Proceeding of International Conference on Advanced Technologies, Innovations and Science
Vol 1, No 1, June 2026, Page 317-326
ISSN XXXX-XXXX (Media Online)
https://journals.adaresearch.or.id/icatis

Performance Analysis of Naive Bayes and SVM in Review Sentiment Classification of Spotify

Mu'ammar Nur Kholis[1,*], Amelia[1], Syaf'a Mubarok Siregar[1], Agus Perdana Windarto[2]

[1] Information System Study Program, Sekolah Tinggi Ilmu Komputer Tunas Bangsa, Pematangsiantar, Indonesia
Jl. Jendral Sudirman Blok A No. 1/2/3, Siantar Barat, Kota Pematangsiantar, Sumatra Utara, Indonesai
[2] Master of Informatics Study Program, Sekolah Tinggi Ilmu Komputer Tunas Bangsa, Pematangsiantar, Indonesia
Jl. Jendral Sudirman Blok A No. 1/2/3, Siantar Barat, Kota Pematangsiantar, Sumatra Utara, Indonesai
Email: [1,*]ammarkholis23@gmail.com, [2]ameliacantik1217@gmail.com, [3,]syafamubbarok0@gmail.com,
[4]agus.perdana@amiktunasbangsa.ac.id
Correspondence Author Email: ammarkholis23@gmail.com
Submitted: **01/06/2026**; Accepted: **12/06/2026**; Published: **30/06/2026**

Abstract−This study analyzes the sentiment of Spotify application users on the Google Play Store using a large dataset containing 100,000 reviews. The primary objective of this study is to compare the effectiveness of Naive Bayes and Support Vector Machine (SVM) algorithms in classifying user opinions into positive and negative categories. The dataset was cleaned through a text preprocessing pipeline that included case folding, cleansing, tokenizing, stopword removal, and stemming using the Sastrawi library. Neutral rating reviews were eliminated to yield a total clean dataset of 95,359 reviews. Numerical feature extraction was performed using the Term Frequency-Inverse Document Frequency (TF-IDF) method. Experimental results on the test data show that the SVM algorithm achieved the highest global accuracy rate of 92.57%, while Multinomial Naive Bayes recorded an accuracy of 92.09%. Interestingly, both algorithms produced an identical recall value of 0.79 in recognizing negative sentiment amidst the dominance of the majority class. The word cloud visualization detected that the main factor driving user satisfaction centered on the completeness of the song library, while critical user complaints were dominated by premium account transition issues and the intensity of advertisements. This research proves that an optimal combination of text preprocessing can mitigate class imbalance bias in both classification model architectures

Keywords: Sentimen Analysis; Spotify; Naive Bayes; Support Vector Machine; Text Classification.

1. INTRODUCTION

The development of information technology in the past decade has massively transformed the landscape of the global music industry, shifting the public consumption paradigm from physical media to digital based audio streaming platforms. Among various service providers, Spotify has successfully solidified its position as one of the most dominant market leaders, with an active user base reaching hundreds of millions worldwide. Through the Google Play Store platform, this application has been downloaded over one billion times and receives a continuous stream of user reviews daily. Every line of review left by users is a highly valuable qualitative asset because it represents a direct perception of service quality (Rahayu et al., 2022). In this study, boundaries regarding the subject and object of research are strictly defined to avoid academic ambiguity. The subject of this research is Spotify application users in Indonesia who voice their opinions, expressions, and personal experiences through the digital comment section on the Google Play Store. Meanwhile, the object of this research is the Indonesian language review text extracted from the platform to analyze its sentiment polarity into positive and negative classes.

To process a large scale textual dataset reaching hundreds of thousands of rows, a manual approach is no longer efficient. Sentiment analysis based on machine learning serves as the most relevant automated text classification methodology. Various algorithms have been proposed by previous researchers, where Naive Bayes and Support Vector Machine (SVM) have become the two supervised learning classification methods most frequently compared due to their efficient track record in processing text structured data (Vincent et al., 2024).

Each method possesses mathematical characteristics that give rise to their respective advantages and disadvantages. The Naive Bayes algorithm has the advantage of very high computational speed and training simplicity because it only involves conditional probability calculations without requiring complex numerical iterations. However, Naive Bayes has a fundamental weakness in its Conditional Independence Assumption, which assumes that every word in a document stands alone without correlation to other words. This assumption is often flawed when faced with the structure of Indonesian review sentences that contain contradictive conjunctions. On the other hand, the SVM algorithm offers high robustness in handling high dimensional feature spaces resulting from text extraction. SVM operates on the principle of Structural Risk Minimization (SRM) to find an optimal separating boundary in the form of a maximal margin hyperplane, making it highly immune to overfitting risks. Nevertheless, SVM carries a drawback in the high complexity of its training time and computational memory requirements, which increase exponentially as the dataset volume grows.

Several previous studies have provided a strong foundation for comparing the performance of these two algorithms in text classification tasks. Vincent et al. (2024) conducted research on sentiment classification on the Twitter platform using Naive Bayes and SVM methods with a multiclass approach. In their study, they emphasized the importance of using Term Frequency-Inverse Document Frequency (TF-IDF) and Information Gain feature extraction to optimize classification accuracy, where the use of N gram techniques proved to provide additional context to the processed words. Another study by Hendrawan & Kusniyati (2024) evaluated the performance of Naive Bayes and SVM in the context of electric vehicle sentiment analysis on Twitter. Their research findings confirmed that SVM frequently demonstrates better

stability and robustness than Naive Bayes when dealing with the complex, diverse, and non standard language styles on social media platforms. The expansion of comparative research on digital application stores is also reinforced by Jatmiko & Dometian (2026), who confirmed that SVM and Naive Bayes algorithms demonstrate high efficiency when processing large scale opinion data binarily. Furthermore, Apip & Kurniawan (2026) added that the reliability of text standardization and feature extraction is highly crucial in maintaining the stability of classification accuracy on mobile application distribution platforms.

The primary challenges and issues that arise when both algorithms are implemented on Spotify reviews are the high level of linguistic noise (noisy data) and the phenomenon of class distribution imbalance (class imbalance). The characteristics of digital reviews in Indonesia are generally non standard, filled with slang, inconsistent abbreviations (such as "bgus", "gk", "yg"), and excessive character repetitions ("baguuuuus"). This dirty text condition can obscure numerical feature weights. In addition, the distribution of review data is imbalanced, where positive reviews (ratings of 4 and 5) heavily dominate over negative reviews (ratings of 1 and 2). This massive data imbalance frequently triggers an orientation bias in the prior probability function of Naive Bayes, causing the algorithm to misclassify critical user complaints into the positive majority class (Wibowo et al., 2026).

To address this issue, this study integrates a rigorous text preprocessing pipeline using the Sastrawi library to purify the text corpus, combined with the TF-IDF weighting method. This preprocessing solution is projected to sharpen feature distinction parameters for each polarity class, thereby reducing the negative impact of class imbalance bias without manipulating the original data. The processing of user data reviews on a large scale can be automated effectively using data analysis techniques based on machine learning algorithms (Alyandi, 2025). By classifying reviews into positive and negative polarity categories, application developers can obtain a solid foundation for making data driven decisions to enhance user experience (Nurfauzi et al., 2025).

In text classification literature, Naive Bayes and SVM are frequently selected due to their efficiency. Naive Bayes is known for its high processing speed and computational simplicity on text data, while SVM is favored for its robustness in handling high dimensional feature spaces. However, the effectiveness of these two algorithms often varies depending on the dataset size and linguistic characteristics, necessitates an empirical comparison using a large scale dataset to find the most optimal model. In the context of evaluating application reviews on digital stores, research by Rhomaningtias et al. (2025) conducted a sentiment analysis on the SMILE Indonesia application. Their study indicated that identifying user complaints through reviews can provide direct recommendations to developers for continuous system improvement. However, reviews on app stores often harbor unique challenges. According to Wibowo et al. (2026) who evaluated algorithm performance on user reviews of the Pinterest application, the main challenge in classifying application review texts is the phenomenon of data imbalance (class imbalance), where the number of positive reviews far outweighs negative reviews. This class imbalance issue is also clearly present in the Spotify dataset within this research. Such a condition can potentially cause prediction bias in the model if not handled properly. As a solution to minimize the impact of class imbalance bias and noisy text data interference (slang and abbreviations), this study applies strict feature conditioning through the integration of a text preprocessing pipeline using the Sastrawi library combined with TF-IDF dimensional weighting.

The primary objective of this research is to perform an in depth comparative study and performance evaluation between the Naive Bayes and SVM algorithms in classifying a massive Indonesian Spotify review dataset reaching 100,000 data rows. Through a rigorous comparison of accuracy, precision, and recall metrics, this research aims to empirically prove which algorithm is more reliable and stable in recognizing user complaints amidst a dominant and imbalanced review data distribution.

2. RESEARCH METHODS

2.1 Dataset

The data used in this research is a secondary dataset obtained from Kaggle in .csv format. This dataset contains 100,000 Spotify user reviews covering various columns such as username, review content, rating, date, number of likes, and application version. The condition of this raw data is still highly noisy, containing extensive punctuation, numbers, emoticons, and non standard language that is not ready for direct processing by classification algorithms. A sample of the raw data can be observed in Table 1.

Table 1. Spotify Review Dataset

Username	Review	Rating	Date	Likes	App Version
Pengguna Google	suka banget sama ni app	5	2025-12-31 17:27:52	0	9.1.6.1145
Pengguna Google	bgus	5	2025-12-31 17:25:25	0	9.1.6.1145
dimas sebastian	Asik tp berbayar wkwkwkk	5	2024-05-23 06:58:21	0	8.9.40.509
…..	…..	…..	…..	…..	…..

Proceeding of International Conference on Advanced Technologies, Innovations and Science
Vol 1, No 1, June 2026, Page 317-326
ISSN XXXX-XXXX (Media Online)
https://journals.adaresearch.or.id/icatis

Username	Review	Rating	Date	Likes	App Version
Xie Suh	Minimal bisa di pake pas make data seluler min gw punya kuota tpi pas mo dengerin lagu di atur ke offline trs situ sehat??	2	2024-05-23 06:54:29	0	

The dataset can be accessed here

2.2 Research Framework

This study is an experimental quantitative research focusing on the comparison of machine learning algorithms for text classification. The object of this research is the application reviews from Spotify users obtained through the official Kaggle site, totaling 100,000 reviews. The primary variables in this study consist of the review text as the independent variable and the sentiment label (positive and negative) as the dependent variable. The research hypothesis states that there is a significant performance difference between Naive Bayes and SVM algorithms when handling large datasets, where SVM is predicted to exhibit a more stable accuracy level for imbalanced data. The following image represents the Research Method Design:

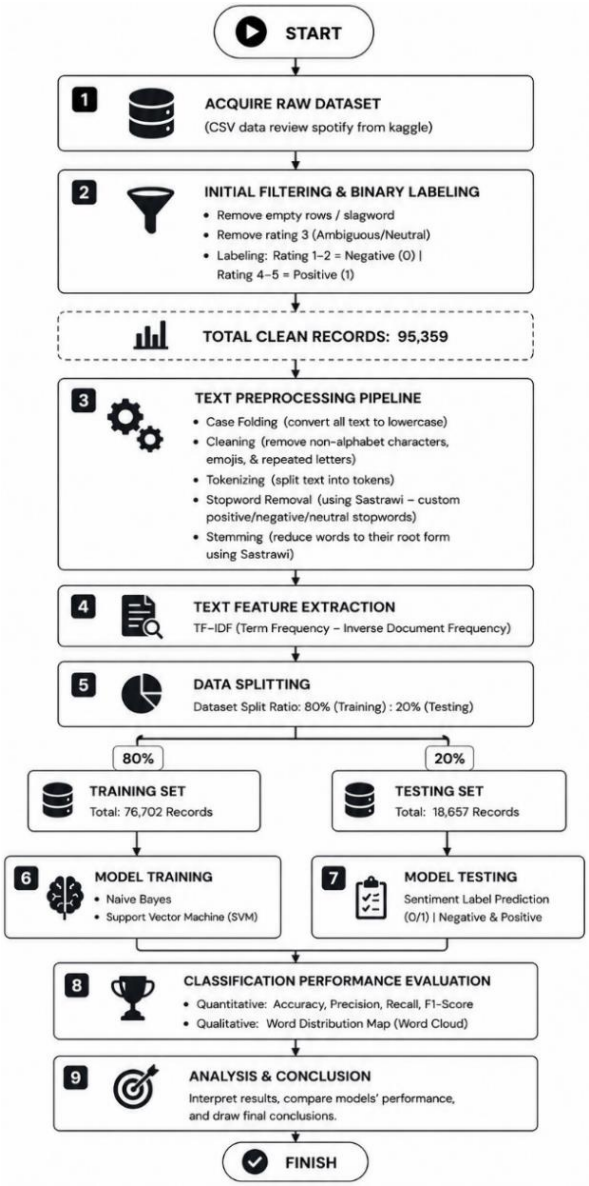

Figure 1. Research Method Design Flowchart

Figure 1 shows the research stages. The initial stage of data processing was data filtering and automated labeling based on user star ratings. Reviews with a rating of 3 were eliminated from the dataset due to their neutral or ambiguous nature, which could reduce the clarity of the classification model's decision boundary. Ratings 1 and 2 were categorized as negative sentiment (label 0), while ratings 4 and 5 were categorized as positive sentiment (label 1). After the filtering process was complete, the final dataset size ready for use was 95,359 reviews, with the distribution details in Table 2.

Proceeding of International Conference on Advanced Technologies, Innovations and Science
Vol 1, No 1, June 2026, Page 317-326
ISSN XXXX-XXXX (Media Online)
https://journals.adaresearch.or.id/icatis

Teble 2. Distribution of Ratings and Sentiment Labels Dataset

Rating	Frequency	Percentage	Label Sentiment	Rating
1	16.088	16,87%	Negative	1
2	3.601	3,78%	Negative	2
4	8.299	8,70%	Positive	4
5	67.371	70,65%	Positive	5
Total	95.359	100%		Total

The data in Table 2 shows quite extreme class imbalance, with the positive class dominating 79.35% of the total corpus data. This data imbalance is a critical parameter for testing the robustness of the Naive Bayes and SVM algorithms in maintaining global accuracy without sacrificing predictive ability in the minority class. Theoretically, this class imbalance has the potential to cause prediction bias in the model if not handled properly. The challenge of class imbalance in app store reviews aligns with previous studies conducted by Dwiansyah & Kurnia (2026) as well as Firmansyah et al., 2026). Both studies emphasize that the extreme dominance of positive reviews on the Google Play Store demands rigid text feature conditioning so that classification algorithms do not experience orientation bias and remain sensitive in recognizing document patterns within the minority class.

To transform the raw text into a form ready for classification, a systematic text preprocessing pipeline is applied (Khaira et al., 2023). Given that user reviews on digital platforms frequently contain non alphabetic characters, abbreviations, and informal language, cleaning actions are absolutely mandatory. The first step is case folding, which flattens all capital letters into lowercase and removes numerical characters so that the data purely contains alphabetic characters (Harmandini & L, 2024). The second step is cleansing, which is implemented to remove punctuation, symbols, and irregular characters that do not carry semantic value in sentiment analysis. The third step is tokenizing, where the review sentences are broken down into single word token units based on spaces (Atmaja, 2025). Next, a stopword removal process is performed to discard common words that have high frequency but do not carry sentiment weight information, such as the conjunctions "yang", "dan", or "itu" (R. Saputra et al., 2025). he final stage of this preprocessing is stemming, which converts inflected Indonesian words into their base forms by stripping prefixes, suffixes, and infixes utilizing the Sastrawi library (Yunanda et al., 2022).

Once the review documents are verified as clean, the next step is to perform feature extraction using the Term Frequency-Inverse Document Frequency (TF-IDF) method to convert text tokens into a numerical vector matrix format. This weighting method operates mathematically to measure the importance of a word within a document based on how frequently the word appears in a specific review (Term Frequency) balanced against the scarcity of that word across all documents in the dataset (Inverse Document Frequency) (Arasy & Agustian, 2025). word tokens that appear frequently in a particular review but are relatively rare in other reviews will be awarded a higher weight because they are considered unique specifiers of that sentiment polarity. The mathematical calculation of this TF-IDF weighting refers to the equation formula:

$$W_{d,t} = tf_{d,t} \times log\left(\frac{N}{df_t}\right) \quad (1)$$

Where $W_{d,t}$ represents the final weight of word t in document d, $tf_{d,t}$ denotes the frequency of the word's appearance in the document, N ndicates the total number of review documents in the experimental dataset corpus, and df_t arepresents the number of review documents containing word t. he high dimensional numerical feature matrix generated from this TF-IDF calculation is then funneled into the modeling stage through a data splitting process. To objectively test the reliability and generalization capability of the models on new data, the entire clean dataset structure totaling 95,359 reviews is randomly split with a proportion of 80% as the training set and 20% as the testing set.

The allocation and purpose of this data structure division are explained as follows:

a. Training Data (80% / 76,702 reviews): Flowed into the algorithm architecture for the training phase. At this stage, the training data is used as the basis for the model to extract patterns, calculate geometric parameter weights in SVM, and refine the probability distribution of word occurrences in Naive Bayes.
b. Test Data (20% / 18,657 reviews): Strictly isolated and not involved in the model training process at all. This test data is used purely as an evaluation instrument to test the model's generalization ability in classifying new, previously unseen reviews.

From the 20% division of the test data, the distribution structure of the actual target classes tested was 18,657 reviews, with 3,871 negative class reviews and 14,786 positive class reviews. This distribution figure served as the basis for compiling the Confusion Matrix components in the final evaluation stage. This experimental sequence concludes by measuring the classification performance reliability of both model architectures comprehensively utilizing the Confusion Matrix instrument. The evaluation of performance quality does not merely focus on the global accuracy metric (Accuracy), but is strictly measured through the Precision metric to observe the proximity of model prediction accuracy for each class, the Recall metric to assess the model's sensitivity in capturing actual data, and the F1-Score metric as the balancing harmonic mean of both. In the context of this application review research, the critical analysis focus will be heavily emphasized on the achievement of the negative class Recall value. This is conducted to evaluate the extent to which the Naive Bayes and SVM algorithms are capable of detecting and filtering technical user complaints precisely under the pressure of the imbalanced positive class review data dominance.

Proceeding of International Conference on Advanced Technologies, Innovations and Science
Vol 1, No 1, June 2026, Page 317-326
ISSN XXXX-XXXX (Media Online)
https://journals.adaresearch.or.id/icatis

2.3 Naïve Bayes

The Naive Bayes algorithm is a machine learning method that applies probability concepts from Bayes' Theorem to perform data classification processes (J. Saputra et al., 2025). The use of this probabilistic model has been widely validated in processing textual opinion data on digital application stores, as demonstrated in the research by Asri et al. (2022) as well as Hasibuan & Heriyanto, (2022). The primary advantage of Naive Bayes lies in its simple, rapid computational process and its capability to provide good performance on high dimensional datasets such as text data. The fundamental principle of Naive Bayes is to calculate the posterior probability of a class based on the prior probability and the feature appearance probability.

This algorithm utilizes a conditional independence assumption, meaning each feature is considered independent of one another once the class is known. Despite being based on the feature independence assumption, Naive Bayes remains capable of achieving strong performance across various complex real world problems (Amaria & Hidayati, 2025). This research utilizes the text data represented in the form of TF-IDF weights. The model is trained using the training data resulting from preprocessing and feature extraction, and is subsequently used to predict the sentiment polarity of Spotify reviews into positive and negative classes. The selection of this algorithm is based on its ability to handle large scale text data with relatively low computational requirements.

2.4 Support Vector Machine

Support Vector Machine (SVM) represents a supervised learning algorithm utilized to perform classification by searching for an optimal hyperplane capable of separating data from two distinct classes. In the case of text classification, each document is represented as a point in a high dimensional feature space formed by the results of TF-IDF weighting. The reliability of this geometric boundary architecture in classifying large scale digital platform review feedback data has been empirically proven in experiments by Kasi et al. (2025) as well as Maldini & Andryana (2025). The principle of Structural Risk Minimization (SRM) guides SVM to discover the decision boundary with the largest margin against the data from both classes. A larger margin indicates that the model can generalize better to new data. The optimal hyperplane is determined based on the data points located closest to the decision boundary, which are referred to as support vectors. Mathematically, the decision function on a linear SVM can be expressed in the equation:

$$f(x) = w^T x + b \quad (2)$$

Where w represents the weight vector, x is the input feature vector, and b is the bias. The value of the decision function is utilized to determine the class of a document based on the document's position relative to the formed hyperplane. n text classification, the document representation resulting from Term Frequency-Inverse Document Frequency (TF-IDF) weighting generates a highly dimensional feature space that is inherently sparse. Theoretically, the characteristics of this high dimensional text data are linearly separable. Therefore, the SVM architecture is configured using a linear kernel function with a standard regularization parameter value of C = 1.0. This choice of architecture makes SVM reliable and efficient in executing high dimensional text data, endows it with superior resistance to overfitting risks, and enables it to construct highly precise decision boundaries between review sentiment polarities (Salsabila et al., 2025).

3. RESULTS AND DISCUSSION

3.1 Experimental Results and Global Performance Evaluation

This section presents the results of the sentiment classification experiments on Spotify application review objects objectively. Model performance evaluation was conducted using a testing set of 20% of the total clean dataset, which had been strictly isolated since the beginning of the data splitting stage. The total independent test data used to evaluate the generalization capability of both model architectures was 18,657 reviews, consisting of 3,871 reviews containing negative sentiment (Class 0) and 14,786 reviews containing positive sentiment (Class 1). To measure the overall effectiveness, robustness, and reliability of the models, the evaluation instrument does not merely rely on the global accuracy metric (Accuracy), but involves detailed evaluation metrics such as Precision, Recall, and F1-Score for each target class. The actual performance computations obtained are comprehensively presented in Table 3.

Table 3. Comparative Performance Evaluation Metrics for Sentiment Classification

Algorithm	Sentiment Class	Precision	Recall	F1-Score	Global Accuracy
Naïve Bayes	Negative (0)	0,82	0,79	0,80	92,09%
	Positive (1)	0,94	0,96	0,95	
SVM	Negative (0)	0,84	0,79	0,82	92,57%
	Positive (1)	0,95	0,96	0,95	

Based on the empirical data presented in Table 3, both machine learning algorithms demonstrate superior classification performance with global accuracy achievements above the 92% threshold. The SVM algorithm excels as the best performing model in this experiment, achieving an accuracy value of 92.57%. However, the Naive Bayes algorithm provides a highly competitive performance matching, recording a global accuracy value of 92.09%. The minute

accuracy margin between the two models, which stands at a mere 0.48%, stands as a highly intriguing theoretical finding in this study. This empirically proves that within the Indonesian mobile application review corpus that has undergone optimal text preprocessing, a simple probabilistic algorithm like Naive Bayes possesses a highly robust memory capacity to generate classification performance almost on par with high dimensional geometric boundary based algorithms such as SVM.

3.2 Microscopic Confusion Matrix Analysis

To microscopically dissect the classification decisions made by both models, an in depth analysis was conducted on the confusion matrix visualizations presented in Figure 2. This matrix evaluation is highly crucial to verify whether the models experienced performance degradation or orientation bias due to the phenomenon of data imbalance (class imbalance), where the number of positive data samples heavily dominates over negative data samples.

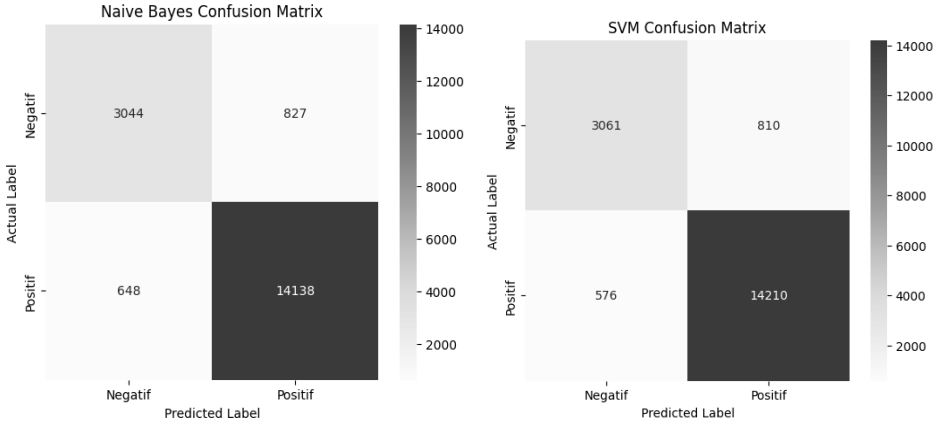

Figure 2. *Confusion Matrix of Naive Bayes Prediction Results & Confusion Matrix of SVM Prediction Results*

One of the greatest scientific findings in this research was detected in the sensitivity of the negative class Recall (Class 0). Both algorithms consistently obtained an identical Recall value of 0.79. Based on actual calculations from the confusion matrix images, Naive Bayes successfully predicted exactly 3,044 negative reviews out of a total of 3,871 actual data points (True Negative), while SVM successfully predicted exactly 3,061 negative reviews. The fact that both models were capable of capturing 79% of the total actual user complaints amidst the influx of dominant positive data indicates that the integrated text preprocessing stage (especially contextual common word elimination and morphological stemming via Sastrawi) successfully purified and highlighted the minority class distinctive features. Neither the probability weights nor the geometric positions of negative words were drowned out by the positive majority class dominance.

The minor performance discrepancy of 0.48% in global accuracy was driven by the superiority of SVM in the negative class Precision metric, where SVM achieved a value of 0.84 while Naive Bayes recorded a value of 0.82. The 2% margin in this metric represents that Naive Bayes possesses a slightly higher False Positive error rate. This means there are several review documents that actually contain positive sentiment but were misinterpreted as complaint reviews (negative) by Naive Bayes. Conversely, the geometric architecture of SVM proved to be more selective and precise in mapping the purity boundary of the negative class through the placement of an optimal separating hyperplane position that segregates the sparse text feature distribution resulting from TF-IDF weighting.

3.3 Comparative Performance Evaluation Metrics for Sentiment Classification

Qualitative information extraction was performed by utilizing word frequency distribution visualization techniques, aiming to visually map the dominant factors driving the satisfaction of research subjects (gain points) as well as the critical points of complaint (pain points) regarding Spotify application feature updates in Indonesia. The comparative distribution of dominant words is presented in Figure 3 and Figure 4.

Figure 3. Visual Word Cloud Representation of Primary Positive Sentiment Topics

Proceeding of International Conference on Advanced Technologies, Innovations and Science
Vol 1, No 1, June 2026, Page 317-326
ISSN XXXX-XXXX (Media Online)
https://journals.adaresearch.or.id/icatis

Figure 4. Visual Word Cloud Representation of Primary Negative Sentiment Topics

Based on the visual analysis in Figure 3 (Positive Sentiment), the lexemes that feature gigantic font sizes and absolutely dominate are the words "bagus" (good), "banget" (very), "suka" (like), "mantap" (excellent), and "lagu" (song). This word distribution pattern reflects high loyalty and satisfaction from users toward the core functionality of the Spotify music streaming platform. The research subjects highly appreciate system stability in presenting crystal clear audio compression quality, the completeness of the song library across various genres, and the artificial intelligence of Spotify's recommendation algorithm system, which is deemed highly accurate in reading personal preferences and listener moods through daily playlists.

Conversely, critical analysis must be directed toward Figure 4 (Negative Sentiment), which exposes the deepest root causes of user dissatisfaction. The keywords that appear massively and repeatedly are "premium", "spotify", "iklan" (ads), "lagi" (again), as well as the emergence of strong negation phrases such as "tidak bisa" (cannot) and "ga bisa" (unable to). This finding delivers a highly valuable practical contribution to digital product developers. The main issue for Spotify users in Indonesia is not triggered by software infrastructure technical constraints such as frequent application crashes, code bugs, or system lagging. The resistance point and anger of the research subjects purely root from product monetization policies and feature restrictions.

Many users complain about the high intensity of audio ads on free accounts, which is felt to be too high and disrupts the comfort of listening to music. Furthermore, the combination of the word "premium" pairing with the phrase "tidak bisa" represents the deep disappointment of users toward Spotify's policy updates, where several basic features that used to be accessible free of charge such as the freedom to choose specific songs, unlimited song skips, or viewing synchronized lyrics are now unilaterally locked by the developer and converted into exclusive features that can only be enjoyed by paying for a premium subscription package.

3.4 Theoretical Discussion of Geometric and Probabilistic Characteristics

The slight edge of the SVM algorithm and the capability of Naive Bayes to closely match the decision boundary performance can be explained from the mathematical structure of both algorithms when processing text feature spaces resulting from TF-IDF weighting. SVM operates based on the principle of Structural Risk Minimization (SRM). This algorithm projects text tokens into a high dimensional vector space and attempts to discover an optimal linear separating boundary (maximal margin hyperplane) that divides positive and negative classes with the furthest distance from the outermost data points (support vectors). Because text features resulting from TF-IDF are generally linearly separable in high dimensions, SVM is capable of minimizing classification errors in sentiment border areas that have high levels of ambiguity, thereby boosting the negative complaint class Precision metric value to 0.84.

On the other hand, the performance leap of Naive Bayes, which managed to pierce a global accuracy of 92.09%, is massively influenced by the dataset volume used in this study. Naive Bayes relies on the Conditional Independence Assumption, which assumes that every word in a document is independent or does not influence one another. Although this assumption is considered less realistic in the reality of human language that possesses complex syntactic bonds, this theoretical weakness was successfully compensated for by the abundant training dataset volume (reaching 76,702 reviews). With a massive supply of training data, the estimation of word appearance probability values for each sentiment class becomes highly mature and accurate. Highly defining keywords like the lexeme "bagus" or "iklan premium" possess highly contrastive probability distributions in both classes, so that the simple probability multiplication function on Naive Bayes is capable of predicting the target class with high precision and high computational speed.

3.5 Case Specimen Analysis and Classification Decision Limits

To deepen the theoretical analysis regarding model decision boundaries, a micro evaluation was conducted on review specimens possessing unique linguistic characteristics. This step aims to dissect how both algorithms respond to sentence structures containing contradictive conjunctions or word ambiguities. Examples of review specimens and the comparative prediction results from both models are presented in Table 4.

Proceeding of International Conference on Advanced Technologies, Innovations and Science
Vol 1, No 1, June 2026, Page 317-326
ISSN XXXX-XXXX (Media Online)
https://journals.adaresearch.or.id/icatis

Table 4. Case Samples of Contextual Model Prediction Differences

Actual Spotify User Review Text	Actual Label	Naive Bayes Prediction	Prediction SVM
"Aplikasi sebenarnya bagus, tapi iklannya terlalu sering muncul jadi malas pakai."	0 (Negative)	1 (Positive)	0 (Negative)
"Kenapa lagu JKT48 tidak bisa diputar lagi? Harus premium terus."	0 (Negative)	0 (Negative)	0 (Negative)
"Suka banget sama fiturnya, cuma kadang nge lag sedikit pas ganti lagu."	1 (Positive)	1 (Positive)	1 (Positive)
"Dulu gratis enak, sekarang apa-apa bayar. Kecewa banget sama update baru."	0 (Negative)	1 (Positive)	0 (Negative)

In the first and fourth case specimens in Table 4, the structural weakness of the Naive Bayes probabilistic function is clearly visible. The Naive Bayes model misclassified those complaint reviews as positive sentiment reviews (Label 1). This phenomenon occurs because Naive Bayes calculates probability values word by word in isolation without considering word order or the presence of contradictive conjunctions. The weights of the words "bagus" (good) and "enak" (pleasant), which have high frequencies in the positive class within the training data, successfully dominated the total probability multiplication, thereby drowning out the negative weights of the lexemes "iklan" (ads), "malas" (lazy), and "kecewa" (disappointed).

Conversely, the SVM model successfully executed the prediction of those reviews correctly as negative sentiment (Label 0). As a model based on geometric vector space, SVM does not compute isolated probabilities, but rather measures the spatial coordinates of the combination of all features present simultaneously. The presence of contradictive word combinations like "tapi" (but), "iklan" (ads), "bayar" (pay), and "kecewa" (disappointed) is geometrically capable of shifting the coordinate position of the review document past the separating hyperplane boundary into the negative class territory. This case confirms the superiority of SVM in handling reviews that possess sentences that reverse direction midway through the document.

3.6 Confrontation of Results with Previous Research (State of the Art)

When compared to the study by Vincent et al. (2024), which conducted a comparative sentiment analysis of Naive Bayes and SVM on the Twitter platform, the global accuracy rate obtained in this research is far higher and more stable (above 92%). On the Twitter corpus, classification model performance generally degrades below the 85% threshold due to the high diversity of conversation topics, the use of unpatterned slang language, and text character length limitations. This research proves that on digital store review corpora like the Google Play Store, the semantic focus of documents is homogeneous because it only revolves around application system performance and service quality, enabling the model to recognize word patterns with a far superior level of precision. Furthermore, these experimental results provide a new perspective that enriches the findings of Wibowo et al. (2026) on Pinterest application review classification.

That study highlighted that data imbalance (class imbalance) is a fatal obstacle that can damage model accuracy and trigger extreme bias in Naive Bayes, causing its performance to drop. However, the experimental findings on Spotify reviews offer a different empirical proof. Through the implementation of an optimal text preprocessing pipeline using a combination of the Sastrawi library and custom stopwords modification, the Naive Bayes algorithm proved not to be destroyed or extremely biased. Naive Bayes was capable of closely matching SVM accuracy with a margin below 0.5% and generated the exact same negative class Recall sensitivity level as SVM, which is 0.79. This finding emphasizes that on a very large dataset scale (reaching the 100,000 data range), an abundant data volume acts as a blessing for Naive Bayes to mature its feature probability calculations, making it no less robust than the geometric boundary model of SVM, which operationally demands far higher and heavier computational costs.

4. CONCLUSION

This research provides an empirical conclusion that the application of an optimal text preprocessing pipeline using a combination of the Sastrawi library and Term Frequency-Inverse Document Frequency (TF-IDF) weighting is capable of significantly reducing class imbalance bias in a massive Indonesian Spotify review sentiment classification task. Quantitative test results prove that the Support Vector Machine (SVM) algorithm generates the best performance with a global accuracy rate reaching 92.57%, followed very closely by the Naive Bayes algorithm with an accuracy of 92.09%. The most crucial theoretical finding in this study is the achievement of an identical negative class recall sensitivity value of 0.79 across both models, confirming that a large training data volume combined with rigid feature cleaning can mature the probability distribution of a simple model to match the reliability of a high dimensional geometric boundary model. Qualitatively through visual word cloud analysis, this research successfully addresses practical issues by revealing that the critical point of user dissatisfaction (pain points) does not stem from software system technical bugs or app crashes, but rather from resistance to product monetization policies via the transition of basic free features to premium accounts and the high intensity of audio ads penayangan. Nevertheless, this research still possesses limitations as it only evaluates architectures with standard default parameters without involving automated hyperparameter optimization. Therefore, the

scientific recommendation for future research development is to integrate metaheuristic optimization methods such as Particle Swarm Optimization (PSO) or Genetic Algorithms (GA) to optimize model parameter placement, as well as exploring advanced architectures based on deep learning or transformers to handle deeper contextual complexities of natural language.

ACKNOWLEDGMENT

The authors express their deepest gratitude to the Informatics Study Program at STIKOM Tunas Bangsa, the research advisors, and all academic peers who have provided computational facility support, academic guidance, and highly valuable constructive feedback during the execution of the experiments through the completion of this research paper writing.

REFERENCES

Alyandi, L. O. (2025). Perbandingan Kinerja Naïve Bayes Dan Svm Dalam Analisis Sentimen Ulasan Free Fire Di Play Store. In *JATI (Jurnal Mahasiswa Teknik Informatika)* (Vol. 9, Issue 3, pp. 5127–5135). LPPM ITN Malang. https://doi.org/10.36040/jati.v9i3.13599

Amaria, S. C., & Hidayati, N. (2025). *Analisis Program Makan Bergizi Gratis Dengan Support Vector Machine (SVM) PADA APLIKASI X*. *12*(2), 16–24.

Apip, A., & Kurniawan, A. (2026). Perbandingan Kinerja Algoritma Naive Bayes dan K-Nearest Neighbors untuk Analisis Sentimen Ulasan Aplikasi Lazada Menggunakan Python. In *Jurnal Publikasi Teknik Informatika* (Vol. 5, Issue 1, pp. 218–234). Politeknik Pratama Purwokerto. https://doi.org/10.55606/jupti.v5i1.6437

Arasy, A., & Agustian, S. (2025). *Sentiment Classification Using Multilayer Perceptron Algorithm with TF-IDF Features Klasifikasi Sentimen Menggunakan Metode Multilayer Perceptron dengan Fitur TF-IDF*. *5*(July), 908–919.

Asri, Y., Suliyanti, W. N., Kuswardani, D., & Fajri, M. (2022). Pelabelan Otomatis Lexicon Vader dan Klasifikasi Naive Bayes dalam menganalisis sentimen data ulasan PLN Mobile. In *PETIR* (Vol. 15, Issue 2, pp. 264–275). Sekolah Tinggi Teknik-PLN. https://doi.org/10.33322/petir.v15i2.1733

Atmaja, F. S. (2025). Peningkatan Layanan Customer Service Melalui Chatbot Menerapkan Algoritma Text Mining dan TF-IDF. In *Bulletin of Data Science* (Vol. 4, Issue 2, pp. 57–69). Forum Kerjasama Pendidikan Tinggi (FKPT). https://doi.org/10.47065/bulletinds.v4i2.6416

Dwiansyah, O., & Kurnia, R. D. (2026). Analisis Sentimen Ulasan Aplikasi Mobile Jkn Menggunakan Naïve Bayes, Knn, Dan Svm Berbasis Smote. In *Rabit : Jurnal Teknologi dan Sistem Informasi Univrab* (Vol. 11, Issue 1, pp. 349–362). LPPM Universitas Abdurrab. https://doi.org/10.36341/rabit.v11i1.6930

Firmansyah, M. D., Setiawan, R. R., & Irawan, Y. (2026). Analisis Perbandingan Kinerja Algoritma Svm, Naïve Bayes, Dan Knn Dalam Klasifikasi Sentimen Ulasan Aplikasi Pinterest Dengan Smote Dan Pso. In *Rabit : Jurnal Teknologi dan Sistem Informasi Univrab* (Vol. 11, Issue 1, pp. 639–652). LPPM Universitas Abdurrab. https://doi.org/10.36341/rabit.v11i1.7095

Harmandini, K. P., & L, K. M. (2024). *Analysis of TF-IDF and TF-RF Feature Extraction on Product Review Sentiment*. *8*(2), 929–937.

Hasibuan, E., & Heriyanto, E. A. (2022). Analisis Sentimen Pada Ulasan Aplikasi Amazon Shopping Di Google Play Store Menggunakan Naive Bayes Classifier. In *Jurnal Teknik dan Science* (Vol. 1, Issue 3, pp. 13–24). Asosiasi Dosen Muda Indonesia. https://doi.org/10.56127/jts.v1i3.434

Hendrawan, G. N., & Kusniyati, H. (2024). Evaluasi Performa Naive Bayes dan SVM dalam Analisis Sentimen Kendaraan Listrik di Media Sosial Twitter. In *CESS (Journal of Computer Engineering, System and Science)* (Vol. 9, Issue 1, pp. 299–313). State University of Medan. https://doi.org/10.24114/cess.v9i1.54086

Jatmiko, S., & Dometian, C. (2026). Analisis Sentimen Ulasan Google Play Store: Studi Komparatif Algoritma SVM, Naïve Bayes, dan Logistic Regression. In *JURNAL FASILKOM* (Vol. 15, Issue 3, pp. 610–620). LPPM Universitas Muhammadiyah Riau. https://doi.org/10.37859/jf.v15i3.10016

Kasi, D. A., Setyawan, Y., & Astuti, F. (2025). Perbandingan Efektivitas Naive Bayes dan Support Vector Machine (SVM) dalam Klasifikasi Sentimen Pengguna Aplikasi Chatgpt di Platform Google Play Store. In *Jurnal Teknologi* (Vol. 18, Issue 2, pp. 110–123). Institut Sains and Teknologi AKPRIND Yogyakarta. https://doi.org/10.34151/jurtek.v18i2.4463

Khaira, U., Aryani, R., & Hardian, R. W. (2023). Komparasi Algoritma Naïve Bayes Dan Support Vector Machine (SVM) Pada Analisis Sentimen Kebijakan Kemdikbudristek Mengenai Kuota Internet Selama Covid-19. In *Jurnal PROCESSOR* (Vol. 18, Issue 2). LPPM STIKOM Dinamika Bangsa Jambi. https://doi.org/10.33998/processor.2023.18.2.897

Maldini, M. A. A., & Andryana, S. (2025). Analisis Sentimen Ulasan Pengguna Aplikasi Perbankan Menggunakan Algoritma Support Vector Machine Dan Naive Bayes. In *JATI (Jurnal Mahasiswa Teknik Informatika)* (Vol. 9, Issue 3, pp. 4098–4105). LPPM ITN Malang. https://doi.org/10.36040/jati.v9i3.13522

Nurfauzi, O. M., Hilabi, S. S., Nurapriani, F., & Huda, B. (2025). Analisis Sentimen, Grab Indonesi Analisis Sentimen Grab Indonesia Pada Ulasan Google Play Store Menggunakan Algoritma Naïve Bayes Dan SVM. In *SMARTICS Journal* (Vol. 11, Issue 1, pp. 8–13). Universitas PGRI Kanjuruhan Malang. https://doi.org/10.21067/smartics.v11i1.11789

Rahayu, A. S., Fauzi, A., & Rahmat, R. (2022). Komparasi Algoritma Naïve Bayes Dan Support Vector Machine (SVM) Pada Analisis Sentimen Spotify. In *Jurnal Sistem Komputer dan Informatika (JSON)* (Vol. 4, Issue 2, p. 349). STMIK Budi Darma. https://doi.org/10.30865/json.v4i2.5398

Rhomaningtias, L., Khairunisa, A., Wara, S. S. M., & Hindrayani, K. M. (2025). Analisis Sentimen Ulasan Aplikasi Smile Indonesia Menggunakan Metode Naive Bayes Dan Support Vector Machine (SVM). In *HOAQ (High Education of Organization Archive Quality) : Jurnal Teknologi Informasi* (Vol. 16, Issue 1, pp. 79–91). Sekolah Tinggi Manajemen Informatika Komputer (STIKOM) Uyelindo Kupang. https://doi.org/10.52972/hoaq.vol16no1.p79-91

Salsabila, S. A., Priyatna, B., Hananto, A., & Tukino. (2025). Komparasi Kinerja Model Naive Bayes, SVM, dan BERT dalam

Klasifikasi Sentimen Ulasan Pada Aplikasi YUMMY. In *STORAGE: Jurnal Ilmiah Teknik dan Ilmu Komputer* (Vol. 4, Issue 2, pp. 42–47). Yayasan Literasi Sains Indonesia. https://doi.org/10.55123/storage.v4i2.5120

Saputra, J., Maryani, L., Rahmaddeni, Wulandari, D., & Eka, W. (2025). Analisis Performa Naive Bayes Dan Svm Terhadap Sentimen Teks Media Sosial Dengan Word2vec Dan Smote. In *Jurnal INSTEK (Informatika Sains dan Teknologi)* (Vol. 10, Issue 1, pp. 143–155). Universitas Islam Negeri Alauddin Makassar. https://doi.org/10.24252/instek.v10i1.54889

Saputra, R., Pristyanto, Y., & Fajri, I. N. (2025). *Generative AI Image Sentiment Analysis on Social Media X using TF- IDF and FastText*. *9*(5), 2509–2520.

Vincent, R., Maulana, I., & Komarudin, O. (2024). Perbandingan Klasifikasi Naive Bayes Dan Support Vector Machine Dalam Analisis Sentimen Dengan Multiclass Di Twitter. In *JATI (Jurnal Mahasiswa Teknik Informatika)* (Vol. 7, Issue 4, pp. 2496–2505). LPPM ITN Malang. https://doi.org/10.36040/jati.v7i4.7152

Wibowo, I. A., Indahyanti, U., Kautsar, I. A., & Dijaya, R. (2026). Evaluasi Performa Naive Bayes Dan SVM Dalam Memprediksi Sentimen Ulasan Pengguna Aplikasi Pinterest. In *JATI (Jurnal Mahasiswa Teknik Informatika)* (Vol. 10, Issue 2, pp. 2933–2940). LPPM ITN Malang. https://doi.org/10.36040/jati.v10i2.17860

Yunanda, G., Nurjanah, D., & Meliana, S. (2022). *Recommendation System from Microsoft News Data using TF-IDF and Cosine Similarity Methods*. *4*(1), 277–284. https://doi.org/10.47065/bits.v4i1.1670