SVM-PSO Algorithm for Tweet Sentiment Analysis #BesokSenin

A D N W Susanto*[1], H Suparwito[2]

[1,2]Department of Informatics, Universitas Sanata Dharma, Yogyakarta, Indonesia

Email: anggitadnws@gmail.com[1], shirsj@jesuits.net[2]

Submitted: 12 April 2023, revised: 2 July 2023, accepted: 8 August 2023

Abstract. The hashtag #BesokSenin is a hashtag that is often trending on Indonesian Twitter on Sunday evenings. Many Indonesian Twitter users expressed their feelings about welcoming Monday using the hashtag #BesokSenin. The tweet containing #BesokSenin is known to be a motivational sentence to welcome Monday full of joy or a disappointed sentence because you have to return to your routine after taking a holiday on Saturday and Sunday. This study conducts sentiment analysis to find out the opinions of netizens on welcoming Mondays. The tweet data used is tweet data with the hashtag #BesokSenin and the keywords school, work, assignments, and college. The classification method used is the Support Vector Machine algorithm, which is optimized using the Particle Swarm Optimization method to optimize the performance of the Support Vector Machine algorithm. Results of 80% accuracy were obtained by applying the Support Vector Machine model based on Particle Swarm Optimization. This accuracy is superior to 1% compared to the results of accuracy using the usual Support Vector Machine model, which equals 79%. This shows that Particle Swarm Optimization can optimize the accuracy of the Support Vector Machine algorithm.

Keywords: #BesokSenin, Particle Swarm Optimization, Sentiment Analysis, Support Vector Machine, Twitter

1. Introduction

Monday is the second day of the week. However, Monday can also be said to be the first day of the week, according to ISO 8601 [1]. Monday is the first day of work and school, often discussed on social media, especially Twitter. Due to its ease of use and accessibility, Twitter has become one of the most popular social media platforms. Until now, it has been estimated that Twitter already has 330 million active users every month and continues to grow every day [2]. The term hashtag (tags and fences) is known in social media networks, symbolized by the fence symbol. The function of this hashtag is to find information on a specific topic.

#BesokSenin is a hashtag that is often discussed on Indonesian Twitter. Uniquely, the #BesokSenin hashtag is a trending topic only on Sunday evenings. Many Indonesian Twitter users have expressed their feelings about welcoming Monday through tweets with the hashtag #BesokSenin [3]. A study says that people who have to go to work after spending two days off at the weekend tend to be more sensitive to stress at the beginning of the week [4]. This drew attention to the sentiment analysis of user tweets based on the #BesokSenin hashtag using a machine learning approach. Sentiment analysis is a process of extracting and understanding information centered on data analysis. It aims to understand the emotion of a text to predict and analyze the public atmosphere, mood, and description of someone's feelings in a case [2, 5].

This study chose the Support Vector Machine (SVM) as the classification algorithm for this research. It used the optimization method to select SVM parameters, namely the Particle Swarm Optimization (PSO) method, hoping to optimize the performance produced by SVM in this sentiment analysis process. Sabrila et al [2] researched tweet sentiment analysis using SVM-PSO to learn more about the opinions of the Twitter user community regarding the Job Creation Law. This study only consisted of positive labels for tweets containing positive thoughts and negative labels for tweets containing negative opinions, with 1000 tweets. The results of this study obtained an accuracy of 92.99%, a precision of 93.24%, and a recall of 93% using the SVM algorithm. While the accuracy generated using the SVM algorithm with PSO optimization produces an accuracy of 95%, a precision of 95.08%, and a recall of 94.97%, These results indicate that the PSO optimization method can overcome the weaknesses of the SVM algorithm in sentiment analysis [2].

Additionally, Que et al [5] researched to ascertain the public opinion of online transportation companies using the SVM classification method and PSO optimization SVM. The researchers analyzed tweets containing positive and negative sentiments with 1,852 tweets from January 1 to October 15, 2019, divided into testing data of 1,130 tweets and training data of 722 tweets. The study's results show that PSO can make up for the weaknesses of the SVM algorithm when choosing the right parameters for sentiment classification. The accuracy went up from 95.46% to 96.04%, which is better than the accuracy of the SVM algorithm without PSO optimization, which was 95.46% [5].

2. Research Methods

The investigation begins with the collection of Twitter data. The next stage is to label the tweet data. In the third stage, each tweet is pre-processed. Next, the data are separated into training data and assessment data. After dividing the data into two subsets, the next stage is feature extraction. The next stage is to conduct the initial modeling using the Support Vector Machine (SVM) algorithm, followed by a performance evaluation. The second modeling was performed using the Support Vector Machine (SVM) algorithm, which was integrated with the Particle Swarm Optimization (PSO) technique and then used to measure performance. The final stage of this research is to analyze the findings. In general, the research process can be seen in the following diagram:

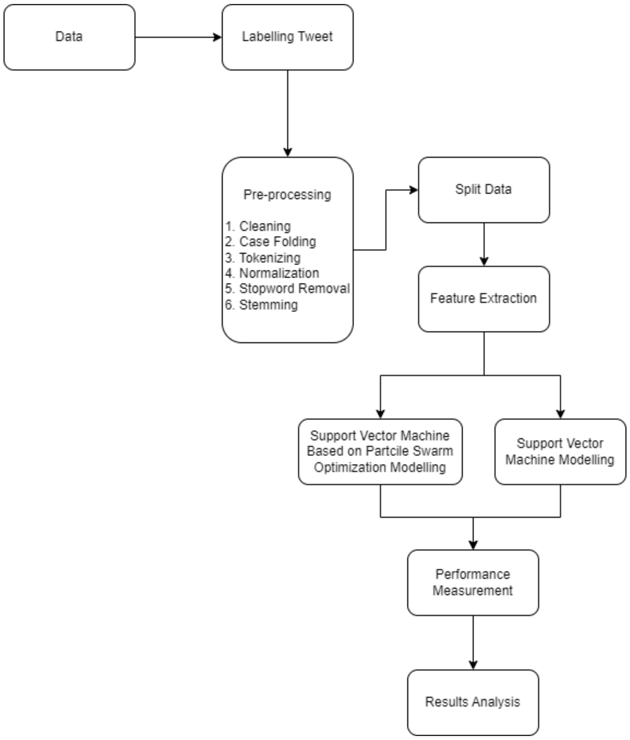


Figure 1. Research Flowchart

2.1. Data Collection

This study used tweet data with #BesokSenin taken from social media Twitter in the period 1 January 2020 to 31 July 2022 with the selected keywords, namely "#BesokSenin kerja", "#BesokSenin tugas", "#BesokSenin kuliah" and "#BesokSenin sekolah". Tweet data was collected using the snscrape library in the Python programming language. The total number of tweets that were obtained was 3,108.

2.2. Labelling Tweet

The tweet labeling process used the Vader module contained in the NLTK (Natural Language Toolkit) library in the Python programming language. The Vader module functions to divide tweets based on their sentiments because the amount of tweet data is so large and always growing. This module was chosen because it can work well when analyzing social media texts sentiment [6].

The Vader module only works in English, so the first step was to translate tweets from Bahasa Indonesia into English and enter them into the Vader module. Because the data's character count exceeded the limits of Python's translator library, Google Document Translator was necessary to assist in the translation process. Before translating tweets into English, a tokenizing process was first carried out to separate sentences into words or terms. Then normalize it so that the words contained in the tweet become more standard.

After the translation process is translated into English, the tweet data will be searched for each polarity value. This polarity value is obtained based on calculating each word weight and emoticon contained in Vader's dictionary. The tweet has a positive sentiment value if the polarity value exceeds 0. The tweet has a negative sentiment value if the polarity value is less than 0. Labeled tweets with positive sentiment are given the symbol 1, and labeled tweets with negative sentiment are given the symbol -1.

2.3. Pre-processing

The data used is tweet data obtained from the social media site Twitter, so it has the form of unstructured data. Therefore, the data needs to be processed to remove non-standard words and noise and make the data more structured through the following pre-processing stages:

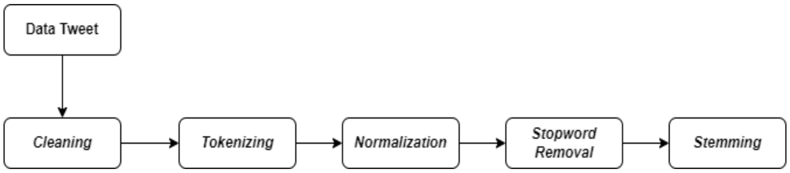


Figure 2. Pre-Processing steps

2.4. Split Data

After the data has gone through the pre-processing stage, it is divided into two subsets: training data and testing data. Data distribution is done in a ratio of 80:20, where 80% is the training data and 20% is the testing data.

2.5. Feature Extraction

Term Frequency-Inverse Document Frequency, or more commonly known as the abbreviation TF-IDF is a method of weighting the relationship between a word or term against a document. The number of occurrences of a term in a document is called the Term Frequency (TF) [7]. Meanwhile, reducing the dominance of a term often appearing in a document containing a word is called the Inverse Document Frequency (IDF). The TF-IDF method works by combining two concepts: the frequency of a word’s occurrence in a document and the inverse of the frequency that contains the word [7].

2.6. Modeling

At this modelling stage used data from feature extraction results. Data validation is carried out on the training set data in each model using the K-Fold Cross Validation method. This stage will be carried out as many as the value of K entered in the K-Fold Cross Validation method. K-Fold Cross Validation is a cross validation method used to evaluate a model by dividing data into two subsets: training data and testing data. This method aims to reduce computation time while maintaining the accuracy obtained [6].

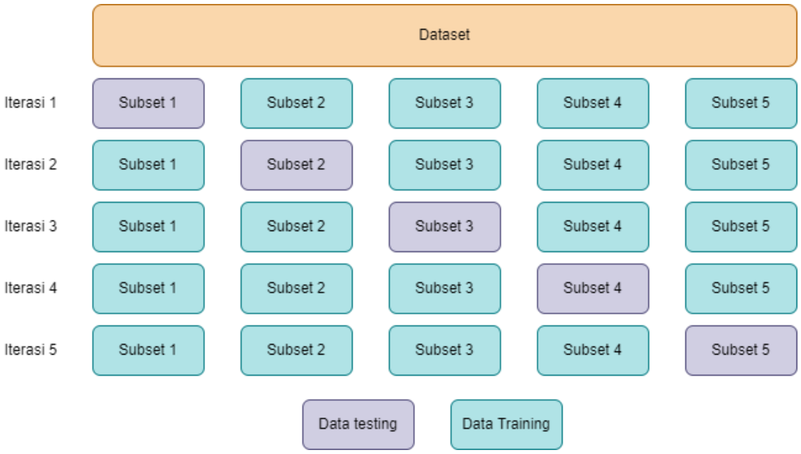


Figure 3. 5-Fold Cross Validation Model Flow

2.7. Support Vector Machine

Support Vector Machine (SVM) is a machine learning method that works based on the Structural Risk Minimization (SRM) principle to find the best hyperplane that separates two classes in the input space [8]. SVM is very effective in solving problems with text data because text data tends to have high dimensions, have several irrelevant features, and generally correlate with each other, which will be arranged in separate categories linearly [9]. The advantage of SVM is being able to classify a pattern accurately, even with limited datasets. However, SVM also has limitations with the number of attributes used; if the number of attributes used is relatively large, it will result in a heavy computational load, making the accuracy value less accurate [10].

2.8. Support Vector Machine Based on Particle Swarm Optimization

Particle Swarm Optimization (PSO) is an optimization technique inspired by the behavior of flocks of birds or fish. PSO has characteristics that are simple in concept, easy to implement, and have efficient computing capabilities [11]. Particle Swarm Optimization works by finding the best position and solution in the search space, called the personal best (Pbest) and global best (Gbest), which will be achieved by the population with the particle index.

The partial swarm optimization method will be used to select the parameters of the Support Vector Machine algorithm. The Support Vector Machine parameters to be optimized are the C and gamma values. The following shows the workflow for implementing SVM optimization using PSO in sentiment analysis.

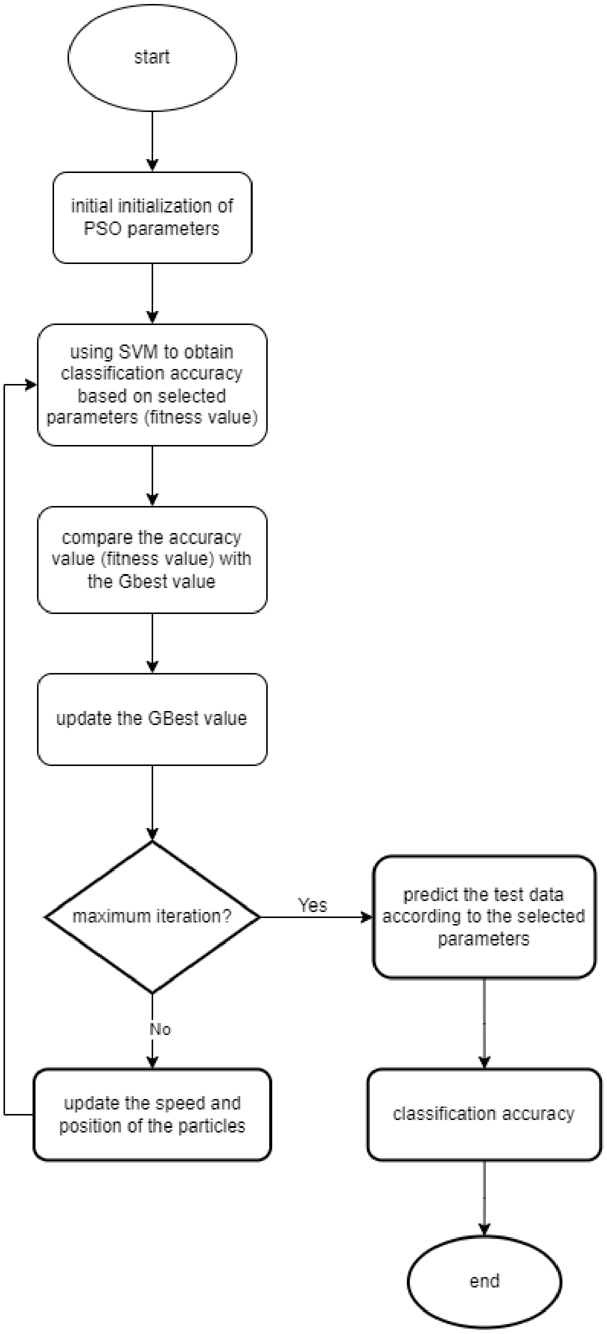
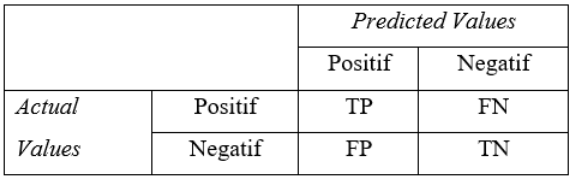

start
initial initialization of PSO parameters
using SVM to obtain classification accuracy based on selected parameters (fitness value)
compare the accuracy value (fitness value) with the Gbest value
update the GBest value
maximum iteration?
Yes
No
predict the test data according to the selected parameters
update the speed and position of the particles
classification accuracy
end

Figure 4. SVM-PSO Flow

2.9. Performance Measurement

The Confusion Matrix method was chosen as a tool to measure the performance of each model. The classification results obtained from each model will form a confusion matrix to calculate accuracy, recall, and precision values.



		Predicted Values	
		Positif	Negatif
Actual	Positif	TP	FN
Values	Negatif	FP	TN

Figure 5. Confusion Matrix

3. Results

3.1. Crawling Tweet

Tweets were crawled four times according to the keywords followed after the hashtag #BesokSenin, *kerja*, *kuliah*, *tugas* and *sekolah*. The data taken includes the post date of the tweet's account owner, the tweet, the tweet owner's user ID, and the tweet's content. The following table is an example of the Twitter crawling stage.

Table 1. Example of Tweet Crawling Results

Date Time	User Id	Tweet	Username
2022-05-08 10:12:41+00:00	1523244491229999105	semangat kerja buat #BesokSenin	arekelek4676825
2022-05-08 09:40:00+00:00	1523236266447233024	Mager Kerja ... Happy WeekEND #BesokSenin	dewiilarra
2022-02-20 15:20:52+00:00	1495418179589263360	Tetap semangat, besok senin, kerja kerja kerja....tetap sehat sehat sehat #BesokSenin	aniskurniawan
2022-02-20 14:14:50+00:00	1495401563871133698	#BesokSenin belum tidur udah capek aja sadar besok kerja	MedsosTm
2022-03-20 14:05:20+00:00	1505546033220136960	BesokSenin males kerja !	Karetmolorpol

3.2. Labelling Tweet

The result of labeling tweets is a polarity value. For polarity values less than zero, they would be labeled -1 or negative, while for polarity values greater than zero, they would be labeled 1 or positive. If there was a tweet with a polarity value equal to zero, then the sentiment label is 0 or neutral. This study only used negative and positive labels, so tweets that have neutral labels would be deleted.

Table 2. Tweet Labeling Results

Date Time	User Id	Tweet	Username	Sentimen
2022-05-08 10:12:41+00:00	1523244491229999105	semangat kerja buat #BesokSenin	arekelek4676825	1
2022-05-08 09:40:00+00:00	1523236266447233024	Mager Kerja ... Happy WeekEND	dewiilarra	-1

Date Time	User Id	Tweet	Username	Sentimen
		#BesokSenin		
2022-02-20 15:20:52+00:00	1495418179589263360	Tetap semangat, besok senin, kerja kerja kerja....tetap sehat sehat sehat #BesokSenin	aniskurniawan	1
2022-02-20 14:14:50+00:00	1495401563871133698	#BesokSenin belum tidur udah capek aja sadar besok kerja	MedsosTm	-1
2022-03-20 14:05:20+00:00	1505546033220136960	BesokSenin males kerja !	Karetmolorpol	-1

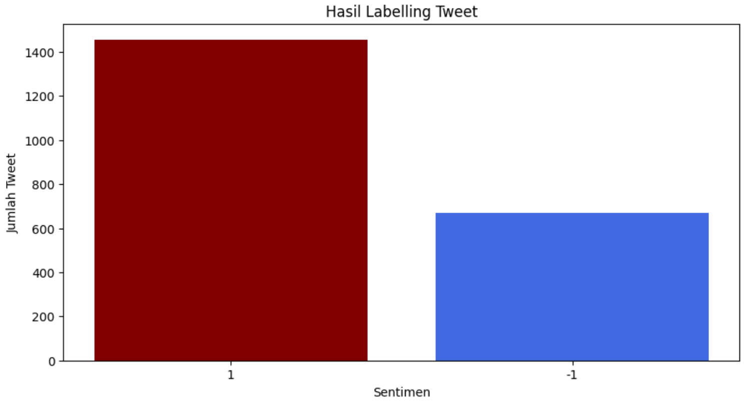


Figure 6. Tweet Sentiment Results Chart

3.3. Pre-processing

The data will be processed at the pre-processing stage to make it more structured. The amount of tweet data that will be pre-processed is 2,126 tweets. The stages that will be carried out at this pre-processing stage are cleaning, tokenizing, normalization, stopword removal, and stemming.

3.3.1 Cleaning

The cleaning stage is the stage that will clean the data from noise. This stage includes the case folding stage, or changing the letter characters contained in the tweet to lowercase. This stage also removes unnecessary characters such as punctuation marks, numbers, hashtags, mentions, and emoticons.

Table 3. Example of cleaning results

Tweet	Hasil Cleaning
semangat kerja buat #BesokSenin	semangat kerja buat
Mager Kerja ... Happy WeekEND #BesokSenin	mager kerja happy weekend
Tetap semangat, besok senin, kerja kerja kerja....tetap sehat sehat sehat #BesokSenin	tetap semangat besok senin kerja kerja kerja tetap sehat sehat sehat

3.3.2 Tokenizing

The next stage is the tokenizing stage, or separating the sentences in the tweet into a word or token.

Table 4. Example of tokenizing results

Hasil Cleaning	Hasil Tokenizing
semangat kerja buat	[‘semangat’, ‘kerja’, ‘buat’]
mager kerja happy weekend	[‘mager’, ‘kerja’, ‘happy’, ‘weekend’]
tetap semangat besok senin kerja kerja kerja tetap sehat sehat sehat	[‘tetap’, ‘semangat’, ‘besok’, ‘senin’, ‘kerja’, ‘kerja’, ‘kerja’, ‘tetap’, ‘sehat’, ‘sehat’, ‘sehat’]

3.3.3 Normalization

At this stage, the spelling of the words contained in the tweet will be corrected. Tweets that contain non-standard words or are in the form of abbreviations will be changed to match the KBBI spelling.

Table 5. Example of normalization results

Hasil Tokenizing	Hasil Normalization
[‘semangat’, ‘kerja’, ‘buat’]	[‘semangat’, ‘kerja’, ‘untuk’]
[‘mager’, ‘kerja’, ‘happy’, ‘weekend’]	[‘malas’, ‘gerak’ ‘kerja’, ‘happy’, ‘weekend’]
[‘tetap’, ‘semangat’, ‘besok’, ‘senin’, ‘kerja’, ‘kerja’, ‘kerja’, ‘tetap’, ‘sehat’, ‘sehat’, ‘sehat’]	[‘tetap’, ‘semangat’, ‘besok’, ‘senin’, ‘kerja’, ‘kerja’, ‘kerja’, ‘tetap’, ‘sehat’, ‘sehat’, ‘sehat’]

3.3.4 Stopword Removal

The stopword removal stage is carried out to remove words that have no meaning so that the remaining words are words that only have meaning.

Table 6. Stopword Removal Results

Hasil Normalization	Hasil Stopword Removal
[‘semangat’, ‘kerja’, ‘untuk’]	[‘semangat’]
[‘malas’, ‘gerak’ ‘kerja’, ‘happy’, ‘weekend’]	[‘malas’, ‘gerak’, ‘happy’, ‘weekend’]
[‘tetap’, ‘semangat’, ‘besok’, ‘senin’, ‘kerja’, ‘kerja’, ‘kerja’, ‘tetap’, ‘sehat’, ‘sehat’, ‘sehat’]	[‘semangat’, ‘sehat’, ‘sehat’, ‘sehat’,]

3.3.5 Stemming

Stemming is the last stage in pre-processing, where this stage is done to convert words that have affixes into basic words.

Table 7. Stemming Results

Hasil Stopword Removal	Hasil Stemming
[‘semangat’]	[‘semangat’]
[‘malas’, ‘gerak’, ‘happy’, ‘weekend’]	[‘malas’, ‘gerak’, ‘happy’, ‘weekend’]
[‘semangat’, ‘sehat’, ‘sehat’, ‘sehat’,]	[‘semangat’, ‘sehat’, ‘sehat’, ‘sehat’,]

3.4. Data Splitting

The data splitting process is carried out by dividing the data 80:20, where 80% is for training data and 20% is for testing data—results in a training set of 1700 tweets and a testing set of 425 tweets. Next, in the training dataset, k-fold cross-validation processes were performed.

3.5. Feature Extractions

The feature extraction stage converts training and testing data into numerical form using the term frequency-inverse document frequency (TF-IDF) method. A matrix will be produced from the feature extraction stage, which will be used to carry out the next stage. The TF-IDF module uses the max-features function, which is filled with the number 700, to get the top 700 terms with the largest term frequency.

3.6. Results Analysis

3.6.1. Model Support Vector Machine

The results of testing using the Support Vector Machine model in the training and testing datasets are carried out using the value of Cost = 1 and Kernel RBF as follows:

Table 8. Test Results for the Support Vector Machine Model

Fold	Gamma	Accuracy (%)	
		Training	Testing
	0,01	70	68
3	0,1	75	75
	1	94	79
	0,01	69	68
5	0,1	74	75
	1	94	79
	0,01	71	68
10	0,1	76	75
	1	95	79

Based on the results of testing the SVM model in the table above, the optimal accuracy results at C = 1, gamma = 1, and fold = 10 produced an accuracy for training of 95% and testing of 79%.

3.6.2. Support Vector Machine Model Based on Particle Swarm Optimization

The results of testing using the PSO-based SVM model on the training and testing datasets were carried out using the Kernel RBF, with the Cost and gamma values obtained based on the best position generated by the PSO.

Table 9. SVM-PSO Model Test Results

Fold	C_1	C_2	Iteration	Particle	C	Gamma	Accuracy (%)	
							Training	Testing
				5	1,6	0,5	91	80
3	3	1	10	10	3,1	0,1	83	79
				20	0,1	0,2	70	68
				5	1,6	0,5	91	90
5	3	1	10	10	3,1	0,1	84	79
				20	0,1	1,6	70	68

				5	1,6	0,5	92	80
10	3	1	10	10	3,1	0,1	85	79
				20	0,1	0,2	71	68

Based on the test results table above, the highest accuracy results are found at fold = 10, for c1 = 3 and c2 = 1, iterations = 5, number of particles = 10, C = 9.5, and gamma = 0.04, resulting in accuracy for training of 92% and testing of 80%. The number of particles, iterations, c1, and c2 values influence C and gamma values in the SVM-PSO model. Based on the analysis above, the number of iterations and particles is very influential in determining the value of C and gamma; if the number of particles increases, the accuracy will decrease. The number of iterations is optimal, with a total of 10 iterations. This is because the iteration peak in this research data is at number 14. If the iteration is higher than 14, the accuracy will decrease.

The best results were obtained in the SVM-PSO model with C = 1.6 and gamma = 0.5 for each fold. Based on these results, it can be assumed that the SVM-PSO model is better than the usual SVM model. To strengthen these results, additional experiments are done by entering the parameter values generated by the SVM-PSO into the regular SVM model as follows:

Table 10. SVM and SVM-PSO accuracy results for parameters C = 1.6, and Gamma = 0.5

Fold	**C**	**Gamma**	**Accuracy (%) Testing**	
			SVM	**SVM-PSO**
3			78	80
5	1,6	0,5	78	80
10			78	80

4. Conclusion

The Particle Swarm Optimization method can increase accuracy by optimizing the Support Vector Machine parameters. The test results obtained the best accuracy of 80% in the application of the Support Vector Machine model based on Particle Swarm Optimization. The accuracy results are 2% superior to those using the usual Support Vector Machine model, which equals 78%. The Particle Swarm Optimization method works based on a number of particles whose speed and position are constantly updated at each iteration. PSO will track the best position and its best solution in the search space.

References

[1] V. Hocken, "Monday: the Moon's Day," timeanddate.com, [Online]. Available: https://www.timeanddate.com/calendar/days/monday.html#:~:text=According%20to%20the%20international%20standard%20ISO%208601%2C%20Monday%20is%20considered,first%20day%20of%20the%20week.. [Accessed 5 September 2022].

[2] ST S, A. Yufis and ACSK, "Analysis of Tweet Sentiment About the Job Creation Law Using the PSO-Based SVM Algorithm," *JISKA*, no. 7(1), pp. 10-19, 2022.

[3] A. Apandi, "Why is the hashtag #BesokSenin Often a Trending Topic on Twitter on Sunday Night?" jabarekspres.com, June 26, 2022. [Online]. Available: https://jabarekspres.com/berita/2022/06/26/mengapa-tagar-besoksenin-sering-jadi-trending-topics-di-twitter-pada-week-malam/. [Accessed September 5, 2022].

[4] P. Shani, Z. Zhiqing E., K. Stacey R, K. Alexandra and PE Spector, "Workdays are not created equal: Job satisfaction and job stressors across the workweek," *SAGE*, vol. 74, no. 9, pp. 1447-

1472, 2020.

[5] VKS Que, A. Iriani and HD Purnomo, "Online Transportation Sentiment Analysis Using Particle Swarm Optimization-Based Support Vector Machine," *National Journal of Electrical Engineering and Information Technology,* vol. 9, no. 2, pp. 162-170, 2020.

[6] R. Wati, S. Enarwati and I. Maryani, "Optimization of SVM-Based PSO Parameters for Sentiment Analysis of English-Speaking Airline Service Reviews," *Evolution: Journal of Science and Management,* vol. 8, no. 2, pp. 64-718, 2020.

[7] N. Musyaffa and B. Rifai, "Support Vector Machine Model Based on Particle Swarm Optimization for Liver Disease Prediction," *Journal of Computer Science and Technology,* vol. 3, no. 2, pp. 189-914, 2018.

[8] BG Osaldi, "Analysis of Sentiment of Online Learning on Social Media Twitter Using Multinomial Naive Bayes and Support Vector Machine," *Thesis, Sanata Dharma University,* 2021.

[9] R. Yulianto, "Sentiment analysis of Lazada E-Commerce reviews on Twitter social media using the Support Vector Machine Algorithm," *Thesis thesis, Sanata Dharma University,* 2022.

[10] NK Wardhani, K. Rezkiani, H. Setiawan, G. Gata, S. Tohari and M. Wahyudi, "Sentiment analysis article news coordinator minister of maritime affairs using algorithm naive bayes and support vector machine with particle swarm optimization," Journal of *Theoretical and Applied Information Technology,* vol. 86, no. 24, pp. 8365-8378, 2018.

[11] YN, "Artist News Sentiment Analysis Using Support Vector Machine and Particle Swarm Optimization Algorithms," *Journal of Information Systems,* vol. 5, no. 2, pp. 104-112, 2017.

[12] TB Sasongko, "Comparison and Performance Analysis of SVM and PSO-SVM Algorithm Models (Case Study of High School Interest Path Classification)," *Journal of Informatics Engineering and Information Systems,* vol. 2, no. 2, pp. 244-253, 2016.

[13] T. Arifin and A. Herliana, "Optimization of the Classification Method Using Particle Swarm Optimization for Identification of Diabetic Retinopathy," *Journal of Computer Science and Informatics,* vol. 4, no. 2, pp. 77-81, 2018.

[14] Samad, A., Basari, H., Hussin, B., Pramudya, I., & Zeniarja, J. (2013). Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization. *Procedia Engineering, 53*, 453-462.

[15] Noor, A. (2018). Comparison Of Ordinary Support Vector Machine Algorithm And Support Vector Machine Based On Particle Swarm Optimization For Earthquake Prediction. *Journal Of Humanities And Technology*, 4(1), 31-37.

[16] Musyaffa, N., & Rifai, B. (2018). Support Vector Machine Model Based On Particle Swarm Optimization For Liver Disease Prediction. Journal Of Computer Science And Technology, 3(2), 189-914.

[17] Gavilnes, M., Lopez, T., Martinez, J., Montenegro, E., & Castano, F. (2016). Unsupervised Method for Sentiment Analysis in Online Texts. *Expert Systems with Applications*, 58, 57-75.

[18] Azies, H., Trishnanti, D., & P.H, E. (2019). Comparison of Kernel Support Vector Machine (SVM) in Classification of Human Development Index (HDI). *IPTEK Journal of Proceedings Series*, 6, 53-57.