Jurnal Teknik Informatika (JUTIF)
Vol. 4, No. 3, June 2023, pp. 521-528

DOI: https://doi.org/10.52436/1.jutif.2023.4.3.881
p-ISSN: 2723-3863
e-ISSN: 2723-3871

FUEL INCREASE SENTIMENT ANALYSIS USING SUPPORT VECTOR MACHINE WITH PARTICLE SWARM OPTIMIZATION AND GENETIC ALGORITHM AS FEATURE SELECTION

Laura Imanuela Mustamu[*1], Yuliant Sibaroni[2]

[1,2]Informatics, School of Computing, Telkom University, Indonesia
Email: [1]elamustamu@student.telkomuniversity.ac.id, [2]yuliant@telkomuniversity.ac.id

(Article received: February 02, 2023; Revision: March 11, 2023; published: June 26, 2023)

Abstract

BBM, or fuel oil, is one of the essential needs of the Indonesian people. The government's policy regarding the increase in fuel prices raises many opinions from the public. Twitter is one of the social media that Indonesian people often use to express opinions on a topic. In this study, sentiment analysis was carried out on public opinion regarding the fuel price increase policy from Twitter social media. This research is expected to help determine public opinion regarding the fuel price increase policy with positive, neutral and negative sentiments. The sentiment analysis method used is the Support Vector Machine (SVM) classification algorithm. The results of the accuracy of SVM were compared with accuracy by adding a feature selection process. The Particle Swarm Optimization (PSO) and Genetic Algorithm (GA) algorithms are used for the feature selection method. After several experiments using the three methods, the SVM method with the Radial Basis Function (RBF) kernel produced the best accuracy of 71.2%. The combination of the SVM method with the RBF and PSO kernels obtained an accuracy of 68.84%, and the combination of the RBF and GA kernel SVM methods obtained an accuracy of 69.52%.

Keywords: *fuel increase, GA, sentiment analysis, SVM, PSO.*

1. INTRODUCTION

BBM, or fuel oil, is a fuel derived from fossil fuels. Currently, fuel is one of the essential needs of the Indonesian people. Based on a report from the Ministry of Energy and Mineral Resources, the use of fuel in Indonesia, especially pertalite, a subsidized fuel, has increased significantly after the pandemic occurred in 2020, from 18.1 million KL to 23 million KL in 2021 [1]. Pertalite is also the fuel most widely used by the public, which is almost 80% compared to other types of fuel such as Pertamax and Solar[1]. On September 3, 2022, the Indonesian government officially announced a new policy regarding fuel price hikes. The price increases occurred for three types of fuel, namely Pertamax, Pertalite and Diesel Fuel. This new policy has generated much public opinion because pertalite, the most widely used fuel type, has also experienced a price increase. Various opinions from the public can be categorized into three, namely those that support, reject, and be neutral towards the policy.

Many people's opinions are expressed on social media, including Twitter. Twitter is one of the social media used by Indonesian people. Twitter users in Indonesia until 2022 have reached 18.45 million users. This number increased by 31.3% from the previous year, with 14.05 million users[2]. In this study, an analysis of the sentiment of Indonesian public opinion regarding the fuel price increase policy was carried out. Sentiment analysis is processing text data that aims to retrieve existing sentiments in an opinion about a topic from various fields[3]. In this study, sentiment analysis is categorized into three: positive, neutral, and negative. From this study, public sentiment regarding the fuel price hike policy can be identified.

The method used in this study for sentiment analysis is the Support Vector Machine (SVM) algorithm. SVM is a supervised learning method that works by looking for hyperplanes where this hyperplane functions to separate data from two or more different types of classes. SVM has been widely used in research on sentiment analysis because it produces good accuracy. However, in text classification, the existing data has high dimensions or too many features. Therefore, feature selection is carried out. Feature selection is carried out to reduce data dimensions or features to improve the classification algorithm's performance[4]. The features intended in this study are the words in tweets that do not affect existing tweets. The feature selection methods used in this study are Particle Swarm Optimization (PSO) and Genetic Algorithm (GA).

In research[5], researchers conducted experiments using SVM, SVM and GA feature selection, SVM and PSO feature selection, as well as SVM and Principal Component Analysis (PCA) feature selection which then produced accuracy

values of 82.00%, 94.00%, 97.00%, and 83.00%. Accuracy value in research[5] shows that adding the feature selection process can improve accuracy.

In research[6], researchers conducted research with Information Gain (IG) feature selection and Naïve Bayes (NB) classification. The accuracy result for NB is 73.2% then the classification result for NB, which is added with IG feature selection, produces an accuracy of 73.2%. From research[6], adding feature selection has no effect on the accuracy value because there is no increase in the accuracy value.

Researchers in research[7] used the PSO feature selection and NB classification algorithms. The accuracy for combining PSO and NB is 86% which is 6% greater than the accuracy using only the NB algorithm. Based on the background and research, this paper will compare the SVM accuracy values without PSO and GA feature selection with the SVM accuracy values with PSO and GA feature selection. The comparisons obtained show the effect of PSO and GA feature selection on the results of SVM accuracy in this study.

This research uses these three methods because in several existing studies, there has been no research using datasets on the topic of increasing fuel prices. In addition, the resulting accuracy values always increase or do not change after adding a feature selection process in previous study. Through this research, it is able to prove whether by adding feature selection, the accuracy value of the SVM classification will increase or decrease. From the research objectives described, questions will arise, including how the data will be obtained, how the pre-processing of the data will be carried out, how to perform feature selection with PSO and GA, and how the results are accurate with the method used.

2. RESEARCH METHODS

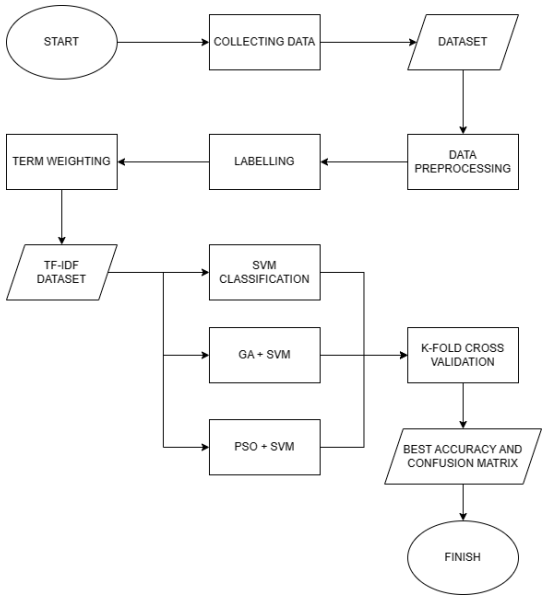

Figure 1. System Architecture

Figure 1 describes the process flow of the research conducted. The research process starts from collecting tweet data which become the dataset in this study. After the dataset are collected, the data then go through a pre-process where the data will be cleaned. The incoming dataset will also be labeled. Next, the data will go through a word weighting process using the Term Frequency–Inverse Document Frequency (TF-IDF) method which will produce a new dataset containing word weights. This new dataset will be classified using the SVM method, SVM with PSO feature selection, and SVM with GA feature selection. The results of the classification will be evaluated using the K-Fold Cross Validation and Confusion Matrix.

2.1. Dataset

In this study, the dataset used is a tweet on Twitter regarding the fuel price hike policy. Data is taken using the python library, namely snscrape. Snscrape is a python library that retrieves data without requiring an Application Programming Interface (API)[8]. By using this library, tweet data is retrieved based on the specified keywords, namely "kenaikan bbm", "bbm naik", or "#kenaikanbbm". Data collection is divided into two, before the occurrence of the policy and after the occurrence of the policy. Table 1 shows an example of the dataset.

Table 1 Dataset

Dataset
@hariankompas @APEKSIid Syukurlah, Pemerintah Jokowi cepat tanggap memberikan dukungan berupa stimulus kenaikan harga BBM dan Pajak PPN 11%.
@dr_koko28 @CNNIndonesia Khan naik BBM biar subsidi dialihkan untuk IKN. 🤣
KAMMI Sulsel Tanggapi Rencana Kenaikan BBM Subsidi: Terkait rencana kenaikan harga BBM bersubsidi oleh Pemerintah, KAMMI Sulsel berharap kebijakan tersebut perlu dikaji ulang dengan alasan… https://t.co/wMugwakEPH #Sulsel #kammisulsel #kenaikanbbm #kammisulselkenaikanbbm

2.2. Data Preprocessing

1. Casefolding

Table 2. Before and After Casefolding

Before Casefolding	After Casefolding
@teguhinvestor Saya cuma bingung , di rezim sebelumnya yang demo BBM naik pada kemana yah ...	@teguhinvestor saya cuma bingung , di rezim sebelumnya yang demo bbm naik pada kemana yah ...
Kan mereka ngaku wong cilik ...	kan mereka ngaku wong cilik ...
Apa wong cilik di negeri ini udah pada kaya raya semua yah ?	apa wong cilik di negeri ini udah pada kaya raya semua yah ?
Yang biasa makan di luar pasti tau kenaikan itu	yang biasa makan di luar pasti tau kenaikan itu
Pura pura hidup senang itu ngk enak padahal	pura pura hidup senang itu ngk enak padahal

At the case folding stage, all capital letters in the tweet will be converted to lowercase. In table 2 it can

be seen the difference in text after the case folding process.

2. Cleaning and Tokenizing

Tweets will be purged of punctuation, hashtags, emoji, URLs, RTs, numbers, mentions of usernames, and newlines. Then, after the tweets that have been cleaned will be tokenized based on the existing spaces. Tokenizing is the process of separating each word in a sentence [9]. Table 3 shows the difference in the text after the cleaning and tokenizing processes.

Table 3. Before and After Cleaning and Tokenizing

Before Cleaning and Tokenizing	After Cleaning and Tokenizing
@teguhinvestor saya cuma bingung , di rezim sebelumnya yang demo bbm naik pada kemana yah ...	['saya', 'cuma', 'bingung', 'di', 'rezim', 'sebelumnya', 'yang', 'demo', 'bbm', 'naik', 'pada', 'kemana', 'yah', 'kan',
kan mereka ngaku wong cilik ...	'mereka', 'ngaku', 'wong', 'cilik', 'apa', 'wong', 'cilik', 'di',
apa wong cilik di negeri ini udah pada kaya raya semua yah ?	'negeri', 'ini', 'udah', 'pada', 'kaya', 'raya', 'semua', 'yah',
yang biasa makan di luar pasti tau kenaikan itu	'yang', 'biasa', 'makan', 'di', 'luar', 'pasti', 'tau', 'kenaikan', 'itu', 'pura', 'pura', 'hidup',
pura pura hidup senang itu ngk enak padahal	'senang', 'itu', 'ngk', 'enak', 'padahal']

3. Stopwords Removal

Stopword removal is a process that aims to remove words that are not important or have no effect on a tweet[10]. Examples of words that will be removed in this process are “wadduh”, “saya”, “sdg”, “engeh”, “ente”, “deket”, “dg”, dan “di”. Table 4 shows the difference in text after the stopwords removal process.

Table 4. Before and After Stopwords Removal

Before Stopwords Removal	After Stopwords Removal
['saya', 'cuma', 'bingung', 'di', 'rezim', 'sebelumnya', 'yang', 'demo', 'bbm', 'naik', 'pada', 'kemana', 'yah', 'kan', 'mereka', 'ngaku', 'wong', 'cilik', 'apa', 'wong', 'cilik', 'di', 'negeri', 'ini', 'udah', 'pada', 'kaya', 'raya', 'semua', 'yah', 'yang', 'biasa', 'makan', 'di', 'luar', 'pasti', 'tau', 'kenaikan', 'itu', 'pura', 'pura', 'hidup', 'senang', 'itu', 'ngk', 'enak', 'padahal']	['cuma', 'bingung', 'rezim', 'sebelumnya', 'demo', 'bbm', 'naik', 'ngaku', 'wong', 'cilik', 'apa', 'wong', 'cilik', 'negeri', 'udah', 'kaya', 'raya', 'semua', 'biasa', 'makan', 'luar', 'tau', 'kenaikan', 'pura', 'pura', 'hidup', 'senang', 'enak', 'padahal']

4. Normalization

Table 5. Before and After Normalization

Before Normalization	After Normalization
['cuma', 'bingung', 'rezim', 'sebelumnya', 'demo', 'bbm', 'naik', 'ngaku', 'wong', 'cilik', 'apa', 'wong', 'cilik', 'negeri', 'udah', 'kaya', 'raya', 'semua', 'biasa', 'makan', 'luar', 'tau', 'kenaikan', 'pura', 'pura', 'hidup', 'senang', 'enak', 'padahal']	['cuma', 'bingung', 'rezim', 'sebelumnya', 'demo', 'bbm', 'naik', 'mengaku', 'wong', 'cilik', 'apa', 'wong', 'cilik', 'negeri', 'sudah', 'kaya', 'raya', 'semua', 'biasa', 'makan', 'luar', 'tahu', 'kenaikan', 'pura', 'pura', 'hidup', 'senang', 'enak', 'padahal']

The normalization that is carried out in this process is to return the abbreviated words to their actual tenses. Table 5 shows the difference in text after the normalization process.

5. Stemming

In the stemming process, words in tweets that have affixes such as "-nya", "be-" and "-me" will be returned to their root words. Table 6 shows the difference in text after the stemming process.

Table 6. Before and After Stemming

Before Stemming	After Stemming
['cuma', 'bingung', 'rezim', 'sebelumnya', 'demo', 'bbm', 'naik', 'mengaku', 'wong', 'cilik', 'apa', 'wong', 'cilik', 'negeri', 'sudah', 'kaya', 'raya', 'semua', 'biasa', 'makan', 'luar', 'tahu', 'kenaikan', 'pura', 'pura', 'hidup', 'senang', 'enak', 'padahal']	['cuma', 'bingung', 'rezim', 'belum', 'demo', 'bbm', 'naik', 'aku', 'wong', 'cilik', 'apa', 'wong', 'cilik', 'negeri', 'sudah', 'kaya', 'raya', 'semua', 'biasa', 'makan', 'luar', 'tahu', 'naik', 'pura', 'pura', 'hidup', 'senang', 'enak', 'padahal']

2.3. Labelling

Labelling in this study was done manually. Text labelling is done based on sentiment with three labels, namely -1, 0, and 1. Negative sentiment is labelled -1, neutral sentiment is labelled 0, and positive sentiment is labelled 1. Table 7 shows an example of the labelling done.

Table 7. Labelling Example

Text	Label
"@detikcom Mestinya menolak, apapun alasan menaikkan harga BBM. Mengingat dan menimbang tangisan terisak-isak jaman pemerintahan SBY. Ayo menangis lagi, harga BBM mau naik."	-1
Ringankan Beban Warga Dampak Kenaikan BBM, Polsek Batang Kuis Bagikan Sembako https://t.co/1jEelHbgH0	0
@zakicha29230025 @M45Broo_ Knp smp demo berjilid² stlh ada kenaikan bbm pd hal sdh ada penjelasan rasional dr menkeu Ibu Sri Mulyani knp hrs menaikkan harga bbm?? Jwbnya ya krn kadrun, knp jd kadrun? Ya krn tak waras...jd hanya sekitar itu pusaran keberpikiran kadrun, tdk meleset.	1

2.4. Term Weighting

Term weighting is one of the processes required in machine learning text classification. Term weighting aims to give weight to words in sentences that are in a dataset. In this study, term weighting was carried out using the Term Frequency – Inverse Document Frequency (TF-IDF) method. The TF-IDF method is a method that has been widely used for word weighting[11]. This method will calculate the value of Term Frequency (TF) and Inverse Document Frequency. TF will calculate the weight of how often a term appears in a document[12]. TF is the frequency of occurrence of a term in a document. TF can be calculated by equation 1.

$$tf_{d,t} \quad (1)$$

Meanwhile, IDF will calculate the proportion of a document containing the term to all documents. In the IDF calculation, words with a small number in the document produce a more excellent IDF value than words with a large number. IDF can be calculated by equation 2.

$$idf_t = log\left(\frac{N}{df_t}\right) \quad (2)$$

Where N is the number of documents and df_t adalah is the number of documents containing the term t[12]. After the TF and IDF values are obtained, they will be multiplied to become the TF-IDF values , as in equation 3.

$$tfidf_{dt} = tf_{d,t}\ x\ idf_t \quad (3)$$

2.5. Support Vector Machine Classification

Support Vector Machine (SVM) is one of the supervised learning classification algorithms. SVM is a reasonable classification algorithm because it can process linear and non-linear data[13]. The basic concept of the SVM algorithm is to find the optimal hyperplane. The hyperplane is a function that works as a separator between data[14]. An existing hyperplane can separate data into two or more classes. The optimal hyperplane search is formulated in equation 4.

$$L_d = \sum_{i=1}^{N} a_i - \frac{1}{2}\sum_{i,j} a_i a_j y_i y_j x_i x_j \quad (4)$$

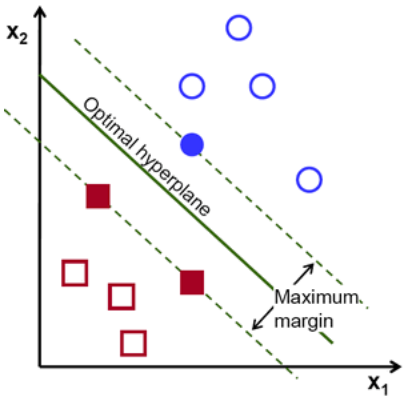

Figure 2. SVM Hyperplane Illustration

Figure 2 illustrates the optimal hyperplane in SVM, namely a straight line that divides the dataset into two different types of classes. However, the hyperplane is not always a linear line that separates two or more classes. Therefore, SVM has kernel tricks. With kernel tricks, data can be projected to a higher dimension so that data with different classes can be separated. An illustration of kernel tricks can be seen in figure 3.

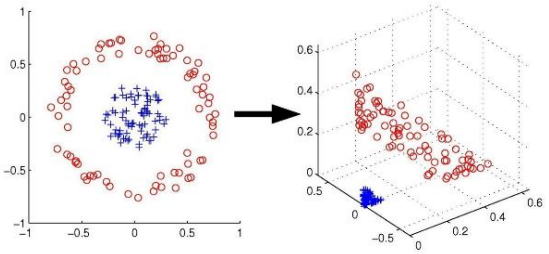

Figure 3. Kernel Tricks Illustration

With the existing kernel tricks, SVM has several kernel functions: Linear, Polynomial, Radial Basis Function (RBF), and Sigmoid. In this study, the kernel used is Linear, Polynomial, and RBF.

2.6. Particle Swarm Optimization

Particle Swarm Optimization (PSO) is a population-based algorithm with several n particles, and the particles move and change their position from time to time. PSO is an algorithm inspired by the social behaviour of flocks of birds and fish that move simultaneously, but no collisions occur[15].

The PSO process starts with a population consisting of several randomly generated particles. Then iteratively, each particle will update its position and velocity to arrive at a new, better solution[15]. Particle Swarm Optimization (PSO) stops when the specified number of iterations is met. Calculating the displacement of the position and velocity of the existing particles can be formulated in formulas 5 and 6[16].

$$v_i^t = w.v_i^t + c_1.r_1(Pbest_i^t - x_i^t) + c_2.r_2(Gbest_i^t - x_i^t) \quad (5)$$

$$x_i^t = x_i^t + v_i^t \quad (6)$$

Where,
v_i^t = velocity of particle i at iteration t
x_i^t = position of particle i in iteration t
w = inertial weight (velocity change impact controller)
$Pbest_i^t$ = the best position of particle i
$Gbest_i^t$ = global best position
c_1 dan c_2 = acceleration coefficient (particle displacement controller)
r_1 dan r_2 = random number 0 - 1

With the existing parameters, PSO can produce different accuracies in classification based on parameter changes.

2.7. Genetic Algorithm

Genetic Algorithm (GA) is an algorithm inspired by or has the same principles as the process of natural selection. The process of natural selection by which the strongest individuals survive has the same concept as GA[13]. The GA process begins by

initializing a specified number of populations, and then the individual selection is carried out based on the fitness value obtained. Thus, the individual's chance to survive depends on the resulting fitness value. After the individual selection process, the surviving individuals undergo a crossover process. At the crossover stage, individuals called parents will exchange genes for producing new solutions called children [17]. After crossover, the following process is mutation. Mutations occur because crossovers cannot produce all good solutions[17]. Therefore, mutations are needed in order to get a good solution. Mutations are carried out by changing genes randomly[17]. After the mutation process, a new population will be generated and go back through the individual selection process. GA will continue until certain conditions are met, or the specified number of generations is met.

2.8. K-Fold Cross Validation

Validation in this study was carried out using the K-Fold Cross Validation. K-Fold Cross Validation validates data by dividing data into several sets based on the specified k number[18]. In this study, the number of k used is 10. So the dataset will be divided into ten parts or folds where the other nine folds will be the train data and one fold will be the test data. Validation will be carried out from the first fold, the test data, to the tenth fold, the test data.

2.9. Confusion Matrix

Confusion matrix is one of the calculation methods used in supervised machine learning[19]. This matrix is used to evaluate the performance of the resulting classification model [20]. This study uses the confusion matrix for more than two classes. The confusion matrix in this study uses three classes. Table 8 shows the confusion matrix used in this study.

Table 8. Confusion Matrix Table

Actual	Prediction		
	-1 (Negative)	0 (Neutral)	1 (Positive)
-1 (Negative)			
0 (Neutral)			
1 (Positive)			

3. RESULTS AND DISCUSSIONS

The dataset used in this text classification is 5000 data with 1012 features. In this study, researchers conducted several scenarios using the SVM method, SVM and PSO feature selection, and SVM and GA feature selection.

3.1. Text classification with Support Vector Machine

The scenario with SVM uses three different kernels, validated with K-Fold Cross Validation. The kernels used in this trial are linear, polynomial, and RBF kernels. The trial results are shown in Tables 9, 10 and 11.

Table 9. K-Fold SVM Linear Kernel

Fold	Accuracy	Precision	Recall	F1-Score
0	63%	64.8%	63.2%	62.8%
1	72%	72.7%	72.2%	72%
2	69.2%	68.7%	68%	68.2%
3	66.8%	65.1%	64.3%	63.7%
4	70.8%	70%	69.4%	69.5%
5	68.2%	71.1%	68.4%	68%
6	74.6%	76.6%	76%	74.9%
7	71.2%	71.5%	71%	71.1%
8	64.8%	65.1%	61.5%	61.1%
9	79%	80.9%	77.4%	77.5

The results of the SVM classification with a linear kernel are in table 9 with 10-fold. The highest score is in the 10th fold, with 79% accuracy, 80.9% precision, 77.4% recall, and 77.5% f1-score.

Table 10. K-Fold SVM Polynomial Kernel

Fold	Accuracy	Precision	Recall	F1-Score
0	62.6%	64.4%	60.8%	60%
1	72.4%	73.7%	71.5%	72%
2	63.6%	64.6%	60.9%	60.5%
3	63.2%	61.7%	59.7%	57.4%
4	68.6%	71.4%	66.3%	65.9%
5	66.6%	70.8%	65.4%	63.9%
6	76%	77.5%	75.6%	75.4%
7	70.6%	72.7%	68.8%	69.5%
8	59.4%	66%	55.1%	53.6%
9	76.8%	82.4%	74.5%	74.4%

The results of the SVM classification with polynomial kernels are in table 10 with 10-fold. The highest score is in the 10th fold, with 76.8% accuracy, 82.4% precision, 74.45% recall, and 74.54% f1-score.

Table 11. K-Fold SVM RBF Kernel

Fold	Accuracy	Precision	Recall	F1-Score
0	65.4%	66.9%	65.4%	65%
1	74.2%	74.8%	74.3%	74.1%
2	69.4%	69%	68.2%	68.4%
3	68.4%	67.2%	65.7%	65%
4	71.2%	70.6%	69.8%	69.9%
5	67.8%	71.1%	67.9%	67.2%
6	75.2%	77.1%	76.4%	75.5%
7	74.2%	74.7%	73.7%	74%
8	66.2%	67.4%	62.9%	62.6%
9	80%	82.1%	78.3%	78.2%

The results of the SVM classification with RBF kernels are in table 11 with 10-fold. The highest score is in the 10th fold, with 80% accuracy, 82.1% precision, 78.3% recall, and 78.2% f1-score.

Table 12. Kernel Average Score Comparison

Kernel	Score Average			
	Accuracy	Precision	Recall	F1-Score
Linear	70%	70.6%	69.1%	68.1%
Polynomial	68%	70.5%	65.9%	65.3%
RBF	71.2%	72.1%	70.2%	70%

The results of the three SVM classification trials in table 12 show that the RBF kernel produced the highest average values for accuracy, precision, recall,

and f1 scores, namely 71.2% and 72.1%, 70.2%, and 70%.

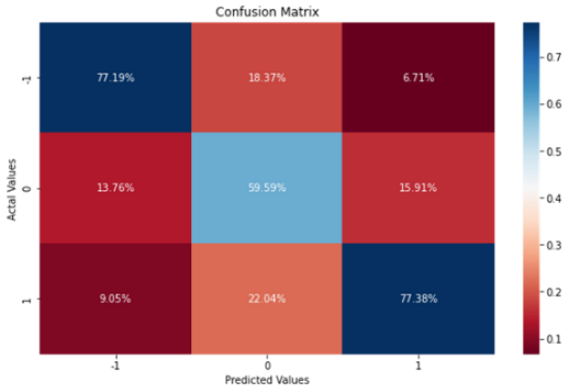

Figure 4. Confusion Matrix SVM with RBF Kernel

Figure 4 shows the confusion matrix for SVM classification with RBF kernel. From this figure, the highest percentage of correct predictions is a positive sentiment at 77.38%, negative sentiment at 77.19%, and neutral sentiment at 59.59%.

3.2. Feature Selection Particle Swarm Optimization with Support Vector Machine Classification

RBF has the highest accuracy value in trials with the three SVM classification kernels. So, in the trial with PSO feature selection, the kernel used is the RBF kernel. Testing at this stage was carried out by changing the number of particles, iterations, weight, c1 and c2, and classification and validation was carried out with k-fold cross-validation. The results of the PSO and SVM trials are shown in Tables 13, 14, 15 and 16.

Table 13. Number of Particle Changes

Number of Particle	PSO Iteration	w	c1	c2	Average Accuracy
20	5	0.1	1	1	67.62%
30	5	0.1	1	1	67.66%
40	5	0.1	1	1	67.7%
50	5	0.1	1	1	66.6%

In table 13, the test results for changing the number of particles with the greatest average accuracy are in the number of particles of 40, namely 67.7%. So the subsequent trial used a number of particles, 40.

Table 14. Weight Changes

Number of Particle	PSO Iteration	w	c1	c2	Average Accuracy
40	5	0.1	1	1	67.7%
40	5	0.3	1	1	67.08%
40	5	0.6	1	1	68.48%
40	5	0.9	1	1	67.4%

Table 14 shows the trial results with a weight change, with the greatest average accuracy at a weight of 0.6, which is 68.48%. So for the subsequent trial, a weight of 0.6 was used.

Table 15. c1 and c2 Changes

Number of Particle	PSO Iteration	w	c1	c2	Average Accuracy
40	5	0.6	1	1	68.48%
40	5	0.6	1	2	66.84%
40	5	0.6	2	1	68.84%
40	5	0.6	2	2	67.68%

Table 15 shows the test results with changes in c1 and c2 with the greatest average accuracy at c1 = 2 and c2 = 1, namely 68.84%. So for the subsequent trial, we used c1 = 2 and c2 = 1.

Table 16. Iteration Changes

Number of Particle	PSO Iteration	w	c1	c2	Average Accuracy
40	5	0.6	2	1	68.84%
40	10	0.6	2	1	67.02%
40	15	0.6	2	1	68.46%
40	20	0.6	2	1	67.68%

Table 16 shows the trial results with a change in the number of iterations with the greatest average accuracy in the number of iterations of 5, namely 68.84%.

Of all the test results on PSO feature selection with SVM classification, the highest accuracy was obtained, namely 68.84% with the number of particles 40, the number of iterations 5, the weight value is 0.6, the c1 value is 2, and the c2 value is 1.

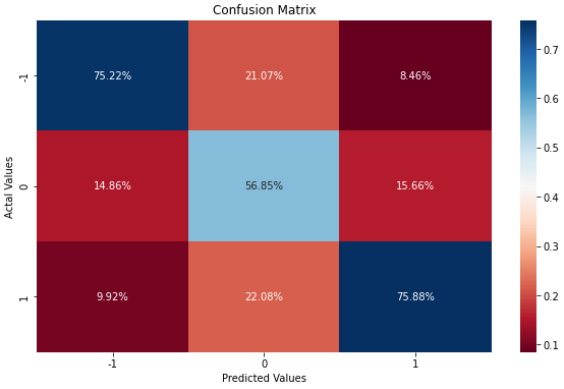

Figure 5. Confusion Matrix PSO + SVM

In figure 5, the confusion matrix for the SVM kernel RBF classification is displayed with the selection of PSO features with the highest accuracy results in several trials. From this figure, the highest percentage of correct predictions is a positive sentiment at 75.88%, negative sentiment at 75.22%, and neutral sentiment at 56.85%.

3.2. Feature Selection Genetic Algorithm with Support Vector Machine Classification

RBF has the highest accuracy value in trials with the three SVM classification kernels. So, in testing with GA feature selection, the kernel used is the RBF kernel. The trials at this stage were carried out by changing the population size and generation, and then classification and validation were carried out with k-

fold cross-validation. The results of the GA and SVM trials are shown in Tables 17 and 18.

Table 17. Population Size Changes

Population Size	Generation	Average Accuracy
30	5	67.9%
40	5	67.58%
50	5	66.78%
60	5	67.72%

Table 17 shows the trial results in population size change with the greatest average accuracy at 67.9%, a population size of 30. So for the generation change trial, a population size of 30 is used.

Table 18. Generation Changes

Population Size	Generation	Average Accuracy
30	5	67.9%
30	10	67.94%
30	15	68.86%
30	20	69.52%

In table 18, the trial results change the number of generations with the highest average accuracy in the number of generations 20, namely 69.52%.

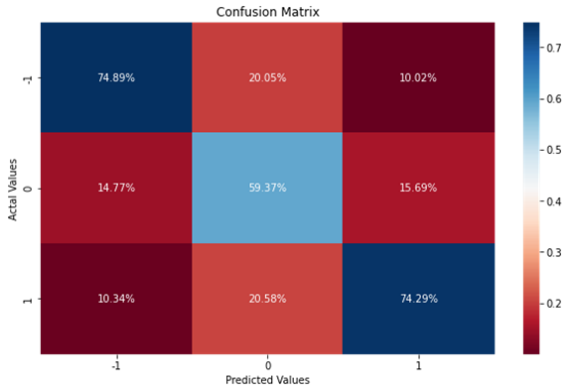

Figure 6. Confusion Matrix GA + SVM

Figure 6 shows the confusion matrix for SVM kernel RBF classification with GA feature selection that has the highest accuracy results in several trials. From this figure, the highest percentage of correct predictions is in negative sentiment at 74.89%, followed by positive sentiment at 74.29% and neutral sentiment at 59.37%.

Based on the results of research that has been conducted, the addition of the PSO and GA feature selection process does not increase accuracy but reduces the accuracy of SVM classification. These results did not prove that by adding a feature selection process, accuracy in a classification can increase in value. As in several previous studies that adding a feature selection process increases or does not change the accuracy value.

4. CONCLUSION

Based on the results of the research that has been done, the accuracy of the SVM with the RBF kernel has the highest accuracy value with an accuracy of 71.2%, followed by using the GA feature selection method and the RBF kernel SVM classification with an accuracy of 69.52%. Furthermore, the last is the use of PSO feature selection and SVM classification, which has an accuracy value of 68.84%. From this study, the use of the feature selection method in this study did not increase accuracy but reduced the value of accuracy. Therefore, using the selection method in this study did not affect increasing the accuracy of the SVM classification. This accuracy does not increase because the important values in the sentence are removed during the feature selection process.

Suggestions that researchers can give for further research are to use different datasets for text classification, and the datasets used are in large numbers so that it can be seen whether the PSO and GA feature selection methods combined with SVM classification provide good accuracy values or not.

REFERENCES

[1] "Konsumsi Pertalite Capai 23 Juta KL, Paling Banyak Digunakan Masyarakat," *Kementerian Energi dan Sumber Daya Mineral*, 2022. https://migas.esdm.go.id/post/read/konsumsi-pertalite-capai-23-juta-kl-paling-banyak-digunakan-masyarakat

[2] M. A. Rizaty, "Pengguna Twitter di Indonesia Capai 18,45 Juta pada 2022," *Dataindonesia.id*, 2022. https://dataindonesia.id/digital/detail/pengguna-twitter-di-indonesia-capai-1845-juta-pada-2022

[3] P. Arsi and R. Waluyo, "Analisis Sentimen Wacana Pemindahan Ibu Kota Indonesia Menggunakan Algoritma Support Vector Machine (SVM)," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 8, no. 1, p. 147, 2021, doi: 10.25126/jtiik.0813944.

[4] M. Taradeh *et al.*, "An evolutionary gravitational search-based feature selection," *Inf. Sci. (Ny).*, vol. 497, pp. 219–239, Sep. 2019, doi: 10.1016/J.INS.2019.05.038.

[5] D. A. Kristiyanti and M. Wahyudi, "Feature selection based on Genetic algorithm, particle swarm optimization and principal component analysis for opinion mining cosmetic product review," *2017 5th Int. Conf. Cyber IT Serv. Manag. CITSM 2017*, 2017, doi: 10.1109/CITSM.2017.8089278.

[6] B. S. Prakoso, D. Rosiyadi, H. S. Utama, and D. Aridarma, "Klasifikasi Berita Menggunakan Algoritma Naive Bayes Classifer Dengan Seleksi Fitur Dan Boosting," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 3, no. 2, pp. 227–232, 2019, doi: 10.29207/resti.v3i2.1042.

[7] E. Purnamasari, D. P. Rini, and Sukemi, "Seleksi Fitur menggunakan Algoritma

Particle Swarm Optimization pada Klasifikasi Kelulusan Mahasiswa dengan Metode Naive Bayes," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 1, no. 3, pp. 469–475, 2020.

[8] I. A. Ogunbiyi, "Web Scraping with Python – How to Scrape Data from Twitter using Tweepy and Snscrape," *freeCodeCamp*, 2022. https://www.freecodecamp.org/news/python-web-scraping-tutorial/ (accessed Jan. 02, 2023).

[9] A. Muzaki and A. Witanti, "Sentiment Analysis of the Community in the Twitter To the 2020 Election in Pandemic Covid-19 By Method Naive Bayes Classifier," *J. Tek. Inform.*, vol. 2, no. 2, pp. 101–107, 2021, doi: 10.20884/1.jutif.2021.2.2.51.

[10] I. T. Julianto, D. Kurniadi, M. R. Nashrulloh, and A. Mulyani, "COMPARISON OF CLASSIFICATION ALGORITHM AND FEATURE SELECTION IN BITCOIN SENTIMENT ANALYSIS," *J. Tek. Inform.*, vol. 3, no. 3, pp. 739–744, 2022.

[11] A. I. Kadhim, "Term Weighting for Feature Extraction on Twitter: A Comparison between BM25 and TF-IDF," *2019 Int. Conf. Adv. Sci. Eng. ICOASE 2019*, pp. 124–128, 2019, doi: 10.1109/ICOASE.2019.8723825.

[12] M. R. A. Utomo and Y. Sibaroni, "Text classification of british english and American english using support vector machine," *2019 7th Int. Conf. Inf. Commun. Technol. ICoICT 2019*, no. July 2019, 2019, doi: 10.1109/ICoICT.2019.8835256.

[13] P. D. W. Mega and Haryoko, "Optimization of parameter support vector machine (SVM) using genetic algorithm to review go-jek's services," *2019 4th Int. Conf. Inf. Technol. Inf. Syst. Electr. Eng. ICITISEE 2019*, vol. 6, pp. 301–304, 2019, doi: 10.1109/ICITISEE48480.2019.9003894.

[14] O. H. Rahman, G. Abdillah, and A. Komarudin, "Klasifikasi Ujaran Kebencian pada Media Sosial Twitter Menggunakan Support Vector Machine," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 1, pp. 17–23, 2021, doi: 10.29207/resti.v5i1.2700.

[15] S. A. Saputra, D. Rosiyadi, W. Gata, and S. M. Husain, "Google Play E-Wallet Sentiment Analysis Using Naive Bayes Algorithm Based on Particle Swarm Optimization," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 3, no. 3, pp. 377–382, 2019.

[16] D. M. BrTarigan, D. P. Rini, and Samsuryadi, "Seleksi Fitur pada Klasifikasi Penyakit Gula Darah Menggunakan Particle Swarm Optimization (PSO) pada Algoritma C4.5," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 1, no. 3, pp. 569–575, 2020.

[17] S. Mirjalili, J. Song Dong, A. S. Sadiq, and H. Faris, *Genetic algorithm: Theory, literature review, and application in image reconstruction*, vol. 811. Springer International Publishing, 2020. doi: 10.1007/978-3-030-12127-3_5.

[18] C. A. Ramezan, T. A. Warner, and A. E. Maxwell, "Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification," *Remote Sens.*, vol. 11, no. 2, 2019, doi: 10.3390/rs11020185.

[19] J. Xu, Y. Zhang, and D. Miao, "Three-way confusion matrix for classification: A measure driven view," *Inf. Sci. (Ny).*, vol. 507, no. xxxx, pp. 772–794, 2020, doi: 10.1016/j.ins.2019.06.064.

[20] E. U. A. C. Khotimah, "Comparison Naïve Bayes Classifier, K-Nearest Neighbor And Support Vector Machine In The Classification Of Individual On Twitter Account," *J. Tek. Inform.*, vol. 3, pp. 673–680, 2022.