Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6543

Submitted : Jul 17, 2025
Accepted : Jun 22, 2025
Published : Jul 23, 2025

Sentiment Analysis on TikTok Discourse Surrounding the 2024 North Sumatra Gubernatorial Election Using Support Vector Machine Algorithm

Istiqomah[1)*], Aidil Halim Lubis[2)]

[1,2)]Universitas Islam Negeri Sumatera Utara, Indonesia
[1)] iqoma607@gmail.com, [2)] aidilhalimlubis@uinsu.ac.id

ABSTRACT

This study aims to analyze public sentiment towards the 2024 North Sumatra gubernatorial election by leveraging social media data, specifically TikTok, which has become a major platform for political discourse in Indonesia. The two competing candidate pairs, Bobby Nasution–Surya and Edy Rahmayadi–Hasan Basri, have sparked widespread online discussions that range from enthusiastic support to harsh criticism. These interactions have a significant impact on public opinion formation and may influence electoral outcomes. To address this phenomenon, this research implements a sentiment classification model using the Support Vector Machine (SVM) algorithm with a polynomial kernel, known for its effectiveness in handling high-dimensional textual data. A total of 2,100 TikTok comments were collected using scraping techniques via Python. The data then underwent several preprocessing stages, including case folding, cleaning, normalization, tokenizing, slangword removal, stopword removal, and stemming. Feature extraction was conducted using the TF-IDF method, followed by lexicon-based sentiment labeling into positive and negative classes. The classification model achieved an accuracy of 82%, with a positive sentiment precision of 0.81, recall of 0.96, and F1-score of 0.88. For negative sentiment, the precision was 0.86, recall 0.51, and F1-score 0.64. These findings indicate that the model performs well in identifying explicit positive sentiments but faces challenges in recognizing complex negative expressions such as sarcasm or implicit criticism. The results provide valuable insights into digital political behavior and demonstrate the potential of machine learning-based sentiment analysis as a tool for monitoring public perception in real time during elections.

Keywords: Sentiment Analysis, TikTok, Support Vector Machine (SVM), TF-IDF, Political Opinion

INTRODUCTION

Democracy is widely regarded as a political system that empowers citizens through mechanisms of representation, participation, and accountability. In the digital era, the dynamics of democracy are increasingly intertwined with technological advancement. Social media platforms have evolved into significant arenas of political discourse, enabling individuals to express their opinions, organize movements, and influence public sentiment in real time. Among these platforms, TikTok has emerged as a particularly influential space, especially among youth, due to its high engagement rates, short-form video content, and viral nature. These characteristics make it an ideal medium for political communication and persuasion, both constructive and manipulative(Nur et al., 2023; Puspitarini & Nuraeni, 2019).

Indonesia, as the third-largest democracy in the world, has undergone substantial democratic transformation since its independence in 1945. The fall of the New Order regime and the beginning of the Reformation Era in 1998 opened a new chapter in Indonesian politics. Direct elections at multiple levels, freedom of speech, and a more active civil society have become central features of post-reform governance. However, the Indonesian democratic landscape continues to face complex challenges. Issues such as identity politics, disinformation, social polarization, and money politics frequently emerge during electoral periods. These issues are amplified by the rapid spread of information—both factual and misleading—through digital platforms (Fathiarahma et al., 2023; Insani et al., 2023).

In the context of the upcoming 2024 regional elections, the gubernatorial election in North Sumatra (Pilgub Sumut) represents one of the most anticipated local political contests. The rivalry between two major candidate pairs—Bobby Nasution–Surya and Edy Rahmayadi–Hasan Basri—has attracted significant attention in both offline and online domains. TikTok, in particular, has been flooded with content that ranges from endorsements and campaign videos to criticisms, satire, and even conflict-laden debates among supporters. The nature of TikTok as a high-engagement platform makes it fertile ground for the formation, reinforcement, and dissemination of political sentiment (Nur et al., 2023). The intensity of discourse on this platform reflects the growing role of social media in shaping

* Corresponding author
Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6543

Submitted : Jul 17, 2025
Accepted : Jun 22, 2025
Published : Jul 23, 2025

public opinion and potentially influencing electoral outcomes at the local level.

Given the volume, variability, and velocity of content on social media, especially on TikTok, there is a need for computational approaches that can analyze and classify political sentiment effectively. In this study, we apply the Support Vector Machine (SVM) algorithm—an established machine learning method in the field of text mining and sentiment classification. SVM has been widely recognized for its robustness in handling high-dimensional data, making it suitable for textual datasets derived from social media (Tarihoran & Manurip, 2018; Utami & Erfina, 2021). The use of SVM in this research focuses on analyzing sentiment polarity (positive or negative) of TikTok comments related to the 2024 Pilgub Sumut, allowing for insights into public perception and emotional dynamics around political figures.

To what extent can the SVM algorithm effectively classify sentiment toward political events, particularly in the case of the 2024 North Sumatra gubernatorial election, as expressed through TikTok content?

By addressing this question, the study aims to contribute to the broader understanding of digital political behavior in Indonesia and the role of machine learning in supporting real-time monitoring of public sentiment. Additionally, the findings are expected to inform future improvements in sentiment classification models, especially in identifying complex language features such as sarcasm, irony, and indirect criticism, which are commonly found in user-generated social media content.

LITERATURE REVIEW

Recent research highlights TikTok's emergence as a political communication arena, particularly among young voters. Its visual storytelling format, short video structure, and personalized content algorithm have made it a powerful platform for shaping political narratives. (Literat & Kligler-Vilenchik, 2023) describe TikTok as a space where youth engage in "civic creativity," producing content that blends entertainment with political messaging. Their study emphasizes TikTok's role in redefining how young people participate in political conversations, outside traditional media or institutional frameworks. Similarly, (Zamora-medina et al., 2023) found that TikTok fosters political persuasion strategies that are emotionally charged and identity-driven, making it an effective tool for reaching the youth electorate during election cycles.

"TikTok is increasingly becoming a political battleground for youth engagement," (Literat & Kligler-Vilenchik, 2023), yet few studies have explored the dynamics of public sentiment within this platform during subnational elections. Most existing literature focuses on national-level campaigns, leaving a gap in understanding how local elections are discussed, critiqued, or supported by citizens on TikTok. As local leaders increasingly utilize this platform for outreach, analyzing sentiment at the regional level—such as in the 2024 North Sumatra gubernatorial election—becomes a timely research focus.

A. Text Mining

Text mining refers to the process of extracting meaningful patterns and information from unstructured textual data. In sentiment analysis, preprocessing is a crucial first step that involves tasks such as case folding, cleaning, normalization, tokenization, slangword removal, stopword removal, and stemming (Aulia et al., 2021; Furqan et al., 2022). These processes help convert unstructured language into structured formats suitable for computational analysis.

Despite its utility, several studies have identified that traditional machine learning models like Support Vector Machine (SVM) tend to perform well on explicit sentiment but struggle with implicit negative sentiment. This includes sarcasm, irony, and metaphorical criticism—linguistic elements often present in social media political discourse (Arsi & Waluyo, 2021; Fathiarahma et al., 2023). The recall rate for negative sentiment is often low due to the subtle nature of language, especially in politically charged or emotionally nuanced contexts. This highlights a critical limitation in current sentiment classification approaches.

The main goal of text mining is to extract important and useful information from unstructured text data. The document text used in the text mining process must first be prepared before it can be used in the main process. The preparation of the document text or raw data set is also called text preprocessing. Text preprocessing serves to convert unstructured or random text data into structured data (Furqan et al., 2022).

B. Support Vector Machine

There are two important types of hyperplane in SVM: linear hyperplane and non-linear hyperplane. If the data can be perfectly separated by a linear hyperplane, the SVM is called a linear SVM. However, if the data cannot

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6543

Submitted : Jul 17, 2025
Accepted : Jun 22, 2025
Published : Jul 23, 2025

be perfectly separated linearly, SVM uses kernel transformation techniques to transform the data into a higher feature dimension, where a linear hyperplane can be created. By using different types of kernel functions, such as linear kernels, polynomial kernels or Gauss kernels, SVM can efficiently deal with data that has non-linear characteristics. A kernel function is a function that transforms data to a higher dimension with the aim of improving the data structure so as to facilitate the separation process (Rabbani et al., 2023).

In this case, the author uses a polynomial kernel, Polynomial Kernel is one type of kernel function in the Support Vector Machine (SVM) algorithm used to map non-linear data into a higher dimensional feature space so that it can be linearly separated. The polynomial kernel formula can be written as follows:

$$K(x, y) = (x, y + c)^d \quad (1)$$

Description:

K(x, y)	: The function or result of a polynomial operation that depends on two variables, xxx and yyy.
x	: The first variable in the polynomial function.
y	: The second variable in the polynomial function.
c	: A constant added to the product of xxx and yyy.
d	: The exponent or power applied to the entire expression in parentheses.
(x, y + c)	: An expression denoting the two elements xxx and y+cy + cy+c, where ccc is added to the variable yyy.

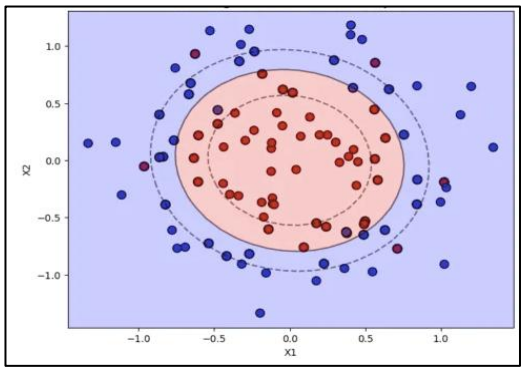


Fig 1. Visualization of SVM model results with polynomial kernel
source : sainsdata.id

1. Determining the Optimal Hyperplane

SVM aims to find the optimal hyperplane that separates two classes with maximum margin. The hyperplane equation in SVM is:

$$w.\, x + b = 0 \quad (2)$$

Description:
w = Weight vector
x = Input vector
b = Bias

To find the best hyperplane, SVM maximizes the margin between two classes by solving the following objective function:

$$\min_a \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j y_i y_j K(x_{i,} x_j) - \sum_{i=1}^{n} a_i \quad (3)$$

with restrictions:

$$\sum_{i=1}^{n} a_i\, y_i = 0, \;\; 0 \leq_{a_i} \leq C \quad (4)$$

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6543

Submitted : Jul 17, 2025
Accepted : Jun 22, 2025
Published : Jul 23, 2025

Description :
a_i : Lagrange multiplier that determines the weight of the support vector.
y_i : Class label (+1 for positive, -1 for negative.
$K(x_i, x_j)$: The previously calculated polynomial kernel.
C : Regulatory parameters to control the balance between margin and misclassification.

2. Determining Support Vector and Decision Function

After the optimization is complete, the SVM finds the support vector, which is the data point that is on the edge of the optimal margin. The weight vector w is calculated as :

$$w = \sum_{i=1}^{n} a_i \, y_i x_i \quad (5)$$

This weight is later used in the decision function to determine the classification of the new data. Then, the decision function to determine the class of the data is :

$$f(x) = \sum_{i=1}^{n} a_i \, y_i \; K(x_i, x) + b \quad (6)$$

Classification rule:
If f(x) > 0 → Data is classified as positive.
If f(x) < 0 → Data is classified as negative.

METHOD

1. Data Collection

This study collected a total of 2,100 comments from TikTok videos related to the 2024 North Sumatra gubernatorial election. The data was retrieved using scraping techniques with the Python programming language, specifically leveraging an unofficial TikTok API endpoint. JSON responses were parsed using the JMESPath library to extract relevant elements such as comment text and user replies.

The selected videos featured content related to both candidate pairs—Bobby Nasution–Surya and Edy Rahmayadi–Hasan Basri—and were chosen based on high engagement levels (likes, comments, shares). The final dataset represents a mixture of positive, negative, and neutral political expressions from users.

2. Preprocessing

a. Cleaning

The preprocessing cleaning stage in sentiment analysis involves cleaning text data to ensure that the data to be used for analysis is of good quality. To clean comments from symbols, numbers, URLs, hashtags (Shevira et al., 2022).

b. Case Folding

Case Folding, this stage will change the form of Text which was originally mixed with uppercase letters to Lowercase from the dataset results that have been obtained through the Crawling stage (Rifaldi et al., 2023).

c. Normalization

The main purpose of this normalization process is to transform non-standard word forms, such as abbreviations or slang words, into a standard language format in accordance with the rules of the Big Indonesian Dictionary (KBBI).

d. Tokenizing

Tokenizing is a process to break sentences into words based on spaces from case folding processing results (Aulia et al., 2021).

e. Stopword Removal

With the Stopword feature, the text before being classified is first removed text that is not related to the sentiment analysis to be studied so that the text dimension will be reduced without reducing the sentiment content of the review (Utami & Erfina, 2021).

f. Stemming

Stemming is the process of converting all words into their basic words (Aulia et al., 2021). This algorithm is based on the morphology of Indonesian sentence structure which consists of prefix, suffix and prefix + suffix.

3. Feature Extraction (TF-IDF)

TF-IDF (Term Frequency-Inverse Document Frequency) weighting is a process to transform data from textual data into numerical data to weight each word or feature. TF-IDF is a statistical measure used to assess the

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6543

Submitted : Jul 17, 2025
Accepted : Jun 22, 2025
Published : Jul 23, 2025

importance of a word in a document. TF is the frequency of occurrence of a word in a particular document. DF describes the number of documents that contain the word. IDF is the inverse of DF, which indicates how unique the word is across documents. The result of word weighting using TF-IDF is the product of TF multiplied by IDF (Septian et al., 2019).

4. Classification Using SVM (Polynomial Kernel)
   The classification task was performed using the Support Vector Machine (SVM) algorithm with a polynomial kernel. A polynomial kernel is a non-linear function used in SVM to handle data that cannot be linearly separated. It works by increasing the dimension of the data space to make the separation between classes easier.
5. Evaluation
   Model performance was evaluated using standard binary classification metrics:
   a. Confusion Matrix: To visualize the distribution of true positives, false positives, true negatives, and false negatives.
   b. Accuracy: Measures how many predictions are correct compared to the entire test data.
   c. Precision: Measures the accuracy of the model in predicting a particular sentiment.
   d. Recall: Measures how well the model detects all relevant data in a class.
   e. F1-Score: Harmonized average of precision and recall, which provides a more balanced assessment.

Prior to training, the dataset was manually annotated into two sentiment classes: positive and negative. The labeling was performed by two independent annotators familiar with the context of the 2024 gubernatorial election and local sociolinguistic nuances.

To assess the consistency between annotators, Cohen's Kappa coefficient was calculated, yielding a value of 0.82, which indicates substantial agreement. Disagreements were resolved through discussion to reach a final consensus label for each comment.

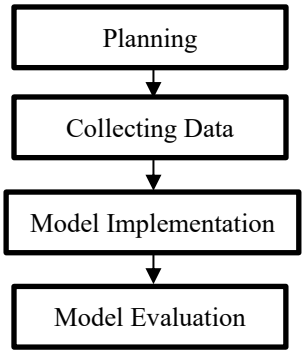


Fig 2. Flowchart

RESULT

To ensure the data is ready for analysis, a series of preprocessing stages are applied. After the cleaning process, the TF-IDF method is used to perform word weighting, which converts the representation of each comment into a numerical format. The processed dataset is then divided into two subsets: training and testing data, which are essential for training and evaluating the performance of the classification model.

Table 1. Sample Data

Komentar
SUDAH LAH JANGAN FRAMING, KALAH KALAH AJA, MANTAN PRAJURIT SEHARUSNYA SIAP KALAH, SIAP MENANG. JANGAN MEMPERMALAUKAN DIRI SENDIRI.
prestasi boby apa selama menjabat walikota? yg ada semua pembangunan mangkrak
PNS dipaksa supaya memilih Boby Surya , bahkan sampai turun tangan ke

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6543

Submitted : Jul 17, 2025
Accepted : Jun 22, 2025
Published : Jul 23, 2025

TPS mengawasi pegawai2 nya 😂😂, sampai org tua murid pun disuruh memilih boby
kalah ya kalah lah bos, namanya juga kontestasi tentu ada yg menang ada yg kalah, kalau gk siap kalah y jgn ikut
edy edy 😂😂😂 kau pikir Sumut tanah nenek moyang mu 😂😂😂
kalo kalah bilangnya musuh maen kotor, giliran menang ngakunya daerah yg paling cerdas 🤣
LOGIKA AJA BRO, EDI GUNERNUR LEBIH LUAS JANGKUAN NYA, SEDANGKAN BOBI CMA WLIKKTA SELEVEL KABUPATEN ,ARTINYA EDI PUN BERMAIN JG CMA KARENA KALAH BARI RIBUT

Data transformation starts with cleaning, a process to eliminate irrelevant characters or elements. After that, case folding is applied to convert the entire text into lowercase format, followed by tokenizing to break sentences into individual words. Next, filtering is performed to filter out words that are not essential for analysis. The process continues with stopword removal, which aims to remove common words without significant meaning (such as “and”, ‘which’, “in”). The final stage is stemming, which returns each word to its base form. The ultimate goal of this entire sequence is to produce clean, consistent, and structured data, thus supporting the effectiveness of sentiment analysis.

Table 2. Final Result of Data Preprocessing

Comment	Cleaning Data	Case Folding	Normalization	Tokenizing	Stopword Removal	Stemming	Sentiment
SUDAH LAH JANGAN FRAMING, KALAH KALAH AJA, MANTAN PRAJURIT SEHARUSNYA SIAP KALAH, SIAP MENANG. JANGAN MEMPERMAL AUKAN DIRI SENDIRI.	SUDAH LAH JANGAN FRAMING KALAH KALAH AJA MANTAN PRAJURIT SEHARUSNYA SIAP KALAH SIAP MENANG JANGAN MEMPERMALA UKAN DIRI SENDIRI	sudah lah jangan framing kalah kalah aja mantan prajurit seharusnya siap kalah siap menang jangan mempermal aukan diri sendiri	sudah lah jangan framing kalah kalah saja mantan prajurit seharusnya siap kalah siap menang jangan mempermal aukan diri sendiri	['sudah', 'lah', 'jangan', 'framing', 'kalah', 'kalah', 'saja', 'mantan', 'prajurit', 'seharusnya', 'siap', 'kalah', 'siap', 'menang', 'jangan', 'mempermalaukan', 'diri', 'sendiri']	['framing', 'kalah', 'kalah', 'mantan', 'prajurit', 'kalah', 'menang', 'mempermalaukan']	framing kalah kalah mantan prajurit kalah menang malau	Positif
prestasi boby apa selama menjabat walikota? yg ada semua pembangunan mangkrak	prestasi boby apa selama menjabat walikota yg ada semua pembangunan mangkrak	prestasi boby apa selama menjabat walikota yg ada semua pembangunan mangkrak	prestasi boby apa selama menjabat walikota yang ada semua pembangunan mangkrak	['prestasi', 'boby', 'apa', 'selama', 'menjabat', 'walikota', 'yang', 'ada', 'semua', 'pembangunan', 'mangkrak']	['prestasi', 'boby', 'menjabat', 'walikota', 'pembangunan', 'mangkrak']	prestasi boby jabat walikota bangun mangkrak	Negatif

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6543

Submitted : Jul 17, 2025
Accepted : Jun 22, 2025
Published : Jul 23, 2025

PNS dipaksa supaya memilih Boby Surya , bahkan sampai turun tangan ke TPS mengawasi pegawai2 nya 😂😂, sampai org tua murid pun disuruh memilih boby	PNS dipaksa supaya memilih Boby Surya bahkan sampai turun tangan ke TPS mengawasi pegawai nya sampai org tua murid pun disuruh memilih boby	pns dipaksa supaya memilih boby surya bahkan sampai turun tangan ke tps mengawasi pegawai nya sampai org tua murid pun disuruh memilih boby	pns dipaksa supaya memilih boby surya bahkan sampai turun tangan ke tps mengawasi pegawai ya sampai orang tua murid pun disuruh memilih boby	['pns', 'dipaksa', 'supaya', 'memilih', 'boby', 'surya', 'bahkan', 'sampai', 'turun', 'tangan', 'ke', 'tps', 'mengawasi', 'pegawai', 'ya', 'sampai', 'orang', 'tua', 'murid', 'pun', 'disuruh', 'memilih', 'boby']	['pns', 'dipaksa', 'memilih', 'boby', 'surya', 'turun', 'tangan', 'tps', 'mengawasi', 'pegawai', 'orang', 'tua', 'murid', 'disuruh', 'memilih', 'boby']	pns paksa pilih boby surya turun tangan tps awas pegawai orang tua murid suruh pilih boby	Negatif
kalah ya kalah lah bos, namanya juga kontestasi tentu ada yg menang ada yg kalah, kalau gk siap kalah y jgn ikut	kalah ya kalah lah bos namanya juga kontestasi tentu ada yg menang ada yg kalah kalau gk siap kalah y jgn ikut	kalah ya kalah lah bos namanya juga kontestasi tentu ada yg menang ada yg kalah kalau gk siap kalah y jgn ikut	kalah ya kalah lah bos namanya juga kontestasi tentu ada yang menang ada yang kalah kalau tidak siap kalah ya jangan ikut	['kalah', 'ya', 'kalah', 'lah', 'bos', 'namanya', 'juga', 'kontestasi', 'tentu', 'ada', 'yang', 'menang', 'ada', 'yang', 'kalah', 'kalau', 'tidak', 'siap', 'kalah', 'ya', 'jangan', 'ikut']	['kalah', 'kalah', 'bos', 'namanya', 'kontestasi', 'menang', 'kalah', 'kalah']	kalah kalah bos nama kontestasi menang kalah kalah	Negatif
edy edy 😂😂😂 kau pikir Sumut tanah nenek moyang mu 😂😂😂	edy edy kau pikir Sumut tanah nenek moyang mu	edy edy kau pikir Sumut tanah nenek moyang mu	edy edy kau pikir sumut tanah nenek moyang mu	['edy', 'edy', 'kau', 'pikir', 'sumut', 'tanah', 'nenek', 'moyang', 'mu']	'edy', 'edy', 'kau', 'pikir', 'sumut', 'tanah', 'nenek', 'moyang']	edy edy kau pikir sumut tanah nenek moyang	Negatif
kalo kalah bilangnya musuh maen kotor, giliran menang ngakunya daerah yg paling cerdas 🤣	kalo kalah bilangnya musuh maen kotor giliran menang ngakunya daerah yg paling cerdas	kalo kalah bilangnya musuh maen kotor giliran menang ngakunya daerah yg paling cerdas	kalau kalah bilangnya musuh main kotor giliran menang ngakunya daerah yang paling cerdas	['kalau', 'kalah', 'bilangnya', 'musuh', 'main', 'kotor', 'giliran', 'menang', 'ngakunya', 'daerah', 'yang', 'paling', 'cerdas']	['kalah', 'bilangnya', 'musuh', 'main', 'kotor', 'giliran', 'menang', 'ngakunya', 'daerah', 'cerdas']	kalah bilang musuh main kotor gilir menang ngakunya daerah cerdas	Negatif

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6543

Submitted : Jul 17, 2025
Accepted : Jun 22, 2025
Published : Jul 23, 2025

LOGIKA AJA BRO, EDI GUNERNUR LEBIH LUAS JANGKUAN NYA, SEDANGKAN BOBI CMA WLIKKTA SELEVEL KABUPATEN ,ARTINYA EDI PUN BERMAIN JG CMA KARENA KALAH BARU RIBUT	LOGIKA AJA BRO EDI GUNERNUR LEBIH LUAS JANGKUAN NYA SEDANGKAN BOBI CMA WLIKKTA SELEVEL KABUPATEN ARTINYA EDI PUN BERMAIN JG CMA KARENA KALAH BARU RIBUT	logika aja bro edi gunernur lebih luas jangkuan nya sedangkan bobi cma wlikkta selevel kabupaten artinya edi pun bermain jg cma karena kalah baru ribut	logika saja bro edi gunernur lebih luas jangkuan ya sedangkan bobi cuma wlikkta selevel kabupaten artinya edi pun bermain juga cuma karena kalah baru ribut	['logika', 'saja', 'bro', 'edi', 'gunernur', 'lebih', 'luas', 'jangkuan', 'ya', 'sedangkan', 'bobi', 'cuma', 'wlikkta', 'selevel', 'kabupaten', 'artinya', 'edi', 'pun', 'bermain', 'juga', 'cuma', 'karena', 'kalah', 'baru', 'ribut']	['logika', 'bro', 'edi', 'gunernur', 'luas', 'jangkuan', 'bobi', 'wlikkta', 'selevel', 'kabupaten', 'edi', 'bermain', 'kalah', 'baru', 'ribut']	logika bro edi gunernur luas jangkuan bobi wlikkta level kabupaten edi main kalah baru ribut	Negatif

1. Data Labeling

The next process is labeling the comment data into two sentiment categories, namely positive and negative. In this research, the labeling process is done with a lexicon-based approach, which is a method that utilizes a dictionary of words (lexicon) that has been given a value or sentiment score. Each word in the lexicon has been categorized as positive or negative based on its meaning.

The cleaned comments are then analyzed based on the number of positive and negative words contained in them. The labeling rules are determined as follows:

a. If the number of positive words is more than the negative words, then the comment is labeled as positive.
b. If the number of negative words is more than the positive words, then the comment is labeled as negative.

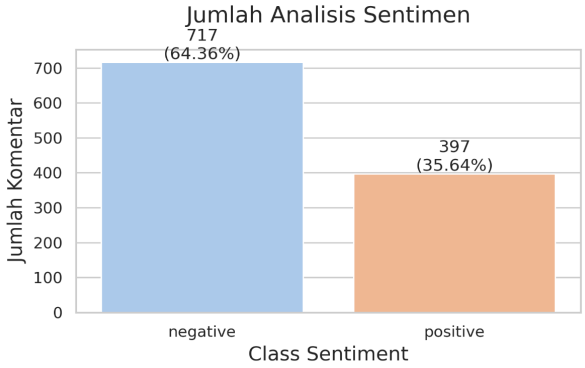


Fig 3. Comment Sentiment Distribution

Based on the results of data labeling using a lexicon-based approach, a total of 1,114 preprocessed comments were obtained, which were distributed into 717 negative comments and 397 positive comments. This distribution quantitatively shows the dominance of negative opinions from the public towards the candidates in the contestation of the North Sumatra Governor Election (Pilgub) compared to positive opinions.

2. TF - IDF weighting

For the purpose of classification analysis, the review data that has gone through a series of text pre-processing stages, although clean and structured, is still textual data. Since classification algorithms are generally designed to process input in numerical format, it is crucial to transform this textual data into a quantitative representation. In this research, such conversion is achieved through a word weighting mechanism using the Term Frequency-Inverse Document Frequency (TF-IDF) method. This TF-IDF weighting process not only converts words into numerical values but also serves to identify and highlight terms (words) that have the greatest significance or weight in the comment data corpus.

After getting the Term Frequency (TF) and Document Frequency (DF) values, the next step is to calculate the Inverse Document Frequency (IDF). After the Term Frequency (TF) and Inverse Document Frequency (IDF)

* Corresponding author
Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6543

Submitted : Jul 17, 2025
Accepted : Jun 22, 2025
Published : Jul 23, 2025

calculation process is complete, the next step is to calculate the TF-IDF weight.

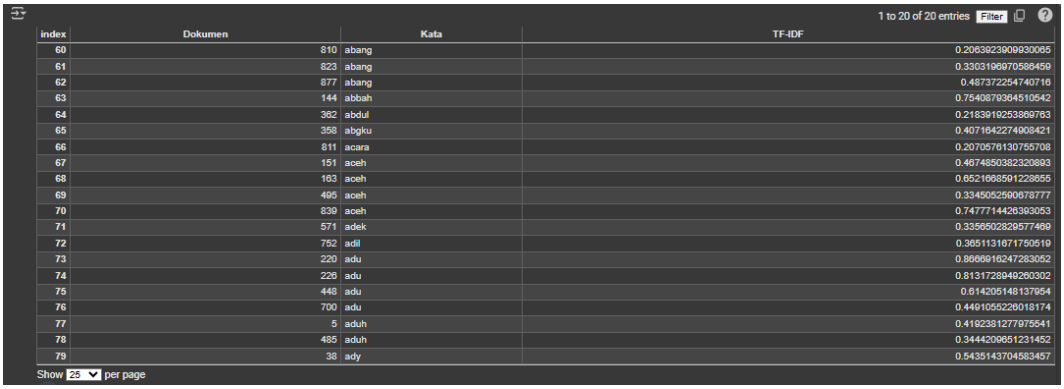
Table 3 TF-IDF Value

TERM	D1	D2	D3	D4	D5
framing	1.477	0	0	0	0
kalah	3.903	0	0	5.204	0
mantan	1.477	0	0	0	0
prajurit	1.477	0	0	0	0
menang	1.301	0	0	1.301	0
malau	1.477	0	0	0	
prestasi	0	1.477	0	0	0
boby	0	1.301	3.903	0	0
jabat	0	1.477	0	0	0
walikota	0	1.477	0	0	0
bangun	0	1.477	0	0	0
mangkrak	0	1.477	0	0	0

After the feature weighting stage through TF-IDF, the next essential step is the application of normalization techniques. This process aims to homogenize the range of values (scale) across the data, thus ensuring each feature contributes proportionally to the classification model.

Table 4. Normalization Value

TERM	D	D2	D3	D4	D5
framing	0.2916	0	0	0	0
kalah	0.7706	0	0	0.8757	0
mantan	0.2916	0	0	0	0
prajurit	0.2916	0	0	0	0
menang	0.2569	0	0	0.2189	0
malau	0.2916	0	0	0	0
prestasi	0	0.4576	0	0	0
boby	0	0.4031	0.5636	0	0
jabat	0	0.4576	0	0	0
walikota	0	0.4576	0	0	0
bangun	0	0.4576	0	0	0
mangkrak	0	0.4576	0	0	0


1 to 20 of 20 entries Filter

index	Dokumen	Kata	TF-IDF
60	810	abang	0.2063923909930065
61	823	abang	0.3303196970586459
62	877	abang	0.487372254740716
63	144	abbah	0.7540879364510542
64	362	abdul	0.2183919253869763
65	358	abgku	0.4071642274908421
66	811	acara	0.2070576130755708
67	151	aceh	0.4674850382320893
68	163	aceh	0.6521668591228655
69	495	aceh	0.3345052500878777
70	839	aceh	0.7477714426393053
71	571	adek	0.335650282957469
72	752	adil	0.3651131671750519
73	220	adu	0.8666916247283052
74	226	adu	0.8131728949260302
75	448	adu	0.614205148137954
76	700	adu	0.4491055226018174
77	5	aduh	0.4192381277975541
78	485	aduh	0.3444209851231452
79	38	ady	0.5435143704583457

Show 25 per page

Fig 4. TF-IDF Value

3. WordCloud Visualization

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6543

Submitted : Jul 17, 2025
Accepted : Jun 22, 2025
Published : Jul 23, 2025

After the sentiment classification process is carried out using the Support Vector Machine (SVM) algorithm, this research produces two main categories, namely positive sentiment and negative sentiment towards the North Sumatra Pilgub 2024 sourced from TikTok user comments. To clarify and strengthen the results of sentiment analysis, a visualization approach is used in the form of wordcloud and sentiment distribution graphs.

Wordcloud presents visualizations based on the frequency of occurrence of words in comments, where the size of the letters reflects how often the words are used. Meanwhile, the sentiment class distribution graph shows the proportion of the number of comments belonging to positive and negative categories quantitatively.

```python
import pandas as pd
from wordcloud import WordCloud
import matplotlib.pyplot as plt

# Pisahkan dataset berdasarkan sentiment
sentimen_positif = data[data['sentiment'] == 'positive']['steming_data'].str.cat(sep=' ')
sentimen_negatif = data[data['sentiment'] == 'negative']['steming_data'].str.cat(sep=' ')

# Fungsi untuk menampilkan dan menyimpan WordCloud
def create_wordcloud(text, title, output_file):
    wordcloud = WordCloud(width=800, height=400, background_color='black').generate(text)

    plt.figure(figsize=(10, 5))
    plt.imshow(wordcloud, interpolation='bilinear')
    plt.axis('off')
    plt.title(title, fontsize=16, color='white', fontweight='bold')

    # Simpan ke file PNG
    plt.tight_layout()
    plt.savefig(output_file, format='png')
    print(f"WordCloud berhasil disimpan ke {output_file}")
    plt.show()

# Buat dan simpan WordCloud untuk sentiment positif dan negatif
create_wordcloud(sentimen_positif, 'WordCloud Sentimen Positif', 'wordcloud_positif_setelah svm.png')
create_wordcloud(sentimen_negatif, 'WordCloud Sentimen Negatif', 'wordcloud_negatif_Setelah Svm.png')
```

Fig 5. Wordcloud Code

1. Negative Sentiment Wordcloud

This wordcloud is generated from data that has passed the preprocessing and classification process using the Support Vector Machine (SVM) algorithm. The words that appear in the wordcloud are the words most often used in comments containing support, hope, or positive responses to candidates or the Pilgub process. The size of the letters in the wordcloud indicates the frequency of occurrence of the word, the larger the size, the more often the word appears in the comments.

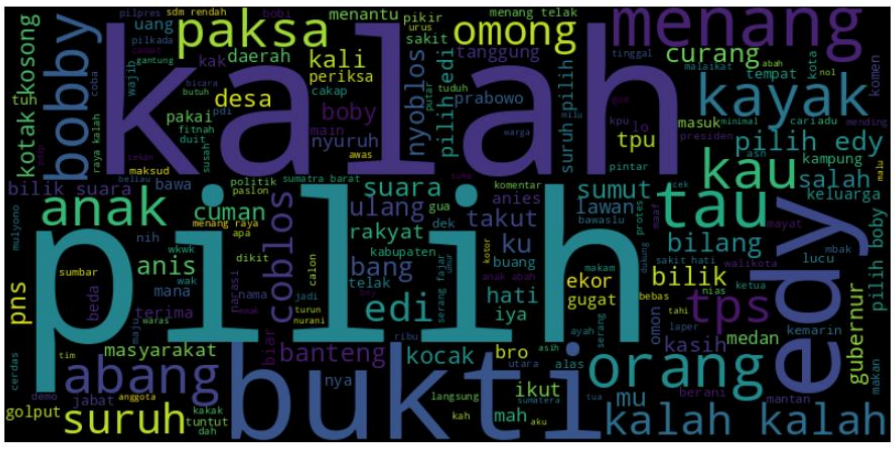


Fig 6. Negative Wordcloud

2. Positive Sentiment Wordcloud

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6543

Submitted : Jul 17, 2025
Accepted : Jun 22, 2025
Published : Jul 23, 2025

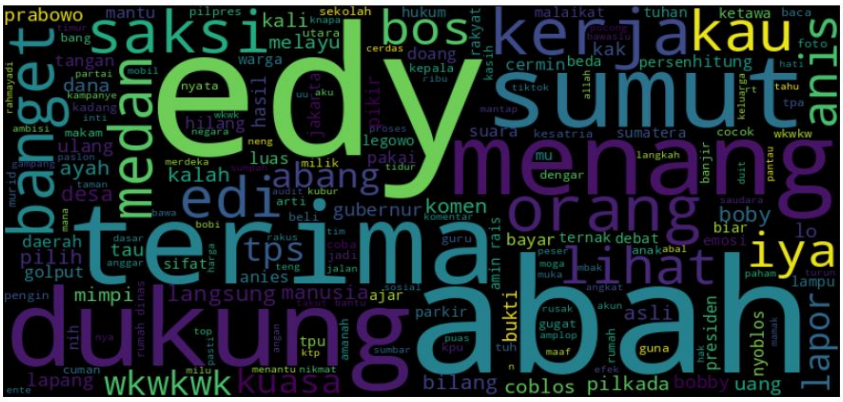


Fig 7. Positive Wordcloud

Overall, the findings from the wordcloud support the classification results of the Support Vector Machine (SVM) algorithm used in the research. The SVM's ability to recognize and classify positive comments seems to be in line with the words that predominantly appear in the visualization. Thus, this sentiment analysis approach not only succeeds in revealing the emotional tendencies of the public, but also opens up space for understanding the dynamics of public opinion more broadly and contextually.

DISCUSSIONS

According to (Fahlevvi, 2022) sentiment analysis is the process of analyzing a person's opinion on a topic to determine whether the opinion is positive or negative. In this study, the performance evaluation of the sentiment classification model for the 2024 North Sumatra Governor Election (Pilgub Sumut) on social media, which was built using the Support Vector Machine (SVM) algorithm, was carried out using a confusion matrix. This confusion matrix summarizes the number of correct and incorrect predictions generated by the model for each sentiment class, namely positive sentiment and negative sentiment.

```python
def plot_confusion_matrix(model_name, y_true, y_pred):
    labels = ['positive',  'negative']
    cm = confusion_matrix(y_true, y_pred, labels=labels)

    plt.figure(figsize=(8, 6))
    sns.heatmap(cm, annot=True, fmt='d', cmap='Blues',
                xticklabels=labels,
                yticklabels=labels)

    plt.title(f'Confusion Matrix - {model_name}')
    plt.xlabel('Predicted Label')
    plt.ylabel('True Label')
    plt.show()

plot_confusion_matrix('SVM', y_test, y_pred_svm)
```

Fig 8. Model Evaluation Code

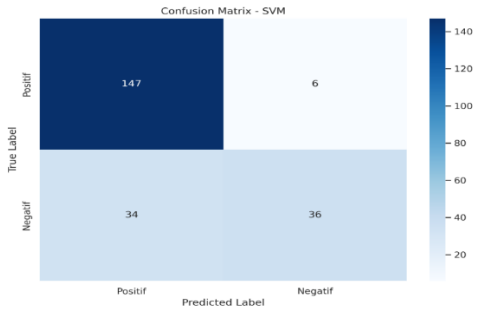


Fig 9. Confusion Matrix Visualization

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6543

Submitted : Jul 17, 2025
Accepted : Jun 22, 2025
Published : Jul 23, 2025

The following are the results of the confusion matrix evaluation values of the SVM classification model used in sentiment analysis of Twitter data for the 2024 Indonesian presidential candidates.

Table 5
Confusion Matrix SVM Model

Predicted	Positive	Negative	Total
Positive	147 (TP)	6 (FN)	153
Negative	34 (FP)	36 (TN)	70
Total	181	42	223

Description:
- True Positive (TP) = 147, Positive data that is correctly classified as positive.
- True Negative (TN) = 36, Negative data that is correctly classified as negative.
- False Positive (FP) = 34, Negative data that is incorrectly classified as positive.
- False Negative (FN) = 6, Positive data that is incorrectly classified as negative.

Visualization of these results is shown in Figure 7 in the form of a confusion matrix heatmap, which shows the distribution of classifications visually, making it easier to understand the prediction errors that occur.

To measure the performance of the sentiment classification model used in this study, an evaluation was conducted using several common metrics in binary classification, namely accuracy, precision, recall, and F1-score. The calculation is done based on the confusion matrix results that have been obtained from testing the SVM model on 223 test data. The following is the formula for each metric:

a. Accuracy

$$Accuracy = \frac{TP + TN}{Total\ Data} = \frac{147 + 36}{223} = \frac{183}{223} = 82.06\%$$

b. Precision Positif

$$Accuracy = \frac{TP}{TP + FP} = \frac{147}{147 + 34} = \frac{147}{223} = 81.21\%$$

c. Precision Negatif

$$Accuracy = \frac{TN}{TN + FN} = \frac{36}{36 + 6} = \frac{36}{42} = 85.71\%$$

d. Recall Positif

$$Accuracy = \frac{TP}{TP + FN} = \frac{147}{147 + 6} = \frac{147}{153} = 96.08\%$$

e. Recall Negatif

$$Accuracy = \frac{TN}{TN + FP} = \frac{36}{36 + 34} = \frac{36}{70} = 51.43\%$$

f. F1 Score Positif

$$Accuracy = 2x\frac{Precission\ x\ Recall}{Precission + Recall} = \frac{1.5601}{1.7729} = 88.01\%$$

g. F1 Score Negatif

$$Accuracy = 2x\frac{Precission\ x\ Recall}{Precission + Recall} = \frac{0.8818}{1.3714} = 64\%$$

Table 6
SVM Model Performance Evaluation Results

Sentimen	Precission	Recall	F1 Score	Total
Positif	0.81	0.96	0.88	153
Negatif	0.86	0.51	0.64	70
Akurasi			0.82	223
Macro Avg	0.83	0.74	0.76	
Weighted Avg	0.83	0.82	0.81	

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6543

Submitted : Jul 17, 2025
Accepted : Jun 22, 2025
Published : Jul 23, 2025

Based on the evaluation results of the classification model using the Support Vector Machine (SVM) algorithm, the following evaluation metrics were obtained: 82% accuracy, precision for the positive class of 0.81, recall of 0.96, and F1-score of 0.88. Meanwhile, for the negative class, the precision reached 0.86, but the recall was only 0.51, with an F1-score value of 0.64.

The high recall value on positive sentiment shows that the model is very good at recognizing comments that are supportive of the North Sumatra Pilgub 2024 candidate pair. This can be interpreted that positive comments on TikTok generally have more explicit and consistent language patterns, making them easier to recognize by SVM-based classification models. Conversely, the low recall value in the negative class indicates the weakness of the model in detecting comments with criticism or dissatisfaction. This could be due to the use of more varied, emotional, or implicit language in negative comments, such as sarcasm, irony, or slang that is common on the TikTok platform.

This phenomenon is in line with the findings of (Fathiarahma et al., 2023), which explain that the SVM algorithm tends to be more effective in classifying text data that has clear linguistic patterns, but is less accurate against data containing complex emotional context. In addition, research by (Arsi & Waluyo, 2021) shows that negative sentiments in the political context on social media tend to be difficult to classify automatically because they are often conveyed in the form of innuendo or metaphors that are not explicit.

In the context of the North Sumatra Pilgub 2024, the model's success in recognizing positive sentiments indicates the dominance of supportive narratives for one candidate pair in the digital space, which most likely comes from active supporter accounts. However, the low recall of negative comments suggests that public perceptions that are critical of issues such as dynastic politics, alleged fraud, or abuse of authority, are less optimally detected by the system.

The classifier's lower performance in detecting negative sentiment may be attributed to the subtleties of sarcasm and indirect criticism often found in TikTok captions or comments. This suggests that SVM with polynomial kernels might require enhancements such as sentiment lexicon integration or ensemble hybrid models.

This study also highlights the dominance of positive sentiment, which may be influenced by algorithmic amplification in TikTok's For You page—indicating not only actual public opinion but also platform-driven visibility.

CONCLUSION

This study successfully implemented the Support Vector Machine (SVM) algorithm with a polynomial kernel to classify public sentiment towards the 2024 North Sumatra gubernatorial election using TikTok comment data. The model achieved an accuracy of 82%, with strong performance in detecting positive sentiments but limited ability to capture negative sentiments particularly those expressed through sarcasm, irony, or informal language. These results affirm that SVM is effective in processing high-dimensional textual data with explicit linguistic patterns but reveals its limitations in handling complex emotional expressions common in political discourse on social media.

Methodologically, the study demonstrates the viability of combining TF-IDF and SVM for sentiment analysis in the context of regional politics. However, the results also highlight a gap in the model's sensitivity toward nuanced language, indicating a need for future research to explore more context-aware algorithms such as LSTM or transformer-based models. Incorporating data from various social media platforms and applying manual or semi-supervised sentiment labeling may further enhance classification accuracy and generalizability.

In terms of practical implications, this research underscores the potential of sentiment analysis as a strategic tool for stakeholders such as election organizers, campaign teams, and policymakers. By monitoring public sentiment in real time, decision-makers can respond proactively to emerging narratives, assess public trust, and promote more inclusive and transparent political communication. Additionally, the findings point to the need for digital literacy initiatives to ensure that online political engagement remains constructive and less susceptible to misinformation or emotional polarization.

While this study provides valuable insight into the digital political climate in North Sumatra, it remains limited by its focus on a single platform and binary sentiment classification. Future work should aim to capture more diverse expressions of opinion and provide richer contextual understanding, thereby supporting a more holistic interpretation of public perception in the digital era.

REFERENCES

Arsi, P., & Waluyo, R. (2021). Analisis Sentimen Wacana Pemindahan Ibu Kota Indonesia Menggunakan Algoritma Support Vector Machine (SVM). *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *8*(1), 147.

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 3, July 2025
https://doi.org/10.47709/cnahpc.v7i3.6543

Submitted : Jul 17, 2025
Accepted : Jun 22, 2025
Published : Jul 23, 2025

https://doi.org/10.25126/jtiik.0813944

Aulia, T. M. P., Arifin, N., & Mayasari, R. (2021). Perbandingan Kernel Support Vector Machine (Svm) Dalam Penerapan Analisis Sentimen Vaksinisasi Covid-19. *SINTECH (Science and Information Technology) Journal*, *4*(2), 139–145. https://doi.org/10.31598/sintechjournal.v4i2.762

Fahlevvi, M. R. (2022). Analisis Sentimen Terhadap Ulasan Aplikasi Pejabat Pengelola Informasi Dan Dokumentasi Kementerian Dalam Negeri Republik Indonesia Di Google Playstore Menggunakan Metode Support Vector Machine. *Jurnal Teknologi Dan Komunikasi Pemerintahan*, *4*(1), 1–13. https://doi.org/10.33701/jtkp.v4i1.2701

Fathiarahma, A., Voutama, A., Ridwan, T., & Heryana, N. (2023). Analisis Text Mining Klasifikasi Kegiatan Keluarga menggunakan Orange dengan Metode Naive Bayes. *Jurnal Teknologi Terpadu*, *9*(1), 35–41. https://doi.org/10.54914/jtt.v9i1.606

Furqan, M., Sriani, S., & Sari, S. M. (2022). Analisis Sentimen Menggunakan K-Nearest Neighbor Terhadap New Normal Masa Covid-19 Di Indonesia. *Techno.Com*, *21*(1), 51–60. https://doi.org/10.33633/tc.v21i1.5446

Insani, S. C., Khuzaimah, N. A. Z., Maryadi, V. Z. D., & Hafizha, T. A. (2023). Meninjau Etika Masyarakat Indonesia Dalam Bermedia Sosial Di Masa Pemilu Menggunakan Etika Media Sosial. *Nusantara: Jurnal Pendidikan, Seni, Sains Dan Sosial Humanioral*, *1:2*(September), 1–25. https://doi.org/10.11111/nusantara.xxxxxxx

Literat, I., & Kligler-Vilenchik, N. (2023). TikTok as a Key Platform for Youth Political Expression: Reflecting on the Opportunities and Stakes Involved. *Social Media and Society*, *9*(1). https://doi.org/10.1177/20563051231157595

Nur, A. A., Dyah, A. G. R., Salima, H. N. S., & Rafi, A. M. A. (2023). Prospek Penggunaan Tiktok Sebagai Instrumen Politik Pada Pemilihan Umum 2024 Penulis. In *Laboratorium Indonesia 2045* (Issue 2).

Puspitarini, D. S., & Nuraeni, R. (2019). Pemanfaatan Media Sosial Sebagai Media Promosi. *Jurnal Common*, *3*(1), 71–80. https://doi.org/10.34010/common.v3i1.1950

Rabbani, S., Safitri, D., Rahmadhani, N., Sani, A. A. F., & Anam, M. K. (2023). Perbandingan Evaluasi Kernel SVM untuk Klasifikasi Sentimen dalam Analisis Kenaikan Harga BBM. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *3*(2), 153–160. https://doi.org/10.57152/malcom.v3i2.897

Rifaldi, D., Abdul Fadlil, & Herman. (2023). Teknik Preprocessing Pada Text Mining Menggunakan Data Tweet "Mental Health." *Decode: Jurnal Pendidikan Teknologi Informasi*, *3*(2), 161–171. https://doi.org/10.51454/decode.v3i2.131

Romzi, M., & Kurniawan, B. (2020). Pembelajaran Pemrograman Python Dengan Pendekatan Logika Algoritma. *JTIM: Jurnal Teknik Informatika Mahakarya*, *03*(2), 37–44.

Septian, J. A., Fachrudin, T. M., & Nugroho, A. (2019). Analisis Sentimen Pengguna Twitter Terhadap Polemik Persepakbolaan Indonesia Menggunakan Pembobotan TF-IDF dan K-Nearest Neighbor. *Journal of Intelligent System and Computation*, *1*(1), 43–49. https://doi.org/10.52985/insyst.v1i1.36

Shevira, S., Suarjaya, I. M. A. D., & Buana, P. W. (2022). Pengaruh Kombinasi dan Urutan Pre-Processing pada Tweets Bahasa Indonesia. *JITTER : Jurnal Ilmiah Teknologi Dan Komputer*, *3*(2), 1074. https://doi.org/10.24843/jtrti.2022.v03.i02.p06

Tarihoran, Y., & Manurip, K. J. (2018). Analisis Sentimen Pemilihan Gubernur Jawa Barat Tahun 2018 Dengan Aplikasi Twitter Menggunakan Metode Naïve Bayesian Classification. *TeIKa*, *8*(1), 99–105. https://doi.org/10.36342/teika.v8i1.2243

Utami, D. S., & Erfina, A. (2021). Analisis Sentimen Pinjaman Online Di Twitter Menggunakan Algoritma Support Vector Machine (Svm). *SISMATIK (Seminar Nasional Sistem Informasi Dan Manajemen Informatika)*, *1*, 299–305.

Zamora-medina, R., Suminas, A., & Fahmy, S. S. (2023). Securing the Youth Vote: A Comparative Analysis of Digital Persuasion on TikTok Among Political Actors. *Media and Communication*, *11*(2), 218–231. https://doi.org/10.17645/mac.v11i2.6348

* Corresponding author

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.