Jurnal Varian
Vol. 7, No. 2, April 2024, Hal. 125–134
e-ISSN: 2581-2017

Naive Bayes Algorithm with Feature Selection Using Particle Swarm Optimization

Iwan Kurniawan[1], **Sri Astuti Thamrin**[1], **Siswanto**[1]
[1]Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Hasanuddin, Indonesia

Article Info

Article history:

Received : 09-26-2022
Revised : 06-12-2024
Accepted : 06-28-2024

Keywords:

COVID-19;
Classification;
Nave Bayes;
Particle swarm Optimization;
Twitter

ABSTRACT

The COVID-19 vaccine in Indonesia has led to the emergence of public opinion which is conveyed on social media such as Twitter. One of the analyses that can be done to produce various information from public opinion is sentiment analysis. Sentiment analysis is used to determine whether an opinion tends to be positive or negative. This study aims to classify the public opinion of the COVID-19 vaccine in Indonesia with sentiment analysis and to visualize the location of the sentiment of the COVID-19 vaccine tweet data in Indonesia. To achieve this aim, the Nave Bayes algorithm with Particle Swarm Optimization (PSO) feature selection was used. This study uses opinions into positive and negative class sentiments towards 2,547 tweets related to the COVID-19 vaccine in Indonesia from January to June 2021. The results show that the distribution of positive and negative class sentiments is 2,328 and 219, respectively. In addition, the positive sentiment for the COVID-19 vaccine was dominated by people on the island of Java based on a random number matrix initialized by the PSO method. The classification of public opinion on Twitter media provides accurate and optimal performance results using a combination of the Nave Bayes algorithm with PSO feature selection. The results of the combination of these methods have accuracy and F1 score values of 91.28% and 95.38%, respectively. The visualization of geo-spatial mapping showed that positive sentiments related to the COVID-19 vaccine exist in almost all regions in Indonesia but are dominated by the Jabodetabek area.

Accredited by Kemenristekdikti, Decree No: 200/M/KPT/2020
DOI: https://doi.org/10.30812/varian.v7i2.2409

Corresponding Author:

Sri Astuti Thamrin,
Department of Statistics, Universitas Hasanuddin
Email: tuti@unhas.ac.id

A. INTRODUCTION

The coronavirus disease 2019 (COVID-19) has infected more than 30 million people and more than 1 million deaths (Bowdle and Munoz-Price, 2020). Vaccine is expected by the community to protect them from contracting COVID-19. Information on the development of the COVID-19 vaccine has emerged, and various opinions have emerged through social media, for example, Twitter. The opinions given by the public through social media will generate knowledge, among others, in the form of sentiments towards the COVID-19 vaccine (Lyu et al., 2021).

Analysis of sentiment can produce opinions of either a positive, negative, or neutral nature (Alsaeedi and Khan, 2019). These opinions can be classified using methods or algorithms, one of which is popular is Nave Bayes. This method uses Bayes' Theorem assuming mutual freedom between classes with other features (Sağlam and Cengiz, 2024).

Feature selection on data will usually make an algorithm work better (Tijjani et al., 2024). Particle Swarm Optimization (PSO) is one of the feature selection methods in the Nave Bayes algorithm with population-based optimization techniques inspired by the social behavior of bird or fish movements (bird flocking or fish schooling) (Hyuningtyas et al., 2023) (Mazdadi et al., 2023). PSO as an optimization tool provides a population-based search procedure with everyone called particle adjusting the speed and searching for the best position based on information obtained from other particles (Papazoglou and Biskas, 2023).

Some of the classification methods that have been used for sentiment analysis include the Lexicon based approach Effendy (2015) on information about waste solutions with an accuracy of 70.68% and the K-Nearest Neighbor (KNN) algorithm (Claudy et al., 2018) regarding the classification of the character of prospective employees based on tweets with an accuracy of 66% as well as the Nave Bayes method (Fanissa et al., 2018) and the selection of the Query Expansion Ranking feature about tourism information in Malang City with an accuracy of 86.6 (Bahri et al., 2022). These related studies showed a comparison of the performance of the support vector machine method without or with PSO on the WhatsApp Review sentiment analysis, which shows that the performance of the PSO-based SVM algorithm has higher accuracy than the performance of the SVM algorithm without PSO. Based on several studies, the Nave Bayes algorithm has a better level of accuracy than other methods. This research also displays a visualization of a map of the sentiment location of COVID-19 vaccine tweet data in Indonesia, which is new to this research, because previous studies only looked at the accuracy of a method. Therefore, this research aims to classify public opinion towards the COVID-19 vaccine in Indonesia with sentiment analysis using the Nave Bayes algorithm with PSO feature selection and also displays a visualization of the sentiment location map of COVID-19 vaccine tweet data in Indonesia. Overall, this research improves classification techniques by integrating optimal feature selection methods, opening opportunities for more effective and efficient implementation in various real-world applications. In addition to using geospatial visualization, Twitter data can be transformed into valuable and easy-to-understand information, enabling better decision-making and faster responses to different situations.

B. RESEARCH METHOD

The data used in this study is a primary data, namely Indonesian tweet data containing the keyword ”covid vaccine” in the form of text format. A total of 12,000 tweets data was obtained from twitter social media from January 1, 2021 to June 30, 2021 using R Studio software. Table 1 shows a sample of the tweet data used.

Table 1. Tweet data results

No	Tweet
1	Pemerintah sdh sgt polan siang & mlm u/ melawan wabah. Kita hrs gimana? Jgn ngeyel, tingkatkan disiplin prokes, segera vaksin Kita hrs gimana? Jgn ngeyel, tingkatkan disiplin prokes, segera vaksin
2	@Bola_Jakarta: Ga usah ngiri sama Hungaria. Mereka lebih dari 55% penduduknya udah d vaksin covid-19, wajar bisa buka stadion buat penonton. Indonesia baru sekitar 7% penduduknya yg udah divaksin, udah gitu masih ribut vaksin ada microchipnya dan bikin badan jadi magnet.
3	Semoga usaha ini akan percepatkan kita untuk diberi suntikan vaksin Covid19.
4	aku udah vaksin 2x juga masih kena, tp gejalanya cuma batuk flu sm ilang penciuman beberapa hari. Kalo ga vaksin ya mungkin gejalaku lebih berat. contohnya temenku sendiri, dia kena covid kmaren sebelum vaksin. Gejalanya lebih parah dibanding aku. intinya sing penting ikhtiar
5	Apresiasi Pak Gub @aniesbaswedan kpd para Ketua RT/RW yg telah menjaga warganya dngn memastikan mereka ikut Vaksin Covid-19. Panutan
6	Syaratnya apa saja nih? Dikutip dari detik.com, Kemenlu Maroko mengatakan kalua syarat masuk ke negaranya adalah membawa bukti vaksin penuh dan bukti negative COVID-19.
7	Menteri Luar Negeri Toshimitsu Motegi mengumumkan pada konferensi pers pada tgl 15 Jun bahwa Jepang akan memberikan vaksin COVID-19 ke Vietnam, Indonesia, Thailand, Filipina dan Malaysia secara gratis. Tanggal 16 Juni ini akan mengirimkan sekitar 1 juta dosis ke Vietnam.
8	Kurva covid naik lg, aing blm jg kebagian vaksin. Tp lusa harus bgkt, yah demi cuan utk bayar cicilan2. Jujur pergi kali ini takut bgt. So klo lusa u pd liat gw & dikira gw liburan, NO,, aing kerja. Kenyataan nanti tidak seindah jepretan lensa.
⋮	⋮
12000	Beberapa minggu setelah mudik itu bunda rada2 cemas sih aku keluar2 karena emang data pasien covid naik terus. Makanya yda disuruh cpt2 vaksin. Gini jd keder mau brgkat Jogja, pdhl 2bulan lagi.

The stages taken to analyze the public opinion on the COVID-19 vaccine in Indonesia can be seen in the flowchart in Figure 1 and are described as follow:

1. Data collection
   The first step of data retrieval is to obtain a token by logging into the twitter account by filling out the registration form specified by twitter. Then through the Application Programming Interface (API) a crawling process is carried out to get data on twitter. The first step of the data crawling process is to enter tweet keywords. From these keywords, it will generate data such as tweet_id, text, and tweet date, favorite count, and retweet count (Khder, 2021).
2. Pre-processing

DOI: https://doi.org/10.30812/varian.v7i2.2409

At this stage, the process of case folding, cleaning, tokenizing, and filtering of tweet data is carried out. The stage will generate a token that is used as machine learning data by the PSO feature selection (Hickman et al., 2022). The feature selection stage uses PSO have several steps in the following:

(a) Randomly initiate speed and starting position.
(b) Evaluate P_{best} each particle by position.
(c) Determine P_{best}, best particle and set it as G_{best}. P_{best} is the result of a comparison of the current position with the previous iteration.
(d) Checks stopping criteria, which are the position of the particles when they reach the maximum iteration. If it does not reach the maximum iteration, return to the step a.

3. Classification

At this stage, the classification of texts is carried out using Nave Bayes to find out the categories of texts based on data. The following are the stages that are carried out:

(a) Manual Classifying

Before determining a tweet falls into a positive or negative sentiment class using the Equations (1), (2) and (3) on Nave Bayes, the determination of the tweet sentiment class is first manually classified one by one.

Some of the equations that can be used in the Nave Bayes Classifier can be seen in the Equation (1) and Equation (2).

$$P(v_j) = \frac{|\text{docs}|}{|\text{training}|} \tag{1}$$

where $P(v_j)$ is the probability of each data against a set of data and $|docs|$ is the data frequency in each category.

$$P(w_k|v_j) = \frac{n_k + 1}{|n + \text{word number}|} \tag{2}$$

where $P(w_k v_j)$ is the probability of occurrence of the word w_k on a data and n_k is the k-th word frequency of each category.

In Nave Bayes classifiers, we need to maximize the probability value of each class, expressed as a Hypothesis Maximum A Posteriori (H_{map}) (Equation (3)) (Ahmed et al., 2024):

$$H_{map} = \underset{\{\text{positive, negative}\}}{\text{argmax}} P(w_k|c) \cdot P(c) \tag{3}$$

where H_{map} is the highest probability value of data from each class, $P(w_k|c)$ is the probability of occurrence of the word w_k in class c and $P(c)$ is the probability of class c.

(b) Data Feature Formation

On the formation of features are selected features that are considered important in a tweet. Features are data parameters in the form of keywords classified into predetermined sentiment classes (positive and negative).

(c) Calculation of Prior Probability of Sentiment Class

After the feature is formed with its appearance on the data, it is then calculated the prior probability value of each class using Equation (1).

(d) Determining Feature Opportunities

After obtaining the probability of each class, the probability of each feature in each class will also be calculated using Equation (2).

(e) Calculation of H_{map}

To determine the class classification of a tweet, the highest data probability from one of the classes will be selected based on the learning process using Equation (3).

(f) Determining the Classification Results

After obtaining the highest score of H_{map} for each tweet, the entire tweet is classified into positive and negative classes.

4. Evaluate the performance of the algorithm with confusion matrix (accuracy and F1 Score)

Accuracy is the most commonly used and easy-to-understand metric. Accuracy measures how often your model makes correct predictions out of the total predictions. Meanwhile, the F1 Score is a measure that combines precision and recall. The F1 score provides a balance between these two measures. This is useful when there is a class imbalance or when the researcher wants to

DOI: https://doi.org/10.30812/varian.v7i2.2409

avoid overdoing or underdoing one of the above metrics that can occur in situations with bias. So, using both together provides a more complete picture of model performance, particularly in imbalanced classes or an imbalance between desired precision and recall (Palanisamy et al., 2022).

5. Visualize the sentiment, word cloud, and geo-spatial classes.

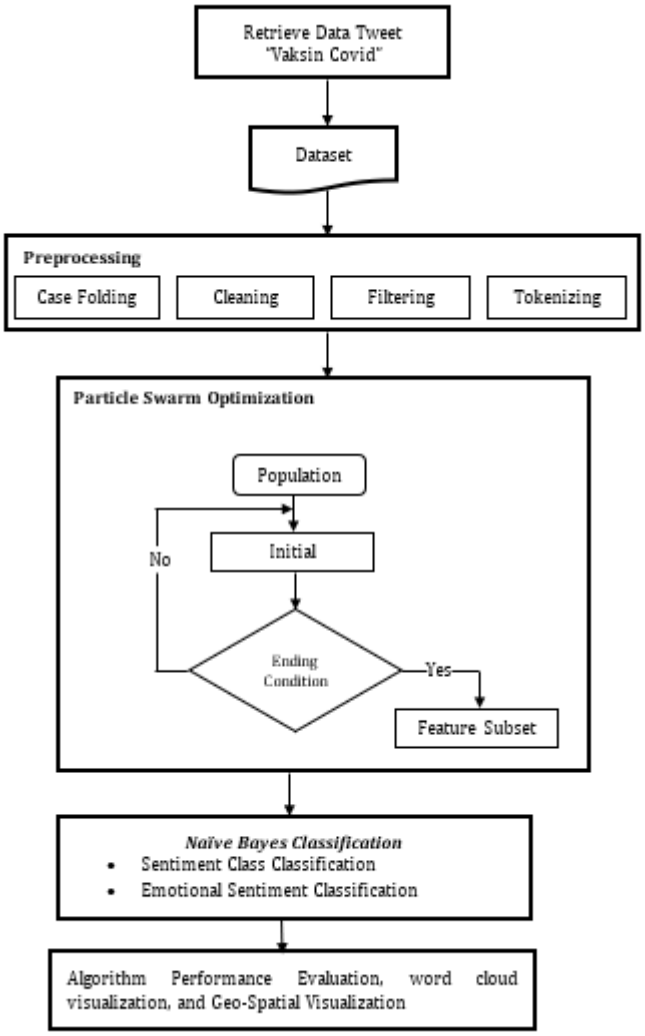

Figure 1. Flowchart of research methodology

C. RESULT AND DISCUSSION

1. Pre-processing Process

The crawling data stored in the database will then go through a preprocessing process to remove words with less influence in the classification process later. This process reads every word in the database. All capital letters in the tweet are changed to lowercase at the case folding stage. The cleaning stage is a word cleaning stage that does not affect the sentiment classification results at all. The tokenizing stage is cutting words based on each word arrange them into single pieces. The filtering stage is the stage of filtering words obtained from the tokenizing stage, which are considered to be able to be used to represent the content of the tweet.

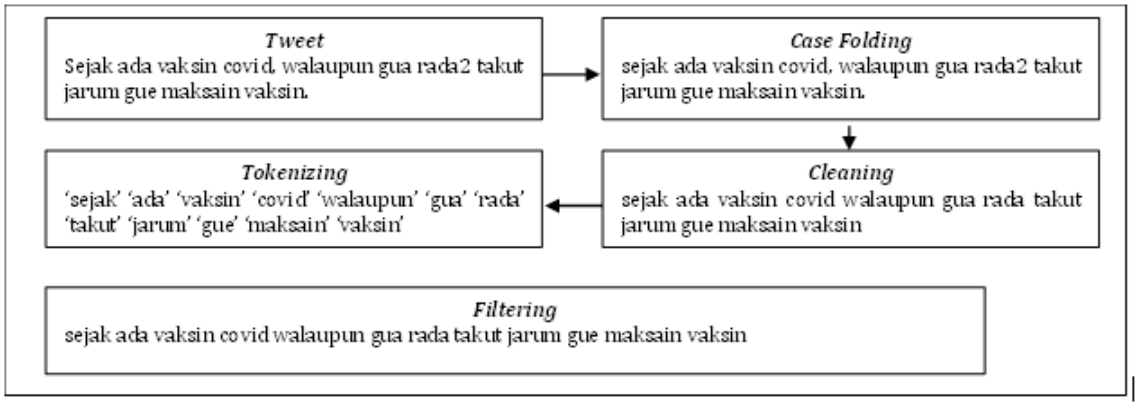

Figure 2. Word Cloud Visualization

This process reads every word that is in the database. The process of pre-processing data for one of the tweets is shown in Figure 2, shows each stage of preprocessing data for one of the tweets. The data preprocessing process also deletes all retweets or words reposted by other users so that each user can only represent 1 opinion. A total of 9,543 retweets were purged out of the existing 12,000 data leaving 2,547 tweet data for analysis.

2. Feature Selection with PSO

PSO feature selection is used to select features (keywords) that will be used in classifying and removing unnecessary features. There are three main procedures in PSO, namely: initialization of particle populations, evaluation of particle fitness values, and updating particle components. Particle population initialization includes feature selection and the generation of the position and initial velocity of the particle. The feature population to be selected is the one that has the highest frequency on the tweet data. A total of 10 features were selected as particle populations.

Based on Table 2, it shows that the ”hit” feature is the feature with the highest frequency of occurrence in the data with a frequency of 244 times. Then followed by the ”program” feature 77 times, the ”dose” feature 73 times, to the ”prokes” feature 62 times. In the 10 features that have been initialized, the calculation of the process of updating the particle component until iteration-3, a change in the value of P_{best} is obtained. Each particle for each iteration can be seen in the Table 2 as follows.

Table 2. Ten Features with Highest Frequency

No	Feature	Frequency
1	Hit	244
2	Program	77
3	Dose	73
4	Positive	70
5	Receive	68
6	Injection	68
7	Citizens	66
8	Virus	64
9	Society	63
10	Prokes	62

Table 3. P_{best} Value Change Particle for Each Iteration

Particle	P_{best}		
	Iteration-1	Iteration-2	Iteration-3
p_1	0,8146	-	-
p_2	0,6504	-	-
p_3	0,4178	0,8146	-
p_4	0,4165	-	-
p_5	0,4945	0,6504	-
p_6	0,8647	0,4178	-

Particle	P_{best}		
	Iteration-1	**Iteration-2**	**Iteration-3**
p_7	0,7124	0,4165	-
p_8	0,8146	0,4945	-
p_9	0,6504	0,8647	-
p_{10}	0,4178	0,7124	-

The features that will be selected in order to enter the classification process are those that are on the particle with a highest value of P_{best} on the latest iterations. In Table 3, it can be seen that the best value of P_{best} owned by p_9, which is at iteration-2 of 0.8647. This value indicates that the accuracy of the PSO method in selecting features is 86.47%.

Table 4. PSO Feature Selection Results

		Feature	**Description**
p_9	f_1	Hit	Not selected
	f_2	Program	Not selected
	f_3	Dose	Not selected
	f_4	Positive	Selected
	f_5	Receive	Not selected
	f_6	Injection	Not selected
	f_7	Citizens	Not selected
	f_8	Virus	Not selected
	f_9	Society	Not selected
	f_{10}	Prokes	Not selected

Based on the results of the PSO feature selection in Table 4, it can be concluded that there is only one feature in p_9 which is selected to enter into the classification process, which is f_4 (positive). Therefore, the word "positive" will remain in the tweet data. Then the other nine features were not selected or issued.

3. Classification with Nave Bayes Algorithm

The determination of a tweet classified into positive or negative classes is carried out through the calculation of the H_{map} value. It aims to determine the highest chances of each class based on the learning process. The highest opportunity value indicates the sentiment from the tweet data.

Table 5. Sentiment Class Classification

Tweet	H_{map}	**Class**
1	0.76145	Positive
2	0.90891	Positive
3	0.11499	Positive
4	0.29571	Positive
5	0.20592	Positive
⋮	⋮	⋮
2,547	0.68295	Positive

Based on Table 5, the number of tweets classified into the positive class was 2,328 and the negative class was 219. It can be seen that, using the PSO-based Nave Bayes algorithm, the data classifications into positive and negative class sentiments by 91.4% and 8.6%, respectively.

4. Algorithm Performance Evaluation

The performance evaluation of the Nave Bayes algorithm was carried out using accuracy and F1 Score. Table 6 shows the distribution of tweet data based on its predictions and actual values. The calculation result resulted in an accuracy value of 0.9128 which means that the accuracy of the algorithm in carrying out the classification was 91.28%. The F1 Score value of 0.9538

means that the feasibility of precision and recall in calcification is 95.38%.

Table 6. Confusion Matrix Results

		Actual	
		Positive	**Negative**
Prediction	Positive	2298	192
	Negative	30	27

5. Word Cloud Visualizatione

Word cloud visualization is used to see the words that most often appear in tweet data after the data preprocessing and PSO feature selection process. The greater the frequency of occurrence of a word in tweet data, the larger the letter size that appears in the word cloud.

Figure 3. Word Cloud Visualization

Figure 3 is an output of the word cloud program on the topic of COVID-19 vaccines in Indonesia. It can be seen that the tweet data is dominated by the words ”positive”, ”president”, and ”believe”. There are also the words ”variant”, ”list, and ”target” with a considerable proportion of visuals. This indicates that the government’s efforts to provide the COVID-19 vaccine can the public believe will provide hope for health protection so that activities can return to normal.

6. Geo-Spatial Visualization

The location of the tweet can be identified on the map by highlighting the respective location of the tweet. Individual tweets can be identified by a point on the map as a marker. The geo-spatial visualization in Figure 4 shows that positive sentiment (blue) exists in almost all regions in Indonesia, but is dominated in the Jabodetabek region. Meanwhile, negative (red) sentiment appears quite widely found in East Java Province. The large number of points on the island of Java and the high number of residents on the island show that the active users of twitter social media are certainly very high in the area.

Figure 4. Geo-Spatial Visualization of Several Regions in Indonesia

This research develops a visualization system to analyze tweet sentiment in various locations. This research uses geospatial mapping techniques to show how positive and negative sentiment is spread across different regions, which helps track the spread of the COVID-19 disease and areas requiring medical attention. This research shows how vital geospatial visualization is in understanding and utilizing social media data.

D. CONCLUSION AND SUGGESTION

Public opinion regarding COVID-19 via Twitter has been classified by using the Nave Bayes algorithm based on PSO feature selection. Of the 2,547 tweets data resulted in a distribution of positive and negative sentiment of 2.328 and 219, respectively. In addition, the positive sentiment for the COVID-19 vaccine was dominated by people on the island of Java based on a random number matrix initialized by the PSO method. The classification of public opinion on Twitter media provides accurate and optimal performance results using a combination of the Nave Bayes algorithm with PSO feature selection. The results of the combination of these methods have accuracy and F1 score values of 91.28% and 95.38%, respectively. From the visualization of geo-spatial mapping, it is known that positive sentiments related to the COVID-19 vaccine exist in almost all regions in Indonesia, but are dominated by the Jabodetabek area. The negative information related to the COVID-19 vaccine is quite common in East Java Province. It is based on a matrix of random numbers initialized in the PSO method. Instead of classifying public opinion towards the COVID-19 vaccine in Indonesia with sentiment analysis using the Nave Bayes algorithm with PSO feature selection, the visualization of the sentiment location map of COVID-19 vaccine tweet data in Indonesia is also considered in our study.

REFERENCES

Ahmed, Z., Issac, B., and Das, S. (2024). Ok-nb: An enhanced optics and k-naive bayes classifier for imbalance classification with overlapping. *IEEE Access*. DOI : https://dx.doi.org/10.1109/ACCESS.2024.3391749.

Alsaeedi, A. and Khan, M. Z. (2019). A study on sentiment analysis techniques of twitter data. *International Journal of Advanced Computer Science and Applications*, 10(2):361–374. DOI : https://dx.doi.org/10.14569/IJACSA.2019.0100248.

Bahri, M. S., Hermawan, A., Kondy, E. P., Semida, R. J., et al. (2022). Performance comparison of supporting vector machine method

without or with particle swarm optimization based on sentiment analysis whatsapp review. *International Journal of Academic and Applied Research (IJAAR)*, 6(6):94–101. www.ijeais.org/ijaar.

Bowdle, A. and Munoz-Price, L. S. (2020). Preventing infection of patients and healthcare workers should be the new normal in the era of novel coronavirus epidemics. DOI : https://doi.org/10.1097/ALN.0000000000003295.

Claudy, Y. I., Perdana, R. S., and Fauzi, M. A. (2018). Klasifikasi dokumen twitter untuk mengetahui karakter calon karyawan menggunakan algoritme k-nearest neighbor (knn). *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 2(8):2761–2765. https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/1967.

Effendy, V. (2015). Analisis sentimen berbahasa indonesia dengan pendekatan lexicon based (studi kasus: Solusi pengelolaan sampah). *Komputa: Jurnal Ilmiah Komputer dan Informatika*, 4(1):49–54. DOI : https://doi.org/10.34010/komputa.v4i1.2411.

Fanissa, S., Fauzi, M. A., and Adinugroho, S. (2018). Analisis sentimen pariwisata di kota malang menggunakan metode naive bayes dan seleksi fitur query expansion ranking. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 2(8):2766–2770. DOI : https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/1962.

Hickman, L., Thapa, S., Tay, L., Cao, M., and Srinivasan, P. (2022). Text preprocessing for text mining in organizational research: Review and recommendations. *Organizational Research Methods*, 25(1):114–146. DOI : https://doi.org/10.1177/1094428120971683.

Hyuningtyas, R. Y., Sari, R., and Yusnaeni, W. (2023). Particle swarm optimization for feature selection in sentiment analysis on the application of digital payments ovo using the algorithm of naive bayes. In *AIP Conference Proceedings*, volume 2714. AIP Publishing. DOI : https://doi.org/10.1063/5.0129011.

Khder, M. A. (2021). Web scraping or web crawling: State of art, techniques, approaches and application. *International Journal of Advances in Soft Computing & Its Applications*, 13(3). DOI : https://doi.org/10.15849/IJASCA.211128.11.

Lyu, J. C., Han, E. L., and Luli, G. K. (2021). Covid-19 vaccine–related discussion on twitter: topic modeling and sentiment analysis. *Journal of medical Internet research*, 23(6):e24435. DOI : https://doi.org/10.2196/24435.

Mazdadi, M. I., Farmadi, A., Kartini, D., et al. (2023). Implementation of particle swarm optimization feature selection on naïve bayes for thoracic surgery classification. *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, 5(3):150–158. DOI : https://doi.org/10.35882/jeemi.v5i3.305.

Palanisamy, T., Sadayan, G., and Pathinetampadiyan, N. (2022). Neural network–based leaf classification using machine learning. *Concurrency and Computation: Practice and Experience*, 34(8). DOI : https://doi.org/10.1002/cpe.5366.

Papazoglou, G. and Biskas, P. (2023). Review and comparison of genetic algorithm and particle swarm optimization in the optimal power flow problem. *Energies*, 16(3):1152. DOI : https://doi.org/10.3390/en16031152.

Sağlam, F. and Cengiz, M. A. (2024). Local resampling for locally weighted naïve bayes in imbalanced data. *Computing*, 106(1):185–200. DOI : https://doi.org/10.1007/s00607-023-01219-0.

Tijjani, S., Ab Wahab, M. N., and Noor, M. H. M. (2024). An enhanced particle swarm optimization with position update for optimal feature selection. *Expert Systems with Applications*, 247:123337. DOI : https://doi.org/10.1016/j.eswa.2024.123337.

.

Vol. 7, No. 2, April 2024, Hal. 125–134
DOI: https://doi.org/10.30812/varian.v7i2.2409