Fari Katul Fikriah

Cek plagiasi.docx

Document Details

20% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Match Groups

- **41** Not Cited or Quoted 20%
  Matches with neither in-text citation nor quotation marks
- **0** Missing Quotations 0%
  Matches that are still very similar to source material
- **0** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

15% Internet sources

12% Publications

5% Submitted works (Student Papers)

Match Groups

- **41** Not Cited or Quoted 20%
  Matches with neither in-text citation nor quotation marks
- **0** Missing Quotations 0%
  Matches that are still very similar to source material
- **0** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

15% Internet sources
12% Publications
5% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	jurnal.itscience.org	3%
2	Student papers	Universitas Prima Indonesia	2%
3	Publication	"Advances in Artificial-Business Analytics and Quantum Machine Learning", Sprin...	2%
4	Internet	ejournal.unkhair.ac.id	1%
5	Internet	www.ioinformatic.org	1%
6	Internet	nemo.asee.org	<1%
7	Internet	join.if.uinsgd.ac.id	<1%
8	Publication	Henna Vartiainen, Thomas Derrick Hull, Erik Nook. "Diverse approaches to senti...	<1%
9	Internet	ejournal.ikado.ac.id	<1%
10	Internet	admin.thescipub.com	<1%

11	Internet	
www.medrxiv.org		<1%
12	Internet	
www.intmaritimeengineering.org		<1%
13	Internet	
export.arxiv.org		<1%
14	Internet	
e-jurnal.lppmunsera.org		<1%
15	Publication	
Sai Kiran Oruganti, Dimitrios A Karras, Srinesh Singh Thakur, Janapati Krishna Ch...		<1%
16	Internet	
ejournal.unama.ac.id		<1%
17	Internet	
github.com		<1%
18	Student papers	
Associatie K.U.Leuven		<1%
19	Internet	
www.nature.com		<1%
20	Publication	
Tanzir Hossain, Ar-Rafi Islam, Md Humaion Kabir Mehedi, Annajiat Alim Rasel, M. ...		<1%
21	Internet	
ijsart.com		<1%
22	Publication	
Ashok Kumar, Geeta Sharma, Anil Sharma, Pooja Chopra, Punam Rattan. "Advanc...		<1%
23	Publication	
Michael Wenceslaus Putong, Suharjito Suharjito. "Classification Model of Contact ...		<1%
24	Publication	
Iacopo Carnacina, Mawada Abdellatif, Manolia Andredaki, James Cooper, Darren ...		<1%

JIKO (Jurnal Informatika dan Komputer) Accredited KEMENDIKBUD RISTEK, No.105/E/KPT/2022
Vol. x, No. x, April 2024, pp. x-x p-ISSN: 2614-8897
DOI: 10.33387/jiko e-ISSN: 2656-1948

NAÏVE BAYES AND SUPPORT VECTOR MACHINE BASED ON OPTIMIZATION FOR PUBLIC SENTIMENT ANALYSIS POST-2024 ELECTION

Fari Katul Fikriah[1], Amelia Devi Putri Ariyanto[2]

[1,2]Information System and Technology Study Program, Widya Husada University
Email: [*1]farichatulfikriyah45@gmail.com, [2]ameliadev@gmail.com

(Received: dd mmm yyyy, Revised: dd mmm yyyy, Accepted: dd mmm yyyy)

Abstract (10pt)

The 2024 Election has generated a variety of public opinions that are widely disseminated across various digital media, making sentiment analysis essential to understand public perceptions quickly and accurately. This study aims to analyze and compare the performance of the optimized Naïve Bayes and Support Vector Machine (SVM) algorithms in classifying public sentiment after the election. The research methods include collecting 10,000 text data entries from various online sources, performing text preprocessing, extracting features using the TF-IDF method, applying both algorithms with parameter tuning, and evaluating their performance using accuracy, precision, recall, and F1-score metrics. The results show that the optimized SVM algorithm delivers superior performance, achieving an accuracy of 88.24% compared to 82.35% for Naïve Bayes. These findings indicate that SVM is more effective for analyzing complex public opinion sentiment and can therefore serve as a valuable reference for post-election policy-making.

Keywords: *Sentiment Analysis, Naïve Bayes, Support Vector Machine, TF-IDF, 2024 Election*

**Corresponding Author: Author1*

1. INTRODUCTION

The 2024 General Election in Indonesia is an important momentum that will not only determine the direction of national policy for the next five years but also trigger massive public opinion dynamics, especially in the digital space. Social media is the main medium for people to voice their opinions, hopes, and disappointments regarding the election process and results. Indonesia is a democratic country, with a system of government in which political power is held collectively by the people, and the democratic system is characterized by periodic general elections [1]. To achieve a democratic nation, the general election mechanism is regarded as a means to realize this goal in alignment with Pancasila and the 1945 Constitution of the Republic of Indonesia [2] [3].

In this context, sentiment analysis is one of the relevant approaches to comprehensively and data-drivenly capture public perception and emotions. The main challenge lies in processing large volumes of text data efficiently and accurately to gain a deep understanding of public responses after the election [4]. Machine learning-based sentiment analysis techniques have proven effective in classifying public opinion into positive, negative, or neutral categories. Among the most commonly used methods are Naïve Bayes and Support Vector Machines (SVM) [5]. Naïve Bayes is referred to as one of the commonly used classification algorithms, and it can be said that this algorithm is also the simplest. Naïve Bayes can estimate the probability of a class based on the distribution of words in a document. Naïve Bayes is often used by researchers because it has several advantages, including high accuracy, speed, and efficiency [6]. Meanwhile, the Support Vector Machine (SVM) algorithm is a type of classification method that utilizes a kernel to map data into a higher-dimensional space, allowing non-linearly separable data to be classified [7] [8].

This study aims to explore and compare the performance of Naïve Bayes and SVM in analyzing public sentiment after the 2024 Election, using a parameter optimization approach to achieve more optimal classification results. The optimization technique applied in this study focuses on improving classification accuracy through optimal parameter selection and enhancements in the feature extraction process. The benefits of this study include providing a scientific and practical foundation for policymakers, academics, and information technology practitioners to better understand public sentiment patterns, particularly in a dynamic socio-political context such as the post-election period. Additionally, this study is expected to enrich the literature and advance the development of machine learning-based sentiment analysis models through the integration of effective optimization techniques.

2. RESEARCH METHOD

This study uses a quantitative approach with an experimental method to compare the performance of the Naïve Bayes and Support Vector Machine (SVM) algorithms in analyzing public sentiment after the 2024 Election. The main data consists of public opinion posts on social media, especially from the Twitter/X platform, which are collected using web scraping techniques based on certain keywords relevant to election topics, such as candidate names, political parties, and popular issues that emerged during and after the election. The steps are as follows:

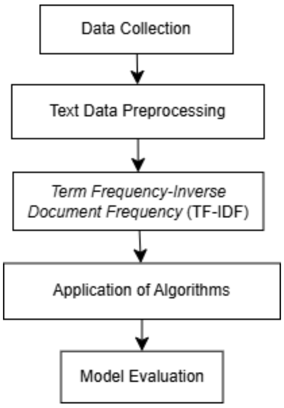

Figure 1. Research Steps

2.1 Data Collection

In this study, data were collected from social media using the Twitter/X API (Application Programming Interface) and crawling tools. The data collection period was determined to be from one week before to one month after the announcement of the official results of the 2024 Election. Data were collected through a data crawling method using the Python programming language.

2.2 Text Data Preprocessing

Data obtained through crawling still contains a lot of noise. Therefore, preprocessing is carried out, which includes:

- Data cleaning: removing URL, emojis, numbers, punctuation, and special characters.
- Tokenization: breaking sentences into words.
- Stopword removal: eliminating common words that do not carry high informational value.

2.3 Feature Extraction TF-IDF

In this step, text features are converted into numeric form using the Term Frequency-Inverse Document Frequency (TF-IDF) technique, which represents the importance weight of a word in a document. The processed text is transformed into a numerical representation by calculating the weight of each word based on its frequency in the document and how common the word is across the entire document collection.

2.4 Application Algoritms

This study uses two algorithms, namely Naïve Bayes and Support Vector Machine (SVM). The Naïve Bayes algorithm uses a combination of values from the dataset and frequency to determine the probability. This algorithm applies Bayes' theorem, which assumes that no attribute of the dataset depends on the value of the class variable . The Naïve Bayes algorithm is formulated as follows:

$$P(x) = \frac{P(H)P(H)}{P(x)} \quad (1)$$

The naive Bayes algorithm, as shown in the image above, starts with the input dataset and then divides it into two parts: training data and testing data. training data is processed by calculating the probability of the testing data to produce a decision [9]. The Support Vector Machine (SVM) algorithm is a classification method that can handle both linear and non-linear data. SVM aims to divide the dataset into classes to find the optimal hyperplane to separate classes in the data with a maximum margin [10].

2.5 Model Evaluation

During the model evaluation phase, this research seeks to assess how well the optimized Naïve Bayes and Support Vector Machine (SVM) algorithms perform in categorizing public sentiment following the 2024 Election. The assessment is carried out by comparing the model's predictions with the actual labels in the test dataset. The evaluation utilizes various performance metrics, including accuracy, precision, recall, and the F1-score. Accuracy reflects the proportion of correctly classified instances. Precision and recall evaluate how relevant and complete the classification outcomes are, while the F1-score, as the harmonic mean of precision and recall, provides a balanced measurement, especially when dealing with imbalanced data classes.

In addition to these metrics, the confusion matrix is examined to illustrate the detailed distribution of true and false predictions for each sentiment category (positive, negative, and neutral). The evaluation findings are further analyzed to determine how much the applied optimization enhances the performance of the Naïve Bayes and SVM algorithms compared to their unoptimized versions. Consequently, this evaluation stage forms the basis for concluding the model's effectiveness in accurately identifying public sentiment, thereby supporting decision-making and sentiment analysis in the post-election context.

3. RESULT AND DISCUSSION

4. 1 Data Collection

The data used in this study were collected from Twitter (now X), discussion forums, and online news comment sections discussing the political situation after the 2024 Election. A total of 10,000 raw text entries were collected. After undergoing cleaning and validation, 8,500 entries were deemed suitable for analysis, with the distribution of sentiment classes shown in Table 1.

Table 1. Sentiment Data Distribution

Sentiment	Amount of Data	Percentage
positive	3400	40%
negative	3060	36%
netral	2040	24%

4. 2 Text Data Processing and TF-IDF Results

In the text preprocessing stage, data cleaning is performed by removing punctuation, eliminating stopwords, and normalizing words. After that, feature extraction is conducted using the Term Frequency–Inverse Document Frequency (TF-IDF) method. The vectorization results show that the words with the highest weights are related to political issues, candidate names, policies, and public opinions. As an illustration, Table 2 presents the 10 words with the highest TF-IDF scores.

Table 2. TF-IDF Score

Text	TF-IDF Score
demokrasi	0.083
protes	0.072
pemimpin	0.072
netralisasi	0.067
kebijakan	0.066
korupsi	0.070

4. 3 Application Algorithm

After the data collection, text preprocessing, and feature extraction using the TF-IDF method, the next stage is the application of classification algorithms to analyze public sentiment. This study employs two popular algorithms, namely Naïve Bayes and Support Vector Machine (SVM), both of which are optimized to enhance classification performance.

a. Naive Bayes

Naïve Bayes is applied as the baseline algorithm because of its strong ability to handle text data with high computational speed. The Naïve Bayes model works by calculating the probability that a piece of text belongs to one of the sentiment categories (positive, negative, or neutral) based on the distribution of words computed from the training data. At this stage, parameter tuning is performed to select the Naïve Bayes variant that best matches the characteristics of the data. After testing several options, the Multinomial Naïve Bayes algorithm was chosen because it proved most effective in handling TF-IDF data. The model was trained on 80% of the dataset with a balanced distribution among the classes. The training process was fast, with an average computation time of less than 5 seconds for the entire training data.

The results show that Naïve Bayes is able to recognize common word patterns that indicate sentiment, especially in texts with explicit words such as *bagus*, *rusak*, *adil*, or *curang*. However, for ambiguous or sarcastic texts, the model tends to show decreased accuracy due to the limitation of the assumption of word independence.

b. Support Vector Machine (SVM)

Unlike Naïve Bayes, the SVM algorithm is applied using a supervised learning approach that builds an optimal hyperplane to separate data into different classes. In this study, the application of SVM focuses on the linear kernel, as experimental results indicate that the linear kernel is more stable than the RBF kernel for high-dimensional data such as TF-IDF vectors.

Additionally, optimization of the **C** and **gamma** parameters is performed to minimize classification errors and avoid overfitting. The parameter tuning process uses the grid search cross-validation method with 5-fold cross-validation on the training data. The tuning results show that a **C** value of 1.0 and the linear kernel produce the best accuracy during validation.

The SVM training time is relatively longer than that of Naïve Bayes, taking approximately 12 seconds for the entire training data. However, SVM is capable of distinguishing more nuanced sentiment patterns, especially in sentences with complex word structures and contextual meanings.

4. 4 Model Evaluation

Both models were tested on 20% of the total data. Based on the test results, SVM demonstrated significant superiority in terms of accuracy, precision, recall, and F1-score. This is due to SVM's ability to build optimal separation margins even when the data have high-dimensional features. Table 3 shows the evaluation metrics that show the performance differences between the two algorithms:

Table 1. Evaluation Metric

Metric	Naive Bayes (%)	SVM (%)
Accuracy	82.35	88.24
precision	81.40	87.5
Recall	80.75	88
F1-Score	81.07	87.75

From the results above, it can be concluded that the SVM model with parameter optimization successfully provides better sentiment classification performance than Naïve Bayes, although it requires more computation time. This result is consistent with the findings of several previous studies, which state that SVM has advantages in text classification for data with complex public opinion contexts, such as post-election political sentiment.

4. CONCLUSION

The 2024 Election has generated various public opinions and sentiments that are widely disseminated through different digital media. Sentiment analysis is an important approach to capture public perceptions quickly and accurately, thereby helping policymakers understand social dynamics after the election.

This study aims to compare the performance of the Naïve Bayes algorithm and the Support Vector Machine (SVM) with parameter optimization in classifying public sentiment. The methods employed include data collection from online platforms, text preprocessing for data cleaning and normalization, feature extraction using TF-IDF, application of both algorithms with parameter tuning, and model performance evaluation using accuracy, precision, recall, and F1-score metrics.

The results indicate that the SVM algorithm with parameter optimization delivers better performance than Naïve Bayes. The SVM model achieved an accuracy of 88.24%, whereas Naïve Bayes achieved an accuracy of 82.35%. Additionally, SVM outperformed Naïve Bayes in terms of precision, recall, and F1-score. This difference demonstrates that SVM is more effective in handling sentiment text data with complex opinion contexts, although it requires relatively more computation time. Therefore, it can be concluded that the application of an optimization-based SVM algorithm is more recommended for post-election public sentiment analysis, as it provides more accurate and reliable classification results to support strategic decision-making in the socio-political domain.