International Journal of Advances in Data and Information Systems
Vol. 2, No. 1, April 2021, pp. 96~104
ISSN: 2721-3056, DOI: 10.25008/ijadis.v2i1.1224

Sentiment Analysis Using Naive Bayes Algorithm with Feature Selection Particle Swarm Optimization (PSO) and Genetic Algorithm

Abi Rafdi[1], Herman Mawengkang[2] , Syahril Efendi[3]
[1,2,3] Faculty of Computer Science and Information Technology, University of Sumatera Utara, Indonesia
[1]abirafdikoto30@gmail.com, [2]hmawengkang@usu.ac.id, [3]syahril1@usu.ac.id *

Article Info

Article history:
Received Jun 26, 2021
Revised Sept 13, 2021
Accepted Oct 22, 2021

Keywords:
Sentiment Analysis
Twitter
Naive Bayes
Feature Selection
Particle Swarm Optimization
Genetic Algorithm

ABSTRACT

This study analyzes Sentiment to see opinions, points of view, judgments, attitudes, and emotions towards creatures and aspects expressed through texts. One of Social Media is like Twitter is one of the most widely used means of communication as a research topic. The main problem with sentiment analysis is voting and using the best feature options for maximum results. Either, the most widely known classification method is Naive Bayes. However, Naive Bayes is very sensitive to significant features. That way, in this test, a comparison of feature selection is carried out using Particle Swarm Optimization and Genetic Algorithm to improve the accuracy performance of the Naive Bayes algorithm. Analyses are performed by comparing before and after testing using feature selection. Validation uses a cross-validation technique, while the confusion matrix is appealed to measure accuracy. The results showed the highest increase for Naïve Bayes algorithm accuracy when using the feature selection of the Particle Swarm Optimization Algorithm from 60.26% to 77.50%, while the genetic algorithm from 60.26% to 70.71%. Therefore, the choice of the best characteristics is Particle Swarm Optimization which is superior with an increase in accuracy of 17.24%.

Corresponding author:
Herman Mawengkang,
Department of Information Technology,
Faculty of Computer Science and Information Technology
University of Sumatera Utara
Medan, North Sumatra,
Email: hmawengkang@usu.ac.id

1. INTRODUCTION

Information. Extraction is the technique of gathering information from an unstructured collection of texts. It is necessary to define target information as structured information to be extracted [1]. Furthermore, information can be obtained from the data regarding social media users' opinions towards certain entities. So that all opinion data can be helpful, then data processing can be done using sentiment analysis [2]. It's also important to pay attention to sentence structure, use of non-formal language, emoticons, or image services. The main task in this analysis is to classify the sign of text at the sentences, document, sentence, feature, or page of level, in case documents, corrections, or positive entity segments.

Rifat et al. In his research, he tried to compare the Support Vector Machine and K-Nearest Neighbor to analyze sentiment. Naive Bayes multinomial algorithm offers the best performance compared to other traditional machine learning algorithms. [3]

Classification is one of the data mining methods used to predict a value or a set of data. Classification methods are use to describe data or forecast data trends in the future. The biggest challenge in classification research in data mining is the class imbalance problem, and one of the solutions proposed by the researcher is feature selection [4].

Feature selection is one way that can affect the level of classification accuracy. Which one, an optimization of features selection will minimize a process of large number for features. A relatively

low subset of characteristics is essential to quickly and effectively improve classification accuracy. [5]. In his research, using 3 correct selection features can make the classifier run well, more effectively and efficiently by minimizing the amount of data analyzed [6].

Sean & Dwi, In his research, he compared the K-Nearest Neighbor and SVM methods into sentiment analysis. Support Vector Machine (SVM) algorithm offers the best performance. [7]. Meanwhile, in your research. Farkhund Iqbal et al., A Hybrid Framework for Sentiment Analysis Using Genetic Algorithm-Based Feature Reduction. With this hybrid approach, we've reduced the size of the feature set size by 42% without sacrificing accuracy. Comparison of our feature reduction technique with Principal Component Analysis (PCA) and the more widespread feature reduction technique based on latent semantic analysis (LSA) has increased precision of up to 15% compared to PCA and increased accuracy. up to 0.2% via LSA [8].

From some of the information above, the authors are motivated to employ the Particle Swarm Optimization (PSO) and Genetic Algorithm (GA) methods to perform features selection to improve accuracy. The results of feature selection in this study are expected to increase the accuracy of Naïve Bayes.

2. RESEARCH METHODS

2.1 DATA COLLECTION

In this study, the author is using qualitative research because the data obtained will be words. These words are obtained from the Twitter Search API crawling process. These words will then be processed to find out the sentiments in them by extracting information to classify the polarity of the text on feature entities, either positive or negative. In this study, we focus on tweets related to vaccine keywords. Because this activity is a hot topic at the moment, this study uses three kinds of data, namely tweet data, stop words, and essential words. In addition, literature studies and references to national and international journals are also needed to gain additional knowledge related to theoretical foundations, analytical concepts, and methods in data classification.

2.2. RESEARCH

In this study, the process is carried out on text mining. Text mining is a text analysis process carried out automatically by a computer, the objective of which is to obtain quality information from the text summarized in a document. The main process in this method is to find words that can represent the content of the manuscript to further analyze the relationship between documents using certain statistical methods such as group analysis, classification, and association. [9].

The steps taken in text extraction are as follows:

1. Word preprocessing The actions taken at this stage include lowercase letters and tokenizing.
2. Feature Selection The action taken at this stage is eliminating stopwords and stemming [10].

The data collection is in tweet data taken from Twitter social media users' crawling with the word vaccine search. In addition, literature studies and references to national and international journals are also need to gain additional knowledge related to theoretical foundations, analytical concepts, and methods in data classification.

2.3. SENTIMENT ANALYSIS

Sentiment analysis or else opinion mining is a subject of study in which evaluation, opinion, even judgment, an attitude, and emotion towards an entity such as product, organization, service, individual, events, problem, and subject. [11] This analysis is used to obtain specific information from the existing dataset. Sentiment analysis focuses on elaborating opinions that contain polarity, which have a positive or negative sentiment value. Sentiment class labeling was performed with Lexicon-based functions. Then find the sentiment value in the sentence with the formula:

$$S_{positive} = \sum_{i=1n}^{n} positive\ score_i$$

$$S_{negative} = \sum_{i=1}^{n} negative\ score_i$$

$S_{positive} =$ number of words of positive opinion

$S_{negative} =$ number of words of negative opinion

So that in one sentence, the number of negative and positive values can be determined by the meaning each word. Next, a comparison will be conduct to see if the sentence has a positive or negative.

$$\text{sentiment class} \begin{cases} \text{positive If } S_{positive} > S_{negative} \\ \text{neutral If } S_{positive} = S_{negative} \\ \text{negative If } S_{positive} < S_{negative} \end{cases}$$

2.4. NAIVE BAYES ALGORITHM

The Naive Bayes is one of the algorithms included in the classification technique. British scientist Thomas Bayes proposed probability and statistical methods that predict future possibilities based on previous experience. The theorem is combined with Naive, assuming that the conditions between attributes are not related. Naive Bayes assumes for presence or absence of a unique characteristic of one class has nothing to do with the features of another type. The equations of Bayes' theorem are the stages in the Naïve Bayes method, namely:

Counting the amount of data

a. Finding the probability value (P)

$$P(x) = \frac{E}{n} \quad (2.2)$$

b. Finding the mean (μ)

$$\mu = \frac{\sum_i^n = 1 X_i}{n} \quad (2.3)$$

c. Finding the standard deviation

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{n-1}} \quad (2.4)$$

d. Classifying continuous data with the Gauss Density formula

$$P(X_i = x_i | Y = y_j = \frac{1}{\sqrt{2\pi\sigma ij}} e^{-\frac{(x_i - \mu_{ij})^2}{2\sigma_{ij}^2}} \quad (2.5)$$

e. Finding the posterior value

$$Posterior(X) = \frac{P(X)P(Atribut1\,|X)\ P(Atribut2\,|X)\ P(Atribut..n|X)}{P(X)P(Atribut1\,|X)\ P(Atribut2\,|X)P(Atribut..n|X)+ P(Y)P(Atribut1\,|Y)\ P(Atribut2\,|Y)P(Atribut..n|Y)}$$

$$Posterior(Y) = \frac{P(X)P(Atribut1\,|Y)\ P(Atribut2\,|Y)\ P(Atribut..n|Y)}{P(X)P(Atribut1\,|X)\ P(Atribut2\,|X)P(Atribut..n|X)+ P(Y)P(Atribut1\,|Y)\ P(Atribut2\,|Y)P(Atribut..n|Y)} \quad (2.6)$$

2.5. PARTICLE SWARM OPTIMIZATION

Particle Swarm Optimization is a branch of changing of algorithm that upon a population of particles that maintains a probability distribution to find the optimal solution. PSO is based on the deportment of a bird or fish community. If a pack doesn't have an alpha to lead them in search of food, they will wander around looking for food locations. This algorithm is based on the social behaviors of animals. Individual actions plus the influence of 4,444 others create social behavior

PSO is the search of solutions that are carried out through a population composed of several particles.

A population is given rise to randomly from an initial value to the maximum value with the least significant value. Some of the particles represent the position and location of the problem in question. Each particle seeks the optimal solution by making adjustments to the work of the best particle. The company offers special particle effects worldwide, which is the best value for their dollar. Search a wide area to explore various paths in the search space. Every solution, it's represented by the particle position. It's evaluated performance by calculating the answers and entering them into the sufficiency function in each iteration. A point in a certain spatial dimension for a particle that is treated as a point in a certain measurement forgives the position of the particle in the search space, the position of X,

The following are equations that describe position and velocity:

$$X_i(t) = x_{i1}(t), x_{i2}(t), x_{i3}(t), \dots, x_{iN}(t) \quad (1)$$

$$\boldsymbol{V_i(t) = v_{i1}(t), v_{i2}(t), v_{i3}(t), \dots, v_{iN}(t)} \quad (2)$$

Which X is the position for particle, V is the speedup of the particle. i and t are particles indices and t- iteration in N-dimensional space. Furtherore, to the explanation of the means for enchancing particle health with the following calculation model.

$$\boldsymbol{V_i(t) = wV_i(t-1) + c_1r_1\left(X_i^L X_i(t-1)\right) + c_2r_2(X^G - X_i(t-1))} \quad (3)$$

$$\boldsymbol{X_i(t) = V_i(t) + X_i(t-1)} \quad (4)$$

X_i^L=,, ..., represents the best locale of the i-th particle. While = ,, ..., represents the whole best herd in the World. And is a constant that has a positive value which is usually called the learning factor. And is the positive random numbers between 1 with 0. is the inertia parameter. Equation (2.20) is used to get the new particle velocity. It is substructure as was pace, the distances among as current position, the best local positions, and the current longitude of the best global job. Then the particles fly according to the equation. 2.21. The PSO workflow can be see in Figure 2.4
$X_{i1}^L\ X_{i2}^L X_{iN}^L X^G X_1^G\ X_2^G\ X_N^G c_1 c_2 r_1 r_2 w$

2.6. GENETIC ALGORITHM

Genetic Algorithm(GA) is a genetic algorithm that belongs to the group of evolutionary algorithms. This algorithm was first introduced by Holland in 1975 and is a method commonly used for search methods and is inspired by population genetics in finding solutions to problems. This algorithm also follows the concept of Carles Darwin with his theory of evolution, where strong individuals will survive the population. The essential elements of natural genetics are natural selection (natural selection), crossover (crossover), and mutation (mutation).

Natural selection is an attempt to retain the best individuals by multiplying the best individuals. So the best is not lost in the next iteration. The interbreeding operator is used to create individuals. To create a new individual, it takes two parents. Two parents are required. The most commonly used parent selection technique is the roulette wheel. The mutation operator is used to replace the individual.

Worst with new individuals. The number of individuals replaced depends on the mutation rate parameter.

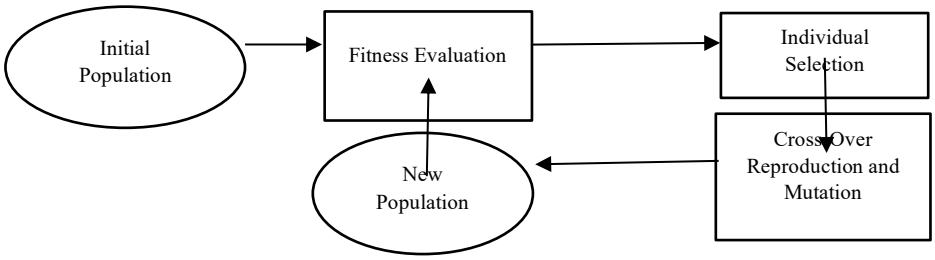


Figure 1. Gentic Cycle Algorithm

3. RESULTS AND DISCUSSION

3.1. DATA SET

In this study, the authors used qualitative research with 1000 respondents. Then the writing is in the sentiment analysis to find out the word is positive or negative with the vaccine keyword. The data has been cleaned to anticipate the occurrence of missing values.

3.2. TESTING PROCESS

The test is carried out following a predetermined scenario, namely the Naive Bayes classification method with feature selection and Particle Swarm Optimization and Genetic Algorithm. The test scenario was carried out three times, each using cross-validation. The first sample was tested with the Naive Bayes algorithm without feature selection. The second test was carried out using the Naive Bayes algorithm with Particle Swarm Optimization feature selection. The last test is done by using a Naive Bayes algorithm with a Genetic Algorithm feature selection. The results of the three algorithm tests will be compared, and the algorithm with the highest accuracy will be used. Before conducting the test, it is necessary to determine the Population Size to increase the accuracy value used. Population size is the number of items in the population from the sample taken. Then the best Population size will be taken as an example to be applied. The accuracy results can be seen in the population size accuracy table below.

Table 1. Population Size Accuracy

Population size	accuracy	ROC curve
5	66.05%	8.6
10	77.50%	0.720
20	75.64%	0.712
40	73.69%	0.711

Based on the above test, it was found that the highest Population Size accuracy value was found in population size 10 with an accuracy value of 77.50% and an AUC of 0.720. Then the population size value will be used to increase the accuracy of feature selection in testing the Naïve Bayes algorithm with a predetermined feature selection. Confusion Matrix values and ROC curves can be seen in the image below:

accuracy: 73.69% +/- 8.14% (micro average: 73.68%)

	true positif	true negatif	class precision
pred. positif	101	37	73.19%
pred. negatif	18	53	74.65%
class recall	84.87%	58.89%	

Figure 2. Confusion Matrix Population Size

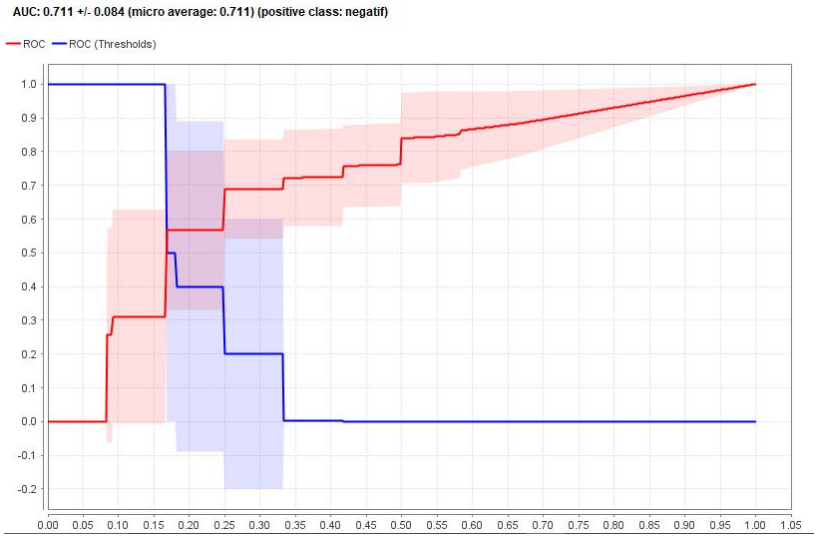


Figure 3. ROC Matrix Population Size Curve

3.2.1 NAIVE BAYES ALGORITHM TESTING PROCESS

In this test, an experiment was done by Naive Bayes without features selection. The results of the test are shown in the table below :

Table 2. Accuracy of Naïve Bayes Algorithm

Test	*Accuracy*	AUC
2	56.00%	0.422
3	57.90%	0.493
4	56.92%	0.523
5	59.33%	0.499
6	57.91%	0.457
7	60.26%	0.519
8	57.96%	0.500
9	55.54%	0.475
10	59.74%	0.539

Based on the above test, it was found that the highest accuracy value was found in tests carried out with k-fold 7 experiments with an accuracy value of 60.26%% and an AUC of 0.519.

3.2.2 PROCESS OF TESTING NAIVE BAYES ALGORITHM WITH FEATURE SELECTION PARTICLE SWARM OPTIMISATION (PSO)

In this test, experiments were carried out using the Naive Bayes algorithm with Particle Swarm Optimization (PSO) feature selection. The test results can be seen in the table below:

Table 3. Accuracy of Naïve Bayes Algorithm with PSO

Test	*Accuracy*	AUC
2	65.05%	0.552
3	69.85%	0.636
4	66.02%	0.604
5	75.10%	0.701
6	74.62%	0.718
7	65.04%	0.621
8	61.27%	0.457
9	69.89%	0.653
10	77.50%	0.720

Based on the above test, it was found that the highest accuracy value was found in the tests carried out with the k-fold 10 experiment with an accuracy value of 77.50% and an AUC of 0.720.

accuracy: 77.50% +/- 6.79% (micro average: 77.51%)

	true POSITIVE	true NEGATIVE	class precision
pred. POSITIVE	106	34	75.71%
pred. NEGATIVE	13	56	81.16%
class recall	89.08%	62.22%	

Figure 4. Confusion Matrix Accuracy PSO

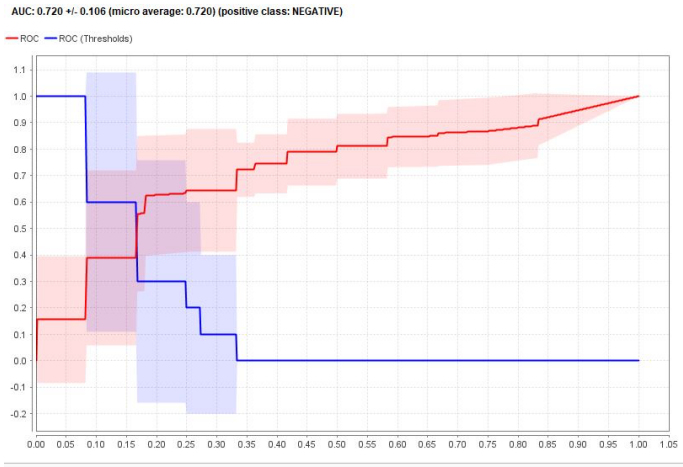


Figure 5. PSO Matrix ROC Curve

3.2.2 NAIVE BAYES ALGORITHM TESTING PROCESS WITH GENETIC ALGORITHM SELECTION FEATURE (GA)

In this test, experiments were carried out using the Naive Bayes algorithm with feature selection Genetic Algorithm (GA). The test results can be seen in the table below:

Table 4. Accuracy of Naïve Bayes Algorithm with GA

Test	*Accuracy*	AUC
2	70.34%	0.657
3	69.39%	0.711
4	69.86%	0.678
5	69.91%	0.624
6	71.78%	0.725
7	66.96%	0.625
8	69.87%	0.706
9	71.78%	0.651
10	70.31%	0.690

Based on the above test, it was found that the highest accuracy value was found in tests carried out with the k-fold 6 experiment. with an accuracy value of 71.78% and an AUC of 0.725.

accuracy: 71.78% +/- 5.18% (micro average: 71.77%)

	true POSITIVE	true NEGATIVE	class precision
pred. POSITIVE	115	55	67.65%
pred. NEGATIVE	4	35	89.74%
class recall	96.64%	38.89%	

Figure 6. Confusion Matrix Accuracy GA

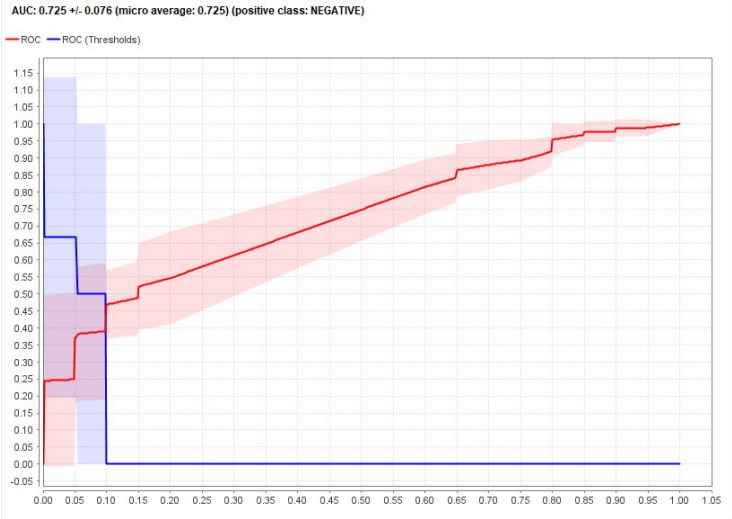
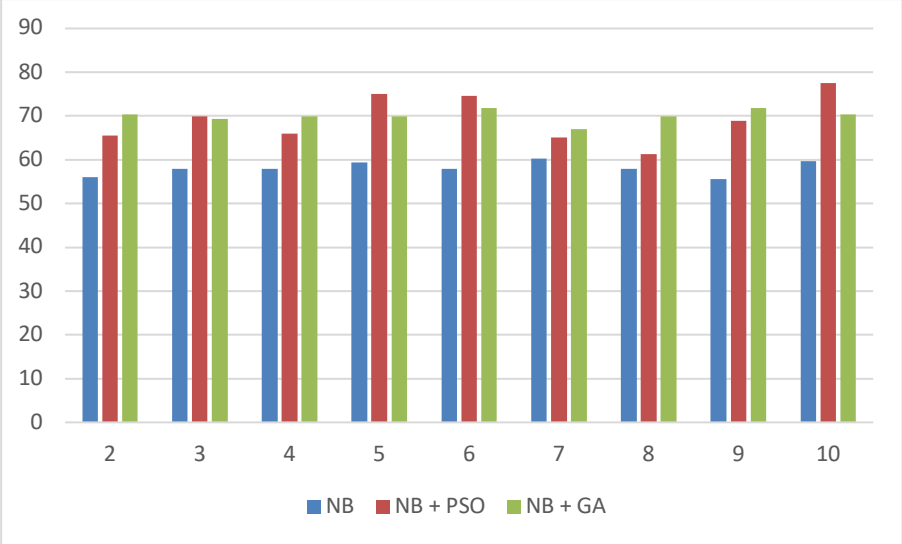


Figure 7. GA Matrix ROCCurve

Based on the tests that have been carried out, the following is a comparison chart table of the tests that have been carried out with Naive Bayes, Naive Bayes combinations of PSO, and Naive Bayes combinations of GA.

Table 5. Graph of Test Comparison



4. CONCLUSION

Based on the testing and analysis, it's been concluded that using the selected feature in the Naïve Bayes algorithm for Twitter sentiment analysis can help improve the performance and accuracy of the tests carried out. The process shows that the Naïve Bayes algorithm model produces the highest level of accuracy of 60.26%. For comparison, the Naïve Bayes algorithm model, which has been combined with the Particle Swarm Optimization (PSO) feature selection, shows the highest accuracy of 77.50%. In contrast to the Naïve Bayes algorithm model, which has been combined with feature selection Genetic Algorithm (GA), which shows the highest accuracy of 71.78%. So the conclusion of the combination between Naive Bayes and PSO is a better result which one an increase in accuracy of 17.24%.

5. REFERENCE

[1] Vikas, BO & Mungara, J. "Enhanced Extraction and Summarization Techniques with User Review Data for Product Recommendations to Customers. " International Journal of Scientific Research in Science, Engineering and Technology, vol 2, p. 25–30, 2016.

[2] A., Pappu Rajan., & SP Victor. Web Sentiment Analysis to Print Positive or Negative Words Using Twitter Data, International Journal of Computer Applications, vol. 96, p. 6, 2014.
[3] Ramadhani, Rif'at Ahdi., Indriani, Fatma & T. Nugrahadi, Dodon., Comparison of Naive Bayes smoothing methods for Twitter sentiment analysis, 3rd International Conference on Information Technology, Information System and Electrical Engineering (ICITISEE) 2016.https://doi.org/10.1109/ICACSIS.2016.7872720
[4] Zhang, X., Shi, Z., Liu, X., & Li, X. A Hybrid Feature Selection Algorithm For Processing Unbalanced Classification Data. IEEE International Conference on the Smart Internet of Things (SmartIoT) 2018, 269–275, 2018.https://doi.org/10.1109/smartiot.2018.00055
[5] Pant, H & Srivastava, R. A Survey of Feature Selection Methods For Unbalance Datasets. International Journal of Computer Engineering and Applications, vol 9 no. 2, 197–204, 2015.
[6] The Lion, Ardiles and Murnawan. Analysis of Decision Support System Models for Proposed Activities at the District Level Development Planning Forum, IEEE Conference on Energy Internet and Energy System Integration (EI2), 2017.https://doi.org/10.1109/CITSM.2016.577522
[7] Semuel Istia, Sean & Dwi Purnomo, Hindriyanto. " A Sentiment Analysis of Law Enforcement Performance Using Support Vector Machine and K-Nearest Neighbor" , 3rd International Conference on Information Technology, Information System and Electrical Engineering (ICITISEE) 2018.https://doi.org/ 10.1109/ICITISEE.2018.8720969
[8] Iqbal, Farkhund., Maqbool Hashmi, Jahanzeb & CM Fung, Benjamin. A Hybrid Framework for Sentiment Analysis Using Genetic Algorithm Based Feature Reduction, IEEE Access. vol 7, pp 14637 - 14652 , 2019.https://doi.org/10.1109/ACCESS.2019.2892852
[9] Amrane, M., Oukid, S., Gagaoua, I. & Ensari, T. 2018. Classification of Breast Cancer Using Machine Learning. IEEE Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), p. 115-116, 2018.
[10] A, Sarlan., C, Nadam., & S, Basri. Twitter sentiment analysis, Conf. proc. - Int.6 conf. Inf. Technology. Multimed. UNITEN Cultivation. Make. Enabling Technology. Through Internet of Things, ICIMU 2014, no. November 2016
[11] Normawati, D., & Winarti, S. Feature Selection Using Data Mining Based on Variable Precision Rough Set (VPRS) for Diagnosis of Coronary Heart Disease. Scientific Journal of Computer Electrical Engineering and Informatics, vol 3 no 2, page 100, 2018.https://doi.org/10.26555/jiteki.v3i2.8072