

Implementation of the Naive Bayes Algorithm for Public Sentiment Analysis Toward Power Plant Development

Vithan Rasyid^{1*}, Ruuhwan²

^{1,2} Informatics Engineering Study Program, Faculty of Engineering, Universitas Perjuangan Tasikmalaya, Tasikmalaya, Indonesia

ARTICLE INFO

Article history:

Initial submission
30-04-2026
Received in revised form
16-05-2026
Accepted 10-06-2026
Available online 30-09-2026

Keywords:

*Sentiment Analysis,
Multinomial Naive Bayes,
Power Plant Development,
Data Preprocessing,
Twitter*

DOI:

<https://doi.org/10.59356/smart-techno.v8i2.216>

ABSTRACT

This study aims to analyze public sentiment toward power plant development using data collected from the social media platform Twitter. The dataset consisted of 2,493 tweets obtained through a crawling process using keywords related to power plant development, such as "PLTS" and "PLTS Cirata." The preprocessing stage included cleaning, case folding, stopword removal, tokenization, and stemming using the Sastrawi library to produce more structured textual data. The dataset was then divided into a training set comprising 85% of the data (2,120 tweets) and a testing set comprising 15% (373 tweets). The classification process was performed using the Multinomial Naive Bayes algorithm, as this method is well suited for text data represented by word-frequency features extracted through CountVectorizer. Model evaluation was conducted using a confusion matrix with accuracy, precision, recall, and F1-score as performance metrics. The results showed that the model achieved an accuracy of 75%, indicating that the Multinomial Naive Bayes method is reasonably effective for text-based sentiment classification. Furthermore, the findings revealed that public opinion regarding power

plant development is influenced by perceptions of renewable energy benefits, environmental impacts, and government policies.

1. INTRODUCTION

The development of digital technology and social media has transformed the way people express their opinions on various public issues. One of the most actively used social media platforms for sharing opinions is Twitter. This platform allows users to express their views openly and in real time, generating a large volume of textual data that can be utilized for various data analysis purposes (Munzir, 2019). In the context of energy infrastructure development, particularly power plant construction, public perception is an important factor that can influence the successful implementation of projects. Public support for or opposition to power plant development is often reflected in comments posted on social media platforms. Therefore, a computational approach is required to systematically process and interpret public opinions.

Sentiment analysis is a text mining technique used to identify positive and negative opinion tendencies within a collection of textual data (Darwis et al., 2021). By utilizing sentiment analysis techniques, public opinions regarding a particular policy or development project can be analyzed automatically and efficiently. One of the classification methods commonly used in sentiment analysis is the Naive Bayes algorithm, as it offers relatively good accuracy and a simple computational process (Naraswati et al., 2021). Several previous studies have applied the Naive Bayes algorithm to sentiment analysis in various domains. A study conducted by Azzahra et al. (2022) on sentiment analysis of the Karang Resik tourism park using the Naive Bayes method achieved an accuracy rate of 84%. Another study by Sianipar et al. (2023) on sentiment analysis of the Jakarta-Bandung high-speed railway development using the Naive Bayes method demonstrated that the method was sufficiently effective in classifying public opinions regarding infrastructure development.

Based on previous studies, the Naive Bayes algorithm has strong potential for application in text-based sentiment classification. Therefore, this study aims to implement the Naive Bayes algorithm to analyze public sentiment toward power plant development based on data collected from Twitter.

2. LITERATURE REVIEW

Sentiment Analysis

Sentiment analysis is a branch of text mining that is used to identify, classify, and evaluate public opinions toward a particular object or topic based on textual data (Darwis et al., 2021). Sentiment analysis is widely applied to social media data because social media has become one of the primary platforms through which people express their opinions freely and rapidly.

Naive Bayes

Naive Bayes is a probabilistic classification method based on Bayes' Theorem, which assumes that all attributes are independent of one another. The main advantages of this method are its simplicity, computational efficiency, and ability to achieve good performance on text-based datasets (Permadi, 2020). In this study, the Multinomial Naive Bayes algorithm was employed because it is more suitable for textual data represented by word-frequency features generated through the CountVectorizer process. Multinomial Naive Bayes is capable of handling discrete data in the form of word occurrence counts within documents, making it one of the most commonly used algorithms for text classification and sentiment analysis. In contrast, Bernoulli Naive Bayes is more appropriate for binary data that only considers the presence or absence of words, while Categorical Naive Bayes is generally applied to non-text categorical data.

Previous Studies

A study conducted by Azzahra et al. (2023) on sentiment analysis of the Karang Resik tourism park using the Naive Bayes method achieved an accuracy rate of 84%. The confusion matrix evaluation produced a precision score of 62%, a recall score of 85%, and an F1-score of 72%. Another study conducted by Sundari et al. (2023) investigated the implementation of Naive Bayes for Twitter sentiment analysis related to stock price movements. The study achieved an accuracy rate of 67% using the Stratified K-Fold Cross-Validation method.

Hidayat and Sugiyono (2023) conducted research on sentiment analysis of movie reviews using a Naive Bayes Classifier combined with Chi-Square feature selection. The study achieved a positive sentiment classification accuracy of 56.40% and a negative sentiment classification accuracy of 64.40%. Based on these previous studies, the Naive Bayes algorithm has demonstrated considerable effectiveness in sentiment classification tasks. Therefore, this study implements the Multinomial Naive Bayes algorithm to analyze public sentiment toward power plant development using data collected from Twitter.

3. METHOD

The research methodology outlines the steps undertaken throughout the study. This process involves a series of stages, including literature review, data collection and exploration, data preprocessing, data labeling, dataset splitting, classification, model evaluation, and result testing. By following these stages, the researchers aim to establish a robust research framework and ensure the accuracy and validity of the interpretation of the research findings.

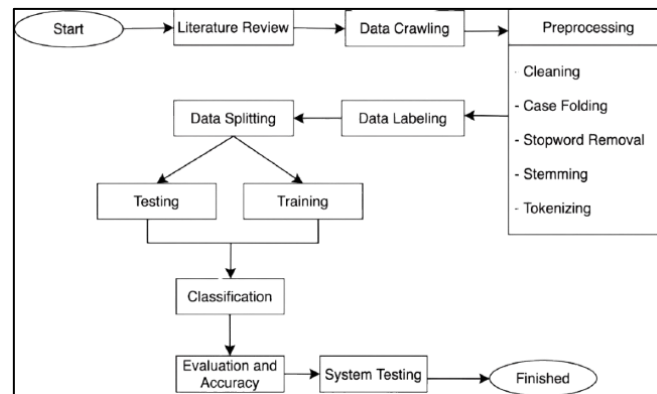


Figure 1. Research Workflow

Literature Review

The literature review involves a comprehensive analysis of various scholarly works relevant to the research topic. This stage enables researchers to gain a deeper understanding of related issues, evaluate previous findings, and identify knowledge gaps that require further investigation. Therefore, a literature review serves not only as a process of collecting references but also as a critical initial stage in developing the theoretical framework and research methodology.

Data Crawling

Data were collected using the Twitter API with the assistance of Python libraries. The crawling process focused on keywords related to power plant projects, such as “PLTS” (Solar Power Plant), “PLTS Cirata”, and other related terms. The collected data were subsequently filtered to ensure their relevance to the research topic.

Data Preprocessing

Preprocessing aims to clean and transform raw data into a format that can improve classification performance. The preprocessing stage in this study consisted of the following steps:

a. Cleaning

Data cleaning is a process intended to remove inconsistencies and improve dataset quality. This stage involves removing punctuation marks, hashtags, URLs, symbols, numbers, and attributes with missing or irrelevant values. As a result, the cleaned dataset becomes more suitable for subsequent analysis (Nofiyani & Wulandari, 2022).

b. Case Folding

Case folding is a text preprocessing technique that ensures consistency in letter capitalization throughout the dataset. This process converts all characters in the text into a uniform format, typically lowercase, regardless of their original capitalization (Azzahra et al., 2022).

c. Stopword Removal

Stopword removal is the process of filtering out common words that contribute little to the semantic meaning of a text. In Indonesian text processing, stopwords removal can be performed using manually created dictionaries or libraries such as Sastrawi (Fitriani & Wibowo, 2023).

d. Tokenization

Tokenization is the process of breaking text into smaller units called tokens, which represent the smallest meaningful elements of textual information. For example, the sentence “Saya sedang belajar pemrosesan bahasa alami” (“I am learning natural language processing”) can be divided into tokens such as “Saya,” “sedang,” “belajar,” “pemrosesan,” “bahasa,” and “alami.” Each token can then be processed independently during text analysis (Sistim et al., 2023).

e. Stemming

Stemming is the process of reducing words to their root forms by removing prefixes and suffixes. In this study, stemming was performed using the Sastrawi library, which is specifically designed for Indonesian-language text processing in Python (Nazief & Adriani, 2021).

Data Labeling

During the labeling process, the system identifies words containing positive and negative sentiments within the collected Twitter data. Labeling was performed automatically using Python by utilizing a sentiment lexicon as the classification reference. Tweets dominated by positive sentiment words were assigned a positive label, whereas tweets dominated by negative sentiment words were assigned a negative label. This approach facilitates the labeling of large-scale datasets, enabling a more efficient and consistent classification process.

Data Splitting

The dataset was divided into two subsets: training data and testing data. The training dataset was used to build and train the classification model, while the testing dataset was used to evaluate the model's performance.

Classification

Before classification, the dataset was split into training and testing sets using an 85:15 ratio, where 85% of the data were allocated for training and 15% for testing. In this study, sentiment classification was performed using the Naive Bayes method. The preprocessed training data were used to calculate probability values required by the algorithm. Subsequently, the trained model was applied to the testing data to predict sentiment categories.

Model Evaluation

Model evaluation was conducted to assess the quality and performance of the developed classification model. Several commonly used evaluation metrics were employed in this study, including:

- a. Accuracy
- b. Precision
- c. Recall (Sensitivity)
- d. F1-Score

These metrics were calculated based on the confusion matrix generated from the classification results.

Testing

The testing phase was carried out to evaluate the ability of the Naive Bayes model to classify previously unseen data. Through this process, the study assessed the effectiveness and predictive performance of the model in sentiment classification tasks based on new Twitter data.

4. RESULT AND DISCUSSION

Data Crawling

In this study, a dataset consisting of 2,493 tweets related to the topic of power plant development was utilized. The data were collected using several relevant keywords, namely "pltscirata" and "plts." Each tweet in the dataset was assigned to one of two sentiment categories, positive or negative, based on the sentiment expressed in the text. Table 1 presents a sample of the data obtained through the crawling process

Table 1. Crawled Data

Created_at	Full_Text	Username
Wed Nov 29 08:52:30 +0000 2023	Presiden Joko Widodo meresmikan operasi Pembrik Tenaga Surya (PLTS) Terapung Cirata berkapasitas 192 MWp yang berlokasi di Purwakarta, Jawa Barat pada Kamis (9/11). #PLTSTerapungCirata https://t.co/NxYxo2Msnp	@KabarBeta
Wed Nov 29 03:35:53 +0000 2023	dalam skala besar akhirnya bisa terlaksana. Indonesia berhasil membangun PLTS terapung terbesar di Asia Tenggara dan nomor 3 di dunia. PLTS Terapung di Cirata ini adalah hasil kerja sama Kementerian ESDM, Kementerian BUMN, bersama PLN dan kolaborasi dengan kekuatan dunia yaitu	@BUMNJABAR _HEBAT

Data Preprocessing

The preprocessing stage was performed on the dataset obtained from the crawling process. This step was carried out to enable the computer system to better understand the structure of the dataset and to transform unstructured data into a more organized and structured format. The preprocessing process consisted of five stages: cleaning, case folding, stopword removal, tokenization, and stemming. Prior to preprocessing, the dataset contained various special characters, symbols, numbers, URLs, hashtags, and unnecessary whitespace. Therefore, preprocessing was necessary to remove these irrelevant elements and improve the quality of the textual data. The cleaning process eliminated special symbols, numerical characters, and other non-essential components, resulting in a more consistent and analyzable dataset. The results of the preprocessing stages are illustrated in the following tables.

Table 2. Preprocessing

Before	After
@Uki23 Gagasan gibran selama ini	Uki Gagasan gibran selama ini
@bonex168 @nantihujan	bonex nantihujan doktermarket
@doktermarket Lumayan gan, tp	Lumayan tp harus nabung juga buat
harus nabung juga buat siap2 ganti	siap ganti baterai PLTS Namanya
baterai PLTS gan 😊.	baterai pasti ada umur nya
Namanya baterai pasti ada umur cycle nya.	

Table 2 presents the results of the cleaning process performed on the tweet dataset. During this stage, unnecessary elements such as mentions (@), emojis, URLs, punctuation marks, numerical characters, and other symbols were removed. The cleaning process aims to produce a cleaner dataset, thereby improving the quality of the data for subsequent analysis. The next step is case folding, which involves converting all uppercase letters into lowercase letters. This process is carried out to ensure consistency in text processing and to prevent words with different letter cases from being treated as separate terms during analysis.

Table 3. Case Folding

Before			After
PLTS	TERAPUNG	TERBESAR	plts terapung terbesar di asia
DI	ASIA	TENGGARA	tenggara resmi beroperasi
BEROPERASI			
DaKh	Perimeter	PAGAR	dakh perimeter pagar pengaman
PENGAMAN	ASET	ANDA	aset anda pagar brc cw cw pintu
PAGAR			kawat durisilet dch
BRC CW	CW	pintu kawat durisilet	
DCH			

Table 3 presents the results of the case folding process, in which all uppercase letters are converted into lowercase letters. This stage is performed to standardize word representations so that words with the same meaning but different capitalization patterns can be recognized as the same feature by the system. The next stage is stopwords removal, which involves filtering out common words that are considered less important and do not contribute significant meaning to the text. This process was carried out using both a manually constructed stopwords dictionary and the Sastrawi library.

Table 4. Stopwords Removal

Before		After
Kami	sangat senang melihat	Senang melihat
perkembangan	proyek	perkembangan proyek
pembangunan	PLTS di	pembangunan PLTS
daerah kami.		daerah.
Semoga	proyek ini dapat	Semoga proyek ini
memberikan	manfaat besar	memberikan manfaat
bagi masyarakat		bagi masyarakat

Table 4 presents the results of the stopwords removal process, which was performed to eliminate common words that do not have a significant influence on sentiment classification. By removing unimportant words, the system can focus more effectively on terms that carry meaningful information and stronger sentiment values. After the stopwords removal stage, the next step is tokenization, in which sentences are split into individual words based on whitespace separators. This process facilitates text processing and classification by transforming the text into smaller units called tokens.

Table. 5 Tokenization

Before	After
plts terapung terbesar di asia tenggara resmi beroperasi	plts,terapung,terbesar,di,asia,tenggara,resmi,beroperasi
pln optimis plts terapung cirata mampu menjadi etalase percepatan transisi energi	pln,optimis,plts,terapung,cirata,mampu,menjadi,etalase,percepatan,transisi,energi

Table 5 presents the results of the tokenization process, in which sentences are divided into individual tokens or single words. This process aims to facilitate feature extraction and the calculation of word frequencies, which are subsequently used during the classification stage. The next step is stemming, which involves removing affixes from words and converting them

into their root forms. In this study, the stemming process was performed using the Sastrawi library, which is specifically designed for Indonesian language text processing.

Table. 6 Stemming

Before		After	
melongok',	'plts',	longok	plts
'terapung',	'sembcorp',	sembcorp	tengeh
'tengeh',	'singapura',	singapura	tara lapang
'setara',	'lapangan',	sepak tempotekno	
'sepak',	'tempotekno		

Table 6 presents the results of the stemming process, in which words containing affixes are transformed into their root forms. This stage aims to reduce variations of words that convey the same meaning, thereby simplifying the feature space and improving the effectiveness of the classification process.

Data Labeling

The data labeling process was carried out automatically using the Python programming language by utilizing a sentiment lexicon containing positive and negative words as the basis for classification. At this stage, the system identifies words with positive and negative sentiment values within each tweet to determine its sentiment category. Tweets containing a predominance of positive sentiment words were assigned a positive label, while tweets containing a predominance of negative sentiment words were assigned a negative label. This approach was adopted to facilitate the labeling of large-scale datasets, enabling the sentiment classification process to be performed more efficiently and consistently.

	stemming_data	sentimen_scores	sentimen
10	hasil plts pasang hidrogen hijau renewable ene...	0.2732	Positif
26	pt usaha listrik negara nusantara power pln np...	0.3400	Positif
30	creatinine mgdl hb gdl gpt ul platelets hb plts	0.1027	Positif
32	dirut pln np target fase plts mw ikn operasi f...	0.3400	Positif
34	kom ridwan pdiperjuangan sulsel musim kemarau ...	0.2732	Positif
46	sumber bangkit listrik tenaga surya plts bangk...	0.2732	Positif
62	nihardesai brownjhm nicogagelmann peteritebimi...	-0.4939	Negatif
68	project plts wp inverter smart inverter solar ...	0.6908	Positif
77	listrik tenaga surya plts bangkit hasil plts p...	0.2732	Positif
82	bangkit listrik tenaga surya plts bangkit hasi...	0.2732	Positif
83	jalan perintah capai zero emissions nze ghp pl...	0.2732	Positif
102	imananda reddingtn iyaa udah papan tanah kai d...	-0.2732	Negatif
114	sumber bangkit listrik tenaga surya plts bangk...	0.2732	Positif
134	kemaren fgd bangun plts sulteng alas bangun si...	-0.2732	Negatif
142	plts cunt coord xoxo httpscoujdberdecl	0.2023	Positif
145	pakai plts pln np bangkit ebt tanah	0.3400	Positif

Figure 2. Data Labeling Process

Figure 2 illustrates the data labeling process performed automatically using positive and negative sentiment lexicons. At this stage, each tweet is analyzed based on the sentiment-bearing words it contains to determine whether it belongs to the positive or negative sentiment category. The resulting labels are then used as input data for the classification process using the Naive Bayes algorithm. Following the data labeling stage, data visualization is conducted to facilitate the analysis and interpretation of sentiment information contained within the textual



Figure 5. Positive Sentiment Word Cloud

Figure 5 presents the word cloud visualization for the positive sentiment category. Words displayed in larger sizes indicate a higher frequency of occurrence within the dataset. This visualization demonstrates that positive public sentiment is generally associated with the benefits of power plant development, support for renewable energy initiatives, and expectations of improved energy supply in the future. The word cloud highlights the most prominent terms used by the public when expressing favorable opinions regarding power plant development.

Data Splitting

The dataset was divided using an 85:15 ratio, where 85% of the data were used for training and 15% were used for testing. The Naive Bayes method was employed as the classification algorithm in this study. After the preprocessing stage, the training data were used to extract probability values required by the algorithm. Subsequently, the testing data were analyzed to determine their sentiment categories based on the probability values generated during the training process.

Classification

The classification process was implemented using a Naive Bayes model within a single pipeline consisting of the following sequence: Vectorizer → Transformer → Classifier. This process was carried out using Python 3 and several supporting libraries. The scikit-learn library was utilized for the classification process, while NumPy and Pandas were used for data handling and processing.

Naive Bayes model accuracy : 0.7459893048128342

Classification Results				
	precision	recall	f1-score	support
0	0.62	0.56	0.59	121
1	0.80	0.83	0.82	253
accuracy			0.75	374
macro avg	0.71	0.70	0.70	374
weighted avg	0.74	0.75	0.74	374

Figure 6. Model Accuracy

Evaluation

The model was analyzed to evaluate the achievement of the research objectives and its effectiveness in addressing the research problem. The evaluation was conducted using a confusion matrix to assess the model's performance in data classification. The performance

metrics employed in this study include accuracy, recall, precision, and F1-score. These metrics provide a comprehensive evaluation of the classification model by measuring its ability to correctly predict sentiment categories and its overall effectiveness in sentiment analysis.

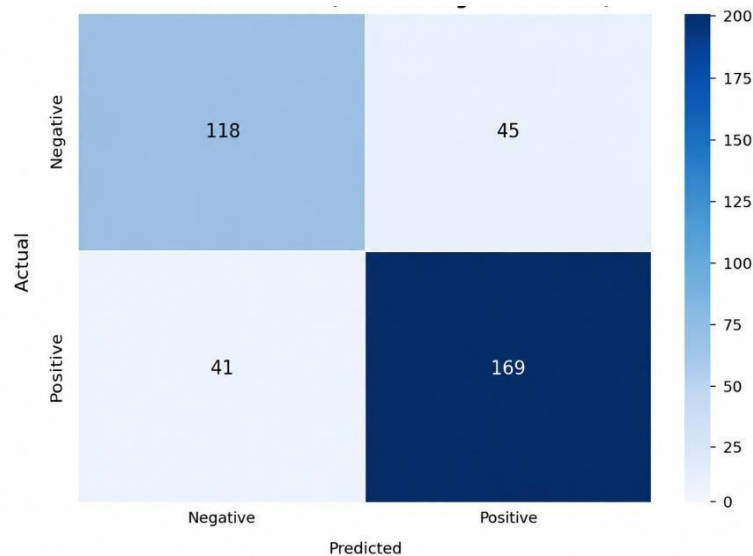


Figure 7. Confusion Matrix

The confusion matrix provides valuable insights into the performance of the classification method. This evaluation considers not only accuracy but also recall, precision, and F1-score. Accuracy measures the model's overall ability to correctly classify data instances. Recall indicates how effectively the model identifies all actual positive instances, while precision represents the proportion of instances predicted as positive that are truly positive. The F1-score, which combines precision and recall, provides a more comprehensive assessment of the model's performance.

The accuracy result of 75% indicates that the Naive Bayes algorithm demonstrates a reasonably good capability in classifying text-based sentiment. However, this performance can still be improved through optimization of the preprocessing stage and feature selection techniques. Compared with previous studies, the results obtained in this research can be categorized as moderate. This may be attributed to the characteristics of Twitter data, which are often unstructured and contain informal language, abbreviations, and symbols that can affect the classification process.

Furthermore, the analysis results indicate that positive sentiment is generally associated with the perceived benefits of power plant development, such as improved energy supply and support for renewable energy initiatives. In contrast, negative sentiment is primarily influenced by concerns regarding environmental impacts and social issues related to power plant development.

5. CONCLUSION

This study successfully implemented the Naive Bayes algorithm to analyze public sentiment toward power plant development based on data collected from the Twitter social media platform. The dataset consisted of 2,493 tweets, which were divided into 2,120 training instances and 373 testing instances using an 85:15 ratio. The experimental results demonstrated that the developed model achieved an accuracy of 75%, indicating that it is reasonably effective in classifying sentiments into positive and negative categories. Furthermore, the analysis provided valuable insights into public perceptions of power plant development, revealing a mixture of positive and negative sentiments expressed by users. These findings suggest that the Naive Bayes method can serve as an effective computational

approach for understanding public opinion and sentiment. Moreover, the results may provide useful information for policymakers and stakeholders in making informed decisions regarding infrastructure development projects.

6. RECOMMENDATIONS

Future studies are encouraged to utilize larger and continuously updated datasets in order to obtain a more comprehensive representation of public opinion. In addition, experiments involving other classification algorithms, such as Support Vector Machine (SVM), Random Forest, and Deep Learning-based methods, should be conducted for comparison purposes to identify the approach that yields the best performance. Furthermore, improvements in the preprocessing stage, feature selection techniques, and feature extraction methods may serve as potential alternatives for enhancing the accuracy and overall quality of sentiment classification in future research. Such developments are expected to contribute to more robust and reliable sentiment analysis models.

7. REFERENCES

- Azzahra, S. N., Ruuhwan, R., & Hikmatyar, M. (2023). Analisis sentimen taman wisata Karang Resik dengan metode Naive Bayes. *E-Jurnal JUSITI: Jurnal Sistem Informasi dan Teknologi Informasi*, 12(2), 200–209.
- Darwis, D., Siskawati, N., & Abidin, Z. (2021). Penerapan algoritma Naive Bayes untuk analisis sentimen review data Twitter BMKG nasional. *Jurnal Tekno Kompak*, 15(1), 131–145.
- Hidayat, F. N., & Sugiyono. (2023). Analisis sentimen masyarakat terhadap perekrutan PPPK pada Twitter dengan metode Naive Bayes dan support vector machine. *Jurnal Sains dan Teknologi*, 5(2), 665–672.
- Munzir, A. A. (2019). Beragam peran media sosial dalam dunia politik di Indonesia. *JPPUMA: Jurnal Ilmu Pemerintahan dan Sosial Politik UMA (Journal of Governance and Political Social UMA)*, 7(2), 173–182.
- Naraswati, N. P. G., et al. (2021). Analisis sentimen publik dari Twitter tentang kebijakan penanganan COVID-19 di Indonesia dengan Naive Bayes classification. *SISTEMASI*, 10(1), 222–238.
- Nofiyani, N., & Wulandari, W. (2022). Implementasi Electronic Data Processing Untuk meningkatkan Efektifitas dan Efisiensi Pada Text Mining. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, 6(3), 1621–1629.
- Permadi, V. A. (2020). Analisis sentimen menggunakan algoritma Naive Bayes terhadap review restoran di Singapura. *Jurnal Buana Informatika*, 11(2), 141–151.
- Sianipar, J. F., Ramadhan, Y. R., & Jaelani, I. (2023). Analisis sentimen pembangunan kereta cepat Jakarta–Bandung di media sosial Twitter menggunakan metode Naive Bayes. *KLIK: Kajian Ilmiah Informatika dan Komputer*, 4(1), 360–367.
- Sundari, S. S., Hidayat, C. R., & Difari, I. (2023). Implementasi Naive Bayes untuk analisis sentimen Twitter terhadap pergerakan harga saham. *INFOSYS (Information System) Journal*, 8(1), 98–110.
- Fitriani, U., & Wibowo, A. (2023). PENERAPAN ALGORITME NAIVE BAYES UNTUK MEMPREDIKSI APPLICATION OF NAIVE BAYES ALGORITHM TO PREDICT GRADUATION OF BUDI LUHUR UNIVERSITY STUDENTS BASED ON. 2(September), 745–753.
- Sistim, J., Hermawati, L., Berland, V., Rahmadiyah, A., Hutabarat, E., & Saputra, D. D. (2023). Komparasi Metode Text Mining Terhadap Masalah Pengklasifikasian Narasi Informative & Non Informative Pada twitter @ PLN _ 123. 5(1), 109–120. <https://doi.org/10.37034/jsisfotek.v4i2.191>