COMPARATIVE SENTIMENT ANALYSIS OF OSS INDONESIA REVIEWS FOR DIGITAL ECONOMY USING SVM AND NAIVE BAYES

Jihan Hasna Iftinan[1]*, Layla Mazidatus Sa'adah[2], Nadin Isna Monica[3], Rifda Nasywatul Affah[4], Safna Faradillah[5]
[1,2,3,4,5]Universitas Pembangunan Nasional "Veteran" Jawa Timur
*) email: jihanhiftinan.0110@gmail.com

Abstract

The rapid growth of Indonesia's digital economy has increased the importance of effective government digital services, particularly for micro, small, and medium enterprises (MSMEs). The Online Single Submission (OSS) Indonesia application was developed to simplify business licensing processes; however, user reviews indicate persistent usability and system performance issues. This study aims to analyze user sentiment toward the OSS Indonesia application and compare the performance of Support Vector Machine (SVM) and Naive Bayes algorithms in classifying sentiment from Google Play Store reviews. A total of 2,208 Indonesian-language reviews collected between 2021 and 2025 were manually labeled into positive and negative categories by three annotators, achieving a Fleiss' Kappa score of 0.95. Text preprocessing and TF-IDF feature extraction were applied, followed by model evaluation using three train test split scenarios (70:30, 80:20, and 90:10). The results show that SVM consistently outperformed Naive Bayes across all evaluation metrics, achieving the best performance with an 80:20 split (accuracy 0.94 and F1-score 0.95). While Naive Bayes demonstrated higher recall, SVM provided a more balanced and reliable classification. These findings indicate that SVM is more suitable for sentiment analysis of Indonesian government digital service reviews and highlight the importance of user feedback in improving public digital platforms.

Keywords: Sentiment Analysis, OSS Indonesia, Support Vector Machine, Naive Bayes, Digital Government Services

1. INTRODUCTION

Digital transformation has become a central pillar in strategies aimed at enhancing competitiveness and economic growth in Indonesia. Digitalization not only improves efficiency in business activities but also expands opportunities for micro, small, and medium enterprises (MSMEs) to access markets and formal economic services. MSMEs play a crucial role in Indonesia's economic structure, contributing significantly to gross domestic product (GDP) and absorbing a large proportion of the national workforce (Sultan et al., 2025). According to a study cited by the Ministry of Industry, MSMEs contributed approximately 60.34% of GDP and accounted for 97.22% of total employment, highlighting their substantial role in the national economy (Sultan et al., 2025).

To support MSMEs and improve the business climate, the Indonesian government developed the Online Single Submission (OSS) Indonesia system as an integrated digital platform for processing business licensing electronically. This application is designed to simplify the acquisition of Business Identification Numbers (Nomor Induk Berusaha or NIB) and other business permits in a faster and more

efficient manner, particularly for MSMEs that were previously constrained by complex administrative procedures (Google Play, 2025).

Despite the high adoption of the OSS Indonesia application, as indicated by more than one million downloads on the Google Play Store, user review data reveal significant challenges related to the quality of user experience. Based on updated data as of 31 December 2025, the application has accumulated approximately 4,000 reviews on the Google Play platform. However, only 2,008 reviews were successfully collected through the web scraping process, with an average rating of 2.7 out of 5, which is relatively low when compared to expectations associated with its adoption level and intended contribution to the digital economy (Google Play, 2025).

This condition of negative or mixed sentiment is important to examine, as user perceptions may reflect real barriers to the adoption of government digital services, particularly when negative experiences discourage compliance or continued use of formal services. Previous studies in the domain of sentiment analysis indicate that mapping public opinion toward policies or digital services can assist policymakers in identifying system limitations that may not be evident through quantitative usage data or download statistics alone (Fransiscus & Girsang, 2022).

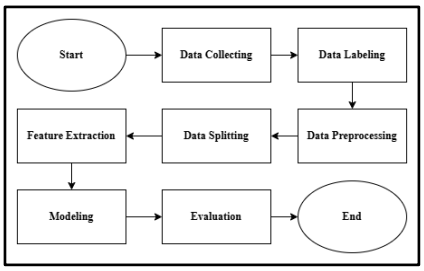
Furthermore, Indonesia's digital economy continues to grow rapidly. Independent reports indicate that Indonesia has one of the largest digital economies in Southeast Asia, with an increasing contribution to national GDP and a growing need for policy innovation to enhance its effectiveness. The digital economy is projected to reach a value of hundreds of billions of US dollars by the end of this decade, with the e-commerce sector projected to contribute around 72 percent of the total digital economy value in Indonesia (Saputra, 2025).

Within this context, sentiment analysis of OSS Indonesia user reviews becomes methodologically and substantively relevant, as it enables a systematic evaluation of public perceptions toward a critical government digital service. A comparative approach using Support Vector Machine (SVM) and Naive Bayes on a dataset collected from the application's initial release until the end of 2025 not only facilitates sentiment classification but also allows an assessment of the effectiveness of classical machine learning methods in capturing public opinion within the context of Indonesian-language digital governance services.

2. METHODOLOGY

Sentiment analysis techniques are applied in this study to compare the performance of Support Vector Machine and Naive Bayes algorithms in analyzing user reviews of the OSS Indonesia application. The research workflow is illustrated in the following flowchart:

Figure 1. Research methodology workflow



Source: Authors' own work (2026)

As illustrated in Figure 1, the research methodology consists of several stages, including data collection, data labeling, data preparation, data splitting, feature extraction, modeling, and evaluation.

2.1 Data Collecting

The data used in this study consist of user reviews of the OSS Indonesia application obtained from the Google Play Store during the period from October 9, 2022, to December 31, 2025. Web scraping techniques were used to automatically extract reviews from the Google Play Store platform (Sandy & Laple Satria Putra, 2025). The application reviews on the Google Play Store are considered to reflect users' perceptions and experiences regarding the application under research (Yuliani et al., 2025).

2.2 Data Labeling

The sentiment labeling process was manually performed by three annotators using two sentiment categories, namely positive and negative. Inter-annotator agreement was evaluated using the Fleiss' Kappa coefficient, which is appropriate for assessing agreement among multiple raters on nominal-scale data. The coefficient was calculated using the following equation (Moons & Vandervieren, 2023):

$$k = (Po - Pe) / (1 - Pe)$$

Where Po represents the proportion of observed agreement, and Pe represents the proportion of agreement expected by chance. The final label for each review was determined using the majority voting method, in which the sentiment category selected by at least two of the three annotators was assigned as the final sentiment (Lindén et al., 2023).

2.3 Data Processing

Before modeling was performed, the text data were subjected to preprocessing stages to improve data quality and model performance (Aufar et al., 2023). These stages included case folding to convert all text into lowercase, cleaning to remove numbers, punctuation, URLs, emojis, and special characters, tokenization to split the text into individual words, stopword removal to eliminate common words with little semantic meaning, and stemming to reduce words to their root forms (Ramadila et al., 2025). This process was intended to reduce noise and normalize the text data so that it could be analyzed accurately (Palomino & Aider, 2022).

2.4 Data Splitting

The prepared data were subsequently divided into training data and testing data using three split scenarios, namely 80:20, 90:10, and 70:30. The use of multiple data split scenarios was intended to evaluate the consistency and stability of model performance across variations in the proportion of training data (Prasetyo et al., 2024).

2.5 Feature Extraction

Feature extraction was performed using the Term Frequency–Inverse Document Frequency (TF-IDF) method. This method was used to transform textual data into numerical representations by considering the frequency of term occurrences in a document and across the entire corpus (Ramadila et al., 2025). The TF-IDF weighting scheme is defined by the following equation:

$$TF\ -IDF\ (t,d)\ =\ TF(t,d)\ \times\ log\frac{N}{DF(t)}$$

In this equation, $TF(t,d)$ represents the frequency of the word t in the document d, $DF(t)$ represents the number of documents containing the word t, and N represents the total number of documents. This method produces a sparse word representation matrix and is widely used in sentiment analysis due to its simplicity and effectiveness (Rahmatullah & Annisa, 2025).

2.6 Modelling

At the modeling stage, two machine learning algorithms were employed, namely Support Vector Machine and Naive Bayes. Both models were trained using the same feature representation and data split scenarios to ensure a fair performance comparison.

2.6.1 Support Vector Machine

The Support Vector Machine (SVM) aims to find an optimal hyperplane that maximally separates data points from different classes. The equation of a SVM can be expressed as (Restiani & Purwadi, 2024):

$$w_i x_i\ +\ b\ =\ 0$$

Where w_i is the weight vector, x_i is the input feature vector, and b is the bias parameter. SVM was chosen due to its ability to handle high-dimensional data and its robustness against overfitting in textual data (Ramadila et al., 2025).

2.6.1 Naive Bayes

Naive Bayes is a probabilistic classifier based on Bayes' theorem, in which the attributes are assumed to be conditionally independent. Samples are classified into the most probable category using the maximum probability principle. The equation of Naive Bayes classification is expressed as follows:

$$P(C_i|X)\ =\ Max\ \{P\ (C_1|X), P(C_2|X), \ldots P\ (C_n|X)\}$$

In this equation, $P\ (C_1|X)$ represents the posterior probability of class C_1 given the sample X. The sample X is assigned to the class with the highest posterior probability among all possible classes $C_1,\ C_2, \ldots, C_n$. This reflects the principle of maximum probability used in Naive Bayes classification (Chen et al., 2021).

2.7 Evaluation

The performance of the classification model was evaluated using a confusion matrix to determine the distribution of prediction errors for each class. The confusion matrix includes four main metrics (Syah et al., 2025):

a. Accuracy, which is the ratio of correctly predicted instances (both positive and negative) to the total number of instances in the dataset.
b. Precision, which is the ratio of true positives to the total predicted positives, including true positives and false positives.
c. Recall, which is the ratio of correctly predicted positive instances to the total actual positive instances.
d. F1-Score, which is the weighted average of precision and recall.

From this process, the algorithm and data split ratio that provide the best performance based on evaluation metrics such as accuracy, precision, recall, and F1-score will be identified. The results allow a comparative analysis of the effectiveness of both algorithms in handling the OSS Indonesia user review dataset.

3. FINDINGS AND DISCUSSION

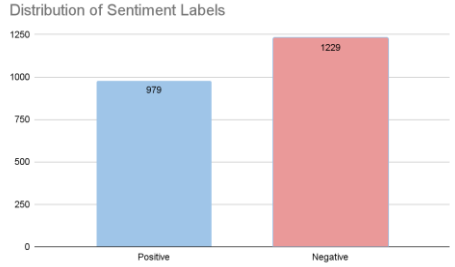
3.1 Data Collection

Based on the results of the data collection process, this study obtained reviews of the OSS application sourced from the Google Play Store. The data collection was conducted using the Python library google-play-scraper. A total of 2,208 reviews were collected between October 9, 2022, to December 31, 2025. All collected data were stored in CSV format to facilitate the sentiment analysis in the subsequent stage.

3.2 Data Labeling

After the data was obtained, the data labeling process was carried out manually by three annotators with two sentiment label categories, namely positive and negative. To measure the level of agreement between annotators, the Fleiss' Kappa method was used. The Fleiss' Kappa value obtained was 0.95, which falls into the almost perfect category. These results show that the level of consistency between annotators is very high, indicating that the labeled data are suitable for use in the modeling stage. The labeling results were then combined by determining the final label based on the majority voting of the three annotators. The data labeling results obtained approximately 979 positive reviews and 1,229 negative reviews. The distribution of the labeling results shown in figure 2.

Figure 2. Sentiment Labeling Distribution



Source: Authors' own work (2026)

3.3 Data Processing

The next stage is data processing, which includes case folding, data cleaning, tokenization, stopword removal, and stemming. The results before and after the processing stage can be seen in table 1.

Table 1. Data Processing

Content	Case Folding	Data Cleaning	Tokenization	Stopword Removal	Stemming
Sudah terdaftar gak bisa login	sudah terdaftar gak bisa login	sudah terdaftar gak bisa login	[sudah, terdaftar, gak, bisa, login]	[terdaftar, gak, login]	[daftar, gak, login]
Tidak bisa masuk ke perizinan, tolong perbaikan yaa.... 😁	tidak bisa masuk ke perizinan, tolong perbaikan yaa.... 😁	tidak bisa masuk ke perizinan tolong perbaikan yaa	[tidak, bisa, masuk, ke, perizinan, tolong, perbaikan, yaa]	[masuk, perizinan, tolong, perbaikan, yaa]	[masuk, izin, tolong, baik, yaa]

Source: Authors' own work (2026)

3.4 Modeling

two classification algorithms were employed to perform sentiment analysis on OSS application reviews: Multinomial Naive Bayes and Linear Support Vector Machine (SVM). Both models were trained using TF-IDF vectorized features with a maximum of 5,000 features and bigram representation (n-gram range of 1-2). The modeling process was conducted across three different train-test split scenarios (70:30, 80:20, and 90:10) to evaluate the impact of training data size on model performance. For each split scenario, both classifiers were trained on the preprocessed and vectorized training data, then tested on the corresponding test set.

Figure 3. Modeling SVM & NB

```
model_results = {}

for name, data in tfidf_data.items():
    # Naive Bayes
    nb_model = MultinomialNB()
    nb_model.fit(data["X_train_tfidf"], data["y_train"])
    nb_pred = nb_model.predict(data["X_test_tfidf"])

    # SVM
    svm_model = LinearSVC()
    svm_model.fit(data["X_train_tfidf"], data["y_train"])
    svm_pred = svm_model.predict(data["X_test_tfidf"])

    model_results[name] = {
        "y_test": data["y_test"],
        "nb_pred": nb_pred,
        "svm_pred": svm_pred
    }
```

Source: Authors' own work (2026)

In figure 3, multinomial Naive Bayes algorithm was selected due to its effectiveness in text classification tasks and computational efficiency, particularly when dealing with discrete features such as word counts. Linear SVM was chosen for its strong performance in high-dimensional spaces and ability to find optimal decision boundaries through maximum margin classification. Both models were trained without hyperparameter tuning to establish baseline performance across different data configurations. The predictions from each model were stored for subsequent evaluation and comparison.

3.5 Evaluation

The performance of both classifiers was evaluated using four standard metrics: accuracy, precision, recall, and F1-score. These metrics provide comprehensive insight into the models' classification capabilities from different perspectives.

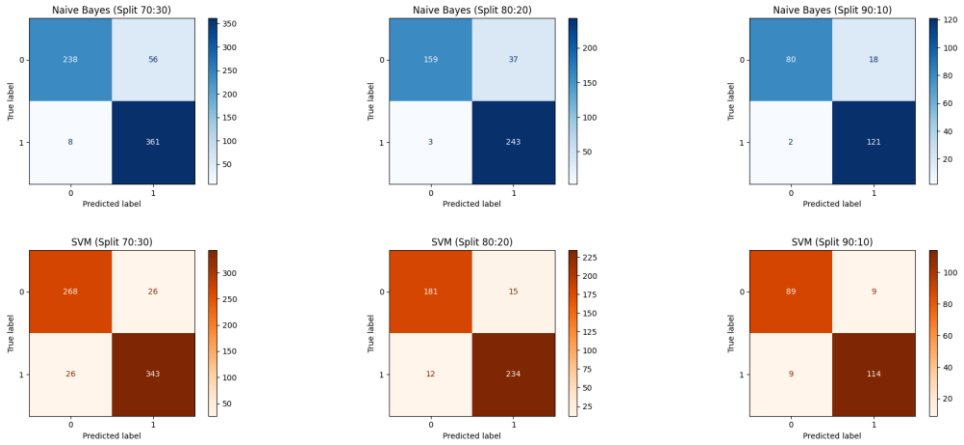
Table 2. Evaluation

Split	Naive Baiyes				SVM			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
70:30	0.903469	0.865707	0.97832	0.918575	0.921569	0.929539	0.9295	0.929539
80:20	0.909502	0.867857	0.98780	0.923954	0.938914	0.939759	0.9512	0.945455
90:10	0.909502	0.870504	0.98374	0.923664	0.918552	0.926829	0.9268	0.926829

Source: Authors' own work (2026)

Table 2 presents the comprehensive evaluation results for both classifiers across all three split scenarios using accuracy, precision, recall, and F1-score metrics. The results demonstrate that SVM consistently outperformed Naive Bayes across all metrics and configurations, with the 80:20 split yielding optimal performance for both algorithms. SVM achieved its peak performance with this split, recording an accuracy of 0.936914, precision of 0.939759, recall of 0.951220, and F1-score of 0.945455, while Naive Bayes achieved accuracy of 0.909502, precision of 0.867857, recall of 0.987805, and F1-score of 0.923954. Naive Bayes exhibited consistently higher recall values (ranging from 0.978320 to 0.987805) compared to precision (0.865707 to 0.870504), indicating a tendency to minimize false negatives at the cost of increased false positives, whereas SVM demonstrated more balanced precision-recall performance, reflecting superior classification equilibrium. Based on these findings, the SVM model with 80:20 split configuration was selected as the final model for deployment due to its highest F1-score and balanced performance across all evaluation metrics.

Figure 4. Confusion Matrix Display



Source: Authors' own work (2026)

To analyze classification errors and prediction patterns, confusion matrices were generated for Naive Bayes and SVM on all three data split scenarios, as shown in Figure 4. In the 70:30 split, Naive Bayes successfully classified 238 positive examples and 361 negative examples, but misclassified 56 false positives and 8 false negatives, while SVM produced 268 true positives, 343 true negatives, 26 false positives, and 26 false negatives, showing a more balanced error distribution. For the 80:20 split, Naive Bayes identified 159 true positives and 243 true negatives with 37 false positives and 3 false negatives, while SVM produced 181 true positives, 234 true negatives, only 15 false positives, and 12 false negatives, showing much better precision in identifying positive sentiment. Then, the 90:10 split shows that Naive Bayes achieved 80 true positives and 121 true negatives with 18 false positives and 2 false negatives, while SVM recorded 89 true positives, 114 true negatives, 9 false positives, and 9 false negatives. Based on the confusion matrix results, SVM is superior to Naive Bayes with minimal prediction errors and is capable of handling class imbalance and achieving better generalization, especially in the 80:20 configuration where SVM minimizes both types of errors most effectively.

4. CONCLUSION

This study demonstrates that sentiment analysis of OSS Indonesia user reviews provides valuable insights into public perceptions of government digital services. Based on the comparative evaluation, Support Vector Machine consistently outperformed Naive Bayes across all data split scenarios, particularly with the 80:20 train test ratio, which produced the most balanced and reliable results. Although Naive Bayes achieved higher recall, its lower precision indicates a tendency toward misclassification, whereas SVM showed superior stability and overall performance. These findings suggest that SVM is more effective for sentiment classification in Indonesian language application reviews. Furthermore, the predominance of negative

sentiment highlights the need for continuous improvement in system usability and performance to support MSMEs and strengthen Indonesia's digital economy. Future studies may explore advanced models and aspect-based sentiment analysis to obtain deeper insights into specific user concerns.

REFERENCES

Aufar, A. F., Rosid, M. A., Eviyanti, A., & Astutik, I. R. I. (2023). Optimizing Text Preprocessing for Accurate Sentiment Analysis on E-Wallet Reviews: Mengoptimalkan Preprocessing Teks untuk Analisis Sentimen yang Akurat pada Ulasan E-Wallet. *JICTE (Journal of Information and Computer Technology Education)*, *7*(2), 42–50. https://doi.org/10.21070/jicte.v7i2.1650

Chen, H., Hu, S., Hua, R., & Zhao, X. (2021). Improved naive Bayes classification algorithm for traffic risk management. *EURASIP Journal on Advances in Signal Processing*, *2021*(1), 30-. https://doi.org/10.1186/s13634-021-00742-6

Fransiscus, & Girsang, A. S. (2022, December 31). *Sentiment Analysis of COVID-19 Public Activity Restriction (PPKM) Impact using BERT Method*. arXiv.Org. https://arxiv.org/abs/2301.00096

Google Play Store. (2025). *OSS Indonesia application*. Google Play Store. https://play.google.com/store/apps/details?id=id.go.oss

Lindén, K., Jauhiainen, T., & Hardwick, S. (2023). FinnSentiment: A Finnish social media corpus for sentiment polarity annotation. *Language Resources and Evaluation*, *57*(2), 581–609. https://doi.org/10.1007/s10579-023-09644-5

Moons, F., & Vandervieren, E. (2023, March 22). *Measuring agreement among several raters classifying subjects into one or more (hierarchical) categories: A generalization of Fleiss' kappa*. arXiv.Org. https://arxiv.org/abs/2303.12502

Palomino, M. A., & Aider, F. (2022). Evaluating the effectiveness of text pre-processing in sentiment analysis. *Applied Sciences*, *12*(17). https://doi.org/10.3390/app12178765

Prasetyo, Y. A., Utami, E., & Yaqin, A. (2024). Pengaruh Komposisi Split Data Terhadap Performa Akurasi Analisis Sentimen Algoritma Naïve Bayes dan SVM. *Journal of Electrical Engineering and Computer (JEECOM)*, *6*(2), 382–390. https://doi.org/10.33650/jeecom.v6i2.9188

Rahmatullah, A., & Annisa, Q. (2025). Application of TF-IDF and Word2vec for feature extraction in sentiment analysis of free nutritious food policies. *Journal of Computer Electronic and Telecommunication*, *6*(2). https://doi.org/10.52435/complete.v6i2.741

Rahmatullah, A., & Annisa, Q. (2026). Application of TF-IDF and Word2vec for feature extraction in sentiment analysis of free nutritious food policies. *Journal of Computer Electronic and Telecommunication*, *6*(2). https://doi.org/10.52435/complete.v6i2.741

Ramadila, H., Syafitri, E. D., & Zamsuri, A. (2025). Sentiment detection of Shopee e-commerce application reviews using natural language processing and support vector machine. *JAISEN: Journal of Advanced Information Systems and Engineering*, *1*(1), 28–38.

Restiani, Y., & Purwadi, J. (2024). Support vector machine for classification: a mathematical and scientific approach in data analysis. *Jurnal Penelitian Pendidikan IPA*, *10*(11), 9896–9903. https://doi.org/10.29303/jppipa.v10i11.8122

Sandy, M., & Laple Satria Putra, R. (2025). View of analisis sentimen komentar pengguna aplikasi Segari di Google Playstrore menggunakan metode Support Vector Machine (SVM). *Jurnal Multimedia Dan Teknologi Informasi*, *07*(03), 642–652.

Saputra, B. (2025, July 31). Minister projects digital sector's eight-percent contribution to GDP. *ANTARA*. https://en.antaranews.com/news/370173/minister-projects-digital-sectors-eight-percent-contribution-to-gdp

Sultan, R., Junus, N., & Elfrikri, N. F. (2025). Legal protection of MSME trademarks as a pillar of local economic justice. *YUDHISTIRA : Jurnal Yurisprudensi, Hukum Dan Peradilan*, *3*(1), 55–65. https://doi.org/10.59966/yudhistira.v3i1.1776

Syah, F. P., Hasanuddin, T., & Kurniati, nia. (2025). Implementasi Naive Bayes Untuk Analisis Sentimen Pada data Twitter Tentang Isu Politik di Indonesia. *LINIER: Literatur Informatika Dan Komputer*, *2*(3), 302–316. https://doi.org/10.33096/linier.v2i3.3142

Yuliani, S. P., Muharani, A. A. P., Fatmawati, R. Q., & Fahmi, F. (2025). Sentiment Analysis in User Reviews of Gojek Application using Natural Language Processing. *Journal of System and Computer Engineering (JSCE)*, *6*(4), 296–305. https://doi.org/10.61628/jsce.v6i4.2062