Journal of Emerging Information Systems and Business Intelligence
ISSN: 2774-3993
Journal homepage: https://ejournal.unesa.ac.id/index.php/JEISBI/

Sentiment Analysis of Public Figures on X Using Naïve Bayes and SVM

M. Hashfiudin Tridharma Putra[1], Aries Dwi Indriyanti[2]
[1,2]*Universitas Negeri Surabaya, Surabaya, Indonesia*
m.hashfiudintridharma.19108@mhs.unesa.ac.id, ariesdwi@unesa.ac.id

ABSTRACT

The rapid growth of social media has created an open public space where users freely express opinions toward public figures, generating positive, negative, and neutral sentiments. Platform X (formerly Twitter) is one of the most widely used media for public discourse in Indonesia. This study analyzes public sentiment toward the Regent of Sidoarjo for the 2021–2024 period, Ahmad Muhdlor Ali, using sentiment classification techniques. The research applies two machine learning algorithms, namely the Naïve Bayes Classifier (NBC) and Support Vector Machine (SVM), to identify and compare their performance in sentiment analysis. Data were collected through web scraping using relevant keywords and processed in Google Colab. A quantitative research approach was employed using the SEMMA framework, which consists of Sample, Explore, Modify, Model, and Assess stages. The process included data cleaning, text preprocessing, sentiment labeling, and classification using both algorithms. Model performance was evaluated using accuracy, precision, and recall metrics. The results show that both NBC and SVM perform well in classifying public sentiment, achieving high accuracy levels. However, differences in performance were observed between the two methods, indicating that algorithm selection influences classification outcomes. This study contributes to the evaluation of public perception toward government officials and provides a reference for the development of sentiment analysis systems based on social media data.

Keyword: Sentiment Analysis, Public Figure, Naïve Bayes, SVM, Platform X.

Article Info:
Article history:
Received February 11, 2026
Revised February 18, 2026
Accepted April 28, 2026

Corresponding Author
M. Hashfiudin Tridharma Putra
Universitas Negeri Surabaya, Surabaya, Indonesia
m.hashfiudintridharma.19108@mhs.unesa.ac.id

1. INTRODUCTION

The increasing use of social media has significantly influenced how people communicate, share opinions, and evaluate public figures. Platforms such as X (formerly Twitter) enable users to express their views openly, resulting in large volumes of user-generated text that reflect public sentiment. These opinions can be positive, negative, or neutral and often influence public perception and policy evaluation.

Public figures, particularly government officials, are frequently discussed on social media. Public reactions toward policies, leadership performance, and political issues are increasingly visible through online platforms. Therefore, analyzing public sentiment on social media can provide valuable insights into societal responses to governance and leadership.

Sentiment analysis, a subfield of text mining and machine learning, has been widely used to classify opinions expressed in textual data. Among the commonly used classification algorithms are Naïve Bayes Classifier (NBC) and Support Vector Machine (SVM). NBC is known for its simplicity and efficiency, especially in text classification tasks, while SVM is recognized for its robustness in handling high-dimensional and complex data.

This study focuses on analyzing public sentiment toward the Regent of Sidoarjo for the 2021–2024 period, Ahmad Muhdlor Ali, using data from Platform X. The research aims to compare the performance of NBC and SVM in classifying sentiment and to provide insights into public perception of local government leadership.

2. METHODS

This study employs a quantitative research design utilizing the **SEMMA** (Sample, Explore, Modify, Model, Assess) framework, which provides a structured and scientifically rigorous methodology for data mining and predictive modeling (Sammut & Webb, 2017). The research is characterized as a comparative-descriptive study, specifically focusing on sentiment analysis of Indonesian-language discourse regarding the Regent of Sidoarjo on Platform X. Data collection was conducted through automated web scraping using Python-based tools to retrieve tweets from the 2021–2024 period, ensuring a longitudinal dataset for analysis. The data analysis method involved a multi-stage preprocessing pipeline—comprising cleaning, case folding, tokenization, stopword removal, stemming, and **TF-IDF** feature extraction—to transform unstructured text into a machine-readable format. For the modeling phase, the study implemented a comparative approach between two supervised learning algorithms: the **Naïve Bayes Classifier (NBC)**, based on probabilistic induction, and **Support Vector Machine (SVM)**, known for its effectiveness in high-dimensional text classification (Azevedo & Santos, 2008). Finally, the models were scientifically validated using an 80:20 data split and evaluated through a performance matrix including accuracy, precision, and recall to determine the most effective algorithm for public policy sentiment detection.

3. RESULTS AND DISCUSSION

3.1 Data Collection Results and Characteristics

The research data was obtained from Platform X (formerly Twitter) through web scraping techniques using Google Colab. Data collection was carried out using keywords related to the public figure of the Regent of Sidoarjo for the 2021–2024 term, Ahmad Muhdlor Ali. The data collected consisted of Indonesian-language tweets containing public opinion on policies, performance, and issues related to this figure.

The distribution of positive and negative sentiments can be seen in Table 1, which shows a comparison of the number of tweets based on sentiment categories.

Table 1 Distribution of Public Sentiment towards the Regent of Sidoarjo on Platform X

Label	Count
Positive	941
Negative	67

The data collection results showed that the number of tweets obtained was dominated by positive public opinion. Based on the keyword-based sentiment labeling process, 941 tweets were labeled as positive and 67 tweets were labeled as negative.

This distribution shows that, in general, public perception of the figure analyzed tended to be positive during the research period.

3.2 Text Data Preprocessing Results

Before classification, the text data underwent several preprocessing stages to ensure data quality and improve model accuracy. These stages included data cleaning, case folding, tokenization, stopword removal, and stemming.

The cleaning process removes URLs, symbols, punctuation marks, and irrelevant duplicate text. Next, all text is converted to lowercase (case folding) to maintain data consistency. Tokenization is performed to break sentences into words, followed by stopword removal to eliminate common words that have no sentiment meaning. The stemming stage converts words to their base form so that word variations do not affect the classification process. The preprocessing results in more structured text data that is ready for use in feature extraction using the TF-IDF method.

3.3 Sentiment Classification Results

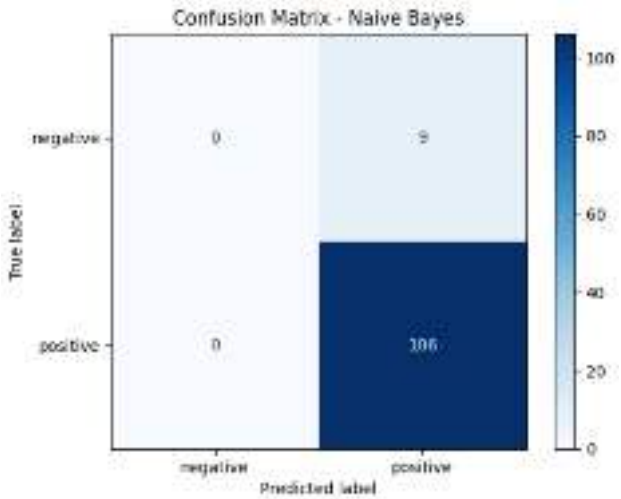
Sentiment classification is performed using two algorithms, namely Naïve Bayes Classifier (NBC) and Support Vector Machine (SVM). The dataset is divided into 80% training data and 20% test data. Feature representation uses the TF-IDF method to convert text into numerical form.

The classification results show that both algorithms are capable of classifying sentiment with a high degree of accuracy. However, there is a difference in performance between NBC and SVM. SVM shows more stable and accurate results in classifying sentiment data compared to NBC, especially in distinguishing data with complex sentiment characteristics.

The results of the model evaluation using a confusion matrix for each algorithm are shown in Table 2 and Table 3.

Table 2 Confusion Matrix for Sentiment Classification Using Naïve Bayes

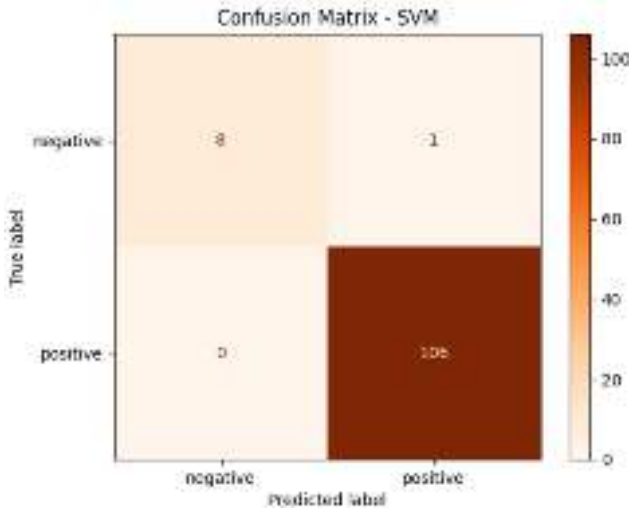
	Negative Prediction	Positive Prediction
Original Negative	0	9
Original Positive	0	106



Picture 1 Naïve Bayes Confusion Matrix Results

Table 3 Confusion Matrix for Sentiment Classification Using Support Vector Machine

	Negative Prediction	Positive Prediction
Original Negative	8	1
Original Positive	0	106



Picture 2 Confusion Matrix Support Vector Machine Result

3.4 Model Evaluation

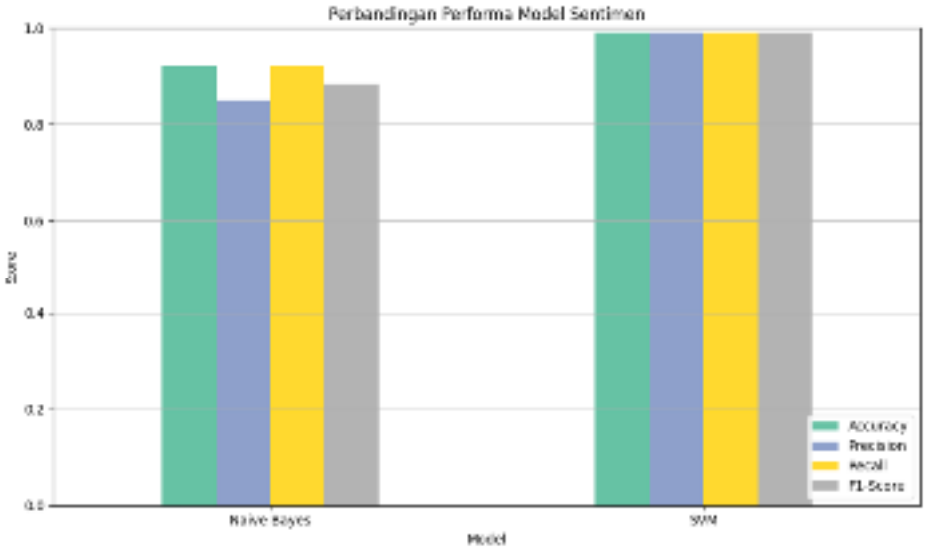
Model performance evaluation was carried out using accuracy, precision, and recall metrics. Based on the test results, both algorithms showed good performance in classifying public sentiment.

The Naïve Bayes Classifier has advantages in terms of speed and model simplicity, making it effective for classifying short texts with a large amount of data. However, SVM has better capabilities in separating sentiment classes that are not linearly separable and is more resistant to data noise.

A comparison of the accuracy, precision, and recall values between NBC and SVM can be seen in Picture 3, which shows that SVM has higher evaluation values overall.

Table 4 Comparison of Naïve Bayes and SVM Accuracy

Classification Method	Accuracy (%)
Naïve Bayes	92,08
Support Vector Machine (SVM)	99,50



Picture 3 Comparison of Accuracy, Precision, and Recall between NBC and SVM

3.5 Discussion

The results of this study indicate that sentiment analysis using social media data can provide a clear picture of public perception of public figures. The dominance of positive sentiment shows that the public tends to respond favorably to the leadership of the Regent of Sidoarjo during the study period, despite fluctuations in sentiment influenced by certain issues.

The difference in performance between NBC and SVM is in line with previous studies which state that SVM is superior in handling high-dimensional and complex text data. NBC remains an efficient alternative for rapid classification, but has limitations in handling the dependence between word features.

Thus, the choice of algorithm greatly affects the results of sentiment analysis, especially when used for social media-based public opinion evaluation.

CONCLUSION

This study analyzed public sentiment toward the Regent of Sidoarjo on Platform X using Naïve Bayes Classifier and Support Vector Machine. The results show that both algorithms perform well in sentiment classification, with SVM achieving better overall performance. The study demonstrates the effectiveness of machine learning techniques in analyzing public opinion on social media and provides insights into public perception of government leadership. Future research may include additional sentiment categories, larger datasets, and advanced deep learning methods.

REFERENCES

[1] N. F. Amin, S. Garancang, and K. Abunawas, "Konsep Umum Populasi dan Sampel dalam Penelitian," *Jurnal PILAR*, vol. 14, no. 1, pp. 15–31, 2023.

[2] I. Cholissodin and A. Soebroto, *Buku Ajar AI, Machine Learning & Deep Learning*, 2019.

[3] R. Febriyanti and A. Mustofa, "Analisis Sentimen Terhadap Tokoh Publik Gus Dur Menggunakan Metode Naïve Bayes dan Support Vector Machine (SVM)," *Jurnal Ekonomi, Ilmu Sosial, dan Bisnis Islam (JEISBI)*, vol. 4, no. 1, pp. 91–104, 2023.
[4] A. M. Husein, M. Harahap, and P. Fernandito, "Pendekatan data science untuk menemukan churn pelanggan pada sektor perbankan dengan machine learning," *Data Sciences Indonesia (DSI)*, vol. 1, no. 1, pp. 8-13, 2021.
[5] M. A. Iskandar and U. Latifa, "Website prediksi customer churn untuk mempertahankan pelanggan pada perusahaan telekomunikasi," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 7, no. 2, pp. 1308-1316, 2023.
[6] A. S. Kusumanegara and I. Rachmawati, "The effect of service quality and price on customer satisfaction and loyalty in Telkomsel cellular operator services," 2023.
[7] P. Lalwani *et al.*, "Customer churn prediction system: A machine learning approach," *Computing*, pp. 1-24, 2022.
[8] H. Rohaeni and W. Yuliyana, "Pengaruh harga dan kualitas produk terhadap loyalitas pelanggan Telkomsel," *Jurnal Sains Manajemen*, vol. 2, no. 1, pp. 37-44, 2020.
[9] S. N. Shukla and B. M. Marlin, "Interpolation-prediction networks for irregularly sampled time series," *arXiv preprint arXiv:1909.07782*, 2019.
[10] J. Sihombing, "Klasifikasi data antropometri individu menggunakan algoritma Naïve Bayes Classifier," *BIOS: Jurnal Teknologi Informasi dan Rekayasa Komputer*, vol. 2, no. 1, pp. 1-10, 2021.
[11] R. A. Siregar and A. Wibowo, "Analisis Kualitas Layanan Jaringan 4G LTE di Indonesia Menggunakan Metode Drive Test," *Jurnal Teknik Telekomunikasi*, vol. 15, no. 2, pp. 45–52, 2023.
[12] Sugiyono, *Metode Penelitian Kuantitatif, Kualitatif, dan R&D*, Alfabeta, 2019.
[13] S. Sugiyono and P. Lestari, *Metode Penelitian Komunikasi (Kuantitatif, Kualitatif, dan Cara Mudah Menulis Artikel pada Jurnal Internasional)*, 2021.
[14] P. C. Susanto, D. U. Arini, L. Yuntina, J. P. Soehaditama, and N. Nuraeni, "Studi Statistik Pengguna Operator Seluler di Indonesia: Tren dan Prediksi," *Jurnal Ilmu Statistik dan Komputasi*, vol. 3, no. 1, pp. 1–12, 2024.
[15] M. Syarif and W. Nugraha, "MWmote dalam mengatasi ketidakseimbangan kelas pada prediksi churn menggunakan klasifikasi C4.5," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 7, no. 1, pp. 54-62, 2023.
[16] Y. T. Utami *et al.*, "Penerapan algoritma C4.5 untuk prediksi churn rate pengguna jasa telekomunikasi," *Jurnal Komputasi*, vol. 8, no. 2, pp. 69-76, 2020.
[17] C. Zai, "Implementasi data mining sebagai pengolahan data," *Jurnal Portal Data*, vol. 2, no. 3, 2022.