Contents lists available at https://csi.unisi.my.id

Computer Science Insights

Journal Page is available to https://csi.unisi.my.id

Research article

Sentiment Analysis of Duolingo App User Reviews on Google Play Store Using Naive Bayes Algorithm

*M.Saleh [1] , Juwardi Wafdan [2], Muhammad Suratman [3]

[123] Sistem Informasi, Fakultas Teknik dan Ilmu Komputer, Universitas Islam Indragiri Kota Tembilahan, Indragiri Hilir, Riau, Indonesia
email: [1,*] msaleh0205@gmail.com, [2] juwardiwafdan@gmail.com, [3] suratmsurat@gmail.com,
* Correspondence

ARTICLE INFO

Article history:
Received November 25, 2025
Revised December 23, 2025
Accepted June 1, 2026
Available Online June 2, 2026

Keywords:
Analisis sentimen, Naive Bayes, Duolingo, ulasan pengguna, klasifikasi

ABSTRACT

The advancement of digital technology has driven the widespread use of mobile-based language learning applications, such as Duolingo, which has become one of the most popular educational platforms globally. As the user base grows, reviews on platforms like Google Play Store serve as a valuable data source for evaluating service quality. This study aims to analyze user review sentiments on Duolingo using the Multinomial Naive Bayes algorithm and assess its performance in classifying positive and negative sentiments. A quantitative approach involving text mining and classification was employed. A total of 1,000 Indonesian-language reviews were collected via web scraping. The analysis involved preprocessing steps including tokenization, stopword removal, stemming, and TF-IDF vectorization, followed by an 80:20 train-test data split. Results showed an accuracy of 77.78%, with a high f1-score for positive sentiment (0.87) but extremely low for negative sentiment (0.05). This imbalance stems from skewed class distribution and the model's limited ability to capture linguistic context. The study highlights the limitations of classical algorithms in Indonesian-language sentiment analysis and recommends advanced approaches such as data balancing, manual labeling, and the use of context-aware models like BERT to produce more accurate and fair opinion analytics.

1. Introduction

In an increasingly digitalized world, the demand for flexible and affordable access to education continues to grow. One prominent solution to address this need is mobile-based learning applications. Among the many applications available, Duolingo has emerged as a pioneer in providing accessible and enjoyable online language learning. Through its gamification-based approach, the application has transformed the way individuals independently learn foreign languages using smart devices.

According to a report by Statista, Duolingo is the most downloaded language learning application globally, with more than 15 million monthly downloads on the Google Play Store in 2024 [1]. This popularity positions Duolingo as an important indicator for evaluating user dynamics and perceptions of digital learning applications. High usage intensity also generates a substantial volume of feedback, primarily in the form of reviews on the Google Play Store. These reviews are not merely comments but represent user experiences that contain elements of emotion, satisfaction, and complaints. Therefore, systematic analysis of user reviews is essential as a means of service evaluation. As stated by Pang and Lee, "Opinion mining, or sentiment analysis, seeks to identify the attitude of a speaker or writer with respect to some topic or the overall contextual polarity of a document" [2]. This indicates that sentiment analysis is concerned not only with language but also with emotions and perspectives.

However, the large volume of user reviews expressed in natural and unstructured language poses challenges in efficiently extracting meaningful information. In this context, text mining and machine learning techniques offer effective solutions—one of which is the Naive Bayes algorithm, which is widely recognized for its effectiveness in text classification due to its simplicity and ability to handle high-dimensional data [3]. Numerous studies support the effectiveness of Naive Bayes in sentiment analysis of educational applications. A study by Sandy, Ghufron, and Muhtadi found that text classification of Duolingo reviews can reveal consistent sentiment patterns that are useful for application development [4]. Similarly, research by Firdaus et al. demonstrated that user emotions expressed in Indonesian-language reviews have a significant influence on user satisfaction [5].

Most previous studies have focused on English-language reviews, thereby overlooking local linguistic and cultural contexts. Consequently, further exploration is required—not only to classify sentiment but also to identify linguistic features that play a significant role in shaping user opinions. This study also aims to add an interpretative layer that is often neglected in conventional approaches. As a continuously evolving platform in terms of both technology and pedagogy, Duolingo benefits from user review analysis using machine learning, which enables developers to make data-driven decisions. Liu emphasized that sentiment analysis has become a key technique in

the era of social media and big data for understanding user opinions at scale [6]. For researchers and academics, review data can reveal user barriers, needs, and preferences that are not captured by traditional survey methods.

The integration of sentiment analysis into real-time feedback systems can foster an adaptive learning ecosystem aligned with the vision of 21st-century education: personalization, adaptability, and a user-centered learning experience. Unfortunately, there is still limited literature that examines Duolingo reviews using probabilistic approaches such as Naive Bayes, particularly in the context of the Indonesian language. Based on this background, this article aims to analyze the sentiment of Duolingo application user reviews on the Google Play Store using the Naive Bayes algorithm. This study also explores linguistic features that contribute to sentiment formation and presents interpretations of classification results as recommendations for application development. The findings are expected to contribute theoretically to computational linguistics and practically to developers in improving the quality of educational application services.

2. Research Methodology

2.1 Research Type and Approach

This study employs a descriptive qualitative research design, as it aims to describe the tendencies of user sentiment toward the Duolingo application. Nevertheless, the data analysis is conducted using a computational and quantitative statistical approach through the implementation of the Naive Bayes algorithm for text classification. This approach falls within the domain of computational social science, an interdisciplinary field that integrates computer science methods such as data mining and machine learning with social science theories based on the analysis of digital data, particularly textual data [7]. Computational social science enables researchers to explore public opinion on a large scale using digital data sources, such as application reviews or social media content, while still considering the social, cultural, and linguistic contexts embedded within the data. This approach is considered highly relevant for understanding the dynamics of users of digital educational applications, as it combines the technical accuracy of statistical methods with the contextual insights derived from social sciences [8].

2.2 Sumber Data dan Teknik Pengumpulan Data

The primary data source in this study consists of secondary data in the form of user reviews of the Duolingo application obtained from the Google Play Store platform. These data take the form of public comments containing users' opinions, criticisms, praises, and suggestions regarding the application. To collect the data, the researchers employed web scraping techniques, which involve the automated extraction of data from web pages using the Python programming language and libraries such as BeautifulSoup and Selenium [9]. The inclusion criteria applied in the data collection process were as follows: (1) the reviews were written in the Indonesian language, (2) each review contained a minimum of five words to reduce noise from short and uninformative comments, and (3) the reviews were published within the last two years to ensure data relevance and timeliness. In total, more than 1,000 review entries were successfully collected. During the data classification stage using the Naive Bayes algorithm, user reviews were categorized into three sentiment classes: Positive, Negative, and Neutral. The overall stages of the research process are illustrated in Figure 1.

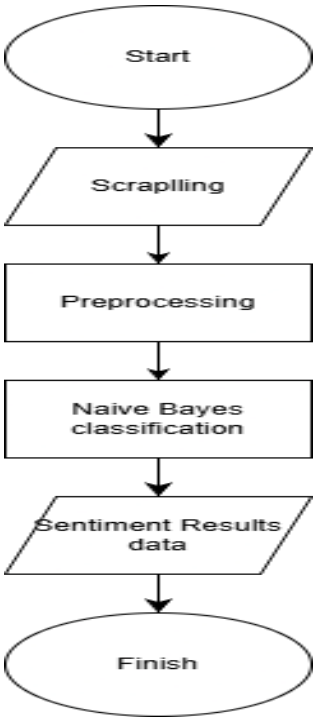


Figure 1. Research Stages

The figure above illustrates the various stages involved in the research process. The initial stage begins with Data Scraping, which involves collecting user review data from the Google Play Store. After the data have been successfully collected, the next step is Data Labeling, in which the reviews are labeled using Python in the Google Colaboratory environment. The labeled data then undergo Preprocessing to clean and organize the dataset into a more structured format. Subsequently, Data Splitting is performed by allocating 20% of the total data for testing purposes. The final stage is Classification using the Naive Bayes algorithm, which aims to group the data into specific sentiment classes based on their characteristics.

a) Data Scraping, Data collection from the app store was conducted through a scraping process using tools such as Instant Data Scraper. This process involved collecting various types of information, including usernames, user comments, ratings, and other relevant attributes, which were then stored in .csv format..

b) Data Labelling, The data used for labeling consisted of user reviews of the Duolingo application obtained from the Google Play Store. The labeling process was carried out in Google Colaboratory using Python, based on the rating scores provided by users. These scores were classified into three sentiment categories: Positive, Negative, and Neutral.

c) Data Preprocessing, This stage aimed to clean and prepare the data to ensure it is more organized and suitable for further analysis. Preprocessing helps make the data more representative and appropriate for the modeling process.

d) Data Splitting, At this stage, the dataset was divided into two subsets: training data and testing data, with 20% of the total data allocated for testing. This division is essential for evaluating the performance of the classification model.

e) Classification Using the Naive Bayes Algorithm The final stage involved classification using the Naive Bayes algorithm, which was applied to categorize the data into specific sentiment classes. This method is effective for solving classification problems based on previously processed textual data.

2.3 Data Preprocessing Techniques (Preprocessing)

The data obtained through the scraping process are generally not ready for direct analysis. Therefore, a series of text data preprocessing steps were performed to improve the accuracy of the classification algorithm. As illustrated in Figure 2, the preprocessing stages include:

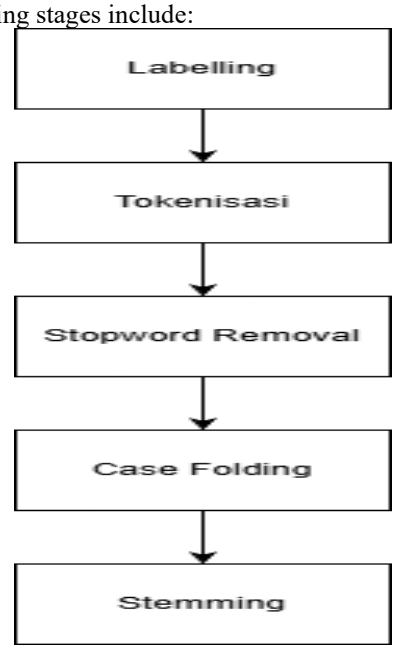


Figure 2. Preprocessing Stages

a) Labeling Labeling is the process of assigning sentiment classes to the data, which are divided into three categories: Positive, Negative, and Neutral.

b) Tokenisasi: Tokenization involves breaking sentences into smaller units, namely individual words, to facilitate further text analysis.

c) Stopword Removal: Stopword removal refers to the elimination of common words that do not carry significant informational value (such as “and,” “that,” and “from”).

d) Case Folding: Case folding converts all text into lowercase characters to ensure data uniformity and reduce variation caused by letter case differences.

e) Stemming: Stemming transforms words into their root or base forms using the Sastrawi library, which is specifically designed for Indonesian language processing.

These preprocessing steps are crucial because user review texts are often informal and may contain noise that can negatively affect classification performance. Overall, this research methodology is designed to achieve high validity and reliability in classifying user opinions toward the application. Moreover, the methodology can be replicated for other educational applications, both at local and global scales. Studies of this nature make a

significant contribution to the development of recommendation systems, user evaluation mechanisms, and the enhancement of real-time, data-driven learning experiences.

3 Results and Discussion

3.1 Results

3.1.1 Praprocessing

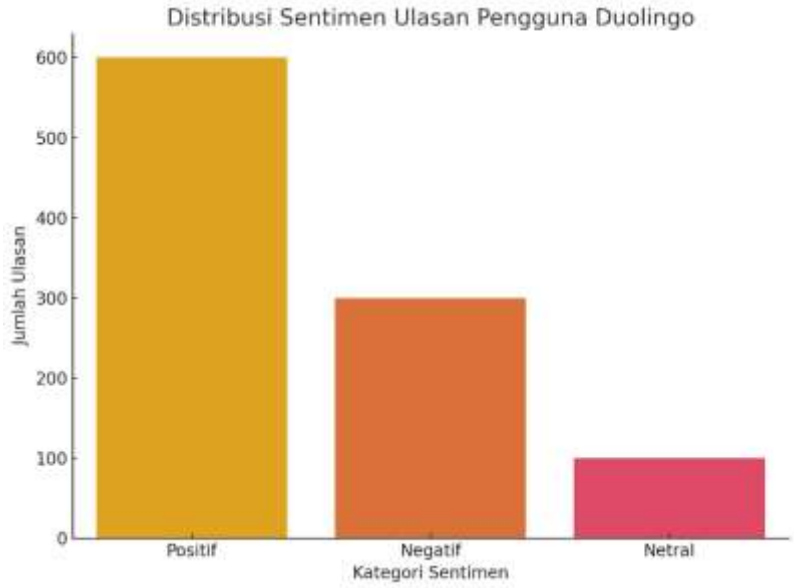
3.1.1.1 Labellling

During the data labeling stage, user reviews were classified into three sentiment categories: Positive, Negative, and Neutral. This classification was based on the ratings provided by users on the Google Play Store. The labeling rules were defined as follows: ratings of 1 or 2 were categorized as Negative sentiment, a rating of 3 was classified as Neutral sentiment, and ratings of 4 or 5 were classified as Positive sentiment.

Table 1. Sentiment Analysis Results

No	Username	Ulasan	Rating	Sentimen
0	Balqis bijawangsa	apk ini sangat seru, gampang dan bisa untuk di…	5	**Positif**
1	Adelia Nashita	sistem pembelajarannya sangat menarik dan muda...	3	**Netral**
2	Muhammad Maulana Alfaris	Bagi pengguna biasa seperti saya, merasa kecew...	2	**Negatif**
999	Siti Zikra Al Zikiro	bagus banget, dibanding saya buang duit buat b...	5	**Positif**

Table 1 presents the results of the sentiment analysis obtained from the data labeling stage. The table illustrates the classification of user reviews from the Google Play Store into three sentiment categories—Positive, Negative, and Neutral—based on the ratings provided in the reviews.



Gambar 3. Hasil Proses Labelling Data

Figure 3 presents the results of the data labeling process, which was conducted using Python and visualized through a bar plot in the Google Colaboratory environment. The sentiment analysis results show that the dataset consists of 605 Positive reviews, 306 Negative reviews, and 89 Neutral reviews.

3.1.1.2 Tokenisasi

Tokenization is a text processing procedure that involves dividing text into smaller units called tokens. These tokens may consist of words, phrases, sentences, or other textual components that carry meaning or significance within a particular context. Through word tokenization, text is segmented into smaller units (words) that can be utilized for further analysis, such as text classification or word grouping. The primary purpose of tokenization is to prepare textual data so that it can be processed more efficiently and accurately in subsequent analytical stages [10].

	content	score	Label	text_clean	text_StopWord	text_tokens
0	apk ini sangat seru, gampang dan bisa untuk di...	5	Positif	apk ini sangat seru gampang dan bisa untuk dij...	apk seru gampang dijadikan pembelajaran menghi...	[apk, seru, gampang, dijadikan, pembelajaran, ...
1	aplikasi ny berguna aku jadi bisa bhs Inggris ...	5	Positif	aplikasi ny berguna aku jadi bisa bhs inggris ...	aplikasi ny berguna bhs inggris kesini ko maks...	[aplikasi, ny, berguna, bhs, inggris, kesini, ...
2	bagus, bikin semangat, animasi nya juga ga bik...	5	Positif	bagus bikin semangat animasi nya juga ga bikin...	bagus bikin semangat animasi nya ga bikin bose...	[bagus, bikin, semangat, animasi, nya, ga, bik...
3	aplikasi nya sangat bagus,bisa untuk melatih k...	5	Positif	aplikasi nya sangat bagusbisa untuk melatih ke...	aplikasi nya bagusbisa melatih kemampuan bahas...	[aplikasi, nya, bagusbisa, melatih, kemampuan, ...
4	sistem pembelajarannya sangat menarik dan muda...	1	Negatif	sistem pembelajarannya sangat menarik dan muda...	sistem pembelajarannya menarik mudah dipelajar...	[sistem, pembelajarannya, menarik, mudah, dipe...

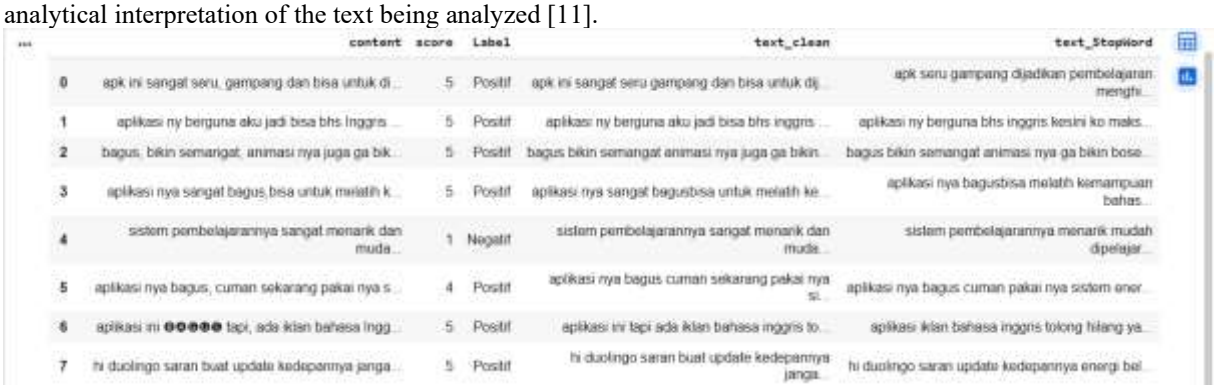
WARNING: Runtime no longer has a reference to this dataframe, please re-run this cell and try again.

Figure 4. Sentence Segmentation Process

This process is essential as an initial step in forming a data structure that can be processed linguistically and statistically in subsequent stages.

3.1.1.3 Stopword Removal

Stopword removal is a text processing procedure that involves eliminating common words that do not carry significant meaning in text analysis. Such words are referred to as *stopwords*. The primary purpose of stopword removal is to clean the text by removing frequently occurring words that contribute little to the understanding or analytical interpretation of the text being analyzed [11].



	content	score	Label	text_clean	text_StopWord
0	apk ini sangat seru, gampang dan bisa untuk di...	5	Positif	apk ini sangat seru gampang dan bisa untuk dij...	apk seru gampang dijadikan pembelajaran menghi...
1	aplikasi ny berguna aku jadi bisa bhs Inggris ...	5	Positif	aplikasi ny berguna aku jadi bisa bhs inggris ...	aplikasi ny berguna bhs inggris kesini ko maks...
2	bagus, bikin semangat, animasi nya juga ga bik...	5	Positif	bagus bikin semangat animasi nya juga ga bikin...	bagus bikin semangat animasi nya ga bikin bose...
3	aplikasi nya sangat bagus,bisa untuk melatih k...	5	Positif	aplikasi nya sangat bagusbisa untuk melatih ke...	aplikasi nya bagusbisa melatih kemampuan bahas...
4	sistem pembelajarannya sangat menarik dan muda...	1	Negatif	sistem pembelajarannya sangat menarik dan muda...	sistem pembelajarannya menarik mudah dipelajar...
5	aplikasi nya bagus, cuman sekarang pakai nya s...	4	Positif	aplikasi nya bagus cuman sekarang pakai nya si...	aplikasi nya bagus cuman pakai nya sistem eror...
6	aplikasi ini ❁❁❁❁❁ tapi, ada iklan bahasa Ingg...	5	Positif	aplikasi ini tapi ada iklan bahasa inggris to...	aplikasi iklan bahasa inggris tolong hilang ya...
7	hi duolingo saran buat update kedepannya janga...	5	Positif	hi duolingo saran buat update kedepannya janga...	hi duolingo saran update kedepannya energi bel...

Figure 5. Stopword Removal Process

The result of the stopword removal process is presented in the form of a table that shows the elimination of non-essential words and the retention of meaningful terms. The primary objective of stopword removal is to clean the text by removing frequently occurring words that do not contribute significantly to the understanding or analysis of the text.

3.1.1.4 Case Folding

Case folding is a text processing procedure that involves converting all alphabetic characters into lowercase or uppercase forms to facilitate the classification process. This procedure is commonly applied as part of text cleaning and normalization during data preprocessing, particularly in applications such as Natural Language Processing (NLP) and text analysis. Case folding helps reduce variations in words caused by differences in letter casing, thereby simplifying text matching and comparison. In this study, all uppercase characters (A–Z) in the dataset were converted into lowercase characters (a–z) [12].

	content	score	Label	text_clean
0	apk ini sangat seru, gampang dan bisa untuk di...	5	Positif	apk ini sangat seru gampang dan bisa untuk dij...
1	aplikasi ny berguna aku jadi bisa bhs Inggris ...	5	Positif	aplikasi ny berguna aku jadi bisa bhs inggris ...
2	bagus, bikin semangat, animasi nya juga ga bik...	5	Positif	bagus bikin semangat animasi nya juga ga bikin...
3	aplikasi nya sangat bagus,bisa untuk melatih k...	5	Positif	aplikasi nya sangat bagusbisa untuk melatih ke...
4	sistem pembelajarannya sangat menarik dan muda...	1	Negatif	sistem pembelajarannya sangat menarik dan muda...
5	aplikasi nya bagus, cuman sekarang pakai nya s...	4	Positif	aplikasi nya bagus cuman sekarang pakai nya si...
6	aplikasi ini ❁❁❁❁❁ tapi, ada iklan bahasa Ingg...	5	Positif	aplikasi ini tapi ada iklan bahasa inggris to...
7	hi duolingo saran buat update kedepannya janga...	5	Positif	hi duolingo saran buat update kedepannya janga...
8	menurut saya aplikasi ini sangat bgs saya jadi...	5	Positif	menurut saya aplikasi ini sangat bgs saya jadi...
9	bagus banget, dibanding saya buang duit buat b...	5	Positif	bagus banget dibanding saya buang duit buat be...

Figure 6. Case Folding Process

The result of the case folding process is presented in the form of a table showing text processing outcomes in which all uppercase characters are converted into lowercase characters. This transformation aims to facilitate the data classification process by ensuring uniformity in text representation.

3.1.1.5 Steamming

Stemming is a text processing technique that involves removing prefixes or suffixes from words to produce their root or base forms. The primary purpose of stemming is to reduce morphological variations in textual data, allowing words that share the same root to be treated as identical, even if they appear in different forms. Stemming is widely used in various text processing applications, such as text classification, string matching, and text indexing. By reducing word variations, stemming helps improve the consistency and accuracy of text analysis [13].

```
...  3535 : menyamakan : sama
     3536 : kolom : kolom
     3537 : uptade : uptade
     3538 : disertai : serta
     3539 : kalinya : kali
     3540 : upgradenya : upgradenya
     3541 : terniat : niat
     3542 : terseru : seru
     3543 : podcastnya : podcastnya
```

Figure 7. Stemming Process

The results of the stemming process are presented in the form of a table showing text processing outcomes in which different word variants are reduced to their base or root forms by removing affixes. This process aims to minimize the number of distinct indices in the dataset and restore words to their fundamental forms.

3.1.2 Splitting Data

Data splitting is the process of dividing a dataset into two or more distinct subsets. The primary purpose of data splitting is to use one subset for training the model (training set) and another subset for evaluating its performance (testing set). This separation is essential to prevent overfitting, a condition in which the model excessively adapts to the training data but fails to generalize patterns to previously unseen data. By using a separate testing subset, the model's performance can be evaluated objectively, ensuring that it can be effectively applied to new data. In this study, the data splitting process allocated 20% of the total dataset for testing purposes [14].

3.1.3 Model Evaluation Discussioni

Based on the results obtained from testing using the Multinomial Naive Bayes algorithm, it was found that the classification performance on Duolingo user reviews from the Google Play Store still exhibits notable weaknesses, particularly in detecting Negative sentiment. Although the model achieved an accuracy of 77.78%, this result can be misleading, as the majority of the accuracy is derived from the model's ability to correctly classify Positive sentiment reviews rather than from balanced performance across sentiment classes.

The data collection process was conducted in real time using web scraping techniques with the google-play-scraper library executed in the Google Colaboratory environment. The collected data consisted of user comments, ratings, and review publication timestamps, which were subsequently analyzed through several preprocessing stages, including tokenization, stopword removal, case folding, and stemming. The labeling process was performed based on user ratings, with sentiment scores classified into three categories: Positive, Negative, and Neutral. After the data were cleaned and labeled, a data splitting procedure was applied, allocating 20% of the dataset for testing to evaluate the trained model

Table 2. Results of the Naive Bayes Algorithm Analysis

Sentiment	Precision	Recall	F1-score	Support
Negative	0.50	0.03	0.05	40
Positive	0.78	0.99	0.87	140

The classification results indicate that the model is highly biased toward the majority class, namely Positive sentiment, with a recall value of 0.99. In contrast, the model's ability to recognize Negative sentiment is extremely weak, with a recall of only 0.03 and an F1-score of 0.05. These results are also visualized through a confusion matrix and a sentiment distribution bar plot, which clearly illustrate the imbalance in prediction outcomes.

Therefore, although the model demonstrates a relatively high numerical accuracy, it is not optimal for use in the context of detecting user complaints or negative reviews. This finding highlights the importance of implementing data balancing strategies, selecting alternative models, and strengthening multi-metric evaluation approaches to obtain more representative and reliable results.

3.2 Discussion

The results of this study show that the application of the Multinomial Naive Bayes algorithm in classifying sentiment from Duolingo user reviews on the Google Play Store achieved an overall accuracy of 77.78%. Although this figure appears high at first glance, it becomes less meaningful when examined more deeply using other evaluation metrics such as precision, recall, and F1-score. The model demonstrates excellent performance in identifying positive sentiment reviews, with a recall of 0.99 and an F1-score of 0.87. However, performance on negative sentiment is extremely poor, with recall reaching only 0.03 and an F1-score of 0.05. This indicates that

the model almost entirely fails to identify negative user opinions, which should be a critical focus in application development. This imbalance is closely related to the fundamental characteristics of the Naive Bayes algorithm, which operates on probabilistic assumptions and assumes independence among features. The performance of Naive Bayes is highly influenced by class distribution in the training data, particularly in text data with complex semantic structures. In this study, the number of positive reviews significantly exceeds the number of negative reviews, causing the model to be more accurate in learning patterns from the majority class. This leads to a severe classification bias and results in very weak generalization performance for negative sentiment. In sentiment classification, high numerical accuracy can often be misleading if it is not supported by balanced performance across classes.

From a practical perspective, these findings have important implications for the use of sentiment classification models in automated feedback systems. If a model tends to ignore negative reviews, application developers may lose critical user input containing complaints or suggestions for improvement. In the context of digital product development such as Duolingo, sensitivity to negative feedback is crucial for enhancing feature quality, user experience, and customer loyalty. An overly "optimistic" model that focuses only on positive opinions can create the illusion that all users are satisfied, while significant issues remain undetected by the system.From a technical standpoint, the model's success in classifying positive reviews cannot be separated from the careful preprocessing steps applied. Processes such as case folding, stopword removal, tokenization, stemming using the Sastrawi library, and word representation using TF-IDF help standardize the text and highlight meaningful terms. However, TF-IDF as a feature representation method also has limitations, as it cannot capture contextual meaning in depth. Many reviews contain sarcasm, ambiguity, or mixed sentiments that are difficult to interpret solely based on word frequency. Consequently, the model's weakness in recognizing negative reviews can also be attributed to the limitations of the text representation technique used.

Another factor influencing the results of this study is the dataset size and labeling method. The dataset consisted of only 1,000 reviews, with an imbalanced distribution between positive and negative classes. In addition, labeling was performed automatically based on star ratings, which may not accurately represent the true sentiment expressed in the review text. Many reviews with high ratings may still contain criticism, and vice versa. This issue can introduce labeling errors that directly affect classification performance. The absence of data balancing techniques such as oversampling or undersampling further exacerbates the model's bias toward the majority class.Another limitation identified in this study is the lack of advanced evaluation techniques, such as ROC-AUC analysis or detailed error analysis, to better understand classification error patterns. The use of more complex algorithms, such as Random Forest, Support Vector Machine (SVM), or transformer-based models like BERT, was also not explored, even though previous studies have shown that such models are more capable of handling natural language variability, semantic context, and imbalanced data distributions.

Nevertheless, this study makes important contributions in two key aspects. Academically, the findings reinforce the limitations of the Naive Bayes algorithm when applied to imbalanced datasets and languages with complex morphological structures such as Indonesian. Practically, the results serve as a reminder that high accuracy does not necessarily indicate good model performance, and that sentiment evaluation systems must be sensitive to negative reviews to ensure that data-driven decision-making truly reflects real user needs and experiences. For future research, it is recommended to expand the dataset size, apply class balancing techniques such as SMOTE, and utilize advanced machine learning models capable of capturing sentence-level context and handling class imbalance more effectively. Manual labeling by human annotators is also recommended to improve data validity. In addition, evaluation should incorporate metrics that better represent minority class performance to ensure that the analysis accurately reflects overall and fair model performance.

4 Conclusion

This study aimed to analyze user sentiment toward the Duolingo application on the Google Play Store platform using the Multinomial Naive Bayes algorithm. The results indicate that the model is capable of providing an initial overview of users' general perceptions, achieving an accuracy of 77.78%. The model demonstrates very strong classification performance for positive reviews, with an F1-score of 0.87, but exhibits a substantial weakness in detecting negative reviews, where the F1-score reaches only 0.05. This imbalance indicates that the model is biased toward the majority class and is unable to capture the linguistic complexity of critical opinions or user complaints.

This uneven performance can be explained by two primary factors. First, the dataset exhibits an imbalanced distribution between positive and negative reviews. Second, the Naive Bayes algorithm has inherent limitations in understanding semantic structures and contextual meaning in review texts. As a classical probabilistic model, Naive Bayes relies on the assumption of feature independence, which does not always align with the characteristics of natural language, particularly Indonesian, which has a complex morphological structure and often conveys nuanced emotional expressions. Theoretically, these findings reinforce previous literature highlighting the limitations of traditional classification algorithms in sentiment analysis, especially when applied to imbalanced and non-English textual data. This study also emphasizes the importance of selecting appropriate feature representations. Although TF-IDF is effective in extracting word frequency information, it has been shown to be insufficient in capturing implicit meaning, contextual usage, sarcasm, or ambiguity in natural language.

The practical contribution of this study lies in emphasizing the need to develop opinion analytics systems that are not only accurate in identifying positive sentiment but also sensitive to constructive negative feedback. In the context of educational application development such as Duolingo, negative opinions often contain valuable critiques and suggestions for feature improvement, enhanced user experience, and increased customer loyalty. Therefore, sentiment analysis systems used as decision-support tools must possess balanced detection capabilities across the full spectrum of user opinions. The limitations of this study include the use of automatic rating-based labeling, which is prone to semantic ambiguity (e.g., high ratings accompanied by critical comments), a relatively small dataset for training more complex models, and the absence of comprehensive evaluation techniques such as AUC-ROC analysis, detailed confusion matrices, or in-depth error analysis.

For future research, it is recommended to expand the dataset to achieve a more balanced sentiment distribution and to apply manual labeling by human annotators to improve ground-truth validity. Furthermore, future studies should explore advanced classification algorithms such as Support Vector Machine (SVM), Random Forest, and transformer-based models like BERT, which have been proven to be more robust in handling natural language variation and understanding sentence-level context. The application of data balancing techniques such as SMOTE (Synthetic Minority Over-sampling Technique) is also strongly recommended to improve classification performance for minority classes. Through these approaches, sentiment analysis systems are expected to become more representative, fair, and capable of comprehensively reflecting user opinions, thereby contributing meaningfully to the development of inclusive and sustainable data-driven educational technologies.

References

[1] Statista, "Most downloaded language learning apps worldwide," *Statista.com*, 2024. [Online]. Available: https://www.statista.com

[2] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

[3] V. Kumar and T. M. Sebastian, "Sentiment analysis: A perspective on Naive Bayes classifier," in *Proc. Int. Conf. on Computer Applications*, 2012, pp. 1–4.

[4] T. A. Sandy, A. Ghufron, and A. Muhtadi, "Text classification of Duolingo reviews on Google Play: Insights for enhancing M-learning applications," *Int. J. Emerg. Technol. Learn. (iJET)*, vol. 20, no. 5, pp. 102–115, 2025.

[5] R. F. Firdaus, O. N. Pratiwi, and I. Y. Mukti, "Sentiment analysis and topic modeling for Duolingo application on Google Play Store," in *Proc. 2024 Int. Conf. on Data Science and Artificial Intelligence (ICDSAI)*, IEEE, 2024.

[6] B. Liu, *Sentiment Analysis and Opinion Mining*. San Rafael, CA: Morgan & Claypool Publishers, 2012.

[7] M. J. Salganik, Bit by Bit: Social Research in the Digital Age. Princeton, NJ: Princeton University Press, 2017.

[8] D. Lazer et al., "Computational social science," Science, vol. 323, no. 5915, pp. 721–723, 2009, doi: 10.1126/science.1167742.

[9] R. Mitchell, Web Scraping with Python: Collecting Data from the Modern Web, 2nd ed., Sebastopol, CA: O'Reilly Media, 2018.

[10] A. McCallum and K. Nigam, "A comparison of event models for Naive Bayes text classification," in *Proc. AAAI 98 Workshop on Learning for Text Categorization*, Madison, WI, USA, Jul. 1998, pp. 41–48.

[11] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sept. 2009, doi: 10.1109/TKDE.2008.239.

[12] A. Hidayatullah and A. P. Wibowo, "Studi komparatif algoritma klasifikasi Naive Bayes dan K-Nearest Neighbor dalam sentiment analysis ulasan produk Tokopedia," *J. RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 4, no. 2, pp. 285–292, Apr. 2020, doi: 10.29207/resti.v4i2.2031.

[13] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, MN, USA, Jun. 2019, pp. 4171–4186, doi: 10.48550/arXiv.1810.04805.

[14] Y. Asri, W. N. Suliyanti, D. Kuswardani, and M. Fajri, "Pelabelan Otomatis Lexicon Vader dan Klasifikasi Naive Bayes dalam menganalisis sentimen data ulasan PLN Mobile," Petir, vol. 15, no. 2, pp. 264–275, 2022, doi: 10.33322/petir.v15i2.1733.

[15] T. A. Sari, E. Sinduningrum, and F. Noor Hasan, "KLIK: Kajian Ilmiah Informatika dan Komputer Analisis Sentimen Ulasan Pelanggan Pada Aplikasi Fore Coffee Menggunakan Metode Naïve Bayes," Media Online), vol. 3, no. 6, pp. 773– 779, 2023, doi: 10.30865/klik.v3i6.884.

[16] R. Wahyudi and G. Kusumawardana, "Analisis Sentimen pada Aplikasi Grab di Google Play Store Menggunakan Support Vector Machine," J. Inform., vol. 8, no. 2, pp. 200–207, 2021, doi: 10.31294/ji.v8i2.9681.

[17] M. Diki Hendriyanto, A. A. Ridha, and U. Enri, "Analisis Sentimen Ulasan Aplikasi Mola Pada Google Play Store Menggunakan Algoritma Support Vector Machine Sentiment Analysis of Mola Application Reviews on

Google Play Store Using Support Vector Machine Algorithm," J. Inf. Technol. Comput. Sci., vol. 5, no. 1, pp. 1–7, 2022.

[18] A. Nurian, "Analisis Sentimen Ulasan Pengguna Aplikasi Google Play Menggunakan Naïve Bayes," J. Inform. dan Tek. Elektro Terap., vol. 11, no. 3s1, pp. 829–835, 2023, doi: 10.23960/jitet.v11i3s1.3348.

[19] T. A. Sandy, A. Ghufron, and A. Muhtadi, "Text classification of Duolingo reviews on Google Play: Insights for enhancing M-learning applications," *Int. J. Emerg. Technol. Learn. (iJET)*, vol. 20, no. 5, pp. 102–115, 2025. [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=18657923&AN=184462183

[20] R. F. Firdaus, O. N. Pratiwi, and I. Y. Mukti, "Sentiment analysis and topic modeling for Duolingo application on Google Play Store," in *Proc. 2024 Int. Conf. Data Sci. Artif. Intell. (ICDSAI)*, IEEE, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10913140