Comparison of Kernels Function between of Linear, Radial Base and Polynomial of Support Vector Machine Method Towards COVID-19 Sentiment Analysis

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@students.ce.undip.ac.id

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

***Abstract---* Support Vector Machine (SVM) is a machine learning algorithm that is generally used to classify data by finding the best hyperplane that separates classes. In SVM algorithm use kernel function when dataset cannot be separated linearly. There are several types of kernel methods, including linear, Radial Basis Function (RBF), and polynomials of the SVM algorithm. In previous research, each kernel has been used. However, the comparison of the three kernel function methods has not been obtained because each kernel function method is used in different datasets. For this reason, this research is proposed to obtain comparative information of the three kernel functions. In this study, we will compare the linear kernel SVM algorithm, RBF, and polynomial using the parameters of accuracy, sensitivity, and specificity. The dataset used is data from Youtube media 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 the linear kernel, 0.90 for the Radial Base Function (RBF) kernel, and 0.91 for the polynomial kernel. The sensitivity values obtained for each model are 0.64 for the linear kernel, 0.48 for the Radial Base Function (RBF) kernel, and 0.20 for the polynomial kernel. While the specificity values obtained for each model are 0.89 for the linear kernel, 0.95 for the Radial Base Function (RBF) kernel, and 0.99 for the 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 voice 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 policy makers.

Sentiment analysis is a field of study that analyzes opinions, feelings, judgments, estimates, attitudes, and human emotions towards products, services, organizations, individuals, problems, events, topics, and other attributes [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 are some common kernel functions used in SVM. Using different kernels will result in different performance.

In a study conducted by Lailil Muflikhah, et al., in 2018, applying the Support Vector Machine (SVM) method with two RBF kernel functions and polynomial kernel on sentiment analysis regarding product reviews. In this study, the accuracy of the RBF kernel was 83.25% while the polynomial kernel was 88.75%, which means that the polynomial kernel is considered to have better performance than the RBF kernel for product review classification [6]. In 2019, Mahendrajaya et al., compared linear kernel and polynomial kernel to analyze Gopay user sentiment on their study. In this study, the accuracy of the linear kernel was 89.17% while the polynomial kernel was 84.38%, which means that the linear kernel is considered to have better performance than the polynomial kernel for sentiment classification in this study [7]. In a study conducted by Neneng and Asep in 2019, the SVM method was applied by comparing RBF kernel and linear kernel in the classification of soybean plant diseases. The results of this study showed that the RBF kernel function was better than the linear kernel function [8].

From previous research, each kernel has been used. However, the comparison of the three kernel function methods has not been obtained because each kernel function method is used in different datasets. For this reason, this research is proposed to obtain comparative information of the three kernel functions. In previous studies, kernel performance was only assessed using accuracy value. This study will compare kernel of SVM algorithm between of linear, RBF, and polynomial using the parameters of accuracy, sensitivity, and specificity. 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 parameter 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 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 the parameters of accuracy, sensitivity, and specificity. 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. First section is introduction, second is literature review, third is research methodology, forth is results and discussion and fifth 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 hyperlane 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. Although it tolerates some errors, linear kernel still tries to find a line that maximizes margins and minimizes misclassification. 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].

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

2. Polynomial Kernel

When data cannot be separated by straight lines, a polynomial kernel is needed. Polynomial kernel can generate nonlinear decision boundary. Polynomial kernel generate new feature by applying polynomial combination of existing features [11].

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

3. Radial Basis Function(RBF) Kernel

Radial Basis Function(RBF) kernel is needed when the data is really unevenly distributed. When training a dataset with RBF kernel, there are two parameters to consider, namely C and gamma. Parameter C aims to tell how much error to avoid in classifying training data, the greater the value of C, the lower the misclassification of training data. The gamma parameter determines how far the influence of a single sample of training data is. This means that the smaller the gamma value, the farther the distance from the data points to be calculated [12].

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

B. Validation

One of the various way 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 [13].

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[14].

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

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) often called recall 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. METHODOLOGY

A. System Planning

In this research methodology section the sequence of research steps, datasets, and methods that were used for classification will be explained.

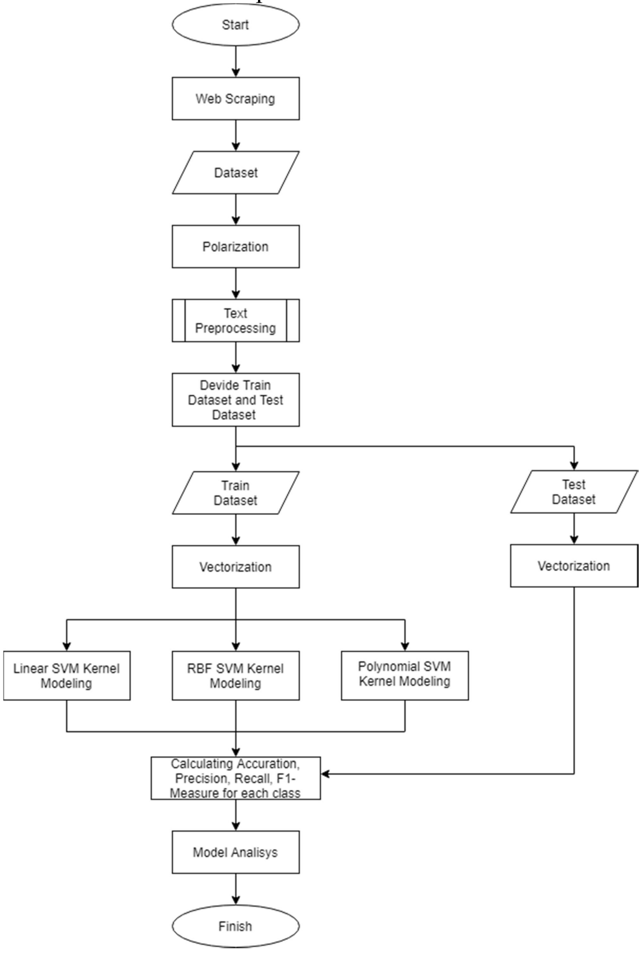
Start
Web Scraping
Dataset
Polarization
Text Preprocessing
Devide Train Dataset and Test Dataset
Train Dataset
Test Dataset
Vectorization
Vectorization
Linear SVM Kernel Modeling
RBF SVM Kernel Modeling
Polynomial SVM Kernel Modeling
Calculating Accuration, Precision, Recall, F1-Measure for each class
Model Analisys
Finish

Figure 1. System Planning Flowchart

1. Retrieving public sentiment data from Youtube comments using scraping techniques and save the data in .csv file format
2. Giving polarity manually
3. Doing text preprocessing so that the data to be processed by the model is more structured
   a. Doing case folding to make all letters lowercase.
   b. Sorting data from punctuation and numbers
   c. Normalizing data so that non-standard words are converted into standard words.
   d. Removing stopwords
   e. Stemming data so that affixed words return to basic words
   f. Breaking sentences into word units through the tokenization process and then saving the preprocessed text file in the form of .csv
4. Dividing the dataset into training data and test data with a ratio of 8:2
5. Vectorizing uses TF-IDF weighting to convert the dataset in the form of words into numbers so that it can be processed by the model to be built
6. Creating a classification model with the Support Vector Machine (SVM) algorithm by applying three kernels, namely linear, Radial Basis Function (RBF), and polynomials.
7. Counting accuracy, sensitivity, and specificity value of each classification models
8. Analizing the performance of the three classification models using the parameter values of accuracy, sensitivity, and specificity.

B. Data

Different from other search before, the data used are comments on the Inodesian 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 was taken using scraping technique and the data obtained were 3848 comments and labeled into negative and positive.

IV. RESULT & DISCUSSION

In this study, three different Support Vector Machine (SVM) kernels are applied, namely linear, polynomial, and Radial Basis Function (RBF). Before it can be processed using the SVM algorithm, the data needs to be validated using K-fold cross validation. In this study, the data 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 [14]. 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 8:2. From 3848 data, 3078 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.

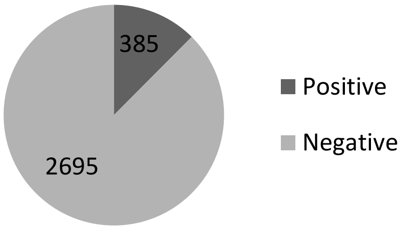

Figure 2. Sentiment Distribution of Training Dataset

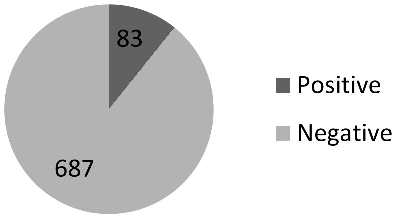

Figure 3. Sentiment Distribution of Test Dataset

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. Confussion 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.

$$\text{Accuracy} = \frac{53+61}{53+612+75+30} = 0{,}86$$

$$\text{Specificity} = \frac{612}{612+75} = 0{,}89$$

$$\text{Sensitivity} = \frac{53}{53+30} = 0{,}64$$

2. Polynomial Kernel

Table 3. Confussion 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.

$$\text{Accuraccy} = \frac{17+68}{17+680+7+6} = 0{,}91$$

$$\text{Specitifity} = \frac{680}{680+7} = 0{,}99$$

$$\text{Sensitivity} = \frac{17}{17+66} = 0{,}20$$

3. RBF Kernel

Table 4. Confussion 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.

$$\text{Accuracy} = \frac{40+655}{40+655+32+43} = 0{,}90$$

$$\text{Specificity} = \frac{655}{655+3} = 0{,}95$$

$$\text{Sensitivity} = \frac{40}{40+43} = 0{,}48$$

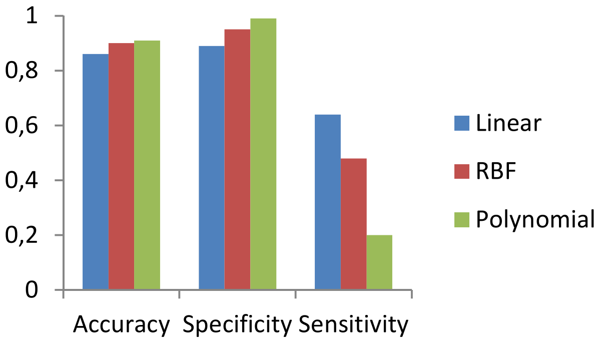

Figure 4. SVM Kernel Performance Comparison Chart

Figure 4 shows the comparison performance of classification model that apply linear, polynomial, and RBF kernel using the parameters of accuracy, sensitivity, and specificity. In the picture above, it can be seen that the three classification models that apply the linear, polynomial, and RBF kernel have good performance for classifying sentiments if only seen from the accuracy value. However, 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 properly classifying data in the negative class. 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.

V. Conclusion

From this research, the accuracy value is 0.86 for the linear kernel, 0.90 for the RBF kernel, and 0.91 for the polynomial kernel. The specificity values obtained are 0.89 for the linear kernel, 0.95 for the RBF kernel, and 0.99 for the polynomial kernel. Meanwhile, the sensitivity values obtained are 0.64 for the linear kernel, 0.48 for the RBF kernel, and 0.20 for the polynomial kernel. In general, the three classification models that apply the linear kernel, polynomial, and RBF have good performance for classifying sentiments if only seen from the accuracy value. However, 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 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 conclude that the increase in accuracy and specificity will be accompanied by a decrease in sensitivity also the accuracy value cannot be used as the only reference to assess the performance of the classification model.

By grouping 3848 public sentiment data related to the entry of COVID-19 at the beginning of the pandemic, people tend to have sentiments because 87.8% of negative comments from the public fall into the negative class. This shows that at the beginning of COVID-19 pandemic in Indonesia, majority of Indonesia 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.

Acknowledments

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 : https:// ppid.jakarta.go.id /siaranpers-covid19. [Accessed : 22-Aug-2021].

[3] L. Bing, *Sentiment Analysis and Opinion Mining*. California : Morgan & Claypool, 2012.

[4] P. Goncalves, et al., “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] L. Muflikhah, et al., “High Performance of Polynomial Kernel at SVM Algorithm for Sentiment Analysis”, *Journal of Information Technology and Computer Science*, volume 3, number 2, pp. 194-201, November 2018.

[7] R. Mahendrajaya, G. A. Buntoro, and M. B. Setyawan, “Analisis Sentimen Pengguna Gopay Menggunakan Metode Lexicon Based dan *Support Vector Machine*”, *Jurnal Teknik Universitas Muhammadiyah Ponorogo*, volume 3, number 2, pp. 52-63, October 2019.

[8] N. R. Feta and A. R. Ginanjar, “Komparasi Fungsi Kernel Metode *Support Vector Machine* Untuk Pemodelan Klasifikasi Terhadap Penyakit Tanaman Kedelai”, *BRITech*, volume 1, number 1, pp. 33-39, July 2019.

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

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

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

[14] R. Kohavi, "A study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection". [Online]. Available: http://ai.stanford.edu/~ronnyk/accEst.pdf. [Accessed 29-Sept-2021].