Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th February 2026. Vol. 5. No. 2; e-ISSN: 2808-4519

Sentiment Analysis of Honkai Star Rail Game Reviews on Google Play Using the Naive Bayes Algorithm

Bagus Doang[1]*, Robi Aziz Zuama [2], Nila Hardi[3]

[1,2,3]*Computer Science Program, Faculty of Engineering and Informatics, Bina Sarana Informatika University, Indonesia*
bagoezdoang.bd@gmail.com[1]*, robi.rbz@bsi.ac.id [2], nila.nad@bsi.ac.id[3]

Abstract

Honkai Star Rail is a turn-based JRPG developed by COGNOSPHERE PTE. LTD. that has attracted a large number of players with over 10 million downloads and 339,000 reviews on the Google Play Store. While most reviews are positive, some users have expressed their dissatisfaction. Sentiment analysis is crucial for extracting insights from reviews without having to read the entire text. The objective of this research is to conduct sentiment analysis on Honkai Star Rail users using the Naïve Bayes algorithm. This research employs the Naïve Bayes method due to its simplicity and ability to produce accurate predictions. Review data was collected through scraping from the Google Play Store and analyzed using the TF-IDF technique for word weighting. The model testing results showed the highest accuracy on test data using split validation with a 75:25 ratio, achieving 84.7%. The evaluation of the confusion matrix for positive sentiment resulted in a precision of 100%, recall of 12%, and an F1-Score of 22%, while for negative sentiment, it resulted in a precision of 84%, recall of 100%, and an F1-Score of 92%. This research contributes by providing feedback for game developers to improve quality and enhance player satisfaction.

Keywords: *Sentiment analysis, Honkai Star Rail, Machine Learning, Naïve Bayes*

1. Introduction

The rapid development of the mobile gaming industry has encouraged increased user interaction through digital distribution platforms, one of which is Google Play Store. This platform not only functions as a medium for application distribution, but also as a means for users to convey their opinions, experiences, and satisfaction levels in the form of reviews [1]. User reviews contain valuable information that can be used to evaluate application quality and understand user perceptions broadly. Honkai Star Rail is a turn-based role-playing game (JRPG) developed by COGNOSPHERE PTE. LTD [2]. his game has attracted global attention with over 10 million downloads and approximately 339,000 reviews on the Google Play Store. Although it has generally received high ratings, user reviews show a variety of sentiments, ranging from appreciation for the gameplay and story to complaints about technical performance, battle mechanics, and the overall gaming experience. The large number of reviews makes manual analysis inefficient and potentially subjective, requiring an automated approach to systematically extract sentiment information.This game has attracted global attention with over 10 million downloads and approximately 339,000 reviews on the Google Play Store. Although it has generally received high ratings, user reviews show a variety of sentiments, ranging from appreciation for the gameplay and story to complaints about technical performance, battle mechanics, and the overall gaming experience. The large number of reviews makes manual analysis inefficient and potentially subjective, requiring an automated approach to systematically extract sentiment information [3]. One of the algorithms widely used in text classification is Naïve Bayes, due to the simplicity of the model, its computational efficiency, and its relatively stable performance on high-dimensional text data, especially when combined with word weighting techniquessuch as TF-IDF. In the context of Honkai Star Rail, sentiment analysis research is still limited, especially on Indonesian-language reviews on the Google Play Store. In addition, the problem of imbalanced datasets often arises in real review data and can have a significant impact on recall and F1-score values, especially in minority classes. Therefore, empirical studies are needed that not only measure accuracy but also evaluate model performance more comprehensively using metrics such as precision, recall, F1-score, and AUC [4].This study aims to analyze the sentiment of Honkai Star Rail user reviews on Google Play Store using the Naïve Bayes algorithm with TF-IDF word weighting. The dataset used consists of Indonesian-language reviews obtained through web scraping, with model evaluation conducted through several testing schemes, specifically split validation and cross-validation, to obtain the most optimal and representative configuration. It is hoped that the results of this study can provide a deeper understanding of user perceptions and serve as input for game developers in improving quality and gaming experience.

2. Research Method

This section presents the methodology used in the research, from dataset collection to model evaluation. The overall research workflow is shown in Figure 1, which illustrates the sequential stages from preparation to final evaluation.

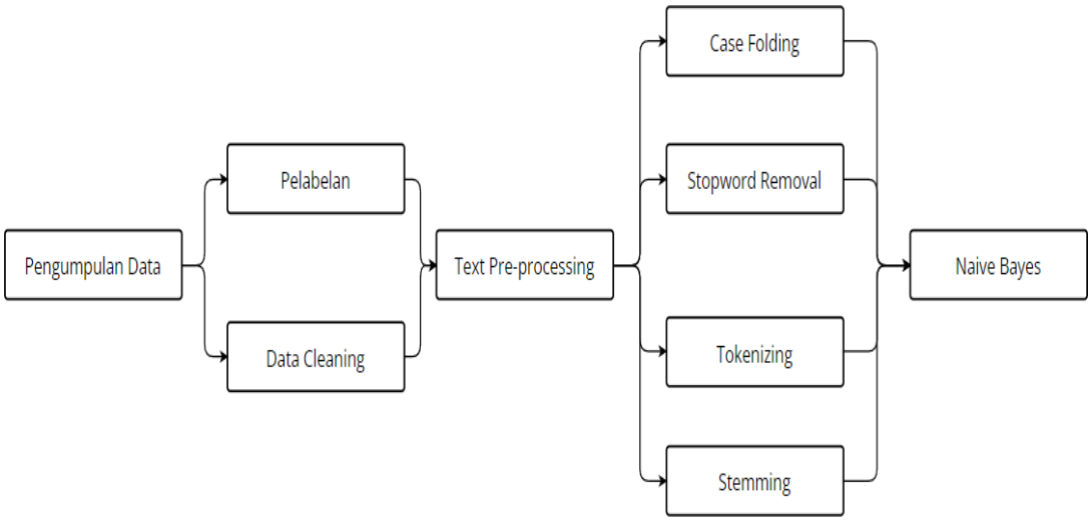

Fig. 1: Research flow diagram

2.1.Data Collection

The data used consists of reviews written in Indonesian. In this study, the researcher used Google Colab to collect and analyze the data. At the data collection stage, the researcher decided to use secondary data available on the Google Play Store website using the data scraping method. Data scraping is the process of automatically collecting data from a specific source [5]. The dataset consists of 200 Indonesian language reviews accompanied by rating scores. The data was selected based on relevance and used as a representation of user opinions about the game.

2.2.Data Labeling

Sentiment labeling is performed automatically based on review ratings. Reviews with high ratings are categorized as positive sentiment, while reviews with low ratings are categorized as negative sentiment. This process produces a labeled dataset that is then used as training data and test data in the classification process. Data labeling is used to divide these sentiment sentences into positive or negative sentiment before entering the data training stage [6]. At this stage, positive or negative sentiment labels are assigned based on the rating value.

2.3.Data Preprocessing

Data preprocessing is a crucial stage in the data mining process because not all data or attributes in the dataset are used in the analysis [7]. This stage aims to simplify the data, identify relationships between variables, perform normalization, remove outliers, and extract relevant features [8]. In this study, data preprocessing is divided into several stages, namely: data cleaning, case folding, stopword removal, tokenizing, and stemming.

2.4.Ekstraksi Fitur

Feature extraction is an important technique in the data reduction process that aims to identify, select, and generate the most relevant features that have a significant contribution to model formation, thereby improving data processing efficiency [9]. In this study, the feature extraction stage was carried out by extracting the pre-processed text represented in the form of numerical vectors using the TF-IDF method so that it could be processed by the classification algorithm.

2.5.Evaluasi Model

In this study, the feature extraction stage was carried out by extracting the pre-processed text represented in the form of numerical vectors using the TF-IDF method so that it could be processed by the classification algorithm learning and test data used to evaluate the performance of the proposed model [10]. Meanwhile, cross-validation is a model validation technique to assess the performance and accuracy of predictive models and reduce bias in the data [11].

3. Results and discussion

3.1.Sentiment Data Distribution

Before the classification process, a preliminary analysis of the sentiment class distribution in the dataset was conducted. This distribution is important for understanding the characteristics of the data and potential model bias. The data used in this study is public opinion data on the game Honkai Star Rail obtained from the Google Play Store platform. A total of 200 data points were used, sorted by relevance**.**

Table 1. Sentiment class distribution of reviews

Sentiment	Number of Data	Percentage
Positif	114	57%
Negatif	86	43%
Total	200	100%

Table 1 shows that the dataset has a relatively unbalanced class distribution, with positive sentiment dominating over negative sentiment. Although the difference is not extreme, this condition still has the potential to affect the performance of the classification model, especially on the recall and F1-score metrics. Datasets with unbalanced distributions often cause models to be more inclined to recognize the majority class and ignore the minority class.

3.2.Sentiment Classification Results

The experiment was conducted using the Multinomial Naïve Bayes algorithm with TF-IDF feature representation and several evaluation schemes, namely split validation (80:20, 75:25, and 70:30) and cross-validation. The test results show that model performance varies depending on the ratio of training data and test data used.

Table 2. Evaluation results of the Naïve Bayes model based on the testing scheme

Evaluation Scheme	Accuracy	AUC
Split 80:20	83.3%	0.75
Split 75:25	84.7%	0.78
Split 70:30	82.0%	0.73
Cross-validation (mean)	81.0%	0.74

Based on Table 2, the split validation scheme with a 75:25 ratio produced the best performance with an accuracy value of 84.7% and an AUC of 0.78, compared to other split configurations. Meanwhile, cross-validation produced an average accuracy value of around 81%, which shows that the model's performance is relatively stable but slightly lower than the best results in split validation. This difference in results indicates that in small datasets, such as 200 reviews, the composition of training and test data has a significant effect on model performance. A 75:25 split provides a better balance between the amount of training data and test data, allowing the model to learn word patterns more representatively without losing too much test data.

3.3. Confusion Matrix Analysis and Evaluation Metrics

Confusion matrix analysis on the best configuration (75:25) shows a noticeable difference in performance between positive and negative classes. The model achieved 100% recall on the negative class, which means that all negative reviews in the test data were classified correctly. However, the recall on the positive class was only 12%, even though the precision reached 100%.

Table 3. Confusion matrix of classification results (75:25)

Actual \ Predicted	Positif	Negatif
Positif	3	22
Negatif	0	25

Table 3 shows that all negative data was classified correctly, while most positive data was misclassified as negative. This indicates the model's tendency to prioritize negative class predictions. This condition shows that the model is very conservative in predicting positive sentiment. In other words, the model only classifies a review as positive when its confidence level is very high, so that many positive reviews are classified as negative. This pattern is common in unbalanced datasets, where one class is more dominant or has more consistent word characteristics.

Table 4. Evaluation metrics per class

Sentimen	Precision	Recall	F1-score
Positif	100%	12%	22%
Negatif	84%	100%	92%

Table 4 shows a clear trade-off between precision and recall. The very high recall value in the negative class indicates that the model is very effective in detecting negative reviews. However, the very low recall in the positive class indicates that many positive reviews are not recognized. This condition often occurs in datasets with class imbalance and greater vocabulary variation in positive sentiments. The high F1-score value in the negative class (92%) and low value in the positive class (22%) reinforce the indication that the model is more optimal in recognizing negative review patterns. This may be due to the use of more explicit and consistent vocabulary in negative reviews, such as complaints related to performance or bugs, compared to positive reviews, which tend to be more varied and subjective.

3.4.ROC Curve and AUC Value

The ROC curve in Figure 2 shows that the model has fairly good discriminatory ability, with an AUC value of 0.78. This value indicates that the model is able to distinguish between positive and negative sentiment at an acceptable level, although it is not yet optimal. This confirms that even though the accuracy is quite high, the model's performance is still influenced by data distribution and feature limitations.

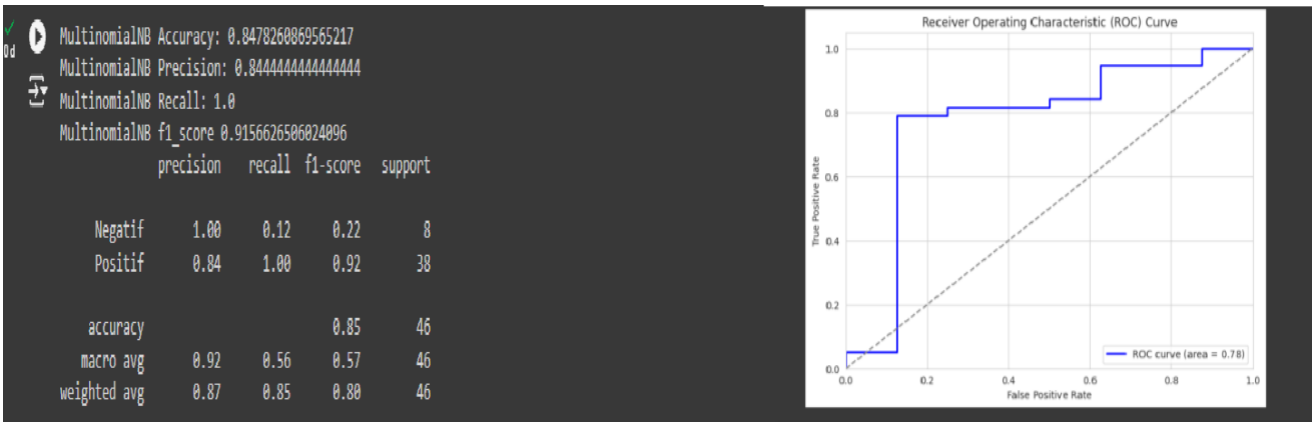

Figure 2.:ROC Curve

The AUC value of 0.78 in the best configuration shows that the model has fairly good discriminatory ability in distinguishing between positive and negative sentiments. This value indicates that the model's probability of giving a higher score to the correct class is still at an acceptable level for text-based sentiment analysis applications. Conversely, testing without data separation (using the entire dataset as training and test data) resulted in a very high accuracy value (0.98), but with an AUC value of 0.00. This result indicates overfitting, where the model fails to generalize to new data. This finding emphasizes the importance of using an appropriate evaluation scheme in sentiment analysis research, especially on limited datasets.

3.5.Implications and discussion

The results show that the Naïve Bayes algorithm with TF-IDF is effective in detecting negative reviews of Honkai Star Rail. This has significant practical implications for game developers, as negative reviews generally contain important information related to technical issues and user experience. However, the low recall on positive sentiment indicates the model's limitation in capturing the diversity of user satisfaction expressions. Methodologically, this finding emphasizes the importance of addressing data imbalance and exploring other more robust classification methods, such as Support Vector Machine or ensemble learning.

4. Conclusion

This study applied the Multinomial Naïve Bayes algorithm with TF-IDF weighting to analyze the sentiment of Honkai Star Rail game reviews on the Google Play Store and achieved a maximum accuracy of 84.7% with an AUC value of 0.78 on a 75:25 split validation scheme. The evaluation results show that the model is very effective in detecting negative sentiment (recall 100%), thus potentially helping developers accurately identify user complaints, although its performance on positive sentiment is still limited. The limitations of this study lie in the relatively small dataset size, class distribution imbalance, and the use of a single classification algorithm without comparing other methods. Therefore, further research is recommended to use a larger and more diverse dataset, apply imbalanced data handling techniques, and compare various classification algorithms or richer labeling approaches to improve generalization and depth of sentiment analysis.

References

[1] Z. F. Ramadhan and A. B. Mutiara, "Sentiment Analysis of Honkai : Star Rail Indonesian Language Reviews on Google Play Store Using Bidirectional Encoder Representations from Transformers Method," vol. 3, no. 3, pp. 1–6, 2023.
[2] PlayStation, "Honkai: Star Rail." [Online]. Available: https://www.playstation.com/en-id/games/honkai-star-rail/?
[3] I. S. K. Idris, Y. A. Mustofa, and I. A. Salihi, "Analisis Sentimen Terhadap Penggunaan Aplikasi Shopee Mengunakan Algoritma Support Vector Machine (SVM)," vol. 5, pp. 32–35, 2023.
[4] J. Wainer, "An empirical evaluation of imbalanced data strategies from a practitioner's point of view," 2024.
[5] E. Dwi, K. Wardani, F. F. Yo, W. N. Meylugita, U. Katolik, and M. Charitas, "IMPLEMENTASI ALGORITMA NAÏVE BAYES UNTUK ANALISIS ULASAN IMPLEMENTATION OF THE NAIVE BAYES ALGORITHM FOR USER REVIEW," vol. 4, no. 1, pp. 13–24, 2025.
[6] R. B. Afandi, T. F. Nurdiansyah, A. N. Ramadhani, and A. P. Sari, "IMPLEMENTASI SUPPORT VECTOR MACHINE UNTUK ANALISIS SENTIMEN APLIKASI ' MPStore - Super App UMKM ,'" pp. 565–570, 2024.
[7] A. Widianti and I. Pratama, "PENANGANAN MISSING VALUES DAN PREDIKSI DATA TIMBUNAN," vol. 9, no. 2, pp. 242–251, 2024.
[8] Suananto and G. Falah, "Penerapan algoritma c4.5 untuk membuat model prediksi pasien yang mengidap penyakit diabetes 1) 1,2)," vol. 7, no. 2, pp. 208–216, 2022.
[9] E. Mulyani, F. Pralienka, B. Muhamad, and K. A. Cahyanto, "Pengaruh N-Gram terhadap Klasifikasi Buku menggunakan Ekstraksi dan Seleksi Fitur pada Multinomial Naïve Bayes," vol. 5, pp. 264–272, 2021, doi: 10.30865/mib.v5i1.2672.
[10] S. Diantika, H. Nalatissifa, R. Supriyadi, N. Maulidah, and A. Fauzi, "IMPLEMENTASI MULTI-CLASS GRADIENT BOOSTING UNTUK MENGKLASIFIKASIKAN JENIS HEWAN PADA KEBUN BINATANG," vol. 17, no. 1, pp. 32–40, 2023.
[11] S. Anwar, I. K. Hasan, and D. Wunggul, "IMPLEMENTASI LEAST SQUARE SUPPORT VECTOR MACHINE DENGAN ALGORITMA ARTIFICIAL BEE COLONY DAN K-FOLD CROSS VALIDATION PADA PERAMALAN HARGA SAHAM PT.PERUSAHAAN GAS NEGARA," vol. 13, no. 03, pp. 59–66, 2025.