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

Sentiment Analysis of 2024 Election Fraud Using SVM and Naïve Bayes Algorithms

Faalih Hibban Hilmi[1], Aries Dwi Indriyanti[2]
[1,2] *Surabaya State University, Surabaya, Indonesia*
faalih.20029@mhs.unesa.ac.id, *ariesdwi@unesa.ac.id*

ABSTRACT

Elections are one of the main pillars of democracy, where the people's voice is the main determinant in government formation. Election fraud not only harms political competitors but also undermines public trust in democracy. The role of social media Twitter in widely disseminating information and disinformation adds to the challenge of maintaining election integrity. Sentiment analysis is the process of collecting and understanding individual opinions related to an event. Support Vector Machine (SVM) and Naïve Bayes algorithms are often used in this analysis due to their effectiveness and efficiency in text classification. This research aims to analyze public sentiment related to the 2024 presidential election fraud and compare the effectiveness of SVM and Naïve Bayes in sentiment classification. The study was conducted quantitatively, involving the stages of data collection, preprocessing, labeling, TF-IDF weighting, classification, and evaluation. The results of the sentiment analysis of public opinion on the 2024 presidential election fraud showed 42.5% negative sentiment, 38.6% neutral, and 18.9% positive. The dominance of negative sentiments reflects the public's concerns about election integrity. The high neutral sentiment indicates public doubt. To overcome this, transparency, strengthening supervisory institutions, electronic election technology, and strict law enforcement are needed. The SVM algorithm with RBF kernel produces 58% accuracy, better than Naïve Bayes with 51%.

Keyword: Sentiment analysis, Naïve Bayes, Support Vector Machine (SVM), Public Opinion, Fraudulent 2024 presidential election.

Article Info:
Article history:
Received December 09, 2024
Revised February 08, 2025
Accepted March 08 , 2024

Corresponding Author
Faalih Hibban Hilmi
Universitas Negeri Surabaya, Surabaya, Indonesia
Faalih.20029@mhs.unesa.ac.id

1. INTRODUCTION

Elections are one of the main pillars of democracy, where the people's voice determines the formation of the government. Elections aim to elect the President, Vice President, DPR, DPD, and DPRD who can realize a democratic state and listen to the aspirations of the people in accordance with the development of national life. Ideally, elections should be conducted with integrity, professionalism, and accountability in accordance with the laws and regulations. The success of fair and honest elections is a measure of the legitimacy of the elected government. However, issues of fraud often arise, reducing public trust in the election process.

Cases of election fraud in Indonesia, from 2004 to 2024, include money politics, voter data manipulation, data falsification, and unethical use of state resources for the campaign interests of those in power. Other forms of fraud that recur in every presidential election show serious challenges in maintaining the honesty and fairness of the election process. This fraud not only harms certain political parties but also undermines public trust in democracy [13].

The 2024 Presidential Election has received sharp attention due to concerns about potential fraud. The rapid development of information technology and social media allows public opinion on election issues to spread quickly and widely. Social media accelerates the spread of information and disinformation, influencing public perception. This condition adds complexity to maintaining the integrity of elections. Therefore, strict supervision and transparency at every stage of the election are very important to ensure fairness and maintain public trust in democracy.

Sentiment analysis helps understand public opinion on certain issues, including election fraud, especially through social media. This process uses machine learning techniques, such as Support Vector Machine (SVM) and Naive Bayes algorithms, which are popular in text classification. Naive Bayes excels in simplicity and efficiency when handling data with many features, while SVM is effective in handling high-dimensional data with the ability to minimize overfitting [7].

This study aims to analyze public sentiment on the issue of election fraud in the 2024 presidential election using Support Vector Machine (SVM) and Naive Bayes algorithms. Naive Bayes and Support Vector Machines (SVM) algorithms are often chosen because of their respective advantages in handling text classification. Naive Bayes, with its probabilistic approach, provides simplicity and speed when processing text data with many features but little training data, and remains effective even with irrelevant features [2]. SVM, on the other hand, is known for its strong performance in classification (especially when there are clear boundaries between classes) and its ability to reduce the risk of overfitting and handle high-dimensional data sets. By using kernel functions, SVM is also able to handle non-linearly separable data. The combination of these two algorithms allows researchers to combine the speed and simplicity of Naive Bayes with the accuracy and flexibility of Support Vector Machines [8], thereby improving the overall performance of sentiment analysis models and supporting better decision-making. This study will collect data from social media during the election period, then analyze the data to identify dominant sentiment patterns. Thus, the results of this study are expected to provide a comprehensive picture of public perceptions related to election fraud and serve as a consideration for policymakers to improve transparency and fairness in the election process.

2. METHODS

The research method used in this study is a quantitative research method to determine how much positive and negative sentiment is reflected in the current data set. The main focus of this research is to explore and understand public opinion related to alleged fraud issues in the presidential election process. There is a research flow that explains the stages that discuss the general description of the research work from the initial stage to the final stage that researchers must go through to achieve the research objectives.

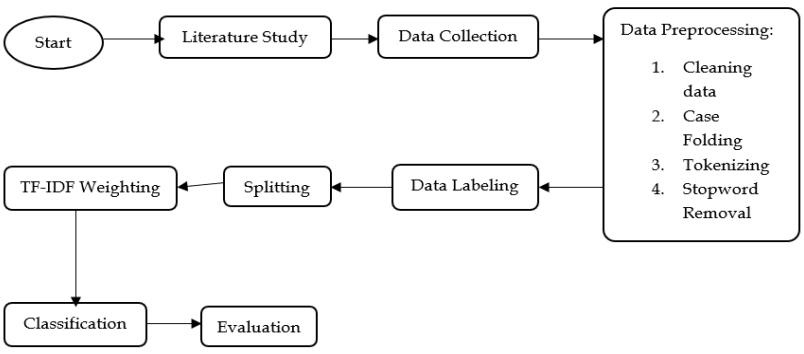

Figure 1 Flow of Research

2.1 Literature Study

In this literature study, researchers collect and analyze various sources relevant to this research topic. In the process, researchers gather various sources such as books, journals, or other sources, then evaluate their quality and relevance. The main purpose of this literature study is to provide a strong theoretical foundation, understanding related to the research, and to identify problems from previous research.

2.2 Data Collection

In the data collection stage, data crawling from Twitter is carried out using the Python programming language implemented on the Google Colab platform. This process includes using Python libraries to access the Twitter API, downloading relevant tweets related to public opinion on the 2024 presidential election fraud, and storing and managing the data for further analysis. Data is collected from July 2023 to July 2024, by entering an auth token for API access and using specific keywords. Data collection is limited to 500 tweets per day, and the results are stored in .CSV file format, utilizing the cloud computing power provided by Google Colab to streamline this process.

2.3 Data Preprocessing

In the data preprocessing stage, the raw dataset obtained from Twitter will be processed to transform the raw data into data ready for classification. The purpose of data preprocessing is to clean the data from unnecessary words, standardize lowercase or uppercase letters, remove numbers and emoticons or punctuation, convert sentences into words, and so on. The stages carried out for data preprocessing are:

1. Cleaning data involves removing unnecessary words or attributes from the dataset for sentiment analysis, such as mentions, hashtags, URLs, characters, or symbols, which will be replaced with spaces.
2. Case Folding is the process of converting all letters in the text to lowercase. This process ensures that capitalization differences do not affect the analysis and is an important step in text normalization.
3. Tokenizing is the process of splitting each word and arranging them into single pieces. The words in the data are separated by spaces. The result of tokenizing can be entered into the database for weighting purposes.
4. Stopword Removal is the process of removing common words that often appear but do not have significant meaning in text analysis.
5. Stemming is the process of converting words with affixes into their base form.

2.4 Data Labeling

In the data labeling stage, after the data from Twitter related to public sentiment towards the 2024 presidential election fraud is collected, it will be grouped into three tendencies: positive, neutral, and negative.

2.5 Data

Splitting data is the process of dividing data into two or more subsets, with one part used for testing data and the other part used for training the model. The purpose of splitting data is

to objectively evaluate the model's performance and ensure that the model can be generalized well [6].

2.6 TF-IDF Weighting

In this stage, the process of weighting words is carried out to assign values to each word. This study uses the TF-IDF method because this technique is considered more effective in calculations. The TF-IDF method is a weighting technique that measures the relationship between a word (term) and a document. This approach integrates two main ideas: how often a word appears in a document and how often the opposite occurs in all documents containing that word. The frequency of word occurrence in a document indicates the importance of that word in the document, while the frequency of word occurrence in various documents shows how common the word is used. Therefore, if a word often appears in a particular document but is rarely found in other documents, that word will have a high weight in that document [11]. In TF-IDF, there is a formula to calculate the weight of each document against the keyword that can be formulated as follows [10].

$$W_{dt} = tf_{dt}\ X\ (log\left(\frac{N}{dft}\right) + 1) \quad (1)$$

Explanation:

d	:	document d
t	:	word t from the keyword
w	:	weight of document d against word t
tf	:	the number of times the word appears in a document
IDF	:	Inverse Document Frequency
N	:	total number of documents
f	:	number of documents containing the token

2.7 Classification

In the classification stage, this study uses two algorithms, namely Naïve Bayes and Support Vector Machine (SVM). The tests will be grouped into three classes: positive, neutral, and negative. The model used in this study is the C-Support Vector Machine (SVC) because it has good capabilities in finding the optimal hyperplane to separate data classes, while the Naïve Bayes algorithm uses the Multinomial Naïve Bayes model because it has the ability to handle text data.

2.8 Evaluation

To evaluate the classification results, a confusion matrix will be used. The purpose of the confusion matrix is to show the identification results between the number of correct predictions and the number of incorrect predictions [11].

After analyzing the results of the confusion matrix, the next step is to calculate the performance matrix. The performance matrix includes measuring the values of accuracy, precision, recall, and F1-score. Through the performance matrix, the model's ability to distinguish positive and negative classes and identify errors in predictions can be evaluated.

3. RESULTS AND DISCUSSION

In this section, the researcher will discuss the results of the research conducted. This discussion will reveal whether public opinion on the 2024 presidential election fraud is predominantly negative or positive. Additionally, the researcher will recommend the best method between Naïve Bayes and Support Vector Machine.

3.1 Data Collection

This research utilizes tweet data collected through a crawling process using the Tweet Harvest library in Python, which facilitates the automatic retrieval of tweets based on keywords related to public opinion on the 2024 Presidential Election fraud. With the Twitter API, Tweet Harvest overcomes the limitations of data retrieval from Twitter and successfully collects more than 1500 tweets in one process. This process begins by entering the twitter_auth_token to enable access to the Twitter API, which is crucial for smooth crawling. Additionally, Node.js installation is required to ensure the library functions optimally and efficiently in data management.

3.2 Preprocessing Data

After obtaining the dataset, the next step is data preprocessing. In this stage, data cleaning and preparation will be carried out for the next stage.

Table 1 Preprocessing Result

Tweet	
Opini publik terhadap kecurangan pemilu presiden 2024	
Processing Data	**Hasil**
Cleaning Data	Hasto Dipolisikan Usai Bahas Dugaan kecurangan pemilu PDIP Pembungkaman Suara Kritis
Case Folding	hasto dipolisikan usai bahas dugaan kecurangan pemilu pdip pembungkaman suara kritis
Tokenizing	['hasto', 'dipolisikan', 'bahas', 'dugaan', 'kecurangan', 'pemilu', 'pdip', 'pembungkaman', 'suara', 'kritis']
Stopword Removal	hasto dipolisikan bahas dugaan kecurangan pemilu pdip pembungkaman suara kritis
Stemming	hasto polisi bahas duga curang pemilu pdip pembungkaman suara kritis

3.3 Data Labeling

In this stage, data labeling will be carried out based on the analysis of Twitter users' tweets expressing their opinions on alleged fraud in the 2024 presidential election. The data will be categorized according to the type of opinion contained in the tweets, namely positive, neutral, and negative opinions.

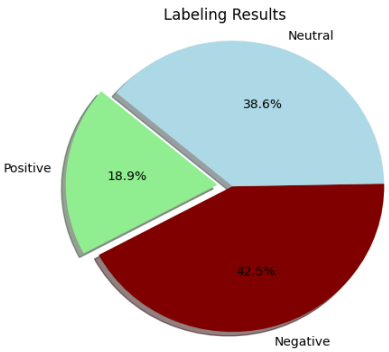

Figure 2 Labeling Data Results

This data labeling process uses the NLTK and TextBlob libraries. The results of data labeling obtained 230 for positive opinion data, 296 for neutral opinion data, and 516 for negative opinion data. There is a visualization in the form of a diagram with results of 18.9% for positive opinions, 38.6% for neutral opinions, and 42.5% for negative opinions.

3.4 Splitting Data

In the data splitting stage, the dataset is divided into two or more subsets. This process aims to enable more in-depth and focused sentiment analysis on certain aspects. By separating the data into different subsets, the analysis can be carried out more specifically and effectively, providing more accurate and relevant results according to the needs of the research or study being conducted. The data splitting stage is carried out by dividing the data using a 70:30 ratio, with the aim of obtaining more optimal results.

3.5 TF-IDF Weighting

In this TF-IDF word weighting module, the value or weight for each word or feature will be calculated. The TF-IDF word weighting process is carried out by multiplying the term frequency (TF) value by the inverse document frequency (IDF) value.

Table 2 TF-IDF Weighting Results

Term	**W = TF * IDF**				
	TF1	TF2	TF3	TF4	TF5
Adil	0	0	0	0	0
Amanah	0	0.3988322	0	0	0
Baswedan	0.222002	0	0	0	0
Benci	0	0	0	0.356107	0
Busuk	0	0	0	0.3299845	0
Count	0	0	0	0.2830776	0
Curang	0	0	0	0.0268782	0.0541621
Damai	0	0.398832	0	0	0
Demokrasi	0	0	0	0.215397	0
Pemilu	0	0.106198	0.2598083	0.0382761	0.077130
Pilpres	0	0	0	0.0454087	0.0915029

3.6 Testing Results

In this testing stage, the Naïve Bayes and Support Vector Machine algorithms are analyzed using evaluation metrics such as Confusion Matrix and Performance matrix, which include measurements of accuracy, precision, recall, and F1-score. By considering these matrices, the performance of both algorithms will be evaluated and compared to determine the most suitable algorithm for sentiment analysis needs.

1. Naïve Bayes

Naïve Bayes, specifically Multinomial Naïve Bayes (MultinomialNB), is used to calculate category probabilities, which is suitable for frequency-based data. The dataset is divided into 70% training data and 30% test data, where the training data trains the model and the test data evaluates its performance.

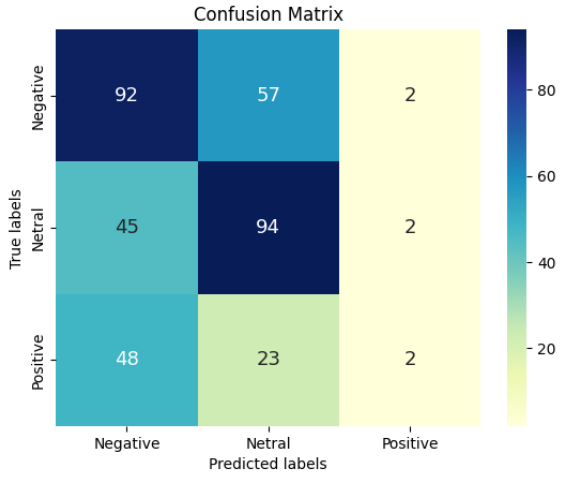

Figure 3 Naive Bayes confusion matrix result

From Figure 3, it is explained that:

a. True Negative (TN) : Data with a negative label that is correctly predicted as a negative label, totaling 92
b. True Positive (TP) : Data with a positive label that is correctly predicted as a positive label, totaling 2
c. True Neutral (TNeu): Data correctly predicted that the sample comes from the neutral class with a neutral prediction, totaling 94
d. False Negative (FN) : Data with a negative label but incorrectly identified as positive, totaling 93
e. False Neutral (FNeu): Data incorrectly predicted that the sample comes from the neutral class but predicted as positive or negative, totaling 80
f. False Positive (FP) : Data with a positive label that is correctly predicted as a negative label, totaling 4

Based on the Confusion Matrix results, model performance evaluation can be carried out by calculating the Performance matrix, which includes measurements of accuracy, precision, recall, and F1-score. This Performance matrix allows researchers to gain a deeper understanding of the model's ability to predict the correct class on test data. The calculation results of Accuracy, Precision, Recall, and F1-Score are presented in Table 3.

Table 3 Naïve Bayes Evaluation Results

Rasio	Naive Bayes			
	Akurasi	Presisi	Recall	F1-Score
30 : 70	51%	53%	51%	47%

2. Support Vector Machine

The classification process using Support Vector Machine (SVM) is applied to separate data based on its characteristics, with the Radial Basis Function (RBF) kernel, which is effective in handling non-linear data patterns. The dataset is divided into 70% training data and 30% test data, where the training data is used to build the model, while the test data is used to evaluate the model's performance.

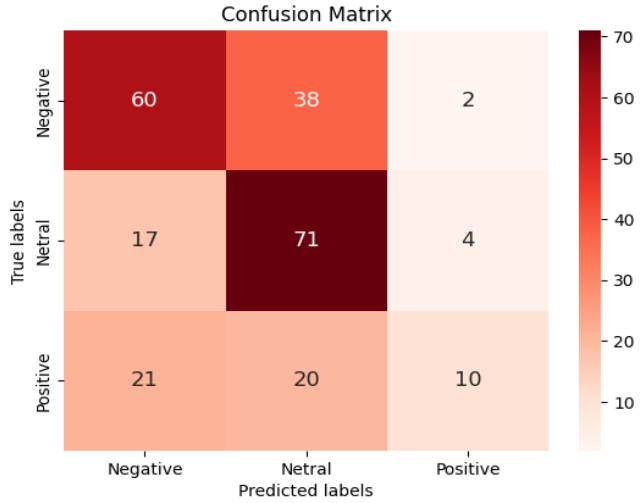

Figure 4 Support Vectore Machine confusion matrix result

From Figure 4 it is explained that:

a. True Negative (TN) : Data with a negative label that is correctly predicted as a negative label, totaling 60
b. True Positive (TP) : Data with a positive label that is correctly predicted as a positive label, totaling 10
c. True Neutral (TNeu): Data correctly predicted that the sample comes from the neutral class with a neutral prediction, totaling 71
d. False Negative (FN) : Data with a negative label but incorrectly identified as positive, totaling 38
e. False Neutral (FNeu): Data incorrectly predicted that the sample comes from the neutral class but predicted as positive or negative, totaling 58
f. False Positive (FP) : f. False Positive (FP) : Data with a positive label that is correctly predicted as a negative label, totaling 6

Based on the Confusion Matrix results, model performance evaluation can be carried out by calculating the Performance matrix, which includes measurements of

accuracy, precision, recall, and F1-score. This Performance matrix allows researchers to gain a deeper understanding of the model's ability to predict the correct class on test data. The calculation results of Accuracy, Precision, Recall, and F1-Score are presented in Table 4.

Table 4 Support Vector Machine Evaluation Results

Rasio	Support Vector Machine			
	Akurasi	Presisi	Recall	F1-Score
30 : 70	58%	60%	58%	55%

3.7 Public Perception and Sentiment Towards the Issue of 2024 Presidential Election Fraud

Public perception and sentiment towards the issue of 2024 Presidential Election fraud can be seen in Figure 4.11. From the analysis of the total tweets analyzed, it was found that 42.5% contained negative sentiment, 38.6% showed neutral sentiment, and only 18.9% contained positive sentiment. The dominance of negative sentiment indicates widespread concern among the public regarding the integrity and transparency of the election process, while the high neutral sentiment reflects many people who are still unsure or waiting for more information before taking a stance. These findings have important implications for the quality of democracy and public trust in election organizing institutions. Therefore, it is recommended that the next election increase transparency through the publication of easily accessible results, strengthen independent oversight institutions, educate voters, and utilize technology and law enforcement to prevent fraud.

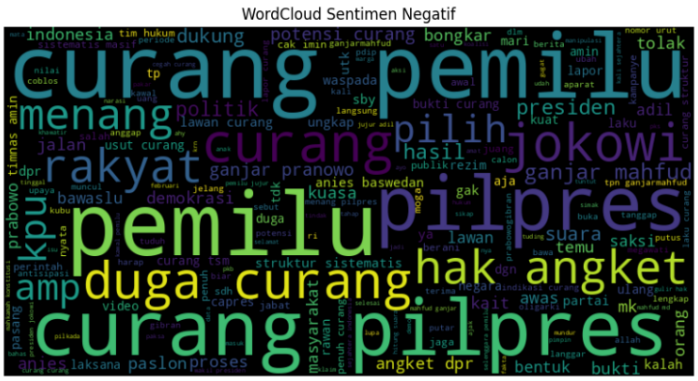

Figure 5 WordCloud Sentiment Negatif

3.8 Comparison of Naïve Bayes and Support Vector Machine Algorithms

In this study, the performance comparison of the Naive Bayes and Support Vector Machine algorithms in analyzing sentiment towards public opinion related to the issue of 2024 Presidential Election fraud was conducted using a dataset of tweets from Twitter. Using a 70:30 ratio of training data to test data, a comprehensive evaluation was conducted to compare the accuracy, precision, recall, and F1-score of the two algorithms, which can be observed in detail in Table V, showing which algorithm is superior in this context.

Table 5 Comparison Results of Naive Bayes and Support Vector Machine

Rasio	Naive Bayes			*Support Vector Machine*		
	Akurasi	Presisi	Recall	Akurasi	Presisi	Recall
70 : 30	51%	53%	51%	58%	60%	58%

Based on the data presented in Table V, it can be seen that with a data split ratio of 30:70, both algorithms, namely Naïve Bayes and Support Vector Machine (SVM), show satisfactory performance in terms of accuracy, precision, and recall. In the accuracy metric, the Support Vector Machine (SVM) algorithm achieved an accuracy rate of 58%, which is 7% higher than the Naïve Bayes algorithm, which has an accuracy of 51%. In the precision metric, the Support Vector Machine (SVM) also shows superiority with a value of 60%, which is 7% higher than Naïve Bayes, which obtained a precision of 53%. Meanwhile, in terms of recall, the SVM algorithm shows better performance with a recall value of 58%, compared to Naïve Bayes, which reached 51%.

Overall, SVM outperforms Naïve Bayes in sentiment analysis related to the issue of 2024 Presidential Election fraud. The advantage of SVM lies in its ability to handle non-linear data, effectively separate classes, and be more resistant to outliers and unbalanced data distributions. Meanwhile, although Naïve Bayes is efficient in some cases, this algorithm tends to be less flexible with complex data and sensitive to outliers, affecting model accuracy when distribution assumptions are not met.

CONCLUSIONS

Based on the results and discussions that have been conducted, the researcher can convey several conclusions as this study analyzes public opinion sentiment towards the 2024 Presidential Election fraud using a combination of Naïve Bayes and Support Vector Machine (SVM) algorithms, with tweet data classified into three sentiment categories: positive, neutral, and negative. The analysis results show that 42.5% of the public expressed negative sentiment, 38.6% showed neutral sentiment, and only 18.9% showed positive sentiment. The dominance of negative sentiment reflects public concern about the integrity and transparency of the election process, while the high neutral sentiment indicates doubt or a wait-and-see attitude for more information. To increase public trust, better transparency, strengthening independent oversight institutions, utilizing technology such as secure electronic voting systems, massive voter education, and strict law enforcement against fraud perpetrators are needed.

The comparison of the Naïve Bayes algorithm using MultinomialNB and SVM with the RBF kernel was conducted to determine the best accuracy level in this sentiment analysis. The results show that the Support Vector Machine with the RBF kernel achieved an accuracy rate of 58%, showing superior performance compared to Naïve Bayes, which produced an accuracy of 51%. This indicates that Naïve Bayes with MultinomialNB is more effective in classifying public opinion related to the issue of 2024 Presidential Election fraud.

ACKNOWLEDGEMENTS

Praise and gratitude be to Allah SWT for His blessings and guidance, enabling the completion of this journal. This journal is titled "Sentiment Analysis of Public Opinion on Electoral Fraud in the 2024 Presidential Election Using Support Vector Machine and Naïve Bayes." The completion of this journal would not have been possible without the support of many individuals. Therefore, the author would like to express sincere gratitude to Allah SWT for His endless blessings, to the author's parents for their unwavering prayers and motivation, to Mrs. Aries Dwi Indriyanti, S.Kom., M.Kom., as the supervising lecturer for her invaluable guidance and support, to Mrs. Monica Cinthya, M.Kom., for her assistance and valuable insights during the journal preparation, to friends for their continuous encouragement, and to all those who have contributed in any way to the completion of this research. May all the kindness and support received be rewarded abundantly..

REFERENCES

[1] Amal, M. (2023). *Tokenization in NLP: Types, Challenges, Examples, Tools*. Retrieved from neptune.ai: https://neptune.ai/blog/tokenization-in-nlp

[2] Cholid, D. (2022). Penerapan Metode Naïve Bayes Classifier Dan Support Vector Machine Pada Analisis Sentimen Terhadap Dampak Virus Corona Di Twitter. *Jurnal Informatika dan Rekayasa Perangkat Lunak (JATIKA)*, 145 - 160.

[3] Dianati, Gigih, & Eko. (2022, September 28). Analisis Sentimen Kinerja Dewan Perwakilan Rakyat (DPR) Pada Twitter Menggunakan Metode Naive Bayes Classifier. *Jurnal Informatika dan Teknik Elektro Terapan (JITET)*, 34-40. Retrieved from Ekrut Media: https://www.ekrut.com/media/sentiment-analysis-adalah

[4] Gillis, A. (2022). *Techtarget*. Retrieved from Data Splitting: https://www.techtarget.com/searchenterpriseai/definition/data-splitting

[5] Hakim, A. L. (2023, 2 10). *Mengenal lebih jauh apa itu Magang dan Studi Independent Bersertifikat (MSIB)*. Retrieved from OSC: https://osc.medcom.id/community/mengenal-lebih-jauh-apa-itu-magang-dan-studi-independent-bersertifikat-msib-4963

[6] Harahap, G. T. (2021). Peran Kepolisian Daerah Sumatera Utara (Polda-Su) Dalam Penegakan Hukum Terhadap Tindak Pidana Pemilihan Umum. *Jurnal Rentum*, 90 - 98.

[7] Hermanto, Mustopa, A., & Kuntoro, A. Y. (2020). Algoritma Klasifikasi Naive Bayes dan Support Vector Machinedalam Layanan Komplain Mahasiswa. *Jurnal Ilmu Pengetahuan Dan Teknologi Komputer*, 211-220.

[8] Kasim, Y. U. (2024). *Sejarah Pemilu di Indonesia: Perjalanan Pesta Demokrasi 1955-2024*. Retrieved from Detiksulsel: https://www.detik.com/sulsel/berita/d-7192898/sejarah-pemilu-di-indonesia-perjalanan-pesta-demokrasi-1955-2024

[9] menzli, a. (2023). *Tokenization in NLP: Types, Challenges, Examples, Tools*. Retrieved from neptune.ai: https://neptune.ai/blog/tokenization-in-nlp

[10] Populix. (2023). *Teknik Analisis Data: Pengertian, Jenis, Metode, Contoh*. Retrieved from info.populix.co: https://info.populix.co/articles/teknik-analisis-data/

[11] Rianto, B. I. (2021). Improving the accuracy of text classification using stemming method, a case of non-formal Indonesian conversation. *Journal Of Big Data*, 2 - 16.

[12] Rina. (2023). *Memahami Confusion Matrix: Accuracy, Precision, Recall, Specificity, dan F1-Score untuk Evaluasi Model Klasifikasi*. Retrieved from esairina.medium: https://esairina.medium.com/memahami-confusion-matrix-accuracy-precision-recall-specificity-dan-f1-score-610d4f0db7cf

[13] Septiani, D., & Isabela, I. (2020). Analisis Term Frequency Inverse Document Frequency (Tf-Idf) Dalam Temu Kembali Informasi Pada Dokumen Teks. *Jurnal Sistem dan Teknologi Informasi Indonesia*, 81-88.

[14] Suharto, A. (2023). *Fundamental Bahasa Pemrograman Python.* Purbalingga: CV.EUREKA MEDIA AKSARA.

[15] Tysara, L. (2024). *10 Kasus Pemilu di Indonesia 2004-2024, Pelanggaran Terjadi Berulang*. Retrieved from Liputan 6: https://www.liputan6.com/hot/read/5528538/10-kasus-pemilu-di-indonesia-2004-2024-pelanggaran-terjadi-berulang?page=4