Performance Comparison of Support Vector Machine Kernel Functions in Classifying COVID-19 Sentiment

Adian Fatchur Rochim[1]
Department of Computer Engineering,
Universitas Diponegoro,
Semarang 50275, Indonesia,
adian@ce.undip.ac.id

Khoirunisa Widyaningrum[2]
Department of Computer Engineering,
Universitas Diponegoro,
Semarang 50275, Indonesia,
khoirunisaw@student.ce.undip.ac.id

Dania Eridani[3]
Department of Computer Engineering,
Universitas Diponegoro,
Semarang 50275, Indonesia,
dania@ce.undip.ac.id

2021 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI) | 978-1-6654-0151-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISRITI54043.2021.9702845

Abstract--- Support Vector Machine (SVM) algorithm is a machine learning algorithm that is used to classify data by finding the best hyperplane that separates classes. In the SVM algorithm there are several types of kernel methods. Linear, Radial Basis Function (RBF), and polynomial kernel are some of the most commonly used SVM kernels. In previous research, each kernel has been used. However, the comparison of the three kernel function methods on the same dataset using accuracy, sensitivity, and specificity parameters has not been obtained. For this reason, this research is proposed to obtain comparative information of the three kernel functions using accuracy, sensitivity, and specificity parameters. The expected results can later be used as a reference for implementing the best kernel functions. The dataset used is comments on Youtube to analyze public sentiment on the increase in cases at the beginning of the entry of the COVID-19 pandemic in Indonesia. In this study, the accuracy values of the classification model were 0.86 for linear kernel, 0.90 for RBF kernel, and 0.91 for polynomial kernel. The sensitivity values obtained for each model are 0.64 for linear kernel, 0.48 for RBF kernel, and 0.20 for polynomial kernel. While the specificity values obtained for each model are 0.89 for linear kernel, 0.95 for RBF kernel, and 0.99 for polynomial kernel..

***Keywords*--Sentiment Analysis, Support Vector Machine, Linear Kernel, Polynomial Kernel, Radial Base Function, COVID-19**

I. INTRODUCTION

In early 2020 a new type of virus called Coronavirus Disease 2019 (COVID-19) entered Indonesia. This COVID-19 case started from an incident where the patient had direct contact with a Japanese citizen who had been infected with the COVID-19 virus [1]. The increase in COVID-19 cases in Indonesia is routinely officially announced by the Ministry of Health through press releases [2]. The Indonesian mass media rebroadcast the press release through various media, one of which is the Youtube channel. The public is free to share their opinions about COVID-19 through the Youtube comment column. These opinions have negative and positive tendencies. The various public sentiments regarding COVID-19 can be used as an alternative source of sentiment information from public opinion for policymakers.

Sentiment analysis is a process to find tendency of opinion towards a problem or object by someone. Sentiment analysis was carried out to find whether an opinion has a positive tendency or negative tendency[3]. Supervised learning is one of the machine learning methods that can be used to analyze sentiment [4]. Support Vector Machine (SVM) is one of the algorithms of the supervised learning method that can be used to analyze sentiment [5]. SVM has several kernel functions that are commonly used to solve training vectors that are not linearly separated or in other words, between classes cannot be separated only by straight lines. Linear, Radial Basis Function (RBF), and polynomial kernel are some common kernel functions used in SVM. Using different kernels will result in different performances.

A study conducted by Maryam Orafaee Azghan in June 2019, linear kernel of Support Vector Machine (SVM) method was used to classify cases of missing children. In this study, the results obtained 89% for accuracy, 84% for sensitivity, and 91% for specificity to predict the possibility of someone missing again. This study also makes a model to predict cases of someone's disappearance due to being involved in gang activity with result 84% for accuracy, 43% for sensitivity, and 90% for specificity [6]. A study conducted in 2019 by Mukti Ratna Dewi applied Radial Base Function (RBF) kernel of SVM to classify internet access by children and adolescents in East Java. In this study, the results obtained were 86.45% for accuracy, 84.64% for sensitivity, and 88.63% for specificity [7]. In a study conducted by Herlina Catur Sulistya Ningrum in 2018, the SVM method was applied by comparing the linear, polynomial, and RBF kernel functions in the classification of advanced study areas chosen by UII alumni. In the study, model that applies the polynomial kernel function has the highest accuracy value which is 95.45%, while the other kernels have an accuracy value, 94.32% for the linear kernel and 93.82% for the RBF kernel [8].

From previous research, each kernel has been used. However, the comparison of the three kernel function methods on the same dataset using the parameters of accuracy, sensitivity, and specificity has not been obtained. For this reason, this research is proposed to obtain information on the comparison of linear, RBF, and polynomial kernel of SVM algorithm using accuracy, sensitivity, and specificity parameter on the same dataset. Sensitivity is a measure of how well a classification algorithm classifies data points in the positive class and specificity is a measure of how well a classification algorithm classifies data points in the negative class [9]. The use of sensitivity and specificity parameters are expected to show how accurately the classification model can classify into each class, not just in general performance. The expected results can later be used as a reference for implementing the best kernel functions.

The dataset used is data from the Youtube media to analyze public sentiment on the increase in cases at the beginning of the entry of the COVID-19 pandemic in

978-1-6654-0151-7/21/$31.00 ©2021 IEEE

Indonesia. The comparison method used is by testing the SVM kernel linear, polynomial, and RBF algorithm with the same dataset and then comparing its performance using accuracy, sensitivity, and specificity parameter. The results obtained can then be used as a consideration for the selection of kernels on the SVM method for sentiment classification.

This paper divides into five sections. The first section is introduction, second section is literature review, third section is research methodology, forth section is results and discussion and fifth section is conclusion.

II. Literature Review

A. Support Vector Machine(SVM)

Support Vector Machine (SVM) is an algorithm in machine learning that aims to find the best hyperplane that separates two classes in the input space. The basic concept of SVM is to find the best hyperplane that separates two different classes. Hyperplane is a separator in the form of d-1 dimensions in d-dimensional space [10].

In SVM there are various kinds of kernel functions. The function of the kernel is to take data as input and convert it into the desired shape. Here are some of the kernels.

1. Linear Kernel

Linear kernel also known as soft margin will try to find a hyperplane that is a straight line but may tolerate one or more data misclassifications. The amount of misclassification tolerance given greatly affects the accuracy of the hyperplane. In Sklearn the tolerance is called C. The larger value of C, the less misclassification tolerance and the narrower the margin [11]. Sklearn or Scikit-learn is a module that can integrate various machine learning algorithms [12]. If x,z are assumed as dot product, then below is the equation formula.

$$K(x,z) = x.z + C \tag{1}$$

2. Polynomial Kernel

Polynomial kernel generates new feature by applying polynomial combination of existing features. In polynomial kernel, there is d parameter which is the value of polynomial degree which will determine how much curvature of the hyperplane[11].

$$K(x,z) = (x.z + C)^d , d > 0 \tag{2}$$

3. Radial Basis Function(RBF) Kernel

In the RBF kernel there is a gamma parameter which affects the data mapping[13]. Gamma is a measure of how much curvature is allowed within the decision limit, the greater the gamma, the more curvature in the hyperplane [14].

$$K(x,z) = \exp(-\gamma\|x - z\|^2), \gamma > 0 \tag{3}$$

B. Validation

One of the various ways to estimate the error of a predictive model is K-fold cross-validation. K-fold cross-validation is similar to the repeated random subsampling method, but sampling is carried out in such a way that no two test sets overlap. In K-fold cross-validation, the training dataset is divided into k subsets of the same size. Here, 'fold' refers to the number of subsets generated. The subset distribution is done randomly from the training dataset without replacement. The model is trained using a subset of k-1 which represents the validation set. Then the model is applied to the remaining subset which is used as the training set and its performance is measured. This process is repeated until each subset k serves as a validation set [15].

C. Model Performance Evaluation

Data imbalance is one of the problems encountered. One class can have numerous differences so that the effect on the calculation of model accuracy will not be ideal. The resulting accuracy will be skewed towards calculations in the majority class. The solution is to use a confusion matrix. The confusion matrix is a matrix that visualizes the performance of the classification algorithm using the data in the matrix. Table 1 is a confusion matrix that stores four values in the actual class of accuracy [9].

Table 1. Confusion Matrix

Predicted/Actual Class	**Positive Class**	**Negative Class**
Positive Class	TP	FP
Negative Class	FN	TN

Accuracy is the amount of data correctly classified by the classification algorithm of the entire data.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP} \tag{4}$$

Sensitivity is a measure of how well a classification algorithm classifies data points in the positive class

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{5}$$

Specificity is a measure of how well a classification algorithm classifies data points in the negative class

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{6}$$

True-positive (TP) is the number of correctly classified data from the positive class. False-positive (FP) is the amount of data that is predicted to be in the positive class but actually belongs to the negative class. True-negative (TN) is the number of correctly classified data from the negative class. False-negative (FN) is the amount of data that is predicted to be in the negative class but actually belongs to the positive class [9].

III. Research Methodology

This section will explain the sentiment classification steps based on the SVM method. The first explain the data used in this study, second explain the data collection method, third explain about sentiment polarization method, forth explain about text preprocessing method, fifth explain about word weighting method, sixth explain about classification method, and seventh explain about data representation.

A. Data

The data used are comments on the Indonesian Health Ministry announcement video regarding the increase in COVID-19 cases in Indonesia on official Indonesian television accounts, namely CNN Indonesia, KOMPASTV, and tvONE. The data obtained were 3850 comments and labeled into negative and positive.

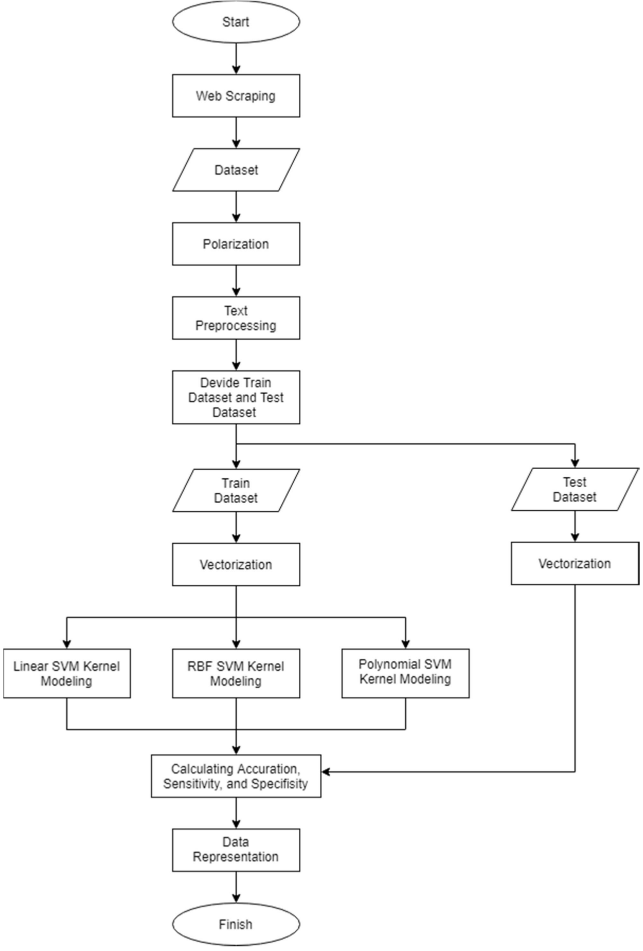

Figure 1. Step by step of res`   earch process

B. Data Collection Method

In this study, the data collection method used is web scraping which will retrieve certain data from a web. In this study, the web scraping method was carried out to collect Youtube comments which in the next stage would be analyzed.

C. Sentiment Polarization Method

In this study, the polarity of sentiment was carried out manually by three people into into positive and negative sentiment to prevent the subjectiveness of the sentiment. Then the majority polarity values from them were taken.

D. Text Preprocessing Method

There are several methods carried out at the text preprocessing stage.

1. Case folding is used to make all letters in the dataset lowercase.
2. Cleaning is used to sort data from punctuation and numbers so the data set is only letters.
3. Normalization is used to change words that are often abbreviated and words that are not standard into standard words.
4. Stopwords removal is used to sort data from words that have no important meaning.
5. Stemming is used to change affixed words back into root words.
6. Tokenization is used to break data in the form of sentences into word for word.

E. Word Weighting Method

Word weighting is done using the Term Frequency Inverse-Document Frequency (TF-IDF) method which will give weight to each word by calculating the frequency of occurrence of a word in a particular document and the inverse of the frequency of documents containing that word.

F. Classification Method

Data classification method in this study uses machine learning methods by applying the Support Vector Machine (SVM) algorithm with three different kernels, namely linear, polynomial, and Radial Base Function (RBF).

G. Data Representation

In this study using a confusion matrix with a format as shown in table 1 to display the results of the three classification models in classifying test dataset. From the confusion matrix obtained, the accuracy, sensitivity, and specificity values of each model can be calculated. The values of accuracy, sensitivity, and specificity of the three classification models are then compared using a bar chart so that the differences of the performance of the three classification models can be seen.

IV. Result & Discussion

In this study, three different Support Vector Machine (SVM) kernels are applied, namely linear, polynomial, and Radial Basis Function (RBF). In this study, K-fold cross-validation was used. The dataset was divided into 10 folds of approximately the same size. The use of these 10 folds recommended because it is the best number of folds for validity test[16]. K-fold cross-validation will use 1 fold for testing and 9 fold for training data where each fold will alternately become a subset for testing and the rest for training data. This study uses a data comparison of 80:20. From 3850 data, 3080 data were used as training data and 770 data were used as test data. The distribution of positive and negative class data can be seen in Figures 2 and 3.

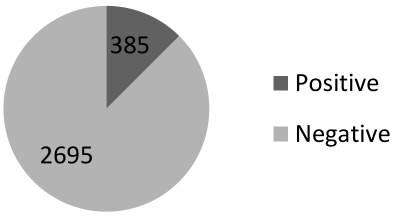

Figure 2. Sentiment Distribution of Training Dataset

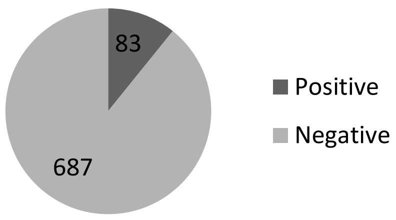

Figure 3. Sentiment Distribution of Test Dataset

The data to be processed with SVM is firstly weighted by Term Frequency-Inverse Document Frequency (TF-IDF) for each word. In this study, all kernels use the value C = 1. The use of the value C = 1 for all kernels because this value is the best C value after a search using the gridsearch function. To determine the classification performance for each kernel using the results of the calculation of accuracy using equation 4, sensitivity using equation 5, and specificity using equation 6.

1. Linear Kernel

Table 2. Confusion Matrix of Linear Kernel Model

		Predicted Class	
		Positive	Negative
Actual Class	Positive	53	30
	Negative	75	612

Table 2 shows that from 83 positive class test data, 53 data were correctly predicted into the positive class or called as True Positive (TP) and 33 positive class data was incorrectly predicted or called as False Negative (FN), also from 687 negative class test data, True Negative (TN) or data with negative polarity that predicted correctly is 612 and False Positive (FP) or data with negative polarity that predicted incorrectly is 75 data. The calculation results for accuracy, sensitivity, and specificity are as follows. After obtaining the value from the confusion matrix, the value of accuracy, sensitivity, and specificity can be calculated so that the accuracy value is 0.86, the specificity value is 0.89, and the sensitivity value is 0.64.

2. Polynomial Kernel

Table 3. Confusion Matrix of Polynomial Kernel Model

		Predicted Class	
		Positive	Negative
Actual Class	Positive	17	66
	Negative	7	680

Table 3 shows that from 83 positive class test data, 17 data were correctly predicted into the positive class or called as True Positive (TP) and 66 positive class data was incorrectly predicted or called as False Negative (FN), also from 687 negative class test data, True Negative (TN) or data with negative polarity that predicted correctly is 680 and False Positive (FP) or data with negative polarity that predicted incorrectly is 7 data. The calculation results for accuracy, sensitivity, and specificity are as follows. After obtaining the value from the confusion matrix, the value of accuracy, sensitivity, and specificity can be calculated so that the accuracy value is 0.91, the specificity value is 0.99, and the sensitivity value is 0.20.

3. RBF Kernel

Table 4. Confusion Matrix of RBF Kernel Model

		Predicted Class	
		Positive	Negative
Actual Class	Positive	40	43
	Negative	32	655

Table 4 shows that from 83 positive class test data, 40 data were correctly predicted into the positive class or called as True Positive (TP) and 43 positive class data was incorrectly predicted or called as False Negative (FN), also from 687 negative class test data, True Negative (TN) or data with negative polarity that predicted correctly is 655 and False Positive (FP) or data with negative polarity that predicted incorrectly is 32 data. The calculation results for accuracy, sensitivity, and specificity are as follows. After obtaining the value from the confusion matrix, the value of accuracy, sensitivity, and specificity can be calculated so that the accuracy value is 0.90, the specificity value is 0.95, and the sensitivity value is 0.48.

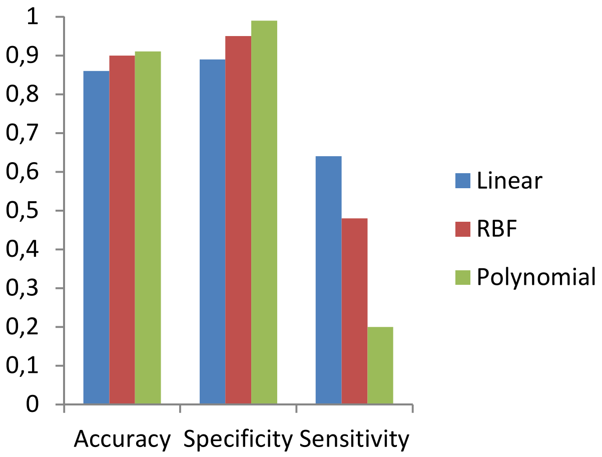

Figure 4. SVM Kernel Performance Comparison Chart

Figure 4 shows the comparison of linear, polynomial, and RBF kernel performance using the parameters of accuracy, sensitivity, and specificity. In the picture above, it can be seen that the three classification models that apply linear kernel, polynomial, and RBF have good performance for classifying sentiments if only seen from the accuracy value. But when viewed from the sensitivity value, the three classification models have an unsatisfactory performance when classifying data in the positive class correctly. However, the three classification models perform well when classifying data in the negative class correctly. Figure 4 shows that the classification model that applies the polynomial kernel has the highest accuracy and specificity values, but the classification model gets the lowest sensitivity value. Similar to previous research conducted by Rajul Parikh et al. in 2008, that the sensitivity value is inversely proportional to the specificity value [17].

In Figure 4 there is a discrepancy between the sensitivity and specificity values of the three kernel models. This can be caused by a significant difference in the amount of data between the positive and negative class data. Similar to previous research conducted by Luqman Assafat in 2016, the more training data used, the better the accuracy obtained [18].

V. Conclusion

From this research, the accuracy value is 0.86 for linear kernel, 0.90 for RBF kernel, and 0.91 for polynomial kernel for classifying sentiment regarding the entry of COVID-19 in Indonesia. The specificity values obtained are 0.89 for linear kernel, 0.95 for RBF kernel, and 0.99 for polynomial kernel. Meanwhile, the sensitivity values obtained are 0.64 for linear kernel, 0.48 for RBF kernel, and 0.20 for polynomial kernel. In general, the three classification models that apply the linear kernel, polynomial, and RBF have good performance for classifying sentiments when seen from the accuracy value. However, from the sensitivity value, the three classification models have an unsatisfactory performance when classifying data in the positive class correctly. However, the three classification models perform well when properly classifying data in the negative class. From accuracy, specificity, and sensitivity values obtained from each model that applies the three kernels, it can be concluded that the increasing specificity value will be accompanied by a decreasing in sensitivity value, also the accuracy value not recommended as the only reference to assess the performance of the classification model.

By grouping 3850 public sentiment data related to the entry of COVID-19 at the beginning of the pandemic, 87.8% from data source fall into the negative class. This shows that at the beginning of the COVID-19 pandemic in Indonesia, the majority of Indonesian people were very worried about the virus. However, this research needs to be studied further because this research only focuses on Youtube comments. For further research, it is recommended to use data from various platforms so that it covers various groups of people.

Acknowledgments

This research was financially supported by the Ministry of Research and Technology, Republic Indonesia through Fundamental Research Grant No. 225-98/UN7.6.1/PP2021.

Reference

[1] Yuliana, “Corona Virus Disease(COVID-19),” *Wellness and Healthy Magazine*, volume 2, number 1, pp. 187-192, February 2020.

[2] PPID Provinsi DKI Jakarta, “Informasi Mengenai Siaran Pers Terkait Pandemi COVID19”, *ppid.jakarta.go.id*. [Online]. Available : ppid.jakarta.go.id/siaranpers-covid19. [Accessed : 22-Aug-2021].

[3] I. F. Rozi, H. S. Pramono and E. A. Dahlan, "Implementasi Opinion Mining(Analisis Sentimen) untuk Ekstraksi Data Publik pada Perguruan Tinggi,," *Electrics, Electronics, Communications, Controls, Informatics, Systems Journal*, volume 6, number 1, pp. 37-43, 2012.

[4] P. Goncalves, M. Araujo, F. Benevenuto, and M. Cha, “Comparing and Combining Sentiment Analysis Methods”, *ACM*, no. May 2014, pp. 27-38, May 2014.

[5] R. Primartha, *Belajar Machine Learning Teori dan Praktik.* Bandung : Informatika Bandung, 2018.

[6] M. O. Azghan, “Classification of Missing Youth Cases using Support Vector Machine”, M. S. thesis, Departement of Computer Science, University Saskachewan, Saskatoon.

[7] M. R. Dewi, “Klasifikasi Akses Internet oleh Anak-anak dan Remaja Dewasa di Jawa Timur menggunakan Support Vector Machine”, *Jurnal Riset dan Aplikasi Matematika*, volume 4, number 1, pp. 17-27, 2020.

[8] H. C. S. Ningrum, “Perbandingan Metode Support Vector Machine (SVM) Linear, Radial Base Function (RBF), dan Polinomial dalam Klasifikasi Bidang Pilihan Alumni UII”, B. S. thesis, Departement of Statistic, Universitas Islam Indonesia, Yogyakarta.

[9] M. Awad and R. Khanna, *Efficient Learning Machines : Theories, Concepts, Applications for Engineers and System Designers*, Apress Open, 2015.

[10] A. S. Nugroho, A. B. Witarto, and D. Handoko, “Support Vector Machine”, *IlmuKomputer.com*. [Online]. Available: asnugroho.net/ikcsvm.pdf. [Accessed: 5-Jun-2021].

[11] L. Cheng. “Support Vector Machine – Simply Explained”, *towardsdatascience.com*. [Online]. Available: towardsdatascience.com/support-vector-machine-simply-explained-fee28eba5496. [Accessed: 5-Jun-2021].

[12] F. Pedegrosa, G. Varoquaux, A. Gramfort, V. Michel, and B. Thirion, “Scikit-learn: Machine Learning in Python”, *Journal of Machine Learning Research*, volume 12, pp. 2825-2830, 2011.

[13] Sickit Learn. “1.4 Support Vector Machines” *scikit-learn.org*. [Online]. Available: scikit-learn.org/stable/modules/svm.html [Accessed: 5-Jun-2021].

[14] R. Felix, “Laporan Praktikum 11 STA582”, *rpubs.com*. [Online]. Available: rpubs.com/RezaFelix/svm. [Accessed : 9-Nov -2021].

[15] D. Berrar, “Cross Validation”, *Encyclopedia of Bioinformatics and Computational Biology*, volume 1, pp. 542-545, 2018.

[16] R. Kohavi. (1995). A study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. Presented at International Joint Conference on Articial Intelligence (IJCAI) 1995. [Online]. Available: http://ai.stanford.edu/~ronnyk/accEst.pdf.

[17] R. Parikh, A. Mathai, S. Parikh, G. C. Sekhar, and R. Thomas, “Understanding and Using Sensitivity, Specificity and Predictive Values”, *Indian Journal of Ophthalmology*, volume 56, number 1, pp. 45-50, 2008.

[18] L. Assafat, “Pengaruh Jumlah Data Latih SVM Pada Peramalan Beban Listrik Bulanan di Sektor Industri”, *Jurnal Fakultas Teknik Universitas Muhammadiyah Purwokerto*,volume 7, number 2, pp. 88-93, 2016.