JINAV: Journal of Information and Visualization *ISSN: 2746-1440 (online)*
Vol. 4 No. 1 (2023)
https://doi.org/10.35877/454RI.jinav1504

Sentiment Analysis on Twitter Social Media towards Shopee E-Commerce through *Support Vector Machine* (SVM) Method

Putri Samapa Hutapea & Warih Maharani

Undergraduate Study Program, Faculty of Informatics, Telkom University, Bandung, Indonesia

Abstract

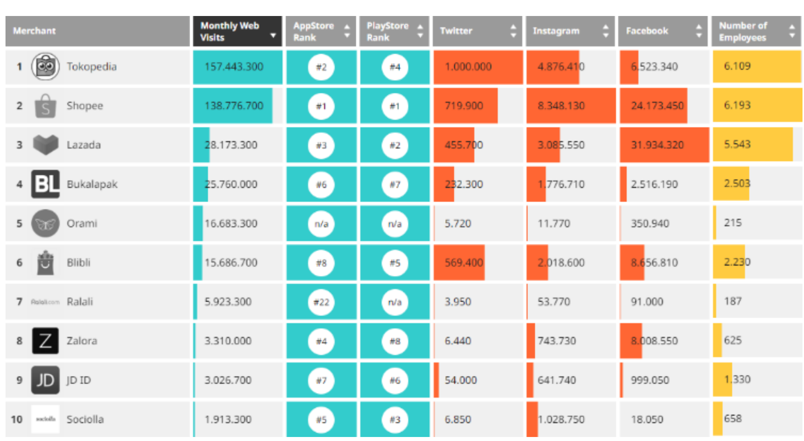
Shopee is e-commerce widely accessed and used in this era. Many people use Shopee because the products offered are cheaper and more affordable. Despite the fact that Shopee is a well-known e-commerce, it still requires responses and suggestions from the public to maintain or improve the features required. In this study, public sentiment analysis was carried out on Twitter social media related to the Shopee marketplace. This study collected data that contained tweets from predetermined keywords and used Word2Vec and Support Vector Machine classification methods. The use of Word2Vec influenced the level of accuracy so that it increased for each SVM kernel. Meanwhile, the best hyperparameter tuning was found in the polynomial kernel, with an accuracy rate of 93.20%.

Keywords: sentiment, Shopee, Twitter, Word2Vec, SVM

1. Introduction

1.1. Background of the Study

The development of technology in this modern era certainly has a great influence on daily life, especially in the business field and the world of competition for the market (Siagian Ade Onny et al., 200 C.E.). Nowadays, promotion and marketing can be done through advertisements on television and banners installed on the streets and digital promotions. Consumers can only view, choose, and buy the needs they want through smartphones, laptops, and computers without traveling in person.



Merchant		Monthly Web Visits	AppStore Rank	PlayStore Rank	Twitter	Instagram	Facebook	Number of Employees
1	Tokopedia	157.443.300	#2	#4	1.000.000	4.876.410	6.523.340	6.109
2	Shopee	138.776.700	#1	#1	719.900	8.348.130	24.173.450	6.193
3	Lazada	28.173.300	#3	#2	455.700	3.085.550	31.934.320	5.543
4	Bukalapak	25.760.000	#6	#7	232.300	1.776.710	2.516.190	2.503
5	Orami	16.683.300	n/a	n/a	5.720	11.770	350.940	215
6	Blibli	15.686.700	#8	#5	569.400	2.018.600	8.656.810	2.230
7	Ralali	5.923.300	#22	n/a	3.950	53.770	91.000	187
8	Zalora	3.310.000	#4	#8	6.440	743.730	8.008.550	625
9	JD ID	3.026.700	#7	#6	54.000	641.740	999.050	1.330
10	Sociolla	1.913.300	#5	#3	6.850	1.028.750	18.050	658

Figure 1. E-Commerce Rating

* Corresponding author.
E-mail address: putrisamapa@student.telkomuniversity.ac.id

Based on Figure 1, it can be seen that Shopee is ranked number one in the application on smartphones. Social media is considered to have an opportunity as a competitor to mass media because it can be a source of information that can also be a medium for socializing and interaction between users. Therefore, sentiment analysis is required to determine the sentiment on social media of Twitter regarding Shopee's e-commerce.

Tweets that were collected and then analyzed are referred to as sentiment analysis. Therefore, sentiment analysis is computational research of sentiment opinions expressed textually (Zulfa & Winarko, 2017). This analysis has the ability to extract public opinion on specific positive and negative values topics. Through sentiment analysis, companies such Shopee can efficiently get core input from the users or consumers. The information can be used to create a better product or service and also relevant to customer needs. In sentiment analysis, there is a feature extraction stage, one of which is TF-IDF which is a calculation of the weight of a frequently used word (Naufal & Setiawan, 2021).

Furthermore, there is also a feature expansion stage to discover the meaning of the sentiment on social media by converting words into numerical forms. One feature expansion aims to use *word2vec* to present words in vector form (Naufal & Setiawan, 2021). Furthermore, a classification process had been carried out there where several methods, namely Naïve Bayes, KNN, and SVM. Among the three ways, SVM has the advantage superior performance because it is able to classify based on the right class (Prastyo et al., 2020). SVM is a learning system using hypothetical spaces in the form of linear functions in a feature with high dimensions based on optimization theory (Monika Parapat & Tanzil Furqon, 2018).

1.2. Topics and Limitations

The topic discussed in this study is sentiment analysis using the classification method of support vector machine and word2Vec for word embedding on people's tweets on Twitter social media which discussed about Shopee e-commerce. The limitation in this study used data sets sourced from Twitter social media and only used "Shopee" category. The data set retrieved only used Indonesian language with .csv data format.

1.3. Objective

The purpose of this study is to conduct a sentiment analysis on Twitter social media that discussed Shopee using the support vector machine classification and the word2vec word embedding methods to determine the performance of model obtained from the SVM and word2vec methods by determining the best combination of word2vec parameter values and to find out the public sentiment situation regarding Shopee e-commerce.

2. Related Studies

Several studies have analyzed sentiment analysis towards an object. Research by (Nugroho et al., 2016) conducted a sentiment analysis of online bike services using the Naïve bayes method based on tweet mentions found on social media of Twitter. The results of the study stated that the system is able to classify sentiment classification using Naïve bayes with a resulting accuracy rate of 80% which was obtained based on a dataset of 800 data *Tweets* that were divided into two, namely 300 train data and 500 test data. The accuracy rate can also be improved by adding to the amount of train data.

Furthermore, research by (Sabrila et al., 2021) conducted a sentiment analysis of the handling Covid-19 by the Indonesian government. The methods used Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) by using word embedding feature extraction to compare the performance of the two methods. The classification process using SVM was carried out by extracting the Word2Vec word embedding feature with 85% accuracy, 85% recall, and 0.92 AUC value. Meanwhile, through KNN and the same feature extraction obtained 78% accuracy value, 77% precision, 76% recall, and 0.87 AUC value. So, it is concluded that the SVM method has better performance rather than the KNN method.

In research by (Tineges et al., 2020), this study conducted a sentiment analysis of Indihome services using the Support Vector Machine (SVM) method, with the aim of being able to find out the level of accuracy generated using the SVM method and also recognizing the level of satisfaction of Indihome users based on uploads found on Twitter social media. After conducting the research using this SVM method, the accuracy rate was obtained at 87%, precision at 86%, error rate at 13%, and f1-score at 90%.

3. Built Systems

In this study, the researcher used the Support Vector Machine and Word2Vec methods. The scheme in this study can be seen in Figure 2.

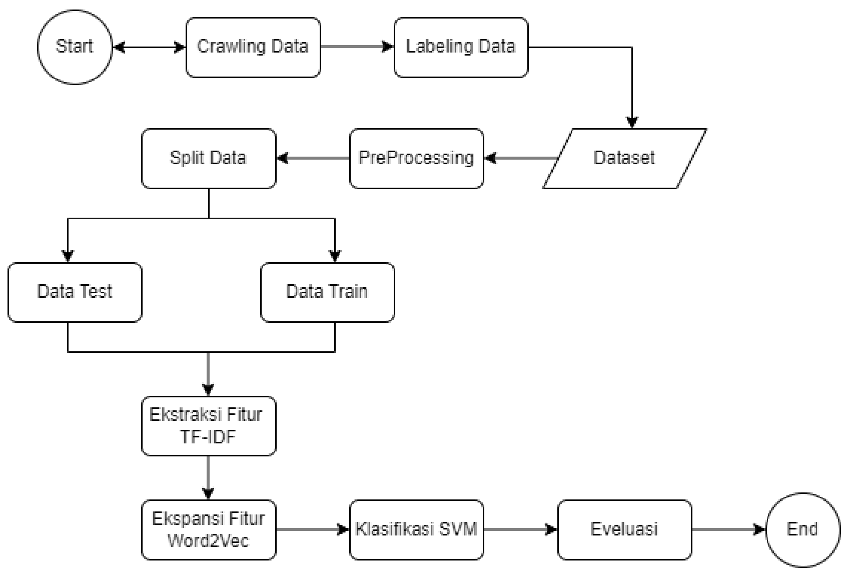


Figure 2. System Design Flowchart

3.1. Crawling Data

The data set was taken from Twitter social media using Shopee keywords.

Data collection used a data crawling method of *a netlytic* website. The data obtained from data crawling became a file with Comma Separated Values (CSV) format which contained tweets in the form of opinions about Shopee e-commerce.

3.2. Data Labeling

The collected data was manually labeled by dividing into three classes or labels, that were positive, neutral, and negative. Labeling was done by three people then selected the most votes in determining labels with the aim of reducing subjectivity in labeling. This process was done by paying attention to the words contained in the tweets data that have been taken. If the tweets data found the harsh words or inappropriate words, so it was labeled negative or -1. If the data did not contain harsh words and had a positive word meaning, then it was given a positive label or 1. Then, the data was given a neutral label if the sentence did not contain a positive meaning or negative.

The rude words have the meaning of swearing, slurs, ridicule, and others. Meanwhile, inappropriate sentences are sentences that contain provocations and have an impact on those who read them or negative sentences that can make things worse.

Table 1. Data Labeling Example

Label	Tweets
Positive	*emang paling bener cuma beli dari shopee dibanding aplikasi lain*
Neutral	*lama dah dia tak bagi weyhh. Sebagai pengguna shopee yang tegar saya amat terasa dengan ketiadaan voucher free shipping tersebut.*
Negative	*SUMPAH SHOPEE JELEK BANGET SKRG?!?! GUE KESEL ITU ONGKIR KENAPA SIH. makin kesini makin jadi dah*

3.3. Preprocessing Data

Preprocessing is the first stage carried out before classifying. Preprocessing aims to clean and transform data that is not needed to make it easier to process the data for the classification process. There are several stages of Preprocessing in this study, namely:

1) Data Cleaning

Data cleaning is the stage to clean or remove punctuation, numbers, *hastags*, and URLs. An example of the results data cleaning can be seen in Table 2.

Table 2. Data Cleaning

Before	After
Menunggu pemenguman pemenang lighstick seventeen dari shopee #shopeeterbaik#LightstickDariShopee# LightstickDariShopee	*Menunggu pemenguman pemenang lighstick seventeen dari shopee*

2) Case Folding

Case folding is the process of changing letters from capital letter prefixes to lowercase letters. An example of case folding results is presented in Table 3.

Table 3. Case Folding

Before	After
Menunggu pemenguman pemenang lighstick seventeen dari shopee	*menunggu pemenguman pemenang lighstick seventeen dari shopee*

3) Normalization

Normalization is a stage to identify excessive word writing and then replaced with words that are based on *KBBI* using a manually created word dictionary. An example of normalization results is presented in Table 4.

Table 4. Normalization

Before	After
menunggu pemenguman pemenang lighstick seventeen dari shopee	*menunggu pengumuman pemenang tongkat cahaya seventeen dari shopee*

4) Stop Word Removal

Stop Word Removal is a filtering process of words that are considered inappropriate or appear frequently. Words that are not unique words usually have no meaning. The process of removing *stopwords* in Indonesian using the *Sastrawi* library. An example of a stop word removal result is shown in Table 5.

Table 5. Stop Word Removal

Before	After
menunggu pengumuman pemenang tongkat cahaya seventeen dari shopee	*menunggu pengumuman pemenang tongkat cahaya seventeen shopee*

5) Stemming

Stemming is the process of converting a word into a base word. This process uses a library on *Python* that is specifically for Indonesian, namely the *Sastrawi* library. An example of stemming results is shown in Table 6.

6) Tokenization

Tokenization is the process of separating words in set data before analysis. An example of Tokenization result is shown in Table 7.

Table 6. Stemming

Before	After
menunggu pengumuman pemenang tongkat cahaya seventeen shopee	*nunggu pengumuman pemenang tongkat cahaya seventeen shopee*

Table 7. Tokenization

Before	After
nunggu pengumuman pemenang tongkat cahaya seventeen shopee	*"nunggu", "pengumuman", "pemenang", "tongkat", "cahaya", "seventeen", "shopee"*

3.4. Weighting of Term Frequency – Inversce Document Frequency (TF-IDF)

The feature extraction used in this study was TF-IDF. The weighting process in the TF-IDF consisted of calculating the value of the Term Frequency (TF) which was in the form of the frequency of occurrence of words transformed in the form of log tf and Inverse Document Frequency (IDF). It was the calculation of a term (sentence) in all documents. The result of the process is in the form of a matrix consisting of rows and columns where the data as rows and features as columns.

For calculations by the TF-IDF method, it can be seen in the following equation:

$$\mathrm{TF}(\mathrm{tk}, \mathrm{dj}) = f(\mathrm{tk}, \mathrm{dj}) \tag{1}$$

$$IDF(tk) = \log\ log\ \frac{N}{df(t)} \tag{2}$$

$$w(t,d) = \mathrm{t}(f,d) * idf(t) \tag{3}$$

Equation (1) is a calculation of the word searched for in a sentence. (tk,dj) is the number of occurrences of the term (k) in the document (j). Equation (2) is the value of the inverse document frequency obtained from the log results of the total document (N) divided by many words containing the searched word (df(t)). Equation (3) is a calculation of weights where the result was obtained by multiplying the number of occurrences of terms in the document with idf values.

3.5. Word2Vec

Feature expansion using the Word2Vec model is a word embedding method that was used to represent a word in vector form (Darliansyah et al., 2019). Word2Vec is used in conjunction with other algorithms to classify sentiment accurately (Al-Saqqa & Awajan, 2019). The principle of Word2Vec comes from the concept of words representation to study the embedding of words and predict between each word and its context so that words that appear in the same context are interrelated (Dirjen et al., 2017). Word2Vec is done before classifying the data set by working to convert words into vectors. Word2Vec results were obtained by searching for similar words so that they have similar vector values. Data in the form of text then turned into numerical data to see the weight of the word so that it can be classified. Here in Table 8 is shown the vector value of the word "shipping" with several other terms related to the word.

3.6. Support Vector Machine (SVM) Classification

Before classifying, data sharing or split data with a ratio of 90:10, 90% data train, and 10% test data were carried out. Then, the classification was carried out using the Support Vector Machine (SVM) method. SVM is one of the classification methods that has the principle of hyperplane which has the largest margin (Rani & Singh, 2017). The SVM method worked by finding the optimal hyperplane that provided a distance (separator) between the two classes (Tineges et al., 2020). The distance between the nearest data point and the hyperplane was referred as a margin (Rong, 2014). The supporting vector became the closest point to the hyperplane (Rong, 2014).

Table 8. Word2Vec

Word	Related Words
shipping	'voucher' , 0.9948447346687317, 'free' , 0.9896515607833862, 'menu' , 0.9547275304794312, 'claim' , 0.9428977966308594, 'minimum', 0.9366272687911987, 'cut' , 0.9333534204722656, 'the postage' , 0.9261620044708252, 'check' , 0.9185121655464172, 'infoin' , 0.9107703566551208

Hyperlane classification SVM can be defined as follows :

$$f(x) = w^T.x + b \tag{4}$$

where:

w = weight parameter,
x = vector input,
b = bias.

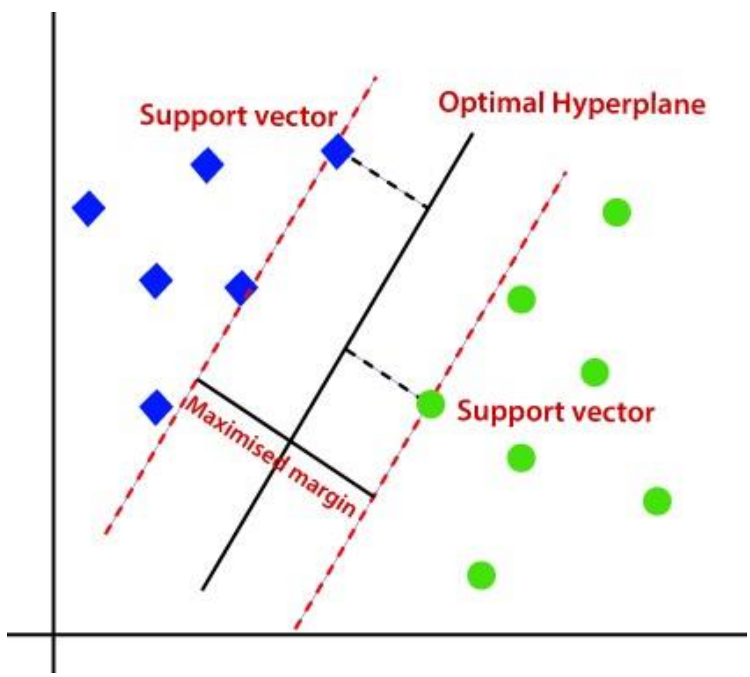


Figure 3. Support Vector Machine

To obtain a hyperlane that separated the data set into two classes, up and down with two optimal margin lines called support vectors. The formulation can be given:

$$w.x1 + b \leq -1 \, for \, Yi = -1 \tag{5}$$
$$w.x1 + b \leq +1 \, for \, Yi = +1 \tag{6}$$

An *x* pattern that includes class-1 is called a negative sample, while an *x* pattern that includes class+1 is called a positive sample (Rong, 2014).

In this study, there were 4 SVM kernel functions used, namely the linear kernel, the *Radial Basis Function* (RBF) kernel, the sigmoid kernel, and the polynomial kernel. Here are some of the kernel functions used:

1) Linear:

$$K(xi, x) = \langle xi^T x \rangle \tag{7}$$

2) Polynomial:

$$K(xi, x) = (\gamma \langle xi^T x \rangle + r)^d \tag{8}$$

where d is interpreted as *the degree* parameter and r is coef0

3) RBF :

$$K(xi, x) = \exp(-\gamma \|xi - x\|^2) \tag{9}$$

where γ is interpreted as a gamma parameter.

4) Sigmoid:

$$K(xi, x) = \tanh(\gamma xi^T + r) \tag{10}$$

Where r is interpreted as coef0.

3.7. Hyperparameter Tuning

Hyperparameter Tuning is a stage to produce an optimal model and improve the performance of a model adapted to a particular model. Four kernel functions on SVM were used in hyperparameter tuning. In this study, hyperparameters was started by initializing optimized hyperparameters and then used grid search to determine the best hyperparameters.

Grid search tests combinations have the aim of determining the combination of parameters for SVM that had the best model performance results by calculating the average value of cross validation for each combination. The hyperparameters that could be optimized in SVM were C parameters and gamma parameters. The C parameter worked by optimizing the SVM and avoiding misclassification in the data train. If the C value was higher, the probability of error would be smaller. The gamma parameter served to determine the influence on the data train. When the gamma value was high, it means that the points around the line would be considered in the calculation.

3.8. Confusion Matrix Evaluation

The evaluation stage was carried out to determine the performance level of the model that has been created. If the performance value owned by the system has a large value, the classification that has been carried out is effective. Performance calculations in this study used a confusion matrix based on accuracy, precision, recall, and f1-score values.

Table 9. Confusion Matrix

Class	**Predictions**	
	Positive	**Negative**
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

Table 9 shows the parameter used to calculate the classification performance value:

1) Accuracy

Accuracy is used to measure the performance of a method. Accuracy has the following equation:

$$\text{Accuracy} = (1\frac{(TP+TN)}{(TP+TN+FP+FN)} \tag{11}$$

2) Precission

Precission is used to calculate the comparison between positive data and the overall predicted results. Precission has the following equation :

$$\text{Precission} = (1\frac{TP}{(TP + FP)} \tag{12}$$

3) Recall

Recall is used to measure the degree of comparison between positive values data that are predicted correctly and all data that is predicted whether to be positive or not. Recall has the following equation:

$$\text{Recall} = (1\frac{TP}{(TP+FN)} \tag{13}$$

4) F1-Score

F1-score is a comparison of values between precission and recall. F1-score has the following comparison:

$$\text{f1-score} = (1\frac{Precission * Recall}{(Precission+Recall)} \tag{14}$$

4. Evaluation

4.1. Data

The tweet data used 5. 000 tweets in Indonesian language with topics related to Shopee which was obtained using the keyword Shopee. The data were divided into three labels, namely positive, negative, and neutral labels. The detail data distribution by label is presented in Table 10 and Figure 4.

Table 10. Data Distribution

Sentiment	Amount of data
Positive	231
Neutral	2627
Negative	2141

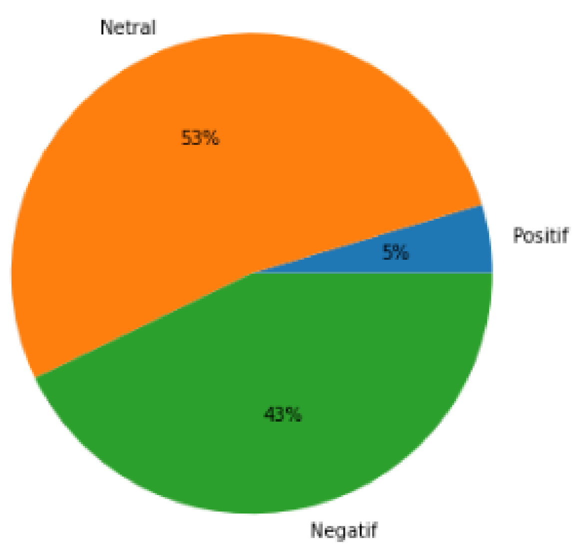


Figure 4. Data Distribution

4.2. TF-IDF weighting

In the TF-IDF weighting process, word weighting was carried out in each sentence in the data with the aim to know the value of the appearance of words in the document using equations (1), (2), and (3).

4.3. Test Scenarios and Results

In this study, two scenarios were carried out:

a. performance testing using Word2Vec on SVM classification model
b. performance testing using hyperparameter tuning

Before the first scenario was conducted, the first was to classify the *Support Vector Machine* without using the *Word2Vec* expansion feature to recognize the effect of adding a feature process of *expansion of Word2Vec* on SVM classification. Results are obtained without using *Word2Vec* as in Table 11.

Table 11. Word2Vec Featureless Testing

Kernel	F1-score	Precision	Recall	Accuracy
RBF	75.32%	84.71%	72.74%	91.58%
Sigmoid	81.44%	88.67%	78.24%	93.31%
Polynomial	61.56%	73.89%	60.03%	79.96%
Linear	81.09%	87.09%	78.24%	93.31%

4.3.1. First Scenario

The first scenario aims to conduct performance testing using *Word2Vec* on the SVM classification model. This test was carried out by finding the vector value of the desired word and related words contained in the data set. After testing using *Word2Vec,* it was found that the accuracy rate of each kernel on SVM increased, as presented in Table 12.

Table 12. Testing with Word2Vec Features

Kernel	F1-score	Precision	Recall	Accuracy
RBF	77.99%	87.88%	74.57%	**92.57%**
Sigmoid	85.25%	93.44%	81.18%	94.30%
Polynomial	61.05%	75.97%	58.71%	**80.19%**
Linear	84.37%	90.87%	80.86%	93.81%

Based on Table 12, it was found that the sigmoid kernel has the highest accuracy value with an accuracy amount of 94.30%.

4.3.2. Second Scenario

The second scenario aims to conduct performance testing using hyperparameter tuning by determining the best parameter on hyperparameter tuning with sigma (C) and a choice of values 0.001, 0.01, 0.1, 1, 10, 100, 1000, and gamma with a choice of values 1, 0.1, 0.01, 0.001, and 0.0001 for each SVM kernel except the linear kernel that did not use gamma values and in polynomial kernel. There were additions to degree grades of 1, 2, 3, 4, 5. In the test, the most optimal value was determined between several choices of C and gamma values. The results of testing with hyperparameter tuning can be seen in Table 13.

Table 13. Testing with Hyperparameter Tuning

Kernel	Best Parameters			Best Score
	C	Gamma	Degree	
RBF	1000	0.001	**-**	**93.01%**
Sigmoid	1000	0.001	-	93.20%
Polynomial	1000	0.001	1	**93.20%**
Linear	1	-	-	93.20%

4.4. Analysis of Test Results

Based on the test results of the two scenarios carried out, it was found that after hyperparameter tuning, there were several SVM kernels that experienced a decrease in the level of accuracy, namely the sigmoid kernel and linear kernel. The linear kernel is a suitable kernel for data already linearly separated and when there are many features. Meanwhile, in the sigmoid kernel, the greater the amount of gamma, so it tends to lowering the level of accuracy on a classification. The other two kernels of RBF and sigmoid experienced an increase, but the biggest increase was in polynomial kernels with changes from 80.19% to 93.20%. This situation was because polynomial kernels were suitable for solving classification problems in normalized data sets.

5. Conclusion

Based on sentiment analysis research using Word2Vec and the Support Vector Machine classification, it can be concluded that Word2Vec is able to provide good results to the level of accuracy of the Support Vector Machine classification with an increase in the accuracy of each SVM kernel used. However, there was no accuracy rate of the SVM kernel in testing all kernels using hyperparameter tuning. The accuracy level of each kernel could be differently influenced by the data set used and the parameter values used for each SVM kernel. The suggestion for future research is to try using imbalance techniques to find out if imbalance techniques can affect the accuracy value of the classification algorithm.

References

Al-Saqqa, S., & Awajan, A. (2019). The Use of Word2vec Model in Sentiment Analysis: A Survey. *ACM International Conference Proceeding Series*, 39–43. https://doi.org/10.1145/3388218.3388229

Darliansyah, A., Naeem, M. A., Mirza, F., & Pears, R. (2019). Sentipede: A smart system for sentiment-based personality detection from short texts. *Journal of Universal Computer Science*, *25*(10), 1323–1352.

Dirjen, S. K., Riset, P., Pengembangan, D., Dikti, R., Irmanda, H. N., & Astriratma, R. (2017). Terakreditasi SINTA Peringkat 2 Klasifikasi Jenis Pantun dengan Metode Support Vector Machines (SVM). *Masa Berlaku Mulai*, *1*(3), 915–922.

Monika Parapat, I., & Tanzil Furqon, M. (2018). Penerapan Metode Support Vector Machine (SVM) Pada Klasifikasi Penyimpangan Tumbuh Kembang Anak. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *2*(10), 3163–3169.

Naufal, H. F., & Setiawan, E. B. (2021). Ekspansi Fitur Pada Analisis Sentimen Twitter Dengan Pendekatan Metode Word2Vec. *E-Proceeding of Engineering*, *8*(5), 10339–10349.

Nugroho, D. G., Chrisnanto, Y. H., & Wahana, A. (2016). Analisis Sentimen Pada Jasa Ojek Online Menggunakan Metode Naive Bayes. *Prosiding SNST Ke-7 Tahun 2016*, 156–161.

Prastyo, P. H., Sumi, A. S., Dian, A. W., & Permanasari, A. E. (2020). Tweets Responding to the Indonesian Government's Handling of COVID-19: Sentiment Analysis Using SVM with Normalized Poly Kernel. *Journal of Information Systems Engineering and Business Intelligence*, *6*(2), 112. https://doi.org/10.20473/jisebi.6.2.112-122

Rani, S., & Singh, J. (2017). Sentiment Analysis of Tweets Using Support Vector Machine. *International Journal of Computer Science and Mobile Applications*, *5*, 83–91.

Rong, X. (2014). *word2vec Parameter Learning Explained*.

Sabrila, T. S., Sari, V. R., & Minarno, A. E. (2021). Analisis Sentimen Pada Tweet Tentang Penanganan Covid-19 Menggunakan Word Embedding Pada Algoritma Support Vector Machine Dan K-Nearest Neighbor. *Fountain of Informatics Journal*, *6*(2), 69. https://doi.org/10.21111/fij.v6i2.5536

Siagian Ade Onny, Martiwi Rini, & Indra Natal. (200 C.E.). Kemajuan Pemasaran Produk Dalam Memanfaatkan Media Sosial Di Era Digital. *Jurnal Pemasaran Kompetitif*, *3*.

Tineges, R., Triayudi, A., & Sholihati, I. D. (2020). Analisis Sentimen Terhadap Layanan Indihome Berdasarkan Twitter Dengan Metode Klasifikasi Support Vector Machine (SVM). *Jurnal Media Informatika Budidarma*, *4*(3), 650. https://doi.org/10.30865/mib.v4i3.2181

Zulfa, I., & Winarko, E. (2017). Sentimen Analisis Tweet Berbahasa Indonesia Dengan Deep Belief Network. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *11*(2), 187. https://doi.org/10.22146/ijccs.24716