KOMTEKINFO JOURNAL
LPPM Universitas Putra Indonesia YPTK Padang
Jl. Raya Lubuk Begalung, Padang, West Sumatera, Indonesia, Zip Code: 25221
Volume: 13, Issue: 1, Page: 27 - 35, March 30th, 2026, e-ISSN: 2502-8758
Available online at website: https://jkomtekinfo.org/ojs/index.php/komtekinfo

Public Sentiment Analysis of Train Services Based on Twitter Opinions Using K-Menas and SVM Methods

Dina Selvia[1]✉, Sumijan[2], Musli Yanto[3]
[1,2,3] Master of Informatics Engineering, Faculty of Computer Science, Universitas Putra Indonesia YPTK, Padang, Padang, 25221, Indonesia
dinaselvia2610@gmail.com

Abstract

The development of social media, particularly Twitter, has become a primary means for the public to express opinions, criticisms, and complaints regarding train services, ranging from delays, facility comfort, to ticket policies. The large number of opinions appearing in short, non-standard characters, and containing slang and emoticons makes manual analysis ineffective, resulting in service providers not optimally utilizing valuable information from the public. This study aims to analyze public opinion sentiment on Twitter regarding train services to systematically and structuredly determine public perceptions. The methods used in this study are K-Means Clustering and Support Vector Machine (SVM). K-Means is used to group public opinion based on similarities in language patterns and sentiments to obtain initial labels, while SVM is used to classify opinions into positive and negative sentiments more accurately. The research data comes from the Twitter platform and is obtained through a crawling technique. The maximum limit of tweets retrieved is set at 2005 tweets. The results show that the K-Means method is able to assist the initial labeling process of sentiment data, while the SVM algorithm can classify public opinion with an accuracy level of 99.02%. The combination of clustering and classification methods has proven effective in processing large-scale, unstructured opinion data. Based on the research results, it can be concluded that the sentiment analysis approach using K-Means and Support Vector Machines can provide an objective picture of public perception of train service quality. The results of this analysis are expected to be used by service providers as evaluation material and a basis for decision-making to improve service quality to the public.

Keywords: Sentiment Analysis, Public Opinion, Twitter, K-Means Clustering, Support Vector Machine (SVM).

1. Introduction

The rapid development of digital technology has impacted communications, industry, and public services [1]. Digital innovation simplifies access to information, speeds up work, and improves service efficiency [2]. Trains have become a popular mode of land transportation in Indonesia due to their large capacity, affordable prices, and extensive reach [3]. As mobility increases, train services need to leverage technology to more effectively understand customer needs and satisfy customers [4].

The development of digital technology has made social media a primary means for the public to express opinions, complaints, and appreciation for public services [5]. Social media also serves as a fast, broad platform for public aspirations, potentially influencing public perception [6]. Twitter has become a popular platform for real-time opinion sharing and is often used as a benchmark for assessing public perception of services, including railways [7]. The sheer number of opinions posted daily makes Twitter a rich and diverse data source for analyzing public satisfaction and perception satisfaction with and perceptions of train services [8].

The facts show that numerous reviews, criticisms, and complaints from train users circulate on Twitter, ranging from delays and facility comfort to ticket policies. However, the large and diverse number of reviews complicates manual analysis, leading to underutilization of valuable information. Opinions on Twitter are typically brief, informal, and often use slang or emoticons. They also change easily with policies or service disruptions, further complicating analysis. The diversity of positive and negative perceptions also complicates opinion grouping, indicating that service providers lack an effective mechanism for processing public data. Therefore, a sentiment analysis approach that can systematically process opinions on a large scale is needed.

The solution presented to overcome this problem is sentiment analysis to systematically understand public opinion from data widely distributed on social media. Sentiment analysis is a collection of random and diverse reviews, criticisms, and appreciations that can be mapped into certain categories such as positive or negative. The results of this mapping provide an objective picture of public perception of train services, so that providers can identify aspects that need to be improved or maintained. Research related to sentiment

Submitted: 18-01-2026 | Revised: 20-02-2026 | Accepted: 12-03-2026 | Published: 30-03-2026

analysis has been conducted by [9] This study proves that text mining-based clustering techniques can be used to monitor and manage negative content on social media platforms. Further research was conducted by [10] the results of this study confirmed that aspect-based sentiment analysis with SVM can provide a more in-depth picture of satisfaction and complaints of KAI Access application users.

The development of artificial intelligence provides the ability to analyze sentiment more accurately, quickly, and efficiently [11]. Artificial intelligence is a branch of computer science that can be utilized in various fields, including analyzing public opinion on social media [12]. One artificial intelligence technology utilized in analyzing public opinion on social media is machine learning [13]. In the context of public sentiment analysis regarding opinion-based train services on Twitter, machine learning plays a crucial role in processing large amounts of unstructured text data [14].

A widely used method in machine learning is clustering, a technique in unsupervised learning that aims to group data into several groups based on the level of similarity or proximity of characteristics [15]. After the public opinion is grouped, the next step is the classification process. The classification algorithm used in this study is the Support Vector Machine [16]. In this study, SVM functions to classify public opinion on Twitter regarding train services into positive and negative sentiment categories with high accuracy. This algorithm learns the best dividing boundary (hyperplane) between each sentiment category, so it can classify public opinion on Twitter regarding train services accurately, quickly, and efficiently.

Research using a combination of K-Means and SVM methods was conducted by [17] The combination of K-Means Clustering and Support Vector Machine methods was able to provide fairly accurate classification results in identifying tweets related to mental health disorders. This method can be a reliable alternative for text analysis on social media, especially in detecting mental health issues automatically and quickly. Further research was conducted by [18] The combination of K-Means Clustering for topic grouping, Support Vector Machine for sentiment classification, and SMOTE to handle data imbalance, resulting in an effective and efficient sentiment analysis of public opinion on Twitter regarding Karapan Sapi. Based on relevant research, researchers are interested in analyzing public sentiment towards train services through Twitter user opinions with the aim of identifying positive and negative sentiments and providing input for improving service quality. The analysis was conducted using K-Means Clustering to group opinions based on feature similarities and Support Vector Machine to classify sentiments accurately, so that the results can be a basis for service managers in improving service quality and user satisfaction effectively.

2. Methods

This research methodology will explain in detail the framework used in this study. A research methodology is a series of stages that encompass the steps for processing and managing data, with the goal of producing high-quality and relevant research. It includes various stages that will be explained sequentially, starting with the needs analysis process, through data collection, and finally data processing. The purpose of this study is to analyze public sentiment toward train services based on opinions on Twitter using the K-Means Clustering and Support Vector Machine methods. The research framework can be seen in Figure 1.

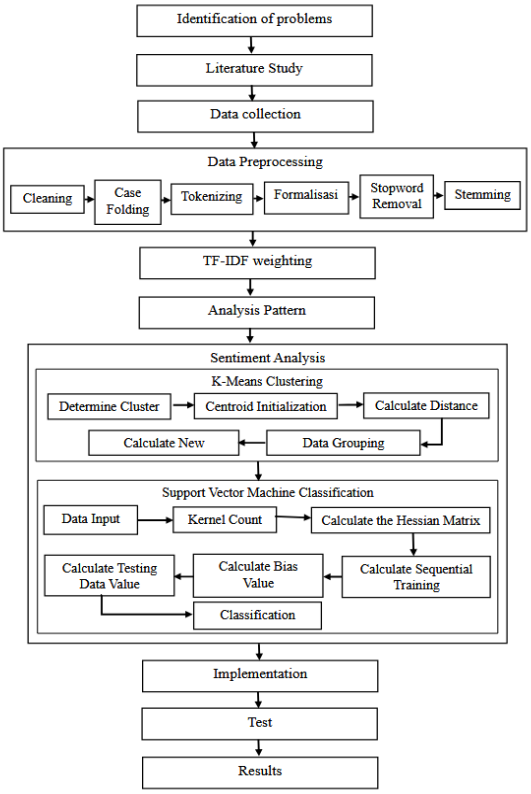


Figure 1. Research Framework

Based on the framework in Figure 1 the research procedure begins with problem identification to understand public opinion regarding train services. The next stage is a literature study to obtain a theoretical basis and relevant methods. Next, data collection and preprocessing are carried out. Next, text data is converted into a weighted numeric representation using TF-IDF so that important words in opinions can be identified more accurately. Next, an analysis pattern is formed as a bridge between the TF-IDF weighted data and the application of the K-Means and SVM algorithms. Data analysis is carried out using K-Means to group opinions based on feature similarities, then the clustering results are input for SVM to build a sentiment classification model. The final stage is model testing using accuracy, precision, recall, and F1-score metrics

to evaluate the model's performance and reliability in predicting public sentiment towards train services.

2.1 K-Means Clustering Method

K-Means Clustering is a data mining method that falls under the unsupervised learning category, a machine learning technique that does not use labeled data [19]. This method is used to group data into several groups or clusters based on the similarity of attributes between the data. Similarity is measured using a distance metric, typically the Euclidean distance, so that data with similar characteristics or attribute values will be grouped into the same cluster [20].

The K-Means algorithm works through an iterative process that involves forming clusters based on the centroid positions of each cluster. This process is repeated until the centroid positions no longer change or converge [21]. The general steps of the K-Means algorithm are as follows [22]:

1. Determining the number of clusters (k): The initial step in the K-Means method is determining the value of k, which is the number of clusters to be formed. This value forms the basis for the data clustering process.
2. Determining the Initial Cluster Center (k-Centroid): The next step is to determine the initial cluster center, or centroid. The selection of this starting point is crucial because it will affect the final results of the clustering process.
3. Calculating the Distance of Each Data Item to Each Centroid: Once the cluster center point is determined, the distance of each data item to the centroid is calculated. This calculation uses the Euclidean Distance, a formula for measuring the distance between objects and the cluster center, as shown in equation 2.1:

$$D_{ij} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

Description:

D (X, Y) = Euclidean distance between data X and Y

xi = value of the i-th variable in data X

yi = value of the i-th variable at the centroid of Y

n = number of variables

4. Data Grouping: Based on the distance calculation results, each data item is then grouped into a cluster based on the closest distance between the data and the centroid.
5. Determining a New Centroid: The new centroid is calculated based on the average value of the data within the same cluster. This process is repeated until there are no further changes in cluster position. This way, data with similar characteristics will be in the same cluster, while data with different characteristics will be separated into different clusters. The new centroid can be determined using equation 2.2:

$$D_{ij} = \frac{x1+x2+x3+\cdots+xn}{\sum x} \quad (2)$$

Description:

Xn = value of the nth record

∑x = number of data records

2.2 Support Vector Machine Method

Support Vector Machine is a reliable method for solving data classification problems. Support Vector Machine (SVM) is also known as the most advanced machine learning technique compared to other machine learning techniques [23]. Support vector machines are capable of recognizing and successfully used in pattern and characteristic recognition. Support vector machines have the ability to handle complex non-linear models accurately.

Support Vector Machine (SVM) is a method in supervised learning that is usually used for classification (Support Vector Classification) and regression (Support Vector Regression). Support Vector Machine can be used for prediction and classification by finding the hyperplane with the maximum margin among an infinite number of hyperplanes, then determining which hyperplane is best. So Support Vector Machine has the basic concept of finding a separating function (hyperplane) that can optimally separate two classes [24]. Hyperplanes with a larger margin are more accurate in classifying data than those with a smaller margin. This is known as the Maximum Hyperplane. Small margins and large margins can be seen in Figure 2.

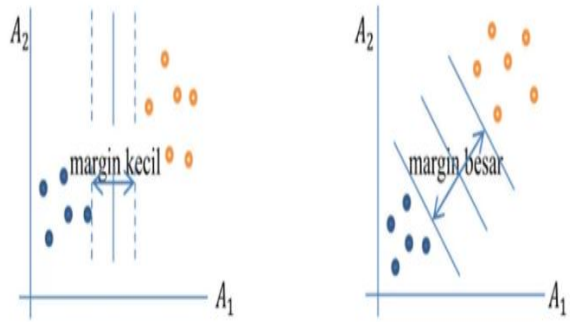


Figure 2. Small Margin and Large Margin

The margin is twice the distance between the hyperplane and the support vector, where the support vector is the point closest to the hyperplane. Support vectors can also be said to be the outermost data objects [25]. This support vector will be calculated by the SVM to find the most optimal hyperplane while other data objects are not taken into account at all, thus the Support

Vector Machine can work more efficiently. The concept of a hyperplane can be seen in Figure 3.

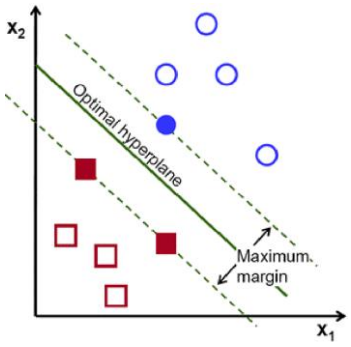


Figure 3. Hyperplane Concept in SVM

Margin is also defined by the shortest distance from the hyperplane to one side of the margin is equal to the shortest distance from the hyperplane to the other side of the margin, where the "side" of the margin is parallel to the hyperplane [26] . The Support Vector Machine (SVM) method in this study begins with kernel calculations using a linear kernel to map data to the feature space and form a Hessian matrix as the basis for the optimization process. Next, a sequential training process is carried out by initializing the parameters α, γ, and C, then calculating the error value (Ei), delta alpha (δα), and updating the alpha value until the optimal parameters are obtained. After the training process is complete, the bias value (b) is calculated based on the data with the highest alpha value from the positive and negative classes to determine the position of the separating hyperplane. The final stage is testing the testing data by inputting the data into the SVM decision function to produce a sentiment class prediction based on the sign of the function f(x).

3. Results and Discussions

This results and discussion section presents an analysis of the application of the K-Means Clustering and Support Vector Machine (SVM) methods in classifying public sentiment towards train services based on Twitter data. Data processing results are presented to demonstrate the model's performance in grouping and classifying public opinion into predetermined sentiment categories. The discussion then focuses on interpreting the classification results and their relevance to the research objective of analyzing public sentiment towards train services.

3.1 Dataset

The data used in this study was sourced from Platform X and obtained through crawling techniques. Data collection was automated using a Python-based script running on Google Collab, enabling efficient data collection of large amounts. For more details see Figure 4.

	created_at	full_text	id_str	username
0	Fri Oct 03 11:26:16 +0000 2025	Teori sotoy wa soal kenapa industrialisasi man...	1.974074e+18	@mflxop55
1	Fri Oct 03 11:17:10 +0000 2025	@KAI121 @budakpsikologi Min mau tanya min @KAI...	1.974071e+18	@KAI121
2	Fri Oct 03 11:13:48 +0000 2025	Halo @KAI121 Sekarang gabisa lagi beli railfoo...	1.974070e+18	@KAI121
3	Fri Oct 03 10:55:07 +0000 2025	4 hari berturut-turut kereta api Jenggala terl...	1.974066e+18	@CommuterLine
4	Fri Oct 03 10:51:44 +0000 2025	ignatius jonan juga investment banker jadi ben...	1.974065e+18	@wijoqd67
...	...	...	...	...
2000	Thu Sep 04 03:06:27 +0000 2025	Dari Madiun untuk dunia! PT Industri Kereta Ap...	1.963438e+18	@eyhll64
2001	Thu Sep 04 02:53:17 +0000 2025	#Lisbon Sebuah kereta api funikular terkenal d...	1.963435e+18	@dcssjn34
2002	Thu Sep 04 02:35:33 +0000 2025	KAI menawarkan promo diskon tiket kereta api s...	1.963431e+18	@ubvgno43
2003	Thu Sep 04 02:35:28 +0000 2025	Laporan situasi arus lalu lintas pagi di seput...	1.963431e+18	@NTMCLantasPolri
2004	Thu Sep 04 02:32:44 +0000 2025	Memasuki September 2025 PT Kereta Api Indonesi...	1.963430e+18	@xspwkwd41

Figure 4. Twitter Comment Data Display Results

The image above shows the process of retrieving X data from Twitter using the tweet-harvest command version 2.6.1. The retrieved data focuses on tweets containing the keyword "Kereta Api" in Indonesian. The search results will be saved in a CSV file named "Kereta Api.csv". The maximum limit of retrieved tweets is set at 2005 tweets. This command uses a Twitter authentication token to legally access the data.

3.2 Data Preprocessing

This subsection presents the outcome of the research activities. Results should be delivered in a clear, factual, and objective manner, without subjective interpretation. Visual aids such as tables, graphs, screenshots, and performance charts should be used to reinforce the data presented. All figures and tables must be numbered, captioned, and referenced in the text Data preprocessing is a crucial part of the sentiment analysis process. This process ensures that the data can be structured and then used in sentiment analysis. As described in the research methodology, the data preprocessing process includes cleaning, case folding, tokenizing, formalization, stopword removal, and stemming.

cleaning	case_folding	tokenizing	Formalisasi	Stopwords Removal	Stemming
hari berturutturut kereta api Jenggala terla...	hari berturutturut kereta api jenggala terla...	['hari', 'berturutturut', 'kereta', 'api', 'je...	['hari', 'berturutturut', 'kereta', 'api', 'je...	['berturutturut', 'kereta', 'api', 'jenggala',...	['berturutturut', 'kereta', 'api', 'jenggala',...
ignatius jonan juga investment banker jadi be...	ignatius jonan juga investment banker jadi be...	['ignatius', 'jonan', 'juga', 'investment', 'b...	['ignatius', 'jonan', 'juga', 'investasi', 'ba...	['ignatius', 'jonan', 'investasi', 'banker', '...	['ignatius', 'jonan', 'investasi', 'banker', '...
naik kereta api cute cute cute gemes bangett...	naik kereta api cute cute cute gemes bangett...	['naik', 'kereta', 'api', 'cute', 'cute', 'cut...	['naik', 'kereta', 'api', 'imut', 'imut', 'imu...	['kereta', 'api', 'imut', 'imut', 'imut', 'gem...	['kereta', 'api', 'imut', 'imut', 'imut', 'gem...
Gue banget mending naik kereta api bisa tidu...	gue banget mending naik kereta api bisa tidu...	['gue', 'banget', 'mending', 'naik', 'kereta',...	['saya', 'banget', 'mending', 'naik', 'kereta'...	['banget', 'mending', 'kereta', 'api', 'tidur'...	['banget', 'mending', 'kereta', 'api', 'tidur'...
naik kereta dari jaman gak ada nomer kursi j...	naik kereta dari jaman gak ada nomer kursi j...	['naik', 'kereta', 'dari', 'jaman', 'gak', 'ad...	['naik', 'kereta', 'dari', 'jaman', 'tidak', '...	['kereta', 'jaman', 'nomor', 'kursi', 'rapi', ...	['kereta', 'jaman', 'nomor', 'kursi', 'rapi', ...

Figure 5. Data Preprocessing Results

Figure 5 shows the preprocessing stages of Twitter opinion text data about train services, which were carried out in stages. The process begins with the

original text (full text), followed by cleaning and case folding to remove irrelevant characters and standardize the letters. Next, the text is broken down into words through tokenization, standardized through formalization, and reduced to zero meaning through stopword removal. The final stage is stemming, which converts words to their basic form so that the data is ready for further analysis using K-Means Clustering and Support Vector Machine methods.

From a total of 2005 tweet data obtained, a data cleaning stage was carried out including removing duplications, removing punctuation, and filtering based on research variables, namely service quality, punctuality, and ticket prices so that the dataset became 509 data.

3.3 TF-IDF weighting

The next process in this research is to weight each word (term) contained in each document or comment using the Term Frequency–Inverse Document Frequency (TF-IDF) method. TF-IDF calculations are carried out based on the research variable categories, namely X1 (Service Quality), X2 (Punctuality), and X3 (Ticket Price). The Term Frequency (TF) value is calculated based on the frequency of occurrence of a word in a document, where the word that appears is symbolized by the value 1 and the word that does not appear is symbolized by the value 0. Furthermore, the Document Frequency (DF) value is obtained by calculating the number of documents that contain the word in at least one document. To overcome differences in document length and reduce frequency value anomalies, a normalization process is carried out on the TF value. The final stage of word weighting is carried out by calculating the TF-IDF value, namely by multiplying the normalized TF value by the Inverse Document Frequency (IDF) value.

The TF-IDF weighting results show the words that have the strongest contribution in describing the contents of the document in that category. The resulting weight pattern shows terms that are truly relevant while distinguishing between categories. These results serve as an important basis for identifying analysis patterns used in the next modeling stage.The TF-IDF weighting results show the words that have the strongest contribution in describing the contents of the document in that category. The resulting weight pattern shows terms that are truly relevant while distinguishing between categories. These results serve as an important basis for identifying analysis patterns used in the next modeling stage. The following are the results of the analysis pattern:

Table 1. Analysis Pattern

Document	Service Quality (X1)	Timeliness (X2)	Ticket Price (X3)
D1	0.825287	0.825287	1.428567
D2	1.064220	1.064220	0.000000
D3	0.242971	0.242971	0.000000
D4	1.088827	1.088827	0.000000
D5	1.154553	1.154553	0.000000
D6	1.597976	1.597976	0.000000
D7	0.763190	0.763190	0.000000
D8	1.064220	1.064220	0.000000
D9	1.311305	1.311305	0.000000
D10	1.005129	1.005129	1.538041
…	…	…	…
D500	1.817642	1.817642	1.405687
D501	1.905688	1.905688	1.512018
D502	1.909933	1.909933	1.651887
D503	1.803485	1.803485	0.000000
D504	1.286903	1.286903	1.786577
D505	1.966771	1.966771	1.752475
D506	1.408337	1.408337	1.803628
D507	1.734932	1.734932	0.706718
D508	1.437493	1.437493	1.335748
D509	0.964206	0.964206	0.000000

With this grouping, each word that has been weighted using TF-IDF is assigned to the appropriate variable. This step makes it easier for researchers to see the pattern of keyword occurrences in each variable, allowing for further analysis in the clustering and classification process using the K-Means and Support Vector Machine (SVM) methods. The results of this stage provide an initial overview of the focus of user comments on each aspect of the research, as well as being the basis for conducting sentiment analysis and evaluating service performance.

3.4 Sentiment Analysis

As described in the previous analysis and design flowchart, the sentiment analysis process in this study was conducted using a combination of the K-Means Clustering and Support Vector Machine (SVM) methods. Public opinions were first grouped using K-Means Clustering based on feature similarities, while also providing initial labeling. Next, a Support Vector Machine was applied to classify sentiment into positive or negative categories.

The first stage in sentiment analysis in this study is the application of the K-Means clustering algorithm to group public opinion data based on the level of similarity of their characteristics. In this study, the number of clusters was set at k = 2, representing positive and negative sentiment. The K-Means process groups the data into two main clusters based on the similarity of the analysis patterns. The centroid values of each cluster were analyzed to determine the sentiment tendencies formed, as presented in Table 2.

Table 2. K-Means Clustering Results

Document	Service Quality (X1)	Timeliness (X2)	Ticket Price (X3)	Cluster	Sentiment
D1	0.8253	0.8253	1.4286	1	Negative
D2	1.0642	1.0642	0.0000	1	Negative
D3	0.2430	0.2430	0.0000	1	Negative
D4	1.0888	1.0888	0.0000	1	Negative
D5	1.1546	1.1546	0.0000	0	Positive
D6	1.5980	1.5980	0.0000	0	Positive
D7	0.7632	0.7632	0.0000	1	Negative
D8	1.0642	1.0642	0.0000	1	Negative
D9	1.3113	1.3113	0.0000	0	Positive
D10	1.0051	1.0051	1.5380	1	Negative
…	…	…	…	…	…
D500	1.8176	1.8176	1.4057	0	Positive
D501	1.9057	1.9057	1.5120	0	Positive
D502	1.9099	1.9099	1.6519	0	Positive
D503	1.8035	1.8035	0.0000	0	Positive
D504	1.2869	1.2869	1.7866	0	Positive
D505	1.9668	1.9668	1.7525	0	Positive
D506	1.4083	1.4083	1.8036	0	Positive
D507	1.7349	1.7349	0.7067	0	Positive
D508	1.4375	1.4375	1.3357	0	Positive
D509	0.9642	0.9642	0.0000	1	Negative

Table 2. shows the results of calculating the distance between each public opinion tweet and two centroids. The sixteenth iteration resulted in the data being divided into two groups, each with 283 and 226 data items. The K-Means Clustering algorithm successfully grouped the data into two clusters with different sentiment characteristics.

1. Cluster 0 contains 283 data points with positive sentiment.
2. Cluster 1 contains 226 data points with negative sentiment.

The iteration process stopped at the sixteenth iteration, where no data transfer between clusters indicated that the centroid position was stable and the K-Means Clustering algorithm had reached convergence. This indicates that the clustering results are optimal and can be used as a basis for further analysis using the Support Vector Machine (SVM) method. These clustering results provide a clear picture of the distribution of public opinion on train services that can be utilized by PT Kereta Api Indonesia (KAI), the Ministry of Transportation, and other transportation policymakers to evaluate and improve the quality of train services based on user perceptions on social media.

After the tweet data was grouped using the K-Means Clustering method, the next step was to apply the Support Vector Machine (SVM) algorithm to build a sentiment classification model. The SVM algorithm was chosen for its ability to optimally separate data into sentiment classes. The classification model was trained using training data and then validated using testing data. This process aimed to evaluate the model's performance in accurately classifying sentiment.

Table 3. SVM Prediction Results on Sample Test Data

	U1	U2	U3
D1	0,3728	0,6268	0,0000
D2	0,3352	0,5636	0,6990
D3	0,1551	0,2607	0,1002
D4	0,1551	0,2607	0,9494
D5	0,4060	0,6826	0,0000
D6	0,8456	0,0371	1,5324
D7	1,0843	0,7123	0,4271
f(x)	-0,5205	1,2431	-0,2046
Prediksi	Negative	Positive	Negative

Based on the calculation results of the test data values on the sample data, each test data was successfully mapped into the appropriate sentiment class using the decision function (f(x)) of the Support Vector Machine. The first and third test data were categorized as negative sentiment, while the second test data included positive sentiment. These results indicate that the SVM model is able to accurately distinguish the test data according to the patterns learned from the training data.

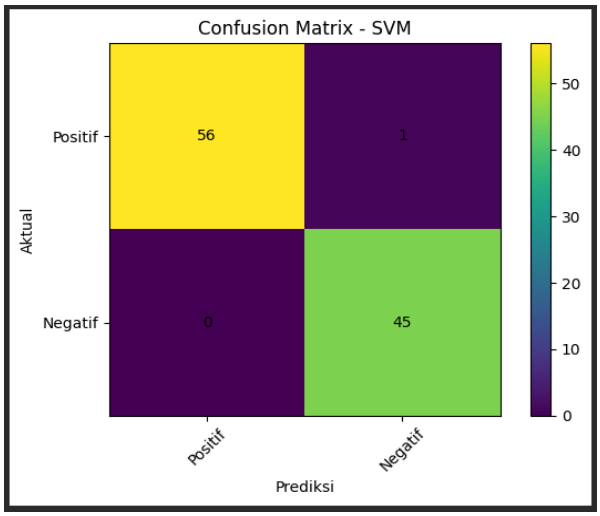
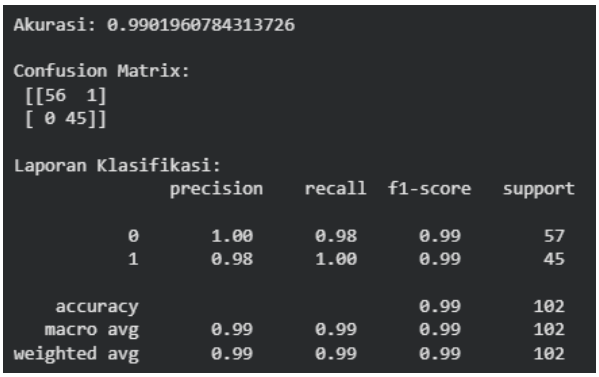


Figure 6. ConfusionMatrix

Figure 7. shows the confusion matrix of the Support Vector Machine (SVM) model on the test data, which compares the predicted labels with the original labels for each sentiment category, namely Positive and Negative. This matrix provides information on how accurately the model classifies the test data according to its respective sentiment category. The results obtained from this matrix help evaluate the overall performance of the SVM model. Furthermore, confusion matrix analysis also allows for the identification of potential misclassifications within each category.



```
Akurasi: 0.9901960784313726

Confusion Matrix:
 [[56  1]
 [ 0 45]]

Laporan Klasifikasi:
              precision    recall  f1-score   support

           0       1.00      0.98      0.99        57
           1       0.98      1.00      0.99        45

    accuracy                           0.99       102
   macro avg       0.99      0.99      0.99       102
weighted avg       0.99      0.99      0.99       102
```

Figure 8. SVM Model Accuracy Performance

Based on the test results, the Support Vector Machine (SVM) model achieved an accuracy of 99.02%, indicating excellent classification performance across the entire data set. The confusion matrix shows that most of the data was correctly classified, with a very low error rate. Precision, recall, and F1-score values, each approaching 1.00 for both classes, indicate that the model is able to distinguish sentiment consistently and equally. These results confirm that the SVM model has a high level of reliability in classifying public opinion sentiment towards train services.

4. Conclusions

This study demonstrates that sentiment analysis of train services can be effectively performed using a combination of K-Means Clustering and Support Vector Machine (SVM) methods based on Twitter data. The developed classification model yielded an overall accuracy rate of 99.02%, demonstrating the model's excellent ability to classify public opinion sentiment. The use of K-Means as the initial clustering process and SVM as a classifier provides an appropriate alternative solution for processing large, unstructured text data. The results of this study contribute to providing structured and objective public sentiment information as a basis for evaluating the quality of train services. Further research can be developed by comparing the performance of other methods, optimizing model parameters, and utilizing data from various social media platforms to obtain more comprehensive results.

References

[1] R. Faris Triana, A. Irma Purnama Sari, A. Bahtiar, and E. Wahyudin, "Implementasi Algoritma Naïve Bayes Untuk Klasifikasi Sentimen Ulasan Pengguna KAI Access," *Jutisi: Jurnal Ilmiah Teknik Informatika dan Sistem Informasi*, 2025.

[2] F. Duta Sanubari, U. Enri, and U. Singaperbangsa Karawang Abstract, "Analisis Sentimen Terhadap Perubahan Rute Krl Commuter Jabodetabek Menggunakan Algoritme Support Vector Machine (Svm)," *Jurnal Ilmiah Wahana Pendidikan*, vol. 2023, no. 15, pp. 155–163, 2023, doi: 10.5281/zenodo.8206986.

[3] A. Wirayudha, M. Murniyati, and R. Rosdiana, "Analisis Sentimen Terhadap Ulasan Access By KAI Pada Google Play Store Menggunakan Metode Indobert," *Portal Riset dan Inovasi Sistem Perangkat Lunak*, vol. 3, no. 1, pp. 9–20, Jan. 2025, doi: 10.59696/prinsip.v3i1.69.

[4] M. D. Pamungkas and H. Februariyanti, "PENERAPAN ALGORITMA K-MEANS CLUSTERING UNTUK MENGELOMPOKAN DATA REVIEW BARANG PADA E-COMMERCE LAZADA," *semanTIK*, vol. 8, no. 2, p. 99, Dec. 2022, doi: 10.55679/semantik.v8i2.29058.

[5] Y. Khoiruddin, A. Fauzi, and A. M. Siregar, "Analisis Sentimen Gojek Indonesia Pada Twitter Menggunakan Algoritme Naïve Bayes Dan Support Vector Machine," *Jurnal Ilmiah Komputer*, 2023.

[6] Antonius Yadi Kuntoro, Hermanto, and Taufik Asra, "KLASIFIKASI KELUHAN PENGGUNA KAI ACCESS UNTUK PEMESANAN TIKET DENGAN ALGORITMA SVM DAN NAÏVE BAYES," *JIKA (Jurnal Informatika) Universitas Muhammadiyah Tangerang*, 2022.

[7] R. Reynaldi, R. J. Faiz Djarot, M. Wahyudi, S. Sumanto, and A. S. Budiman, "Analisa Pola Penyebaran Pengguna Layanan Transjakarta dengan Metode K-Means Clustering," *Journal Software, Hardware and Information Technology*, vol. 5, no. 2, pp. 128–138, Jun. 2025, doi: 10.24252/shift.v5i2.205.

[8] K. N. F. Ariyanti and A. Susanti, "Identifikasi Pengguna Aplikasi Transportasi Access by KAI dengan Ulasan dan Rating Menggunakan Analisis Sentimen," *Mitrans: Jurnal Media Publikasi Terapan Transportas*, vol. 2, 2024.

[9] M. Refa *et al.*, "Analisis Sentimen Terhadap Komentar Negatif (Hate Speech) Di Twitter Dengan Algoritma K-Means Clustering Menggunakan Rapidminer," *Journal of Information Technology and Informatics Engineering*, vol. 1, no. 1, pp. 57–61, 2025.

[10] G. Radiena and A. Nugroho, "ANALISIS SENTIMEN BERBASIS ASPEK PADA ULASAN APLIKASI KAI ACCESS MENGGUNAKAN METODE SUPPORT VECTOR MACHINE," 2023.

[11] A. Octaviani and P. Dewi, "Kecerdasan Buatan sebagai Konsep Baru pada Perpustakaan," *ANUVA*, vol. 4, no. 4, pp. 453–460, 2021.

[12] O. H. Rahman, G. Abdillah, and A. Komarudin, "Klasifikasi Ujaran Kebencian pada Media Sosial Twitter Menggunakan Support Vector Machine," *Jurnal RESTI*, vol. 5, no. 1, pp. 17–23, Feb. 2021, doi: 10.29207/resti.v5i1.2700.

[13] E. S. Eriana, S. Kom, M. Kom, and D. A. Zein, *ARTIFICIAL INTELLIGENCE (AI) PENERBIT CV. EUREKA MEDIA AKSARA*. Purbalingga: PENERBIT CV.EUREKA MEDIA AKSARA, 2023.

[14] M. Indah Ramadhani, "IJIRSE: Indonesian Journal of Informatic Research and Software Engineering Implementation Of K-Means Algorithm For Palm Oil Productivity Data Clustering Implementasi Algoritma K-Means Untuk Klustering Data Produktivitas Kelapa Sawit," *IJIRSE: Indonesian Journal of Informatic Research and Software Engineering*, vol. 3, no. 1, pp. 56–64, Mar. 2023, Accessed: Feb. 25, 2025. [Online]. Available: Journal Homepage: https://journal.irpi.or.id/index.php/ijirse

[15] M. F. Shahzad, S. Xu, W. M. Lim, X. Yang, and Q. R. Khan, "Artificial intelligence and social media on academic performance and mental well-being: Student perceptions of positive impact in the age of smart learning," *Heliyon*, vol. 10, no. 8, Apr. 2024, doi: 10.1016/j.heliyon.2024.e29523.

[16] Y. Sibaroni and S. S. Prasetiyowati, "Buzzer Detection on Indonesian Twitter using SVM and Account Property Feature Extension," *Jurnal RESTI*, vol. 6, no. 4, pp. 663–669, Aug. 2022, doi: 10.29207/resti.v6i4.4338.

[17] K. Kusumaningtyas *et al.*, "Tweet Analysis of Mental Illness Using K-Means Clustering and Support Vector Machine Analisis Tweet Gangguan Kesehatan Mental Menggunakan K-Means Clustering dan Support Vector Machine," *Jurnal Informatika dan Teknologi Informasi*, vol. 20, no. 3, pp. 295–308, 2023, doi: 10.31515/telematika.v20i3.9820.

[18] Z. Rani and B. K. Khotimah, "ANALISIS SENTIMEN TERHADAP KARAPAN SAPI DI TWITTER MENGGUNAKAN METODE K-MEANS DAN SUPPORT VECTOR MACHINE (SVM)," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 13, no. 1, Jan. 2025, doi: 10.23960/jitet.v13i1.5685.

[19] A. A. Arrosyad, A. I. Purnamasari, and I. Ali, "IMPLEMENTASI ALGORITMA K-MEANS CLUSTERING UNTUK ANALISIS PERSEBARAN UMKM DI JAWA BARAT," 2024.

[20] S. Trisya Amanda, K. Berlianindita, A. Jahid Alfarizi, W. A. Arifin, L. Kelautan, and K. Daerah Serang, "ANALISIS KOMPARATIF SVM DAN K-MEANS DALAM DATA MINING UNTUK PROMOSI PERGURUAN TINGGI," *JURNAL REKAYASA INFORMASI SWADHARMA (JRIS)*, vol. 05, Jul. 2025.

[21] T. M. Ghazal *et al.*, "Performances of k-means clustering algorithm with different distance metrics," *Intelligent Automation and Soft Computing*, vol. 30, no. 2, pp. 735–742, 2021, doi: 10.32604/iasc.2021.019067.

[22] K. Kusumaningtyas *et al.*, "Tweet Analysis of Mental Illness Using K-Means Clustering and Support Vector Machine Analisis Tweet Gangguan Kesehatan Mental Menggunakan K-Means Clustering dan Support Vector Machine," *Jurnal Informatika dan Teknologi Informasi*, vol. 20, no. 3, pp. 295–308, 2023, doi: 10.31515/telematika.v20i3.9820.

[23] A. Z. Fuadi, I. N. Haq, and E. Leksono, "Support Vector Machine to Predict Electricity Consumption in the Energy Management Laboratory," *Jurnal RESTI*, vol. 5, no. 3, pp. 466–473, Jun. 2021, doi: 10.29207/resti.v5i3.2947.

[24] Styawati, A. Nurkholis, Z. Abidin, and H. Sulistiani, "Optimasi Parameter Support Vector Machine Berbasis Algoritma Firefly Pada Data Opini Film," *Jurnal RESTI*, vol. 5, no. 5, pp. 904–910, Oct. 2021, doi: 10.29207/resti.v5i5.3380.

[25] I. Salsabila and Y. Sibaroni, "Multi Aspect Sentiment of Beauty Product Reviews using SVM and Semantic Similarity," *Jurnal RESTI*, vol. 5, no. 3, pp. 520–526, Jun. 2021, doi: 10.29207/resti.v5i3.3078.

[26] M. Zalukhu and K. Kunci, "Analsis dan Implementasi Metode Naïve Bayes dan SVM Pada Sentimen Pemilihan Calon Presiden RI Info," *Jurnal Informatika Faatuatua Media Karya*, 2023, [Online]. Available: https://jurnal.faatuatua.com/index.php/KETIK

Author Biography

Dina Selvia

She was born in Pasaman Baru on October 26, 2001, is a graduate of the Bachelor of Education (S.Pd.) in the Informatics and Computer Engineering Education Study Program from the Sjech M. Djamil Djambek State Islamic University Bukittinggi, Indonesia, which was completed in 2024. Currently, she is continuing her Masters in Informatics Engineering (M.Kom) studies at UPI-YPTK Padang University starting in 2024. She has an interest in the field of information technology and education, and is known as a diligent, responsible person, and has a high enthusiasm for learning.

	**Sumijan** he is born in Nganjuk on May 7, 1966, is an academic, researcher, and educator who is active in the development of science, especially in the field of Information Technology. He serves as a Lecturer at Universitas Putra Indonesia "YPTK" Padang, with NIDN 0005076607 and can be contacted via email soe @upiyptk.org or HP/WhatsApp number 08126607355. Having a strong academic track record, he earned a Bachelor's degree in Informatics Management from Universitas Putra Indonesia YPTK, a Master's degree in Information Systems from University Technology Malaysia, and a Doctoral degree in Information Technology from Gunadarma University. As a productive researcher, Dr. Sumijan has Scopus ID 57194787076, ORCID ID 0000-0002-9932-4325, and has produced various reputable international publications. His expertise includes digital image processing, information technology auditing, computer vision, data mining, and information systems engineering. He has mentored thousands of students to graduation (7,558 undergraduate students and 445 postgraduate students) and is active in research, community service, and scientific seminars at both national and international levels. Known as a figure who consistently pursues knowledge and makes real contributions to the advancement of education, Dr. Sumijan adheres to the principle that mastery of knowledge requires hard work, dedication, and a willingness to always develop..
	**Musli Yanto** he is a lecturer at Universitas Putra Indonesia YPTK Padang. Educational background since 2014 has completed postgraduate studies in the field of informatics engineering. His skills include data science analysis, algorithms and programming, big data and artificial intelligence. He can be contacted at email: musli_yanto@upiyptk.ac.id.