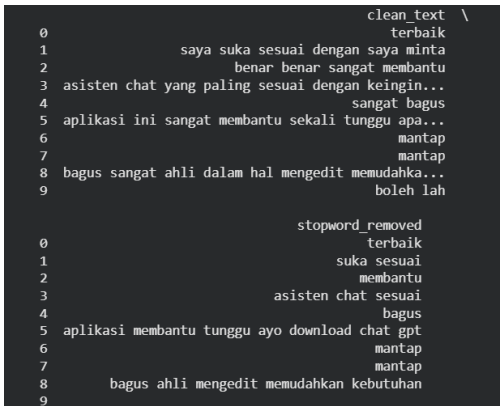
```
   HASIL DATA CLEANING
                                                content  \
... 0                                          terbaik..
    1                  saya suka sesuai dengan saya minta
    2                       benar benar sangat membantu..
    3  asisten chat yang paling sesuai dengan keingin...
    4                                       sangat bagus
    5  aplikasi ini sangat membantu sekali, tunggu ap...
    6                                            mantap
    7                                            MANTAP
    8  bagus sangat ahli dalam hal mengedit memudahka...
    9                                         boleh lah

                                         clean_text
    0                                        terbaik
    1                                   suka sesuai
    2                                      membantu
    3                           asisten chat sesuai
    4                                         bagus
    5  aplikasi membantu tunggu ayo download chat gpt
    6                                         mantap
    7                                         mantap
    8       bagus ahli mengedit memudahkan kebutuhan
    9
```

Figure 5. Data cleaning results

E. Stopword Removal

The stopword removal process was then applied, during which commonly used words such as *"lama," "bisa," "di,"* and *"ke"* were removed because they belong to the category of stopwords. These words frequently appear in everyday language but generally contribute little meaningful information in the context of sentiment analysis. As a result, the text became more concise and focused on terms with greater informational value. By eliminating less relevant words, the stopword removal process enhanced the overall quality of the dataset, thereby improving the effectiveness and accuracy of subsequent analytical procedures. The results of this process are shown in Figure 6.



| | clean_text |
|---|---|
| 0 | terbaik |
| 1 | saya suka sesuai dengan saya minta |
| 2 | benar benar sangat membantu |
| 3 | asisten chat yang paling sesuai dengan keingin... |
| 4 | sangat bagus |
| 5 | aplikasi ini sangat membantu sekali tunggu apa... |
| 6 | mantap |
| 7 | mantap |
| 8 | bagus sangat ahli dalam hal mengedit memudahka... |
| 9 | boleh lah |

| | stopword_removed |
|---|---|
| 0 | terbaik |
| 1 | suka sesuai |
| 2 | membantu |
| 3 | asisten chat sesuai |
| 4 | bagus |
| 5 | aplikasi membantu tunggu ayo download chat gpt |
| 6 | mantap |
| 7 | mantap |
| 8 | bagus ahli mengedit memudahkan kebutuhan |
| 9 | |

Figure 6. Stopword removal process results

## F. Stemming

The next stage was stemming, a process that transforms words resulting from the previous preprocessing steps into their root or base forms. In this study, stemming was performed using the Sastrawi library, a tool specifically designed for Indonesian language processing. Sastrawi operates by removing affixes, including prefixes and suffixes, allowing words to be reduced to their fundamental forms. The use of Sastrawi is highly beneficial in Indonesian text mining and Natural Language Processing (NLP) tasks, as it helps standardize word variations into more consistent representations. This normalization process contributes to improved text analysis and classification performance. Examples of data after the stemming process are presented in Figure 7.

| content | score | clean_text | stopword_removed | tokens | stemming |
|---|---|---|---|---|---|
| terbaik.. | 5 | terbaik | terbaik | ['terbaik'] | ['baik'] |
| saya suka sesuai dengan saya minta | 3 | saya suka sesuai dengan saya minta | suka sesuai | ['suka', 'sesuai'] | ['suka', 'sesuai'] |
| benar benar sangat membantu.. | 5 | benar benar sangat membantu | membantu | ['membantu'] | ['bantu'] |
| asisten chat yang paling sesuai dengan keinginan saya | 5 | asisten chat yang paling sesuai dengan keinginan saya | asisten chat sesuai | ['asisten', 'chat', 'sesuai'] | ['asisten', 'chat', 'sesuai'] |
| sangat bagus | 5 | sangat bagus | bagus | ['bagus'] | ['bagus'] |
| aplikasi ini sangat membantu sekali, tunggu apalagi ayo segera download chat gpt | 5 | aplikasi ini sangat membantu sekali tunggu apalagi ayo segera download chat gpt | aplikasi membantu tunggu ayo download chat gpt | ['aplikasi', 'membantu', 'tunggu', 'ayo', 'download', 'chat', 'gpt'] | ['aplikasi', 'bantu', 'tunggu', 'ayo', 'download', 'chat', 'gpt'] |
| mantap | 5 | mantap | mantap | ['mantap'] | ['mantap'] |
| MANTAP | 4 | mantap | mantap | ['mantap'] | ['mantap'] |
| bagus sangat ahli dalam hal mengedit memudahkan kebutuhan | 5 | bagus sangat ahli dalam hal mengedit memudahkan kebutuhan | bagus ahli mengedit memudahkan kebutuhan | ['bagus', 'ahli', 'mengedit', 'memudahkan', 'kebutuhan'] | ['bagus', 'ahli', 'edit', 'mudah', 'butuh'] |

Figure 7. Stemming Results

## G. Labelling

The data labeling process was conducted to identify the sentiment expressed in each review by assigning it to a specific sentiment category. This step plays an important role in enabling the computer system to understand users' perceptions, opinions, and evaluations of the reviewed product or service. After completing the preprocessing stage, each review was assigned a sentiment label according to predefined classification criteria. These labels served as the basis for subsequent sentiment analysis and model training. The results of the labeling process are presented in Figure 8.

| content | score | clean_text | stopword_removed | tokens | stemming | label |
|---|---|---|---|---|---|---|
| terbaik.. | 5 | terbaik | terbaik | ['terbaik'] | ['baik'] | positif |
| saya suka sesuai dengan saya minta | 3 | saya suka sesuai dengan saya minta | suka sesuai | ['suka', 'sesuai'] | ['suka', 'sesuai'] | netral |
| benar benar sangat membantu.. | 5 | benar benar sangat membantu | membantu | ['membantu'] | ['bantu'] | positif |
| asisten chat yang paling sesuai dengan keinginan saya | 5 | asisten chat yang paling sesuai dengan keinginan saya | asisten chat sesuai | ['asisten', 'chat', 'sesuai'] | ['asisten', 'chat', 'sesuai'] | positif |
| sangat bagus | 5 | sangat bagus | bagus | ['bagus'] | ['bagus'] | positif |
| aplikasi ini sangat membantu sekali, tunggu apalagi ayo segera download chat gpt | 5 | aplikasi ini sangat membantu sekali tunggu apalagi ayo segera download chat gpt | aplikasi membantu tunggu ayo download chat gpt | ['aplikasi', 'membantu', 'tunggu', 'ayo', 'download', 'chat', 'gpt'] | ['aplikasi', 'bantu', 'tunggu', 'ayo', 'download', 'chat', 'gpt'] | positif |
| mantap | 5 | mantap | mantap | ['mantap'] | ['mantap'] | positif |
| MANTAP | 4 | mantap | mantap | ['mantap'] | ['mantap'] | positif |
| bagus sangat ahli dalam hal mengedit memudahkan kebutuhan | 5 | bagus sangat ahli dalam hal mengedit memudahkan kebutuhan | bagus ahli mengedit memudahkan kebutuhan | ['bagus', 'ahli', 'mengedit', 'memudahkan', 'kebutuhan'] | ['bagus', 'ahli', 'edit', 'mudah', 'butuh'] | positif |

Figure 8. Labeling Results

Sentiment labels were assigned based on the rating scores provided by users when submitting their reviews. These ratings serve as numerical representations of users' levels of satisfaction and overall evaluations of the

application. In this study, the review data were classified into three sentiment categories: positive, neutral, and negative. Reviews with ratings of 4 and 5 were categorized as positive sentiment, reviews with a rating of 3 were classified as neutral sentiment, and reviews with ratings below 3 were categorized as negative sentiment. The distribution of these sentiment categories is illustrated in the pie chart presented in Figure 9.

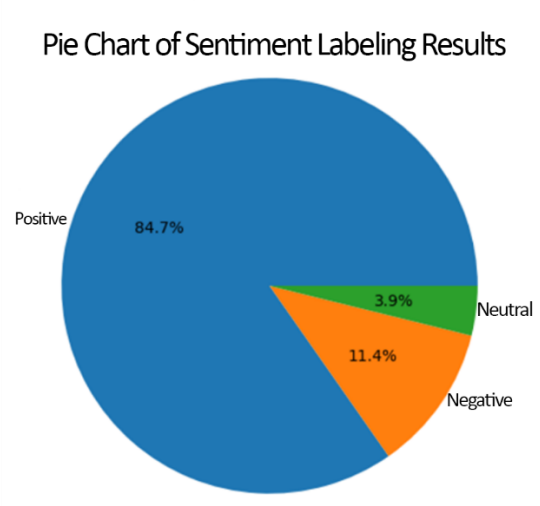



Figure 9. Pie Chart Of Labeling Results

Figure 9 presents the results of the sentiment labeling process for the 10,000 user reviews analyzed in this study. The findings indicate that 84.7% of the reviews were classified as positive sentiment, 3.9% as neutral sentiment, and 11.4% as negative sentiment. These results suggest that the majority of users expressed positive perceptions and experiences regarding the ChatGPT application, while only a relatively small proportion of reviews reflected neutral or negative opinions.

## H. TF-IDF

During the feature extraction stage, the Scikit-Learn library was utilized, as it provides a wide range of modules and functions for text data processing and machine learning tasks. One of the key functions employed in this stage was TfidfVectorizer, which converts a collection of review documents into a TF-IDF (Term Frequency–Inverse Document Frequency) feature matrix. Through this process, each word in the documents is transformed into a numerical value that reflects its relative importance within the dataset. The TF-IDF weighting scheme calculates scores for all terms appearing across the document collection to identify words with meaningful occurrence patterns. Words that are more relevant and distinctive within a document receive higher TF-IDF weights, while commonly occurring words across many documents receive lower weights. This approach helps highlight informative terms that contribute more significantly to the sentiment classification process. The results of the TF-IDF weighting process are presented in Figure 10.

| | Term | Average TF-IDF Rank |
|---|---|---|
| 428 | bagus | 0.156074 |
| 591 | bantu | 0.075110 |
| 3254 | mantap | 0.045864 |
| 2729 | keren | 0.034362 |
| 534 | banget | 0.032809 |
| 290 | aplikasi | 0.028748 |
| 3880 | nya | 0.020450 |
| 1996 | good | 0.019158 |
| 451 | baguss | 0.019004 |
| 5162 | suka | 0.015281 |

Figure. 10. TF-IDF weighting results

I. Naïve Bayes

The Naïve Bayes method was employed in the classification stage using the Multinomial Naïve Bayes algorithm, which is widely used in text classification tasks due to its effectiveness in handling data represented as multinomial distributions. In this study, two data-splitting scenarios were implemented: 90:10 and 80:20. Under these scenarios, 90% or 80% of the dataset was allocated for training, while the remaining 10% or 20% was reserved for testing. After the model was trained using the training dataset, an evaluation phase was conducted to assess its performance. The evaluation was carried out using a confusion matrix, which provides insights into the model's classification performance by measuring metrics such as accuracy and the correctness of predictions across different sentiment classes. The evaluation results for each data-splitting scenario are presented in Figure 11 and Figure 12, respectively.

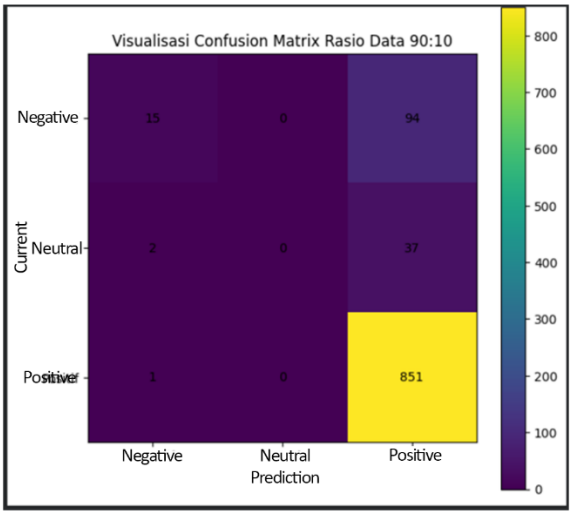



Figure 11. Confusion Matrix Visualization Of 90:10 Data Ratio

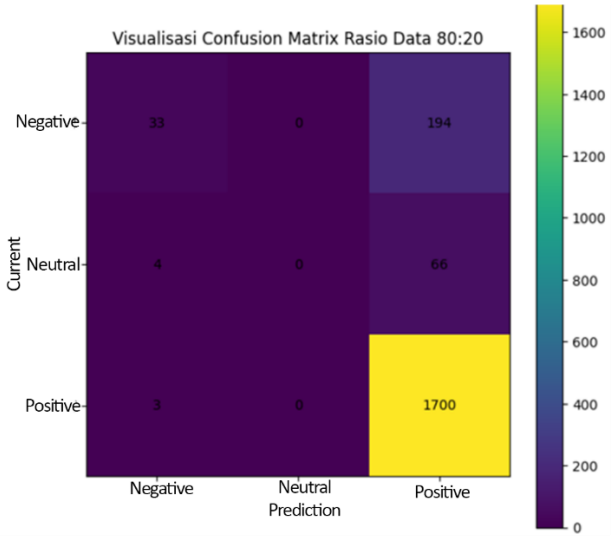



Figure 12. Confusion Matrix Visualization Of 80:20 Data Ratio

The next stage involved performing cross-validation using the Naïve Bayes Classifier (NBC) algorithm. In this process, the dataset was repeatedly partitioned into training and testing subsets to build and evaluate the classification model across multiple iterations. This approach helps provide a more reliable assessment of the model's generalization capability and reduces the risk of performance bias caused by a single train–test split. The model was evaluated using several performance metrics, including accuracy, precision, recall, and F1-score, which are commonly employed to assess the effectiveness of machine learning algorithms in classification tasks. These metrics were calculated based on the results obtained from the confusion matrix generated during the testing phase. The classification results achieved by the Naïve Bayes algorithm combined with TF-IDF weighting are presented in Figure 13.

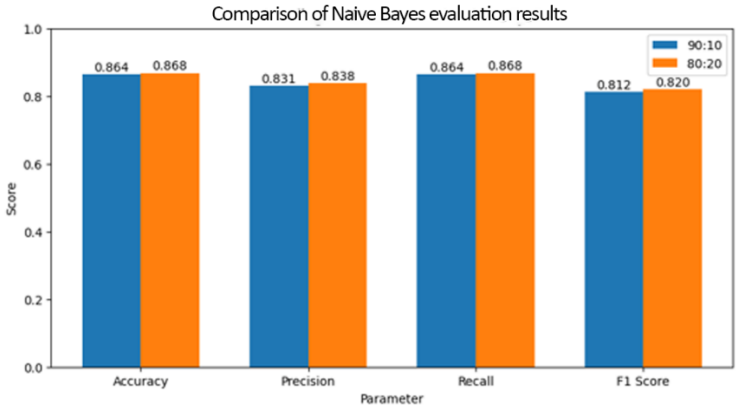



Figure 13. Model Evaluation Metric Results

Based on Figure 13, the evaluation results of the classification model using the Multinomial Naïve Bayes (MultinomialNB) algorithm indicate that the 80:20 data split ratio achieved better performance than the 90:10 ratio. Under the 80:20 scenario, the model attained an accuracy of 86.8%, precision of 83.8%, recall of 86.8%, and an F1-score of 82.0%. These results demonstrate that the model performs effectively in classifying sentiment data and exhibits a strong ability to correctly identify instances belonging to the positive sentiment category. Overall, the obtained performance metrics suggest that the MultinomialNB algorithm is capable of producing reliable classification results for sentiment analysis tasks on the collected user review data.

## J. Word Cloud

The final stage of this study involved data visualization using WordCloud, which was employed to display and identify the most frequently occurring words within each sentiment category. Through this visualization technique, dominant word patterns associated with positive, neutral, and negative sentiments can be observed more clearly, facilitating a deeper understanding and interpretation of the sentiment analysis results. By highlighting the most prominent terms in each category, the WordCloud provides valuable insights into the key topics, opinions, and expressions reflected in users' reviews.

Figure 14 illustrates the Word Cloud generated from reviews classified as positive sentiment. The visualization is dominated by words such as *helpful*, *good*, *easy*, *fast*, *recommend*, and *useful*, indicating that users generally expressed satisfaction with the ChatGPT application. These frequently occurring terms suggest positive emotional tendencies, including appreciation, trust, satisfaction, and confidence in the application's ability to assist users in learning, obtaining information, and improving productivity. The predominance of these positive expressions is consistent with the sentiment distribution results, in which 84.7% of the reviews were classified as positive.

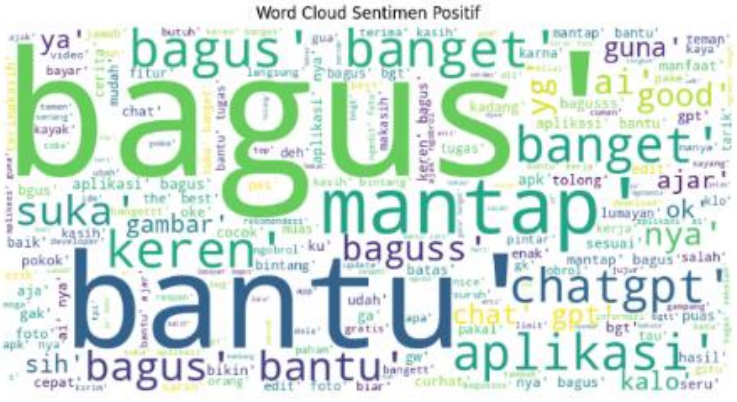



Figure 14. Positive Sentiment Word Cloud

Figure 15 presents the Word Cloud of reviews classified as neutral sentiment. Compared with the positive and negative categories, the neutral reviews contain fewer dominant words and are characterized by descriptive rather than evaluative expressions. Frequently occurring words mainly refer to the application's general features, updates, or user experiences without expressing strong positive or negative opinions. This finding indicates that neutral reviews primarily represent informational comments or balanced observations, reflecting neither satisfaction nor dissatisfaction. Consequently, the emotional tendency within this category can be interpreted as objective or indifferent.

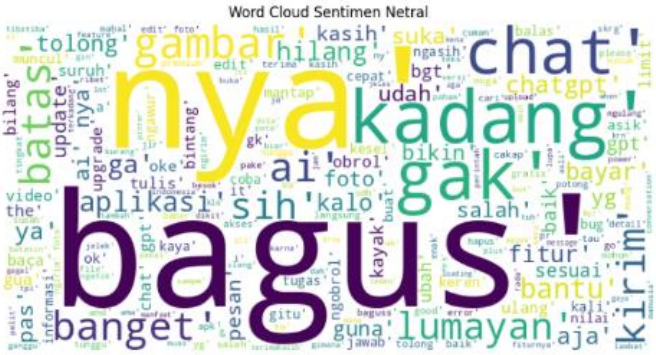



Figure 15. Neutral Sentiment Word Cloud

Figure 16 displays the Word Cloud for reviews classified as negative sentiment. Frequently appearing words include terms associated with *error*, *problem*, *subscription*, *login*, *slow*, and *failed*, indicating that users experienced technical issues, performance limitations, or dissatisfaction with premium service policies. These dominant words reflect negative emotional tendencies such as frustration, disappointment, and dissatisfaction. The visualization also provides developers with valuable insights into the aspects of the application that require further improvement, particularly system stability, accessibility, and subscription-related features.

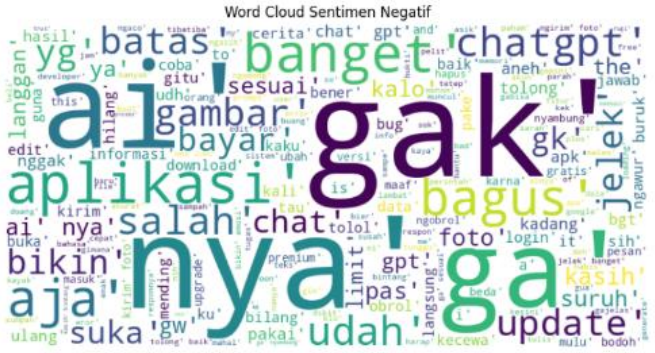



Figure 16. Negative Sentiment Word Cloud

Overall, the Word Cloud analysis complements the quantitative sentiment classification results by providing qualitative insights into users' emotional tendencies. While the Naïve Bayes classifier identifies the polarity of user opinions (positive, neutral, or negative), the Word Cloud reveals the dominant themes and emotional expressions underlying each sentiment category. Therefore, the combination of sentiment classification and lexical interpretation offers a more comprehensive understanding of user perceptions toward the ChatGPT application.

## 4. CONCLUSION

The experimental results indicate that the Multinomial Naïve Bayes model achieved its best performance using an 80:20 train–test split, obtaining an accuracy of 86.8%, precision of 83.8%, recall of 86.8%, and an F1-score of 82.0%. In addition, the sentiment distribution revealed that 84.7% of the reviews were classified as positive, 3.9% as neutral, and 11.4% as negative. These results demonstrate that the proposed approach is effective for sentiment classification while also providing meaningful insights into users' emotional tendencies through lexical analysis. The predominance of positive sentiment suggests that users are generally satisfied with the performance and benefits provided by ChatGPT. Many positive reviews indicate that the application effectively supports a variety of user activities, including enhancing learning experiences, facilitating information retrieval, and improving efficiency in completing daily tasks. These findings demonstrate that ChatGPT has delivered a favorable user experience for most of its users and has been positively received by the public.

Based on the findings of this study, several recommendations can be proposed for future research and application development. Future studies may consider utilizing larger and more diverse datasets to obtain more comprehensive insights into user sentiment. In addition, comparing the performance of Naïve Bayes with other machine learning or deep learning algorithms could provide a deeper understanding of the most effective approaches for sentiment classification. Further research may also incorporate more advanced emotion detection techniques and explore additional factors influencing user perceptions of AI-based applications. Such efforts could contribute to improving both the accuracy of sentiment analysis models and the overall quality of AI-powered services.

1. For ChatGPT application developers, the findings of this study may serve as valuable feedback for improving service quality, addressing system errors, and developing features that better align with users' needs and expectations.
2. Future research is encouraged to utilize larger datasets collected from multiple platforms, such as X (formerly Twitter), Reddit, and Apple App Store, in order to obtain more comprehensive and representative sentiment analysis results.
3. Future studies may also compare the performance of the Naïve Bayes algorithm with other machine learning approaches, such as Support Vector Machine (SVM), Random Forest, and Deep Learning models, to identify classification methods that yield superior predictive performance.
4. Subsequent research could incorporate aspect-based sentiment analysis and more advanced emotion detection techniques to gain deeper insights into the specific factors that influence user satisfaction and dissatisfaction with the application.

**REFERENCES**


[1] G. F. Febrian and G. Figueredo, "KemenkeuGPT: Leveraging a Large Language Model on Indonesia's Government Financial Data and Regulations to Enhance Decision Making," 2024, [Online]. Available: http://arxiv.org/abs/2407.21459

[2] S. A. R. Rizaldi, S. Alam, and I. Kurniawan, "Analisis Sentimen Pengguna Aplikasi JMO (Jamsostek Mobile) Pada Google Play Store Menggunakan Metode Naive Bayes," *STORAGE J. Ilm. Tek. dan Ilmu Komput.*, vol. 2, no. 3, pp. 109–117, 2023, doi: 10.55123/storage.v2i3.2334.

[3] mukrodin mukrodin and N. Mega Sasmita, "rtificial Inteligence Dalam Apilkasi Chatbot Sebagai Helpdesk Obyek Wisata Dengan Permodelan Natural Language Processing (Studi Kasus: Kabupaten Cilacap)," *Smart Comp Jurnalnya Orang Pint. Komput.*, vol. 10, no. 1, pp. 7–14, 2021, doi: 10.30591/smartcomp.v10i1.2135.

[4] N. I. F. Angga Ardiansyah, Suleman, Eva Argarini Pratama, "Analisis Sentimen Pengguna Terhadap Aplikasi ChatGPT Di Google Play Store: Penerapan Algoritma Support Vector Machine," *J. Tek. Inform. dan Sist. Inf.*, vol. 11, no. 2, pp. 247–254, 2024, [Online]. Available: https://jurnal.mdp.ac.id/index.php/jatisi/article/view/7945

[5] Khairul Marlin, Ellen Tantrisna, Budi Mardikawati, Retno Anggraini, and Erni Susilawati, "Manfaat dan Tantangan Penggunaan Artificial Intelligences (AI) Chat GPT Terhadap Proses Pendidikan Etika dan Kompetensi Mahasiswa Di Perguruan Tinggi," *Innov. J. Soc. Sci. Res.*, vol. 3, no. 6, pp. 5192–5201, 2023.

[6] A. Samiaji, B. Hananto, and S. Kom, "Analisis Sentimen Review Aplikasi Berita Online Pada Google Play Menggunakan Metode Naïve Bayes Studi Kasus: Tribunnews.Com," *Semin. Nas. Mhs. Ilmu Komput. dan Apl.*, no. 2, pp. 733–743, 2022.

[7] S. K. Lubis, M. H. Dar, and F. A. Nasution, "Analisis Sentimen Ulasan Pengguna Aplikasi pada Google Play Store Menggunakan Algoritma Support Vector Machine," *Informatika*, vol. 11, no. 2, pp. 120–128, 2024, doi: 10.36987/informatika.v11i2.5860.

[8] B. A. Maulana, M. J. Fahmi, A. M. Imran, and N. Hidayati, "Analisis Sentimen Terhadap Aplikasi Pluang Menggunakan Algoritma Naive Bayes dan Support Vector Machine (SVM)," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 2, pp. 375–384, 2024, doi: 10.57152/malcom.v4i2.1206.

[9] B. Purbayanto and T. N. Suharsono, "Analisis Sentimen Pengguna X terhadap Chatgpt dengan Algoritme Naive Bayes," *J. Telemat.*, vol. 18, no. 2, pp. 63–71, 2024, doi: 10.61769/telematika.v18i2.614.

[10] Syafrizal, M. Afdal, and R. Novita, "Sentiment Analysis of PLN Mobile Application Review Using Naïve Bayes Classifier and K-Nearest Neighbor Algorithm," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. January, pp. 10–19, 2024.

[11] S. M. Salsabila, A. Alim Murtopo, and N. Fadhilah, "Analisis Sentimen Pelanggan Tokopedia Menggunakan Metode Naïve Bayes Classifier," *J. Minfo Polgan*, vol. 11, no. 2, pp. 30–35, 2022, doi: 10.33395/jmp.v11i2.11640.

[12] Y. Akbar and T. Sugiharto, "Analisis Sentimen Pengguna Twitter di Indonesia Terhadap ChatGPT Menggunakan Algoritma C4.5 dan Naive Bayes," *J. Sains dan Teknol.*, vol. 5, no. 1, pp. 115–122, 2023, [Online]. Available: https://doi.org/10.55338/saintek.v4i3.1368

[13] K. D. Indarwati and H. Februariyanti, "Analisis Sentimen Terhadap Kualitas Pelayanan Aplikasi Gojek Menggunakan Metode Naive Bayes Classifier," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 10, no. 1, 2023, doi: 10.35957/jatisi.v10i1.2643.

[14] A. Muzaki, V. Febriana, and W. N. Cholifah, "Analisis Sentimen Pada Ulasan Produk di E-Commerce dengan Metode Naive Bayes," *J. Ris. dan Apl. Mhs. Inform.*, vol. 5, no. 4, pp. 758–765, 2024, doi: 10.30998/jrami.v5i4.9647.

[15] A. Nurian, M. S. Ma'arif, I. N. Amalia, and C. Rozikin, "Analisis Sentimen Pengguna Aplikasi Shopee Pada Situs Google Play Menggunakan Naive Bayes Classifier," *J. Inform. dan Tek. Elektro Terap.*, vol. 12, no. 1, 2024, doi: 10.23960/jitet.v12i1.3631.


.