Improved Accuracy of Naïve Bayes Algorithm and Support Vector Machine Using Particle Swarm Optimization for Menstrual Cup Sentiment Analysis on Twitter

Dini Shalikha [1,*], Alamsyah [1]

[1] Department of Computer Science, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Semarang, Indonesia
*Corresponding author: shalikhadini@gmail.com

ARTICLE INFO

Article history
Received 8 October 2022
Revised 19 October 2022
Accepted 25 October 2022

Keywords
Text Mining
Sentiment Analysis
Naïve Bayes
Support Vector Machine
Particle Swarm Optimization

ABSTRACT

Menstrual cup is a menstrual hygiene sanitation tool that replaces disposable sanitary napkins for women that reaps many pros and cons in its use. From this, it is necessary to analyze the public's views regarding the use of menstrual cups, which is called sentiment analysis. Sentiment analysis is a process that aims to determine the polarity of the sentiment of a text. This paper performs a classification of menstrual cup sentiment analysis on Twitter using Naïve Bayes and Support Vector Machine algorithm. Particle Swarm Optimization is applied to improve the accuracy of both classification algorithms. The final result of the accuracy obtained by the Naïve Bayes algorithm is 92.72% and the Support Vector Machine algorithm is 96.13%. While the accuracy results after Particle Swarm Optimization is applied, for Naïve Bayes it produces an accuracy rate of 95.87%, and Support Vector Machine is 96.68%.

1 Introduction

Today, social media plays an important role in providing feedback (MacReadie et al., 2011). One of the social media that is most often used to provide opinions and express opinions is Twitter. Twitter is considered as a social media for users to send messages in real time. This feedback can certainly be felt both for individuals and groups. Social media also provides a space for expressing various thought and ideas, as well as conveying various opinions. According to the website Databoks.katadata.co.id which is a website with online media companies and research in the economic and business fields and accessed on February 1, 2022, The average Indonesian who uses Twitter is 59% of users,This makes Twitter the 5th most used social media in Indonesia in 2020. This indicates that Twitter is one of the social media that is quite influential for the social media user community in Indonesia.

Of the many public opinions have entered trending Twitter topics, the use of menstrual cups as menstrual sanitation products for the health of women's organs doesn't escape from discussion. Menstrual cup is a silicone device used as a substitute for disposable sanitary napkins for women. Menstrual cup are also considered an environmentally friendly product because they can be used repeatedly. However, in Indonesia, the use of menstrual cups is still considered taboo and has many pros and cons. From the many pros and cons related to the use of menstrual cups, it is necessary to analyze the public's views regarding the use of menstrual cups, which is called sentiment analysis.

Sentiment analysis included in text mining. Text mining is an activity to analyze a document or data with one another to find new data. Text mining includes things like category information and text grouping (Betesda, 2020). Sentiment analysis is an opinion exploration with the aim to analyze and evaluate a topic, product, or service refers to the broad field of natural language processing

(Kristanto et al., 2019). Sentiment analysis is a process of classifying opinions or opinions of a text into positive or negative opinion sentiments (Larasati et al., 2019). Public opinion, especially on social media, is very important to make a decision that will be beneficial for individuals and organizations. Currently, the sentiment of product reviews has become important information for producers and potential customers (Hamzah, 2021). The purpose of sentiment analysis is to determine the extent of public or public understanding related to the use of menstrual cups.

In sentiment analysis there are several data classification algorithms including the Naïve Bayes method and the Support Vector Machine. Both algorithms are considered to be able to work well to analyze public sentiment. The Naïve Bayes algorithm is said to be able to calculate the possibility of each factor, then choose the result with the highest probability (Wisnu et al., 2020). This algorithm is considered suitable for the classification process of sentiment analysis because it can produce a fairly high level of accuracy. While the Support Vector Machine is a classification algorithm that is able to produce a good classification model even though it is trained with little data and only with simple parameters (Hasan & Wahyudi, 2018). However, the Support Vector Machine algorithm has several weaknesses, one of which is the problem of selecting appropriate features (Ratino et al., 2020).

From these weaknesses, it is necessary to add Particle Swarm Optimization feature selection to improve its performance. Particle Swarm Optimization is an optimization algorithm with the aim of producing an optimum response value by determining process parameters (Sateria et al., 2019). Particle Swarm Optimization is considered quite easy to use because it doesn't require many lines of programming code and complicated mathematical operators. Therefore, Particle Swarm Optimization can streamline the required memory and speed. Based on the description of the problem above, the research focuses on increasing the accuracy of two classification algorithms, it is Naïve Bayes and Support Vector Machine using the Particle Swarm Optimization feature selection for sentiment analysis of menstrual cup usage on Twitter.

2 Methods

The process of increasing the accuracy consists of several main stages, including the preprocessing stage, labeling stage, normalization stage, feature extraction stage, feature selection stage, data splitting stage, classification stage, and model testing stage. Each of these stages has a different result. The stages of the process can be seen in Figure 1.

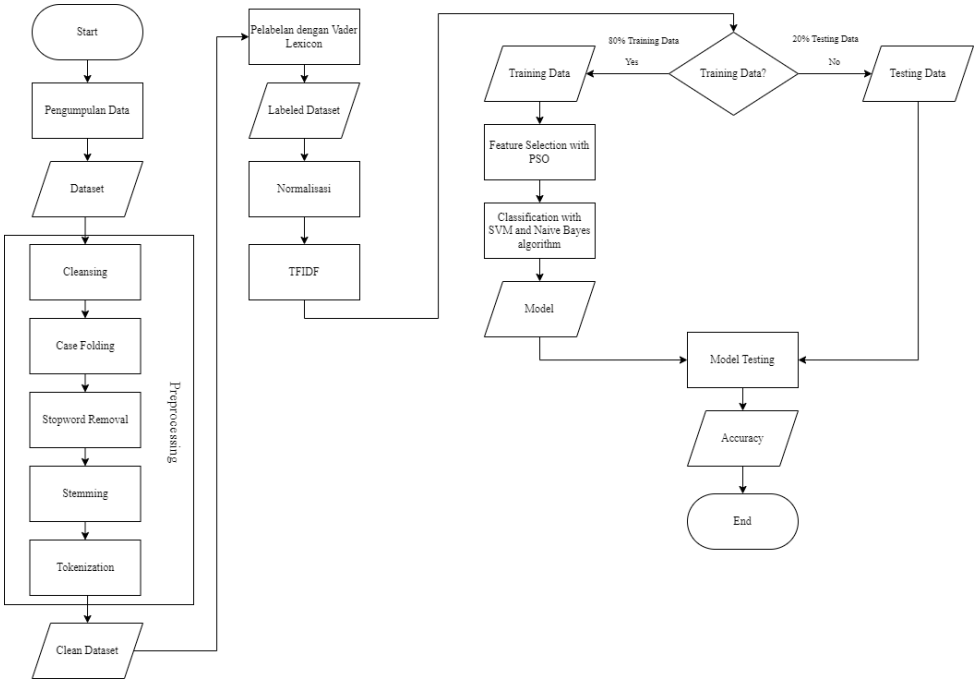

Figure. 1 The stage of process

2.1 Dataset

The dataset used is based on research by Fauziah, (2020). The study retrieved the dataset by crawling tweets on Twitter. The dataset was taken from April 26, 2020 to May 25, 2020, with the keyword menstrual cup totaling 1,108 tweets in English. It has two attributes, namely created_at which is the time the tweet was created, and text which is the content of the tweet. The dataset used in this study can be seen in Table 1.

Table 1. Dataset

Created_At	Text
Sat May 02 06:56:18 +0000 2020	i just learned how to use ampons and menstrua cup.....
Mon May 25 03:01:53 +0000 2020	Omg i use menstrual cup
Tue May 12 09:38:16 +0000 2020	Been using menstrual cup for almost 1 year now! And save a lot of money. No need o buy pad anymore https://t.co/mx2aa6sjhe

2.2 Preprocessing

The preprocessing stage aims to eliminate noise so that the sentiment analysis process becomes more accurate and can be used in general. The preprocessing stage is also carried out in order to produce more structured data for further processing in the next stage (Jumeilah, 2017). Data pre-processing consists of the following processes:

- Cleansing is a process used to remove all unused characters and serves to reduce noise in the data (Bayhaqy et al., 2018).
- Case Folding is a process used to change all letters in a sentence into lower case.
- Stopword Removal is a process to remove unimportant words such as conjunctions, time, and others (Nooryuda Prasetya & Winarso, 2021).
- Stemming is the process of converting tokens into basic words. This word conversion is carried out to ensure that every word that is the same but has a different suffix can be recognized as the same value to avoid bias in the transformation stage (Mariel et al., 2018).
- Tokenization is the stage for splitting text data into tokens.

After that, clean data is obtained which is ready to be processed in the next stage.

2.3 Labelling

This stage is the stage for labeling sentiments on the text using lexicon based. Lexicon based is a method for classifying a sentence into positive or negative sentiments. In this study, text labeling uses Vader (Valence Aware Dictionary And Sentiment Reasoner) as a labeling library. Vader Lexicon is a lexicon based library that is used for automatic labeling in text analysis. Lexicon based has several stages such as determining word polarity, handling negation, and scoring each tweet entity (Mustofa & Prasetiyo, 2021). The labeling process with this method begins after knowing which words contain positive and negative sentiments, then each word containing each of these sentiments is calculated by calculating the opinion value (Mahendrajaya et al., 2019). Vader is considered to work well for sentiment, especially on social media, and is available in an NLTK package that can be directly applied to unlabeled datasets.

2.4 Normalization

The normalization stage is the stage for the process of converting linear data to the original data (Nurjanah et al., 2017). Normalization is used to balance the data in research, so that the purpose of this normalization process is to produce a balance of comparison values between the data before and after the process and to form data with the same range value.

2.5 Feature Extraction

The feature extraction stage is the process of converting tokens into numeric vectors. The method used in this stage is TFIDF (Term Frequency Inverse Document Frequency). TFIDF is a technique to count the number of times a word appears in a document. Term Frequency (TF) is the number of words that appear in a document or text. While Inverse Document Frequency (IDF) is the level of importance of a word in the document. To calculate the weight of TF used Equation 1.

$$\frac{D}{tf_i} \tag{1}$$

Meanwhile, to calculate the IDF weight, Equation 2.

$$\frac{D}{tf_i} \tag{2}$$

Thus, the equation for calculating TFIDF is used Equation 3.

$$tfidf = tf \times \log log \frac{D}{tf_i} \tag{3}$$

2.6 Feature Selection

In this study, Particle Swarm Optimization is used for the feature selection algorithm that is used to increase the accuracy of the classification algorithm, namely Naïve Bayes and Support Vector Machine. The search using Particle Swarm Optimization is based on a population in a number of particles. Then the flight speed of each particle is updated to find the best new solution. Particle Swarm Optimization will stop when a condition has been reached.

Particle Swarm Optimization is likened to the behavior of a flock of birds in a habitat. Each particle is like a bird. This bird behaves using its own intelligence similar to the behavior of its collective flock. When a bird finds a proper path or a shorter path to a food source, the rest of the group will also follow that path even though they are far apart. Each bird or particle is treated like a point in a certain dimension of space. The steps taken by Particle Swarm Optimization in selecting features according to Sabrila et al., (2022) can be seen in Figure 2.

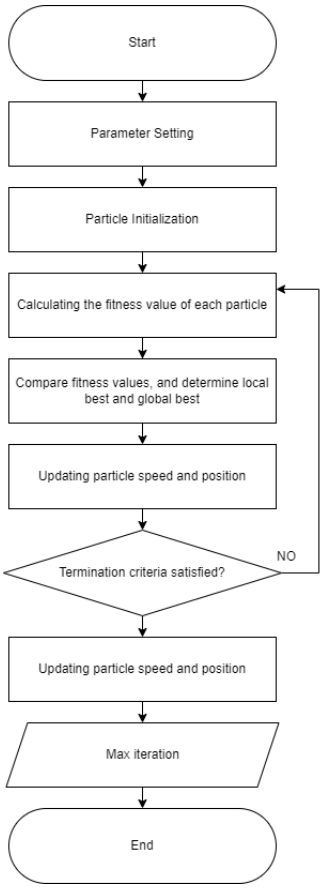

Figure 2. PSO process

2.7 Splitting Data

The splitting data stage is the data sharing stage with the aim of dividing the data into training data and testing data. In this study, the proportion of the distribution of training data and testing data is 80:20, respectively. This is based on research conducted by Gholamy et al., (2018) which states that the distribution of the proportion of data as much as 80:20 is empirically the best ratio for the distribution of training data and testing data. The distribution of split data with the proportion of training data and testing data of 80:20 is also based on the Pareto Principle that Pareto makes observations and gets the results that 20% of the factors determine 80% of success (Harvey & Sotardi, 2018).

2.8 Classification

At this stage of classification consists of two classification processes by using the Naïve Bayes algorithm and Support Vector Machine. Classification is a technique that can be used to predict data or describe data classes (Alamsyah & Fadila, 2021). In this study, two classification algorithms are used, namely Naive Bayes and Support Vector Machine. Method of Naïve Bayes has 3 stages such as previous research, find the probability value, and looking probabilitasi end of posterior (Insani, et al., 2018). While the Support Vector Machine has a goal to provide a value for the number of occurrences of a word and can classify sentences into positive or negative labels (Giovani et al., 2020). The stages of Naïve Bayes classification and Support Vector Machine with Particle Swarm Optimization in this study can be seen in Figure 3.

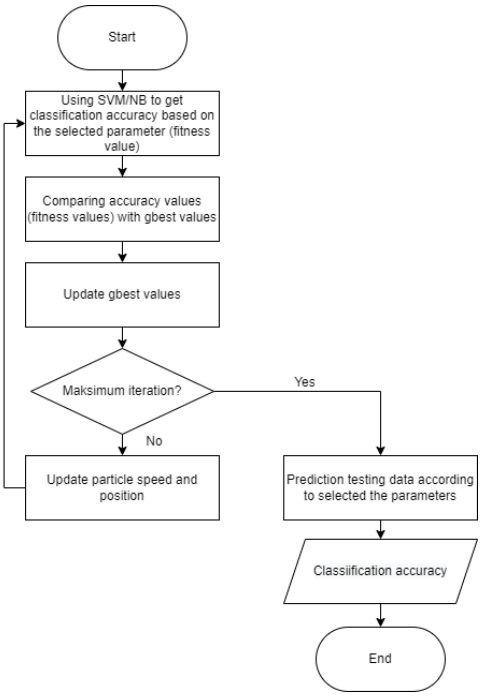

Figure 3. Classification flowchart with PSO

The first stage is to initialize the Particle Swarm Optimization parameters. Then evaluate the fitness value with the Support Vector Machine classification algorithm based on the parameters of each selected particle to get the classification accuracy value. Furthermore, if the optimization process has not reached its maximum iteration, then the velocity and position of each particle is continuously updated until it reaches its maximum iteration.

2.9 Model Testing

In this study, model testing was carried out using a confusion matrix and three experiments using k-Fold with a value is 5, so in one experiment it would produce five accuracy results. The confusion matrix is considered to be able to evaluate the performance of the system that has been built (Sabrila et al., 2022). Calculation of accuracy in the confusion matrix is calculated by Equation 4.

$$Accuracy = \frac{T_P}{T_p + F_p + T_N + F_N} \quad (4)$$

Accuracy is used to measure and determine the value and level of similarity between the measured value and the actual value. Where T_P is the value of True Positive and F_p is the value of false Positive.

3 Results and Discussion

This study applies feature selection with Particle Swarm Optimization for menstrual cup sentiment analysis on Twitter using the Naïve Bayes classification algorithm and Support Vector Machine. The higher the accuracy value generated, the better algorithm in predicting the resulting sentiment. In this study, three experiments were conducted using the k-Fold value model test, which is 5, meaning that there will be five accuracy results in one experiment.

After the data is collected, the next process is pre-processing, labeling, and normalization. The results of the normalization stage are in the form of datasets that have been labeled positive and negative with the same scale range. The results of the normalization stage can be seen in Table 2.

Table 2. Normalization result

Text	Score Vader Result	Labelling Result
know need hear forget emov menstrual cup forget	-0.4215	negatif
hey friend come posse ilicon menstrual cup don need box say size	0.4939	positif

After that, the dataset is processed at the feature extraction stage with TFIDF. Then to increase accuracy using Particle Swarm Optimization, several parameters are used in it including cognitive learning factor (c1) and social learning factor (c2) which is 1.49, inertia weight (w) is 0.72, the number of iterations is 10, particle size is 10. The cost results from Particle Swarm Optimization can be seen in Table 3.

Table 3. Cost value PSO

k-fold	Cost Value		
	1st try	2nd try	3rd try
1	0.1135804	-0.69908551	0.8782786
2	0.97884806	0.18515851	0.22099704
3	1.19913752	-0.21286452	0.88879046
4	0.42351067	0.28835495	-3.17126388
5	1.77319395	-2.90054524	1.49026702
Final cost	0.1135804	-2.90054524	-3.17126388

Then, the accuracy results of the Naïve Bayes algorithm and Support Vector Machine before applying Particle Swarm Optimization can be seen in Table 4.

Table 4. Accuracy results before applying PSO

Try to-	Accuracy NB	Accuracy SVM
1	91.90%	95.85%
2	90.57%	96.13%
3	92.72%	95.76%
Best Accuracy	92.72%	96.13%

While the results of the accuracy of Naïve Bayes and Support Vector Machine after applying Particle Swarm Optimization can be seen in Table 5.

Table 5. Accuracy results after applying PSO

Try to-	Accuracy NB PSO	Accuracy SVM PSO
1	95.41%	94.66%
2	93.57%	96.22%
3	95.87%	96.68%
Best Accuracy	95.87%	96.68%

From these results, a graph of the increase in accuracy of each algorithm is obtained. The graph of the results of increasing the accuracy of Naïve Bayes with and without Particle Swarm Optimization can be seen in Figure 4.

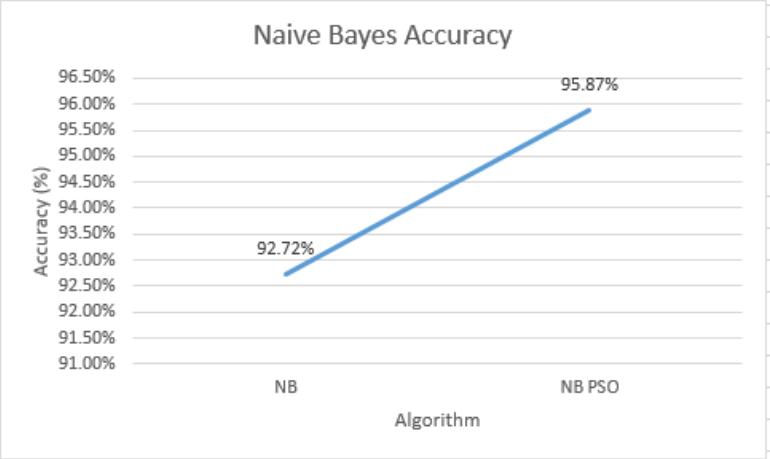

Figure 4. Naïve Bayes accuracy improvement graph

Then the graph of the results of increasing the accuracy of Support Vector Machine with and without Particle Swarm Optimization can be seen in Figure 5.

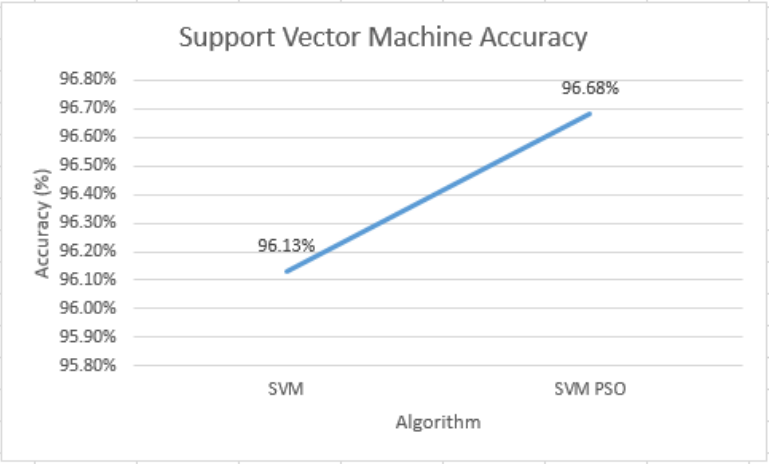

Figure 5. Support Vector Machine accuracy improvement graph

Based on the accuracy values obtained from each algorithm, the results of the accuracy of the application of Naïve Bayes and Support Vector Machine show an increase in accuracy after Partcile Swarm Optimization is applied. Thus, Particle Swam Optimization has proven to be able to work

well to improve the accuracy of the Naïve Bayes classification algorithm and Support Vector Machine for menstrual cup sentiment analysis on Twitter. Then, the comparison of this study with previous studies can be seen in Table 6.

Table 6. The relate research

Research	Algorithm	Accuracy Result
Fauziah, (2020)	NB + IG	83.7%
	SVM + IG	84.2%
Septianingrum & Irawan, (2021)	NB + IG	80.20%
	NB + N-gram	85.10%
	NB + Chi square	87.77%
	NB + PSO	89.09%
Rosadi et al., (2021)	Lexicon based + NB	71%
Windasari et al., (2017)	SVM	86%
Ratino et al., (2020)	NB	72.02%
	SVM	80.23%
	NB + PSO	79.07%
	SVM + PSO	81.16%
Bayhaqy et al., (2018)	KNN	78%
	Naïve Bayes	77%
	Decision Tree	80%

The advantage of this research is there is an increase in the accuracy of the classification algorithm before Particle Swarm Optimization is applied, so the feature selection process with Particle Swarm Optimization is considered to be able to improve the accuracy of the Support Vector Machine and Naïve Bayes classification algorithms when compared with previous research.

While the weakness of this research is the use of Particle Swarm Optimization depends on the iteration value and the number of particles used during system execution. The larger the number of iteration values and the particles used, the longer it will take to execute.

4 Conclusion

This paper examines the Naïve Bayes classification algorithm and Support Vector Machine by applying Particle Swarm Optimization for menstrual cup sentiment analysis on Twitter. The aim is to find out the extent of public understanding related to menstrual cups. Particle Swarm Optimization is used to improve the accuracy of the classification algorithm. The results obtained indicate that the Naïve Bayes classification algorithm and Support Vector Machine have a higher level of accuracy when Particle Swarm Optimization is applied compared to using only the classification algorithm. The accuracy results obtained are for Nave Bayes of 92.72%, Support Vector Machine of 95.87%, Nave Bayes with Particle Swarm Optimization of 96.13%, and Support Vector Machine with Particle Swarm Optimization of 96.68%.

5. References

Alamsyah, A., & Fadila, T. (2021, July). Increased accuracy of prediction hepatitis disease using the application of principal component analysis on a support vector machine. In *Journal of Physics: Conference Series* (Vol. 1968, No. 1, p. 012016). IOP Publishing.

Bayhaqy, A., Sfenrianto, S., Nainggolan, K., & Kaburuan, E. R. (2018). Sentiment Analysis about E-Commerce from Tweets Using Decision Tree, K-Nearest Neighbor, and Naïve Bayes. *2018 International Conference on Orange Technologies, ICOT 2018*, 1–6.

Betesda. (2020). Peningkatan Optimasi Sentimen Dalam Pelaksanaan Proses Pemilihan Presiden Berdasarkan Opini Publik Dengan Menggunakan Algoritma Naïve Bayes Dan Paricle Swarm Optimization. *JSI (Jurnal sistem Informasi) Universitas Suryadarma, 7*(2), 101-114.

Fauziah, D. (2020). *Menstrual Cup Analysis Sentiment in Twitter Using Support Vector Machine and Naïve Bayes Classifier*. *Scientific Journal of Informatics, 6*(1), 1–10

Gholamy, A., Kreinovich, V., & Kosheleva, O. (2018). Why 70/30 or 80/20 Relation Between Training and Testing Sets : A Pedagogical Explanation. *Departmental Technical Reports (CS)*, 1–6.

Giovani, A. P., Ardiansyah, A., Haryanti, T., Kurniawati, L., & Gata, W. (2020). Analisis Sentimen Aplikasi Ruang Guru Di Twitter Menggunakan Algoritma Klasifikasi. *Jurnal Teknoinfo*, *14*(2), 115.

Hamzah, M. B. (2021). Classification of Movie Review Sentiment Analysis Using Chi-Square and Multinomial Naïve Bayes with Adaptive Boosting. *Journal of Advances in Information Systems and Technology*, *3*(1), 67-74.

Harvey, H. B., & Sotardi, S. T. (2018). The Pareto Principle. *Journal of the American College of Radiology*, *15*(6), 931.

Hasan, F. N., & Wahyudi, M. (2018). Analisis Sentimen Artikel Berita Tokoh Sepak Bola Dunia Menggunakan Algoritma Support Vector Machine dan Naive Bayes Berbasis Particle Swarm Optimization. *Jurnal Akrab Juara, 3*(4), 42-55.

Insani, M. I., Alamsyah, A., & Putra, A. T. (2018). Implementation of Expert System for Diabetes Diseases using Naïve Bayes and Certainty Factor Methods. *Sci. J. Informatics*, *5*(2), 185-193.

Jumeilah, F. S. (2017). Penerapan Support Vector Machine (SVM) untuk Pengkategorian Penelitian. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *1*(1), 19–25.

Kristanto, S. P., Prasetyo, J. A., & Pramana, E. (2019). Naive Bayes Classifier on Twitter Sentiment Analysis BPJS of HEALTH. *Proceedings - 2019 2nd International Conference of Computer and Informatics Engineering: Artificial Intelligence Roles in Industrial Revolution 4.0, IC2IE 2019*, 24–28.

Larasati, U. I., Muslim, M. A., Arifudin, R., & Alamsyah, A. (2019). Improve the accuracy of support vector machine using chi square statistic and term frequency inverse document frequency on movie review sentiment analysis. *Scientific Journal of Informatics*, *6*(1), 138-149.

MacReadie, P. I., Bishop, M. J., & Booth, D. J. (2011). Implications of climate change for macrophytic rafts and their hitchhikers. *Marine Ecology Progress Series, 443,* 285–292.

Mariel, W. C. F., Mariyah, S., & Pramana, S. (2018). Sentiment analysis: A comparison of deep learning neural network algorithm with SVM and naïve Bayes for Indonesian text. *Journal of Physics: Conference Series*, *971*(1).

Mustofa, R. L., & Prasetiyo, B. (2021). Sentiment analysis using lexicon-based method with naive bayes classifier algorithm on #newnormal hashtag in twitter. *Journal of Physics: Conference Series*, *1918*(4).

Nooryuda Prasetya, Y., & Winarso, D. (2021). Penerapan Lexicon Based Untuk Analisis Sentimen Pada Twiter Terhadap Isu Covid-19. *Jurnal Fasilkom*, *11*(2), 97–103.

Nurjanah, W. E., Perdana, R. S., & Fauzi, M. A. (2017). Analisis Sentimen Terhadap Tayangan Televisi Berdasarkan Opini Masyarakat pada Media Sosial Twitter menggunakan Metode K-Nearest Neighbor dan Pembobotan Jumlah Retweet. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer (J-PTIIK) Universitas Brawijaya*, *1*(12), 1750–1757.

Ratino, Hafidz, N., Anggraeni, S., & Gata, W. (2020). Sentimen Analisis Informasi Covid-19 menggunakan Support Vector Machine dan Naïve Bayes. *Jurnal JUPITER*, *12*(2), 1–11.

Sateria, A., Saputra, I. D., & Dharta, Y. (2019). Penggunaan Metode Particle Swarm Optimization (PSO) pada Optimasi Multirespon Gaya Tekan dan Momen Torsi Penggurdian Material Komposit Glass Fiber Reinforce Polymer (GFRP) yang ditumpuk dengan Material Stainless Steel (SS). *Manutech : Jurnal Teknologi Manufaktur*, *10*(01), 1–7.

Wisnu, H., Afif, M., & Ruldevyani, Y. (2020). Sentiment analysis on customer satisfaction of digital payment in Indonesia: A comparative study using KNN and Naïve Bayes. *Journal of Physics: Conference Series*, *1444*(1).