Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

June 2026. Vol. 5. No. 3; e-ISSN: 2808-4519

# Performance Evaluation of Machine Learning Algorithms in Sentiment Analysis of Spotify Reviews

**Frizi Olivian[1], Sahrul Bariyah[2], Grant Christo Budiyanto[3], Riski Annisa[4], Lady Agustin Fitriana[5], Weiskhy Steven Dharmawan[6*]**

[1,2,3,4]*Program Studi Informatika,*
[5]*Program Studi Sistem Informasi*
[6]*Program Studi Sistem Informasi Akuntansi*
*Universitas Bina Sarana Informatika*
Frizi190305@gmail.com[1] ,15235006@bsi.ac.id[2], grantchristo03@gmail.com[3], riski.rnc@bsi.ac.id[4], lady.lag@bsi.ac.id[5], Weiskhy.wvn@bsi.ac.id*[6]

**Abstract**

The rapid growth of digital music streaming platforms has generated a massive volume of user reviews on the Google Play Store, making manual analysis practically infeasible. This study evaluates and compares the performance of three machine learning algorithms Support Vector Machine (SVM), Neural Network (Multilayer Perceptron), and Random Forest in classifying sentiments from Spotify user reviews written in Indonesian. A total of 10,000 reviews were collected from the Google Play Store using the google-play-scraper library and processed through a text preprocessing pipeline comprising cleaning, case folding, word normalization, tokenization, stopword removal, and stemming using the Sastrawi library. Sentiment labeling was performed automatically using the InSet lexicon, categorizing reviews into three classes: Positive (56.63%), Neutral (30.60%), and Negative (12.76%). Feature extraction was conducted using the TF-IDF method, with an 80:20 train-test split strategy and stratified sampling to maintain class distribution. Model performance was evaluated based on accuracy, precision, recall, and F1-score metrics. The results demonstrate that SVM and Neural Network achieved equivalent and superior accuracy of 0.937, with macro F1-scores of 0.908 and 0.907, respectively, outperforming Random Forest which recorded an accuracy of 0.853 and a macro F1-score of 0.777. These findings indicate that SVM and Neural Network are more optimal and reliable for sentiment classification of Indonesian-language Spotify reviews, while Random Forest requires further improvement, particularly in recognizing minority classes.

***Keywords***: *Sentiment Analysis; Machine Learning; Support Vector Machine; Neural Network; Random Forest; Spotify; Natural Language Processing*

## 1. Introduction

Digital music streaming platforms have experienced rapid growth over the past decade as part of the transformation of the internet-based entertainment industry. Spotify, as one of the world's largest music streaming platforms, now serves more than 600 million active users across various countries, including Indonesia, which is one of the largest emerging markets in Southeast Asia [1]. This significant user growth has driven increased digital interaction, one of which is through the provision of reviews and application ratings on distribution platforms such as the Google Play Store [2]. This phenomenon makes user review data a valuable source of information that has not yet been fully and optimally utilized by application developers.

The large volume of Spotify user reviews on the Google Play Store presents its own challenges in the process of data analysis and decision-making. Programmatic retrieval of review data using libraries such as the google-play-scraper allows the acquisition of thousands to tens of thousands of reviews in a short time, which is practically impossible to analyze manually [3]. Each review contains multidimensional information in the form of free text, star scores (ratings), usernames, and timestamps, all of which have the potential to reveal patterns of user satisfaction and dissatisfaction comprehensively. This limitation of manual analysis at scale underscores the urgency of developing automated methods capable of processing and classifying user opinions efficiently and accurately [4].

Sentiment analysis, as a branch of Natural Language Processing (NLP), has emerged as a proven and effective solution for automating the understanding of user opinions at scale. By definition, sentiment analysis aims to identify and extract subjective information from text, such as opinions, assessments, and emotions, and then classify them into specific categories in this study encompassing positive, negative, and neutral sentiments [5]. Sentiment labeling can be performed using a lexicon-based approach, where words in the text are matched against a sentiment dictionary such as InSet (Indonesian Sentiment Lexicon), which contains positive and negative word weights in the

Indonesian language [6]. The application of sentiment analysis to application reviews provides strategic benefits for developers in understanding user perceptions, prioritizing feature improvements, and continuously enhancing service quality.
The development of machine learning algorithms for text classification has progressed from traditional statistical methods toward more sophisticated and adaptive approaches. Classic methods such as Naive Bayes and Decision Trees provided an important early foundation, but their limitations in handling the complexity of language drove the exploration of more advanced algorithms [7]. Support Vector Machine (SVM), Neural Network based on Multilayer Perceptron (MLP), and Random Forest are now the leading representatives in text classification, thanks to their ability to handle high-dimensional data produced from the text vectorization process using methods such as Bag-of-Words (BoW) [8]. This development is also accompanied by advances in text preprocessing techniques, including normalization of non-standard words, stopword removal, and the stemming process using libraries such as PySastrawi for the Indonesian language [9].

The three algorithms that are the focus of this research SVM, Neural Network, and Random Forest each have comparative characteristics and advantages that make them relevant for sentiment classification tasks. SVM works by finding the optimal hyperplane that separates data classes in a high-dimensional feature space, making it highly effective for vectorized text data that tends to be sparse [10]. Neural Network based on MLP, with its hidden layer architecture, is capable of capturing complex non-linear patterns in text feature representations, making it superior in understanding the context and nuances of language [11]. Random Forest, as an ensemble method that combines predictions from a large number of decision trees, is known to be highly robust against overfitting and capable of delivering stable performance on datasets with imbalanced class distributions [12].

Research conducted by Ginabila and Ahmad Fauzi examined the effectiveness of the Support Vector Machine (SVM) and Naive Bayes algorithms for classifying user review data from a platform with a broad music catalog. Through a comprehensive text data labeling process, the results revealed that both classification methods were highly competitive in performance. The SVM algorithm recorded an accuracy rate of 82.42%, while the Naive Bayes algorithm showed a slightly superior result with an accuracy of 84.73% [13].

Subsequent research was conducted by Habibillah et al. (2025), analyzing Spotify reviews on the Google Play Store using the VADER method and the Random Forest and Multinomial Naive Bayes algorithms. The results showed that user satisfaction (positive sentiment) centered on recommendation and playlist features, while complaints (negative sentiment) were dominated by issues with advertisements and application updates. By applying the Random Forest and Multinomial Naive Bayes algorithms, this research successfully developed a prediction model with competitive performance [14].

Prior research by Syahra Audiyani Fitra and Dwika Ananda Agustina (2025) on sentiment analysis of Spotify reviews on the Google Play Store showed that the use of digital technology has transformed music consumption patterns,but also presents challenges in maintaining user satisfaction. Through a comparison of classification methods, it was found that the Neural Network algorithm had superior performance in terms of accuracy, F1-score, and recall, while Naive Bayes showed better results on the AUC, precision, and MCC metrics. The data findings revealed a dominance of negative sentiment at 52.8%, indicating an urgent need for developers to make technical improvements and enhance the user experience. Overall, the study affirmed that sentiment analysis is a strategic instrument for extracting valuable insights from massive review data to support the development of applications that are more adaptive to market needs [15].

Based on prior research, algorithms such as Support Vector Machine (SVM), Naive Bayes, Random Forest, and Neural Network demonstrate competitive performance in sentiment analysis of Spotify user reviews, each offering distinct advantages. Sentiment analysis is also capable of identifying user satisfaction factors, such as application features, as well as dissatisfaction factors, such as advertisements and technical issues. Therefore, this study aims to compare the performance of SVM, Neural Network (MLP), and Random Forest algorithms in sentiment analysis of Spotify reviews from the Google Play Store, using the InSet lexicon method for labeling sentiments as positive, negative, and neutral, and evaluating the model using accuracy, precision, recall, and F1-score to determine the best-performing model.

## 2. Theoretical Foundation

### 2.1. Sentiment Analysis

Sentiment analysis is an approach in the field of Natural Language Processing (NLP) that aims to extract and classify emotional nuances in text in order to understand the dynamics of public perception over time [16].

### 2.2. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a classification algorithm that relies on finding an optimal hyperplane to divide data into specific classes. The main strength of this method lies in maximizing the distance or margin between the separating line and the nearest data points (support vectors), with the goal of improving prediction accuracy on new data [17].

### 2.3. Neural Network

A Neural Network can be understood as a machine learning algorithm that mimics the workings of the human neural network, processing data through several layers of neurons to identify complex patterns. This method excels at capturing non-linear relationships between features through a training process using weights and activation functions, enabling accurate predictions across various types of data, including sentiment analysis [18].

### 2.4. Random Forest

The application of Random Forest in sentiment analysis is based on its ability to produce stable predictions through the integration of many decision trees. This method has proven effective in improving classification accuracy because it uses a random approach in processing data and features, making the results more reliable [19].

## 3. Research Method

This research applies a comparative experimental approach to test and compare the performance of three algorithms: Support Vector Machine (SVM), Neural Network (NN), and Random Forest. The main focus is on examining the extent to which these three models are capable of classifying sentiment from user reviews of the Spotify application. All research stages, from beginning to end, have been systematically organized in a flowchart diagram in the attached figure.

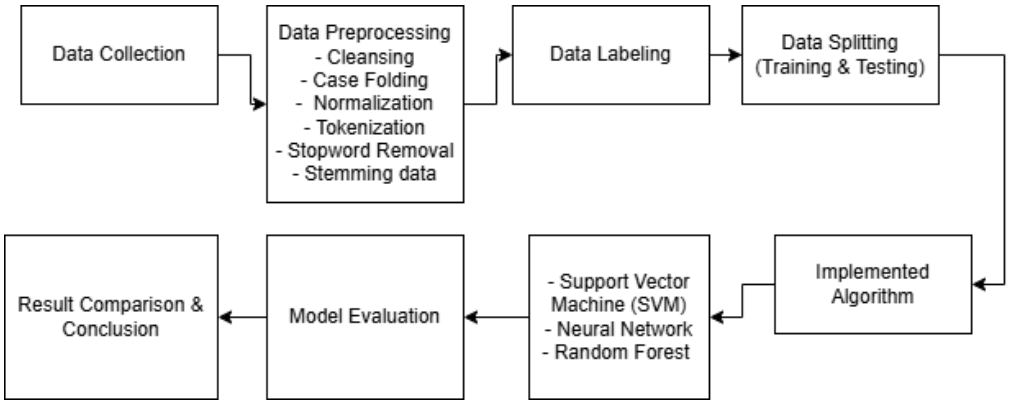


**Fig. 1.** Research Stages

### 3.1. Data Collection

Data collection was carried out by extracting the 10,000 most recent reviews of the Spotify application on the Google Play Store using the google-play-scraper library. The extraction focused on users in the Indonesian region who write in Indonesian. The collected data was then organized into a Pandas DataFrame and saved in CSV format to ensure that the analysis and data exploration process in Google Colab runs more optimally.

```
from google_play_scraper import reviews, Sort

app_id = 'com.spotify.music'

def get_reviews(app_id, lang='id', count=10000, sort=Sort.NEWEST, filter_score_with=None, filter_device_with=None, continuation_token=None):
    try:
        result, continuation_token = reviews(
            app_id,
            lang=lang,
            country='id',
            sort=sort,
            count=count,
            filter_score_with=filter_score_with,
            filter_device_with=filter_device_with,
            continuation_token=continuation_token
        )
```

**Fig. 2:** Scraping Process

### 3.2. Preprocessing

Following the data collection stage, the next step is data preprocessing. This stage aims to clean the review data of noise so that the SVM, Neural Network, and Random Forest algorithms can work optimally. The steps carried out are as follows:

#### 3.2.1 Cleaning

The data filtering stage was conducted to ensure the reliability of analytical results by repairing or removing corrupted and irrelevant data. This step involves cleaning the text of non-informative components, such as excess whitespace, numerical values, and certain symbols. By optimizing the cleanliness of the raw data, subsequent processing stages can run more accurately and in a structured manner.

#### 3.2.2 Case Folding

The case folding process in this study involves converting all review text to lowercase. This is crucial for ensuring data standardization and preventing semantic ambiguity that often arises from differences in the use of uppercase and lowercase letters. Through this standardization, every word with the same meaning will be counted as an identical entity by the model.

#### 3.2.3 Word Normalization

The word normalization stage was conducted to standardize text representation by converting non-standard terms, slang, and abbreviations into their standard forms. This procedure is crucial for minimizing lexical variation and maintaining data consistency, thereby improving text quality and readiness for processing in subsequent analysis stages.

#### 3.2.4 Data Labeling and Tokenization

In this sentiment analysis, data labeling was performed by classifying user reviews into positive, negative, or neutral categories to sharpen the understanding of the implied perceptions. Furthermore, to dissect the text into units of information, a tokenization process was applied that breaks sentences into individual words. This procedure was carried out using the Natural Language Toolkit library through the word_tokenize function. For preprocessing efficiency, a special word_tokenize_wrapper(text) function was developed to transform input text into a list of ready-to-process tokens.

In addition, a lexicon-based approach was also used in the data mining process to support sentiment analysis. The lexicon method works by utilizing a collection of words that already have specific weights or sentiment values, such as words with positive, negative, or neutral values. Each token produced by the tokenization process is matched against the lexicon dictionary so that the system can identify the sentiment tendency of a review. For example, words such as 'good', 'fast', and 'satisfying' are categorized as positive sentiment, while words such as 'bad', 'slow', and 'disappointed' fall under negative sentiment. This approach helps improve the accuracy of sentiment classification because the system is able to understand the emotional meaning contained in the text based on the vocabulary used by the user.

#### 3.2.5 Stopword Removal

Stopword removal is a technique for eliminating common words such as 'and', 'or', or 'is' that carry low semantic weight in sentiment analysis. By discarding these less informative terms, the analysis process can focus on more relevant key words. This study uses a list of Indonesian and English stopwords from the NLTK library combined with an additional list produced through manual curation in CSV format. The entire filtering process is carried out through the stopword_removal(review) function to ensure more comprehensive word coverage appropriate to the context of user reviews.

### 3.3. Stemming

Stemming is a stage for reducing every word to its base form or root. In this study, this process was implemented using the Sastrawi library, which has been optimized for Indonesian text processing. Through the stemming(review) function, every token in the data is systematically processed into a sequence of root words, thereby producing a more standardized text format for the needs of analysis in the next stage.

### 3.4. Lexicon

The lexicon method is one of the research methods used to analyze specific words, terms, or vocabulary based on the meaning, category, or sentiment contained within them. This method is widely used in linguistic research, text analysis, social media, sentiment analysis, and language-based qualitative research.

### 3.5. Split Data

Data splitting was carried out using the train_test_split function from Scikit-learn, allocating 80% of the data for the training stage and the remaining 20% for testing. This strategy aims to measure the model's ability to generalize to independent data, so that the analytical results obtained are more credible and measurable.

### 3.6. Classification

This stage focuses on the implementation of three machine learning algorithms: Support Vector Machine (SVM), Neural Network (NN), and Random Forest (RF). The SVM algorithm operates by determining the optimal hyperplane as a separator between data classes, while Neural Network adopts a neural computational structure to recognize complex patterns, and Random Forest works by combining results from a number of decision trees. All three methods are evaluated comparatively to identify the algorithm with the highest effectiveness in classifying user review sentiments.

#### 3.6.1. Support Vector Machine (SVM)

The SVM algorithm is relied upon in classification tasks due to its ability to identify the most effective separating hyperplane between data categories. By applying the SRM principle, SVM can reduce the risk of classification errors and remain stable even when facing high-dimensional data. SVM's adaptability to non-linear data through kernel functions makes it one of the robust models in processing complex data [20].

#### 3.6.2. Neural Network

The Neural Network algorithm is often the primary choice in text data analysis due to its ability to process features in depth and adapt to the dynamic characteristics of data. By adopting the functional structure of the brain's neural network, this method is capable of mapping non-linear input-output relationships through various processing layers, thereby producing high accuracy in handling complex data problems[21].

#### 3.6.2. Random Forest

The Random Forest algorithm is relied upon in text analysis due to its ability to produce more stable predictions by combining various decision tree units. By adopting a random approach to feature and data sample selection, this method is effective in enhancing classification power while maintaining consistency of results. This is what underlies the effectiveness of Random Forest in accurately handling customer opinion data [22].

### 3.7. Evaluation

To assess the level of classification accuracy, each model is evaluated using a set of performance metrics including accuracy, precision, recall, and F1-score. The technical implementation relies on the Scikit-learn library to produce detailed classification reports. This stage also involves confusion matrix analysis to map the distribution of prediction errors across different sentiment categories. This systematic evaluation process is crucial for comparing the performance of different algorithms to find the best solution for user review data.

## 4. Result and Discussion

| | Review ID | Username | Rating | Review Text | Date |
|---|---|---|---|---|---|
| 0 | 345c026a-d534-4deb-94be-0e01d1f1bd4c | Pengguna Google | 4 | gratis nya gak full | 2026-04-19 12:03:41 |
| 1 | ac70515e-de30-443b-ba88-da2aec29f54a | Pengguna Google | 5 | bgussss | 2026-04-19 12:00:09 |
| 2 | c3bc73e9-fef4-4e98-8e77-e346e5d21a5d | Pengguna Google | 5 | 👍 👍 👍 | 2026-04-19 11:58:03 |
| 3 | 9bdc3e46-6a92-4d95-aeb0-c9600ac39365 | Pengguna Google | 5 | tidak tau suka aja | 2026-04-19 11:54:27 |
| 4 | 11013933-0eaa-4c5d-8c12-a250e6edc9c8 | Pengguna Google | 5 | bagus | 2026-04-19 11:50:45 |

### 4.1. Data Collection

This study collected user reviews of the Spotify application consisting of profile names, star ratings, publication times, and review text content. Data retrieval was focused on users in the Indonesian region who commented in Indonesian, with the data sorted based on the most recent reviews. Once collected, the dataset was converted into a Pandas DataFrame format and saved as a CSV file. Data display was performed through Google Colab to ensure data quality before analysis was carried out in the next stage.

**Fig. 3:** Scraping Results of the Spotify

### 4.2. Prepocessing Data

To produce clean and structured data, this study applied a series of text preprocessing steps. The procedure includes the removal of disruptive characters, conversion of text to lowercase, and vocabulary standardization through the normalization stage. After the text is broken into tokens through the tokenization process, stopwords are eliminated to highlight meaningful terms. As a final stage, the Sastrawi stemmer is used to reduce variations of affixed words to a consistent base form for subsequent analysis.

#### 4.2.1. Cleaning dan Case Folding

The cleaning and case folding stages are important parts of the text data preprocessing process in sentiment analysis. The cleaning stage was carried out by eliminating various non-textual elements such as URLs, HTML tags, mentions, emojis, symbols, and numbers using Regular Expression (Regex) techniques. This procedure aims to minimize noise in the data so as to produce a cleaner text corpus. Subsequently, the case folding stage was performed by transforming all characters in the reviews to lowercase to ensure data format uniformity and avoid ambiguity caused by differences in capitalization. Through both of these stages, the text data becomes cleaner, more consistent, and optimal for processing in the next sentiment analysis stage.

| Review Text | cleaning | case_folding |
|---|---|---|
| gratis nya gak full | gratis nya gak full | gratis nya gak full |
| bgussss | bgussss | bgussss |
| 👍 👍 👍 | | |
| tidak tau suka aja | tidak tau suka aja | tidak tau suka aja |
| bagus | bagus | bagus |

**Fig. 4:** Results of the Cleansing and Case Folding

#### 4.2.2. Word Normalization dan Tokenization

The normalization and tokenization stages are further parts of the text preprocessing process in sentiment analysis. The normalization stage was performed by converting non-standard terms in the reviews into standard forms based on a prepared glossary or reference dictionary.

This step aims to minimize variations in the writing of words with the same meaning so that the linguistic integrity of the data is maintained. Subsequently, the tokenization stage was performed by decomposing the review text into the smallest linguistic units in the form of individual words using a space-based splitting method. This procedure converts sentences into a collection of separate tokens, facilitating subsequent processes such as feature extraction and the development of sentiment classification models. Through both of these stages, the text data becomes more consistent and structured for more optimal sentiment analysis.

| normalisasi | tokenize |
|---|---|
| gratis ya tidak full | [gratis, ya, tidak, full] |
| bagus | [bagus] |
| | [] |
| tidak tau suka saja | [tidak, tau, suka, saja] |
| bagus | [bagus] |

**Fig. 5:** Result of the Word Normalization and Tokenization

#### 4.2.3. Stopword Removal dan Stemming

Tahap stopword removal dan stemming merupakan bagian penting dalam proses preprocessing untuk meningkatkan kualitas data teks pada analisis sentimen. Tahap stopword removal memanfaatkan koleksi kata pengisi bahasa Indonesia dari pustaka NLTK untuk mengidentifikasi dan mengeliminasi istilah-istilah umum yang memiliki kontribusi minimal terhadap makna sentimen, seperti konjungsi dan preposisi. Dengan menghilangkan kata-kata tersebut, sistem dapat mengurangi redundansi bahasa dan memfokuskan analisis pada kosakata yang memiliki bobot semantik lebih kuat. Selanjutnya, tahap stemming dilakukan dengan memanfaatkan pustaka Sastrawi serta modul terkait dari NLTK untuk mereduksi berbagai variasi kata berimbuhan menjadi bentuk dasar atau akar kata. Proses ini bertujuan untuk menstandarisasi istilah yang memiliki makna serupa sehingga meningkatkan konsistensi data. Melalui kedua tahapan ini, pemrosesan morfologi bahasa Indonesia menjadi lebih optimal dan efektif dalam mendukung kinerja model analisis sentime.

| stopword removal | steming_data |
|---|---|
| [gratis, ya, full] | gratis ya full |
| [bagus] | bagus |
| [] | |
| [tau, suka] | tau suka |
| [bagus] | bagus |

**Fig. 6:** Result of the Stopword Removal dan Stemming

### 4.3. Labelling Data

Sentiment label determination for Spotify application reviews was performed automatically through a lexicon-based approach using InSet (Indonesian Sentiment Lexicon). After going through a series of preprocessing steps, each review was evaluated based on the frequency of occurrence of positive and negative vocabulary found in the dictionary. Sentiment classification was determined through aggregate comparison: reviews dominated by positive words were categorized as 'Positive', reviews with a tendency toward negative words were categorized as 'Negative', while reviews with a balanced number of both were designated as 'Neutral' sentiment [23].

| | Rating | steming_data | Score | Sentiment |
|---|---|---|---|---|
| **0** | 4 | gratis ya full | 2 | Positif |
| **1** | 5 | bagus | 0 | Netral |
| **2** | 5 | tau suka | -1 | Negatif |
| **3** | 5 | bagus | 0 | Netral |
| **4** | 5 | hembat | 0 | Netral |

**Fig. 7:** Results of the Labeling Process

### 4.4. Split Data

The dataset was split systematically by implementing the train_test_split function using an 80:20 ratio. This proportion allocates 6,300 training data and 1,575 testing data. This strategy was applied to ensure the model has optimal learning capacity from the training data, while also enabling objective performance evaluation on the test data. Furthermore, the use of the stratify parameter ensures that the proportional distribution across the positive, negative, and neutral sentiment classes is consistently maintained across both subsets The dataset was split systematically by implementing the train_test_split function using an 80:20 ratio. This proportion allocates 6,300 training data and 1,575 testing data. This strategy was applied to ensure the model has optimal learning capacity from the training data, while also enabling objective performance evaluation on the test data. Furthermore, the use of the stratify parameter ensures that the proportional distribution across the positive, negative, and neutral sentiment classes is consistently maintained across both subsets.

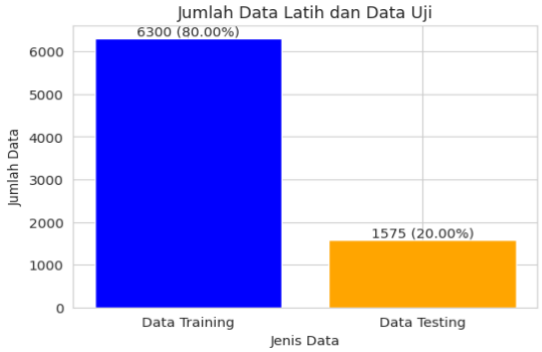


**Fig. 8:** Split Data

### 4.5. Total Sentiment Analysis

Based on the sentiment count analysis results, it is known that the data distribution is dominated by positive sentiment with a total of 4,460 data or 56.63% of the entire dataset. Neutral sentiment is in second place with 2,410 data (30.60%), while negative sentiment has the least amount at 1,005 data or 12.76%. This dominance of positive sentiment indicates that the majority of responses or opinions analyzed tend to give favorable assessments of the research object. Meanwhile, the relatively small proportion of negative sentiment indicates that the level of user dissatisfaction or criticism is not very significant. However, the considerable difference in distribution between classes also indicates the presence of data imbalance (imbalanced dataset), which could potentially affect the performance of the classification model, particularly in recognizing classes with fewer data such as negative sentiment.

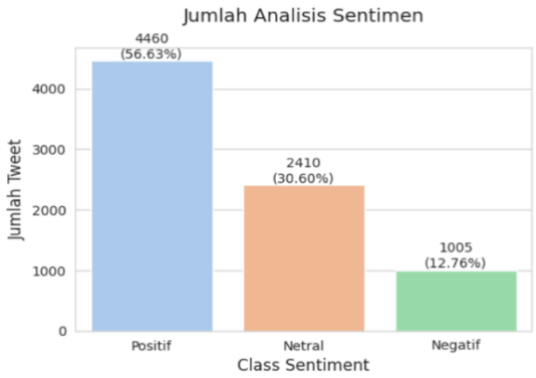


**Fig. 9:** Sentiment Analysis Results

### 4.6. Algorithm

### 4.6.1 Support Vector Machine (SVM)

Based on the classification report results for the Support Vector Machine (SVM) model, an overall accuracy rate of 0.937 was obtained, indicating very good model performance in classification. The Positive class had the best performance with a precision of 0.972, recall of 0.979, and F1-score of 0.975, indicating the model's very high ability to accurately identify positive data. Meanwhile, the Neutral class also showed stable performance with an F1-score of 0.904, followed by the Negative class with an F1-score of 0.846, which is still considered good despite being lower than the other classes. A macro average of 0.908 indicates balanced performance across classes, while a weighted

average of 0.937 indicates that the model works very optimally considering the data count distribution (support) dominated by the Positive class. Overall, the SVM model is able to provide accurate and consistent classification results on the dataset used.

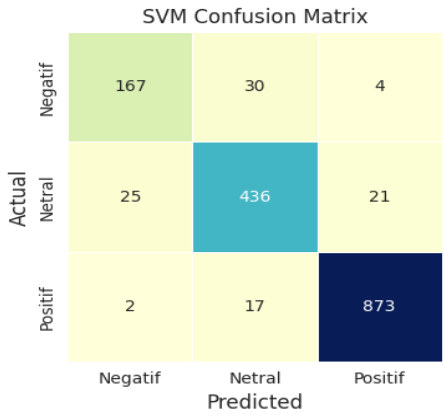


**Fig. 10:** Confusion Matrix SVM.

### 4.5.2 Random Forest

Based on the classification report results for the Random Forest model, an accuracy rate of 0.853 was obtained, indicating fairly good model performance but still below the previous SVM model. The Positive class had the most dominant performance with a high recall of 0.967 and F1-score of 0.909, indicating the model's very good ability to detect positive data. Conversely, the Negative class showed performance imbalance with high precision (0.967) but low recall (0.438), indicating that many negative data were not well detected. The Neutral class had relatively stable performance with an F1-score of 0.820. A macro average of 0.777 reflects performance disparities across classes, particularly in the Negative class, while a weighted average of 0.843 indicates that model performance is considerably influenced by the larger data distribution in the Positive class. Overall, the Random Forest model is capable of classification fairly well, but still requires improvement especially in recognizing the Negative class more optimally.

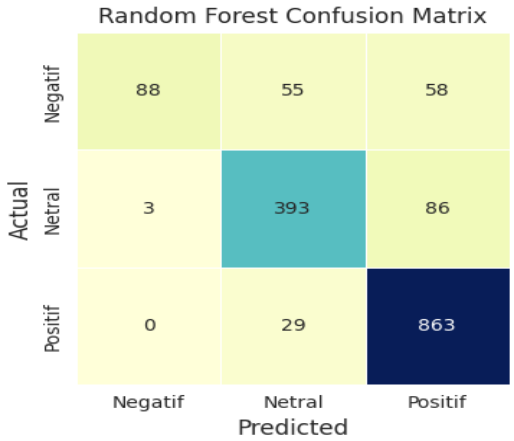


**Fig. 11:** Confusion Matrix Random Forest.

### 4.5.3 Neural Network

Based on the classification report results for the Neural Network model, an accuracy rate of 0.937 was obtained, indicating very good model performance comparable to SVM. The Positive class again proved most optimal with precision of 0.970, recall of 0.979, and F1-score of 0.974, indicating the model's very high ability to identify positive data. The Neutral class also showed stable performance with an F1-score of 0.904, while the Negative class had fairly good performance with an F1-score of 0.842, although slightly lower than the other classes. A macro average of 0.907 reflects balanced performance across classes, while a weighted average of 0.936 indicates the model works

consistently considering the data distribution. Overall, the Neural Network model is able to provide accurate and relatively balanced classification results across all classes.

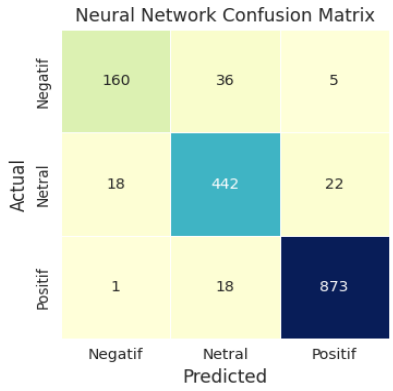


**Fig. 12:** Confusion Matrix Neural Network

### 4.7. Evaluasi

Based on the model accuracy comparison chart, it is evident that the SVM and Neural Network algorithms have equivalent and most superior performance with an accuracy value of 0.937 (93.7%), while Random Forest is below them with an accuracy of 0.853 (85.3%). These results indicate that SVM and Neural Network are more effective in capturing data patterns and performing accurate classification compared to Random Forest on the dataset used. These findings are also consistent with the previous classification report results, where both models had better F1-score values and inter-class balance, while Random Forest tended to experience performance degradation especially in certain classes. Therefore, it can be concluded that SVM and Neural Network are the most optimal models to use in this classification case, both in terms of accuracy and performance consistency.

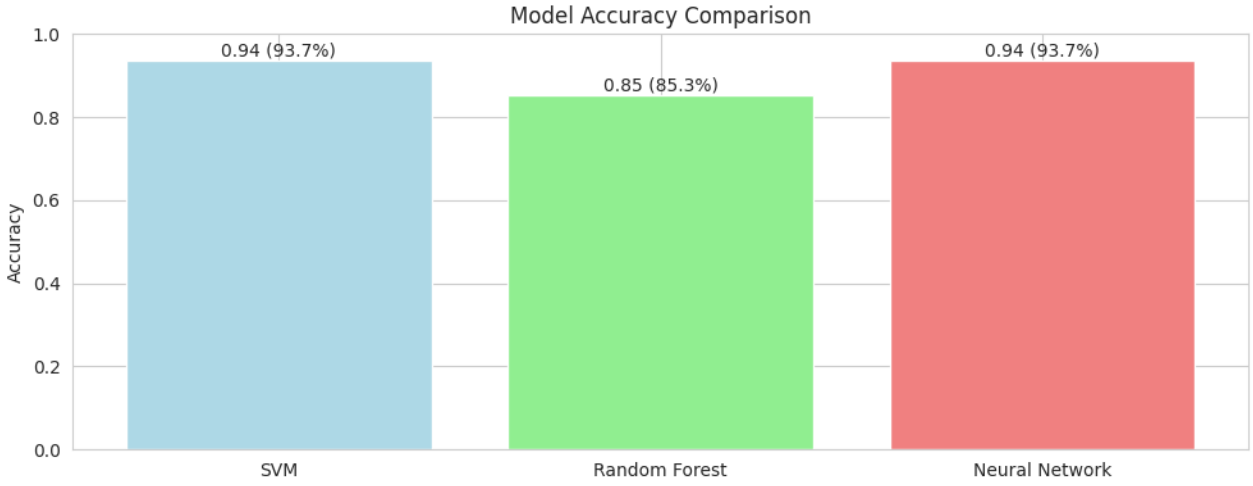


**Fig. 13:** Performance Results of SVM, RF, and NN

Table 1 shows that the Support Vector Machine (SVM) and Neural Network models outperform the Random Forest model on all evaluation metrics. Both models achieved the highest accuracy rate of 0.937 and demonstrated more balanced performance based on macro average and weighted average values, while Random Forest produced relatively lower performance with smaller F1-score values, particularly due to performance degradation in certain classes.

**Table 1.** Model Performance Comparison

| Metric | SVM | Random Forest | Neural Network |
|---|---|---|---|
| Accuracy | 0.937 | 0.853 | 0.937 |
| Macro Precision | 0.912 | 0.883 | 0.918 |
| Macro Recall | 0.905 | 0.740 | 0.897 |
| Macro F1 | 0.908 | 0.777 | 0.907 |
| Weighted F1 | 0.937 | 0.843 | 0.936 |

## 5. Conclusion

Based on the comparison of the three classification algorithms used Support Vector Machine (SVM), Random Forest, and Neural Network it can be concluded that SVM and Neural Network showed superior performance compared to Random Forest. Both models achieved the

highest accuracy of 0.937 and had more balanced macro and weighted F1-score values, indicating consistent classification capability across all classes. Meanwhile, Random Forest only achieved an accuracy of 0.853 and showed performance degradation on several metrics, particularly macro recall and macro F1, indicating imbalance in recognizing certain classes. Therefore, SVM and Neural Network can be considered the most optimal algorithms in this study, while Random Forest still requires improvement to achieve comparable performance.

As a recommendation, future research can perform parameter optimization (hyperparameter tuning) on each algorithm to improve model performance, particularly for Random Forest so that it can recognize all classes more evenly. In addition, handling data imbalance through resampling techniques (oversampling or undersampling) also needs to be considered to improve the model's ability to classify minority classes. The use of ensemble methods or the combination of multiple algorithms can also be explored to obtain more optimal results. Furthermore, increasing the amount of data and improving the quality of text preprocessing are expected to improve the accuracy and stability of the model overall.

## References


[1] H. Muhibin and Z. Fitriyah, "The Effect of Price Perception and Electronic Word of Mouth on the Purchase Intention of Spotify Premium Service in Surabaya City," vol. 3, no. 5, pp. 1605–1618, 2023.

[2] G. R. Ramadhan *et al.*, "ANALISIS SENTIMEN ULASAN APLIKASI DANA DI GOOGLE PLAY STORE MENGGUNAKAN ALGORITMA NAÏVE BAYES," vol. 8, no. 5, pp. 9849–9857, 2024.

[3] W. Ramadlan, M. Arifin, D. L. Fithri, and P. Setiaji, "ANALISIS SENTIMEN ULASAN PENGGUNA APLIKASI SPOTIFY DI GOOGLE PLAY STORE MENGGUNAKAN ALGORITMA NAIVE BAYES," vol. 9, no. 2, pp. 3600–3607, 2025.

[4] C. Wulandari and L. Sunardi, "Analisis Sentimen Aplikasi Spotify Pada Ulasan Pengguna di Google Play Store Menggunakan Metode Support Vector Machine," vol. 4, no. 5, pp. 2588–2595, 2024, doi: 10.30865/klik.v4i5.1762.

[5] H. Wijaya and N. Hayati, "NATURAL LANGUAGE PROCESSING ( NLP ) UNTUK ANALISIS SENTIMEN ULASAN SEBLAK BANDUNG PEDAS KUDUS Natural Language Processing ( NLP ) for Sentiment Analysis of Seblak Bandung Pedas Kudus Reviews," vol. 8, no. 1, pp. 13–22, 2025.

[6] A. Nadira, N. Y. Setiawan, and W. Purnomo, "ANALISIS SENTIMEN PADA ULASAN APLIKASI MOBILE BANKING MENGGUNAKAN METODE NAÏVE BAYES DENGAN KAMUS INSET," pp. 35–47, 2023.

[7] D. Saputro, I. Arwani, and D. E. Rahmawati, "ANALISIS SENTIMEN PADA ULASAN PENGGUNA APLIKASI JAKLINGKO DI GOOGLE PLAY STORE DENGAN METODE SUPPORT VECTOR MACHINE ( SVM )," vol. 9, no. 9, pp. 1–10, 2025.

[8] R. Saputra, A. Purno, W. Wibowo, J. C. No, K. C. Kidul, and K. Bandung, "Analisis Komparasi Algoritma SVM , Random Forest dan MLP-NN," vol. 11, no. 1, pp. 211–224, 2026.

[9] B. Classifier, M. R. Syafapri, E. Haerani, I. Iskandar, and L. Afriyanti, "Jurnal Computer Science and Information Technology ( CoSciTech ) Sentiment classification of interfaith marriage ban using Naive Bayes Classifier method," vol. 5, no. 1, pp. 10–18, 2024.

[10] V. N. Mei *et al.*, "Implementasi Model Support Vector Machine Dalam Analisa Sentimen Masyarakat Mengenai Kebijakan Penerapan Aplikasi Mypertamina Program Studi Sistem Informasi , Universitas Jambi , Indonesia Aplikasi MyPertamina merupakan aplikasi yang diluncurkan oleh PT Pertamina," no. 2, pp. 176–193, 2024.

[11] K. F. Sugiantari *et al.*, "PENDEKATAN MLP DALAM KLASIFIKASI BAHASA ISYARAT : ANALISIS JARAK EUCLIDEAN LANDMARK TANGAN," vol. 9, no. 2, pp. 209–218, 2025.

[12] O. Y. Inonu and K. Magda, "Analisis Kinerja Algoritma Random Forest Dengan Model Machine Learning Pada Dataset Penyakit Diabetes," vol. 15, no. 1, pp. 1–7, 2025.

[13] A. Fauzi, P. Studi, S. Informasi, U. Bina, and S. Informatika, "Analisis Sentimen Terhadap Pemutar Musik Online Spotify Dengan Algoritma Naive Bayes dan Support Vector Machine," vol. 6, no. 2, pp. 111–122, 2023.

[14] N. H. Habibillah, R. Nurlaela, R. D. Lestari, F. Mosyarrina, and N. Najma, "Analisis Sentimen Terhadap Aplikasi Spotify Berdasarkan Ulasan di Google Play Store," vol. 2, no. 1, pp. 300–306, 2025.

[15] S. Audiyani and F. Syahra, "Sentimen t analysis spotify applications on google play store with naïve bayes and neural network methods," vol. 3, no. 2, pp. 62–74.

[16] M. Suhendra, B. Lailiah, and L. A. Fitriana, "Evaluation of Machine Learning Algorithms in Sentiment Analysis of the Satu Sehat Application," vol. 5, no. 2, 2026.

[17] K. D. Iris, "PERBANDINGAN KINERJA ALGORITMA SVM DAN KNN UNTUK," pp. 472–478, 2026.

[18] J. Antares, T. Infomasi, and U. Dharmawangsa, "Artificial Neural Network Dalam Mengidentifikasi Penyakit Stroke Menggunakan Metode Backpropagation ( Studi Kasus di Klinik Apotik Madya Padang )," vol. 1, no. 1, pp. 6–14, 2020.

[19] A. Sagita *et al.*, "PENERAPAN METODE RANDOM FOREST DALAM MENGANALISIS," vol. 7, no. 6, pp. 3307–3313, 2023.

[20] M. M. Siregar, R. Hizria, and D. Pardede, "Perbandingan Kinerja Kernel SVM dalam Klasifikasi Kategori Kanker Kulit Menggunakan Transfer Learning," vol. 4, no. 2, pp. 83–90, 2025.

[21] K. L. Tan and C. P. Lee, "applied sciences A Survey of Sentiment Analysis : Approaches , Datasets , and Future Research," 2023.

[22] F. Teknik, U. Muhammadiyah, P. Pekalongan, F. T. Informasi, and I. W. Pratama, "IMPLEMENTASI ALGORITMA RANDOM FOREST DALAM ANALISIS," vol. XX, no. 2, pp. 38–41, 2025.

[23] F. Koto, "InSet Lexicon : Evaluation of a Word List for Indonesian Sentiment Analysis in Microblogs," pp. 391–394, 2017.