UNIVERSITAS BHAYANGKARA JAKARTA RAYA
FAKULTAS ILMU KOMPUTER
Kampus I: Jl. Harsono RM No. 67, Ragunan, Pasar Minggu, Jakarta Selatan 12550
Telepon: (021) 27808121 – 27808882
Kampus II: Jl. Raya Perjuangan, Marga Mulya, Bekasi Utara, Jawa Barat, 17142
Telepon: (021) 88955882, Fax.: (021) 88955871
Web: fasilkom.ubharajaya.ac.id, E-mail: fasilkom@ubharajaya.ac.id

SURAT TUGAS

Nomor: ST/1556/I/2022/FASILKOM-UBJ

1. Dasar: Kalender Akademik Universitas Bhayangkara Jakarta Raya Tahun Akademik 2021/2022.

2. Dalam rangka mewujudkan Tri Dharma Perguruan Tinggi untuk Dosen di Universitas Bhayangkara Jakarta Raya maka dihimbau untuk melakukan Penelitian.

3. Sehubungan dengan hal tersebut di atas, maka Dekan Fakultas Ilmu Komputer Universitas Bhayangkara Jakarta Raya menugaskan:

NO.	NAMA	NIDN	JABATAN	KETERANGAN
1.	Dr. Rakhmi Khalida, S.T., M.M.S.I.	0304099201	Dosen Tetap Prodi Informatika	Sebagai Penulis Keempat

Membuat Artikel Ilmiah dengan judul **"*Sentiment Analysis on Social Media (Twitter) about Vaccine-19 Using Support Vector Machine Algorithm*"** pada *media Institute of Electrical and Electronics Engineers (IEEE) Xplore*, 11 Februari 2022, DOI: 10.1109/ISRITI54043.2021.9702775, dengan tanggal Konferensi: 16 s.d 17 Desember 2021 di *2021 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*.

4. Demikian penugasan ini agar dapat dilaksanakan dengan penuh rasa tanggung jawab.

Jakarta, 27Januari 2022
DEKAN FAKULTAS ILMU KOMPUTER

Dr. Dra. Tyastuti Sri Lestari, M.M.
NIP. 1408206

International Seminar on Research of Information Technology and Intelligent Systems

PROCEEDING

#4th ISRITI 2021

Yogyakarta - Indonesia
16 December 2021

MACHINE LEARNING for DATA SCIENCE

isriti.akakom.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/ISRITI54043.2021.9702882

STMIK AKAKOM YOGYAKARTA

IEEE

IEEE INDONESIA SECTION

Sentiment Analysis on Social Media (Twitter) about Vaccine-19 Using Support Vector Machine Algorithm

Agus Sulistyono
Computer Science Department, BINUS Online Learning
Bina Nusantara University
Jakarta, Indonesia, 11480
Agus.sulistyono@binus.ac.id

Sri Mulyani
Computer Science Department, BINUS Online Learning
Bina Nusantara University
Jakarta, Indonesia, 11480
Sri.mulyani@binus.ac.id

Emny Harna Yossy*
Computer Science Department, BINUS Online Learning
Bina Nusantara University
Jakarta, Indonesia, 11480
emny.yossy@binus.ac.id

Rakhmi Khalida
Computer Science Department
Gunadarma University
Depok, Indonesia, 16424
Rakhmikhalida7@gmail.com

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/ISRITI54043.2021.9702775

***Abstract*—Currently the world is experiencing a Corona Virus Disease (Covid-19) pandemic which attacks the respiratory tract and spreads very quickly to various countries including Indonesia, so the World Health Organization (WHO) has declared Covid-19 as a pandemic. To overcome this pandemic, experts in the medical field also intervened by making vaccinations to strengthen human immunity against the Covid virus. This sentiment analysis was carried out to see opinions on the object, namely the existence of a Covid-19 vaccine. Data collection by crawling data with the keyword 'Covid Vaccine'. The method that will be used is the Support Vector Machine (SVM). The analysis was carried out by comparing the classification accuracy values of the two SVM kernel functions, namely linear and Radial Basic Function (RBF). The results of the study obtained positive sentiment of 43.5%, negative of 19.1%, and neutral of 37.4%. Then the evaluation of the system using the confusion matrix obtained an accuracy value for the linear kernel of 79.15%, a precision value of 77.31%, and a recall value of 78.09%. While the RBF kernel has an accuracy of 84.25%, a precision value of 83.67%, and a recall value of 81.99%. While the cross validation obtained the optimum value at k = 1 with an accuracy value of 80.18% for the linear kernel and 85.88% for the RBF kernel. So the RBF kernel has a higher accuracy than the linear kernel.**

Keywords—Covid-19, Vaccine, Support Vector Machine, Linear, Radial Basis Function.

I. INTRODUCTION

Advances in technology, one of which is the internet, makes it easy for us to exchange information or express thoughts, opinions, and respond to events through online media [1]. Online media to exchange information with others can be called social media. One type of social media that is widely used is Twitter. Twitter is a microblogging service where the flow of interaction is faster than blogs [2]. Microblogging can be optimized as a channel for fast interaction, so that concise and important information can be known by other users. This makes Twitter often used as a medium to provide experiences, share opinions, and respond to events. This response or experience can be classified to determine sentiment on a topic or can be called sentiment analysis [3]. Therefore, sentiment analysis is one solution to the problem of classifying opinions or reviews into positive or negative opinions. The opinion that will be reviewed in this study is the Covid Vaccine in Bahasa.

Currently the world is experiencing a Covid-19 pandemic (Corona Virus Disease-19) where this virus attacks the respiratory tract and spreads very quickly. In addition to the implementation of strict health protocols, other effective interventions are needed to break the chain of disease transmission, namely through vaccination efforts. Covid-19 vaccination aims to reduce the transmission / transmission of Covid-19, reduce morbidity and mortality due to Covid-19, achieve group immunity in the community (herd immunity) and protect the community from Covid-19 in order to remain socially and economically productive [4].

Research conducted is the application of the SVM algorithm for sentiment analysis on the twitter data of the Corruption Eradication Commission of the Republic of Indonesia [5]. Classification is divided into negative, neutral, or positive responses. The results of the testing and evaluation of the research are the accuracy of the test results by 82% and precision testing by 90%, as well as recall by 88% and f1-score by 89%. Another sesearch conducted is applying the SVM method in the classification of public figure tweet sentiment [6]. Classification is done by using the RBF kernel and polynomials on the SVM method. In this study to see the level of accuracy produced. From these results, it was found that the RBF kernel provides a better accuracy rate than the polynomial kernel with the accuracy for the RBF kernel on unigram features of 72.5% and the accuracy of the polynomial kernel only 72%.

In addition, the SVM algorithm for sentiment analysis reviews ruang guru applications [7]. Classification is done to see the positive or negative response. The SVM kernel used in this study are linear, RBF, and polynomial. From the results of this study, it was found that a high accuracy value was in the range of 90% with a linear kernel giving a better accuracy value than RBF and polynomials. A Study on the Implementation of Support Vector Machines for Sentiment Analysis of Twitter Users towards Telkom and Biznet Services [8]. With the aim of analyzing the sentiments of twitter users towards Telkom and Biznet services. Tests using the Confusion Matrix and K-Fold Cross validation are intended to share training information and testing information. K-Fold Cross validation and Confusion Matrix share the results of an accuracy value of 79.6%, precision 76.5%, recall 72.8%, and F1-score 74.6% for Telkom, and

*Corresponding Author

accuracy 83.2%, precision 78, 8% recall 71, 6%, and F1-score 75% for biznet.

Based on the description above, this research is about sentiment analysis on social media (Twitter) to classify Twitter user responses to the keyword in Bahasa is "Vaksin Covid "into positive, negative or neutral responses. In this study the classification method used is the Support Vector Machine (SVM).

II. Literature Review

A. Sentiment Analysis

Sentiment analysis is a branch of text mining research. Specifically, the purpose of text mining can be divided into two, namely text data categorization and text clustering. In categorization, text mining is used as a tool to find categories that match the specified class (supervised learning), while grouping in text mining functions as a tool to group text data based on similar characteristics, and clustering can be used to label unknown classes [9].

Text mining is a process of mining data in the form of text where the data source is usually obtained from documents and the goal is to find words that can represent the contents of the document so that an analysis of the connectivity between documents can be carried out. Data mining (Pattern Discovery) is the process of seeking knowledge or patterns that are interesting/valuable. Evaluation is the interpretation of patterns found [10]. The purpose of text mining is to extract useful information from data sources. So, the data source used in text mining is a collection of documents that have an unstructured format through the identification and exploration of interesting patterns. The stages of text mining are tokenization, lower case, removing tad abaca, stemming, and filtering.

B. Twitter

Twitter is a microblogging service that was officially released on July 13, 2006. Twitter's main activity is posting short things (tweets) via the web or mobile. The maximum length of a tweet is 140 characters, about the character length of a newspaper title. Twitter being an almost unlimited source used in text classification, there are many characteristics of twitter tweets. Meanwhile, in Indonesia, the number of Twitter users reached 14.05 million as of January 2021 [11].

C. Support Vector Machine

Supervised Learning Method is a learning method to find the relationship between input attributes and target/class attributes from the training data to be used as a model and can be used to predict the value of the target attribute. In the supervised learning method, the attribute already has a label and is then used as a model. The model is used for classification at the next test stage. In sentiment analysis, the supervised learning method is useful in determining the opinion of a product that is more likely to be positive or negative [12].

The Support Vector Machine (SVM) algorithm is a type of supervised learning method [13]. The general characteristics of SVM are summarized as follows: SVM is a linear classifier, pattern recognition is done by transforming the data in the input space to a higher-dimensional space (feature space) and optimization is carried out on the new vector space, implementing a Structural Risk Minimization strategy (SRM), basically only able to handle the classification of two classes, but has been developed for the classification of more than two classes with pattern recognition.

SVM technique aims to find the optimal hyperplane. Hyperplane that can divide the two classes with the farthest margin between classes. Margin is the distance between the hyperplame and the closest pattern of each class. The instance closest to this is called the support vector. In real-world problems, data sets are generally non-linearly separated. To solve this problem, you can use the kernel trick on space. In general, the kernel function is to convert the power set in the input space into a feature space with a higher dimension. By using the kernel trick, you only need to know the kernel function used to determine the support vector and you don't need to know the shape of the nonlinear function. In general, there are 3 types of kernel functions, namely kernel linear, kernel Gaussian / Radial Basis Function (RBF), dan kernel polynomial.

D. K-Fold Cross Validation

K-fold cross validation is a technique for validating datasets to find good accuracy [14]. This technique divides the dataset into k subsets. One of these subsets will be used as test data and the remaining k-1 subsets will be used to process training data. This process is carried out k times so that each subset will be the test data of the model. This process will get k performance scores from the learning process. All these performance values will be averaged and the value with the highest average will be selected as the model. K-fold cross validation has the advantage of being able to classify datasets more efficiently, but this method has a weakness in that the computational process used will be larger because it processes k times.

E. Confussion Matrix

Confusion matrix is a matrix that displays a visualization of the performance of the data classification algorithm in the matrix [15]. It compares the predicted classification to the actual classification. The confusion matrix table can be seen in table 2.1.

Table 1. Confusion Matrix

Actual	Prediction	
	Positive	Negative
Positive	True Positive	False Negative
Negative	False Positive	True Negative

From the confusion matrix table, the accuracy, precision, and recall values can be calculated. Accuracy value is a value that describes how accurate the method used in classifying is correctly from the entire existing data. The precision value describes the number of correctly classified positive category data divided by the total data classified as positive. The recall value shows what percentage of the positive category data is classified correctly. It can be seen in the following formula:

$$\text{Acuracy} = \frac{TP+TN}{TP+TN+FN+FP}\ x\ 100\%$$

*Corresponding Author.

$$\text{Precision} = \frac{TP}{TP+FP}\ x\ 100\%$$

$$\text{Recall} = \frac{TP}{TP+FN}\ x\ 100\%$$

III. Research Methods

In this study, a sentiment analysis system will be built on social media, namely Twitter. The use of Twitter as a data source is because Twitter uses more text in it and Twitter is widely used to exchange opinions or express opinions related to a topic. So that sentiment can be analyzed based on opinions or opinions from Twitter users. Data retrieval or data crawling uses the Python language with a workspace, namely Jupyter Notebook.

Data collection based on opinions related to a topic on Twitter. The topic that will be taken to discuss the sentiment analysis is "Covid Vaccine". This topic had become a hot discussion and controversy in the community because the vaccine was not yet clinically tested and there were side effects. After specifying keywords, in order to access data and collect data or crawling requires a developer twitter account. So a twitter developer account needs to be created in order to get the twitter API which will be used to collect data. After that, the data will be processed at the preprocessing and labeling stages. The processed data will then be analyzed using a support vector machine (SVM) algorithm with a linear kernel and a radial basis function (RBF). The flow of the research method can be seen in Figure below.

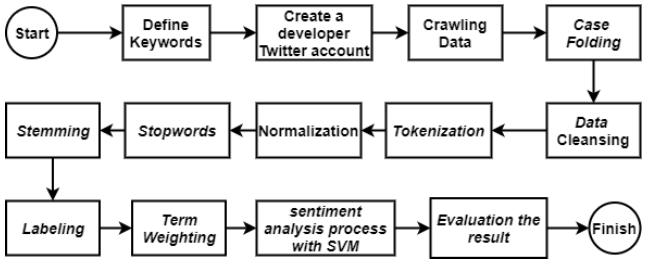

Figure 1. Research Methods.

IV. Result and Discussion

A. Data Collection

The retrieval of data from tweets is known as the crawling process. The crawling process is carried out in the python programming language using the API provided by twitter with the keyword in Bahasa is “Vaksin Covid-19”. Before the program is run, it is necessary to register with the Twitter developer first to get the API token. After registering, the tokens used are consumer key, consumer secret, access token and access token secret, which are APIs for access to Twitter data. In the crawling process, only use attributes as needed that will be used in the processing of sentiment analysis, namely id and text. Id is the identity of each tweet, and the text contains the tweet itself. The crawled data is shown in table 2.

Table 2. Crawling Data Result.

Id	*Tweet*
1411539632014060000	4 July - Number of COVID-19 vaccines administered to 3 July 'The one that's valid on RTM News' #RTM #RTM75 #TerusUnggul #BeritaRTM #COVID19Malaysia #PKP #LindungDiriLindungAll #PEMULIH https://t.co/0Yo1jbApRO

B. Pre-Processing

Before carrying out the sentiment analysis process on tweets, it is necessary to clean the data first, which aims to reduce words that have no effect on the results of data processing, so that the system can process accurately. The stages of pre-processing are as follows:

1. Case Folding

   The crawled text contains a variety of uppercase and lowercase letters. In case folding will change all letters to lowercase. The results of case folding are shown in table 3.

Table 3. Case Folding Result.

Before case folding	After case folding
4 July - Number of COVID-19 vaccines administered to 3 July 'The one that's valid on RTM News' #RTM #RTM75 #TerusUnggul #BeritaRTM #COVID19Malaysia #PKP #LindungDiriLindungAll #PEMULIH https://t.co/0Yo1jbApRO	4 july - the number of covid-19 vaccines given to 3 july 'valid on rtm news #rtm #rtm75 #terusunggul #beritartm #covid19malaysia #pkp #lindungdiridindungall #pemulih https://t.co/0yo1jbapro

2. Cleansing Data

   In cleansing the data will be deleted such as hashtags, numbers, tags, users, and other characters other than letters will be removed. This process aims to reduce random errors (noise) in the tweet data to be classified. The results of data cleansing are shown in Table 4.

Table 4. Cleansing Data Result.

After Case folding	Data Cleansing
4 july - the number of covid-19 vaccines given to 3 july 'valid on rtm news #rtm #rtm75 #terusunggul #beritartm #covid19malaysia #pkp #lindungdiridindungall #pemulih https://t.co/0yo1jbapro	july the number of valid covid vaccines so far on the news rtm

3. Tokenization

   At this stage it will break the paragraph or sentence in the tweet into smaller parts, namely words that stand alone. Tokenization results are shown in table 5.

Table 5. Tokenization Result

Data Cleansing	Tokenization
july the number of valid covid vaccines so far on the news rtm	['july', 'amount', 'administration', 'vaccine', 'covid', 'so far', 'july', 'which', 'valid', 'at', 'news', 'rtm']

4. Normalization

   In the normalization process, words originating from tweets will be corrected for spelling to conform to the KBBI. The results of normalization are shown in table 6.

Table 6. Normalization Result.

Tokenization	Normalization
['july', 'amount', 'administration', 'vaccine', 'covid', 'so far', 'july', 'which', 'valid', 'at', 'news', 'rtm']	['july', 'amount', 'administration', 'vaccine', 'covid', 'agree', 'july', 'which', 'valid', 'at', 'news', 'rtm']

5. Stopwords

This step aims to eliminate words that are considered meaningless such as conjunctions, prepositions, and others. The results of the stopwords are shown in Table 7.

TABLE 7. STOPWORDS RESULT.

Normalization	Stopwords
['july', 'amount', 'administration', 'vaccine', 'covid', 'agree', 'july', 'which', 'valid', 'at', 'news', 'rtm']	['july', 'gift', 'vaccine', 'covid', 'agreed', 'july', 'authentic', 'news']

6. Stemming

The stemming process is looking for basic words from the stopwords results. Each word will be changed to its root word, this change includes the removal of affixes and derivatives of the word. The stemming process uses the Sastrawi library. The results of stemming can be seen in table 8.

TABLE 8. STEMMING RESULT.

Stopwords	Stemming
['july', 'gift', 'vaccine', 'covid', 'agreed', 'july', 'authentic', 'news']	['july', 'berry', 'vaccine', 'covid', agreed, 'july', 'authentic', 'news'']

7. Merge Word and Translation into English

After the data goes through the preprocessing stage, it needs to be recombined. After that, the words were translated into English. This change was made because the data still used foreign words. So to make the word uniform, it needs to be changed into English, in addition to making the labeling process easier. The process of converting to English is carried out with the function of the google spreadsheet, namely "=GOOGLETRANSLATE("text","id","en")" where the word text is the cell that will be changed from Indonesian to English. Examples of translated data can be seen in Table 9 below.

TABLE 9. TRANSLATION RESULT.

Text Bahasa	Text English
july gave the covid vaccine agreed july is valid news	July give valid takat covid vaccine news

C. *Labelling and Sentiment algorithm SVM*

The labeling classification process is carried out using python with the TextBlob library to determine the polarization of each tweet data so that the tweet data will automatically get the polarization value. After getting the polarization value, the tweet data will be categorized into 3, namely positive, negative, and neutral. If a tweet with a polarization greater than 0 (>0) will be labeled positive, a tweet with a polarization less than 0 (< 0) will be labeled negative and a tweet with a polarization equal to 0 (= 0) will be labeled neutral. Examples of polarization results are shown in table 10.

TABLE 10. LABELLING RESULT.

Text English	Polarization	Label
July give valid takat covid vaccine news	0.05	positive
covid drug dock drink let sick vaccine vaccine vaccine	-0.714285714	negative
Siti Fadilah said Nidom prof tire advocacy team leader covid vaccine vaccination appointment	0	neutral

From the results of the labeling process, there were 4087 positive classifications, 1808 negative classifications, and 3314 neutral classifications. The results can be seen in the following figure.

```
In [9]: print("Netral :", (sum(df['Score']=='neutral')))
        print("positif :", (sum(df['Score']=='positive')))
        print("Negatif :", (sum(df['Score']=='negative')))

        Netral : 3314
        positif : 4087
        Negatif : 1808
```

Figure 3. Labeling Process Results.

The results of the labeling obtained will be visualized in the form of a word cloud. In making WordCloud using the WordCloud library and matplotlib.pyplot. The results of the visualization of all data which is a combination of positive, negative, and neutral labeling are shown in Figure 2.

Figure 2. WordCloud of All Tweet.

From the labeling classification process that has been carried out, the percentage of positive sentiment is 44.38%, the percentage of negative sentiment is 19.64%, and the percentage of neutral sentiment is 35.98%. Based on the percentage results obtained, it is known that the positive response to the covid vaccine is good because the number is greater than the negative response. However, there are still some people who are neutral about the existence of a covid vaccine.

Reported from the official website of the Ministry of Health's vaccine on 28 Oktober 2021 that 56 per 100 Indonesian population targets had received one dose (the total target of the vaccination target until the final stage). So people tend to vaccinate.

D. *Evaluation Classification*

Evaluation of the classification algorithm uses two methods, namely the confusion matrix and k-fold cross validation.

1. Evaluation Using Confusion matrix

The evaluation is done by knowing the accuracy value on the SVM linear kernel and RBF. The method used is the confusion matrix. After classifying the SVM algorithm with a linear kernel, the accuracy value is 79.15% with the confusion matrix table shown in table 11.

TABLE 11. CONFUSION MATRIX LINEAR KERNEL.

		Actual		
		Negative	Neutral	Positive
Predict	Negative	258	49	55
	Neutral	49	568	46
	Positive	74	111	632

To prove the calculation of the accuracy value obtained from the confusion matrix table with the formula:

$$\text{Accuracy} = \frac{Tpos + TNet + TN}{Tpos + FPos + TNet + FNet + TN + FN}$$

$$\text{Accuracy} = \frac{632 + 568 + 258}{632 + (55+46) + 568 + (49 + 111) + 258 + (49+74)}$$

$$\text{Accuracy} = \frac{1458}{1842} = 0{,}7915309 = 79{,}15\%$$

Also obtained a precision value of 77.3198% or 77.32%. Furthermore, the recall value obtained in the linear SVM classification is 78.09%. The evaluation of the classification of the SVM algorithm with the RBF kernel obtained an accuracy value of 84.25% with the confusion matrix table shown in table 12.

TABLE 12. CONFUSION MATRIX RBF KERNEL.

		Actual		
		Negative	Neutral	Positive
Predict	Negative	248	54	60
	Neutral	16	629	18
	Positive	42	100	675

To prove the calculation of the accuracy value obtained from the confusion matrix table with the formula:

$$\text{Accuracy} = \frac{Tpos + TNet + TNeg}{Tpos + FPos + TNet + FNet + TNeg + FNeg}$$

$$\text{Accuracy} = \frac{675 + 629 + 248}{675 + 142 + 629 + 34 + 248 + 308}$$

$$\text{Accuracy} = \frac{1552}{1842} = 0{,}84256 = 84{,}25\%$$

The results of the evaluation carried out on the classification of the SVM kernel RBF algorithm can be shown in the following figure.

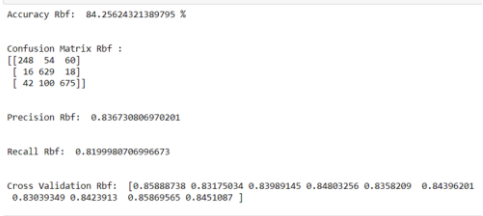

Figure 4. SVM RBF Classification Evaluation.

The precision value of the SVM RBF classification was also obtained at 83.67%. Furthermore, the recall value obtained in the SVM RBF classification is 81.99%. The comparison of the evaluation results of linear SVM classification with SVM RBF can be seen in table 13.

TABLE 13. COMPARISON OF EVALUATION RESULT.

Kernel	Accuracy	Precision	Recall
Linear	79,15%	77,31%	78,09%
RBF	84,25%	83,67%	81,99%

The results of this study obtained less than optimal results because the amount of labeling classification data from the results of positive, negative, and neutral sentiments was not balanced or not the same amount of data. In addition, there are still foreign words from regions in Indonesia. So that in the preprocessing process these foreign words are also processed. Then in the process of translating words into English these words cannot be translated. Thus affecting the results of the accuracy value.

2. Evaluation Using K-Fold Cross Validation

At this stage, an evaluation of the sentiment classification of the SVM kernel linear algorithm and RBF is carried out using the k-fold cross validation method. The fold size used is ten (k=10) because 10-fold cross validation is one of the recommended k-fold cross validations for selecting the best model because it tends to provide less biased estimates. This is to determine the composition of the kernel that has the most optimum performance. The results of the 10-fold cross validation test can be seen in table 14.

TABLE 14. CROSS VALIDATION SVM.

Fold	SVM (%)	
	Linear	RBF
1	80.18	85.88
2	76.25	83.17
3	77.74	83.98
4	79.64	84.80
5	78.42	83.58
6	77.06	84.39
7	77.20	83.03
8	78.26	84.23
9	78.53	85.86
10	77.98	84.51

From table 15 the form of the graph is shown in Figure 15.

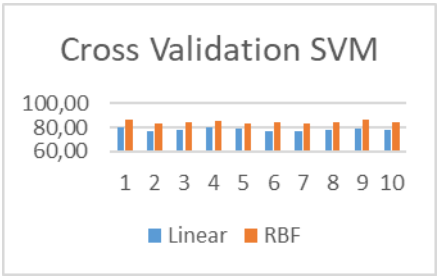

Figure 5. Cross Validation Chart

From the results of the cross-validation test, it is known that the SVM with the linear kernel and the RBF kernel has the optimum result or score at k=1, with an accuracy of 80.18% for the linear kernel and 85.88% for the RBF kernel.

The results of the performance comparison and validation test with cross validation show that the SVM method with the RBF kernel has the best results compared to the linear kernel.

V. CONCLUSIONS AND SUGGESTION

A. *Conclusion*

The conclusions that can be drawn from the results of sentiment analysis research with the SVM algorithm are as follows:

1. Sentiment analysis is carried out using the crawling method for data collection, then preprocessing or data processing is carried out. After the data is processed, it is then labeled based on positive, negative, and neutral sentiments. Furthermore, classification is carried out using the SVM algorithm with linear kernel and RBF. After classification, positive sentiment is 43.5%, negative sentiment is 19.1%, and neutral sentiment is 37.4%. It can be concluded that with the existence of this COVID-19 vaccine program, there is a greater positive response in the community.
2. Testing the performance of sentiment classification about Vaccine-19 with the SVM algorithm, namely the confusion matrix and cross validation. In the confusion matrix, the accuracy value for the linear kernel is 79.15%, the precision value is 77.31%, and the recall value is 78.09%. While the RBF kernel has an accuracy of 84.25%, a precision value of 83.67%, and a recall value of 81.99%. In cross validation, the optimum value is obtained at k=1 with an accuracy value of 80.18% for the linear kernel and 85.88% for the RBF kernel. So that the RBF kernel has a higher accuracy than the linear kernel. The results of the classification evaluation are less than optimal because there are still Indonesian regional languages that are also processed. The results of the classification evaluation are not optimal because in the labeling classification process the results for positive, negative, and neutral labels have different results that affect the accuracy of the algorithm.

B. *Suggestion*

The suggestions that can be given by the author are:

1. For preprocessing at the normalization stage, a more complete regional language dictionary is needed because there are still tweet users about Vaccine-19 who use abbreviations or regional languages in Indonesia.
2. To maximize the accuracy of each kernel, the amount of data from the labeling classification process for each sentiment for Vaccine-19 should be balanced.
3. Further sentiment analysis about Vaccine-19 can be done by comparing other SVM kernels or with other algorithms such as Naïve Bayes, or Random Forest.

ACKNOWLEDGEMENT

Thanks to Binus Online Learning Computer Science for guiding and facilitating the author in completing the research.

REFERENCES

[1] J. Serrano-Guerrero, J. A. Olivas, F. P. Romero, and E. Herrera-Viedma, "Sentiment analysis: A review and comparative analysis of web services," *Inf. Sci. (Ny).*, vol. 311, pp. 18–38, 2015, doi: https://doi.org/10.1016/j.ins.2015.03.040.

[2] K. Rudra, A. Sharma, N. Ganguly, and M. Imran, "Classifying Information from Microblogs during Epidemics," in *Proceedings of the 2017 International Conference on Digital Health*, 2017, pp. 104–108, doi: 10.1145/3079452.3079491.

[3] S. M. Mohammad, "9 - Sentiment Analysis: Detecting Valence, Emotions, and Other Affectual States from Text," in *Emotion Measurement*, H. L. Meiselman, Ed. Woodhead Publishing, 2016, pp. 201–237.

[4] P. Pemerintah, "KEPUTUSAN MENTERI KESEHATAN REPUBLIK INDONESIA NOMOR HK.01.07/MENKES/4638/2021," 2021.

[5] D. Darwis, E. S. Pratiwi, and A. F. O. Pasaribu, "Penerapan Algoritma SVM untuk Analisis Sentimen Pada Data Twitter Komisi Pemberantasan KorupsiRepublik indonesia," *J. Ilm. Edutic*, vol. 7, no. 1, pp. 1–11, 2020, [Online]. Available: https://journal.trunojoyo.ac.id/edutic/article/view/8779.

[6] M. A. Rizaty, "Siapa Tokoh Terpopuler di Twitter pada 2021?," *databoks*, 2021. https://databoks.katadata.co.id/datapublish/2021/07/09/siapa-tokoh-terpopuler-di-twitter-pada-2021.

[7] F. F. Irfani, "Analisis Sentimen Review Aplikasi Ruangguru Menggunakan Algoritma Support Vector Machine," *J. Bisnis,Manajemen dan Inform.*, pp. 258–266, 2020.

[8] B. W. Sari and F. F. Haranto, "Implementasi Support Vector Machine Untuk Analisis Sentimen Pengguna Twitter Terhadap Pelayanan Telkom dan Biznet," *J. PILAR Nusa Mandiri*, vol. 15, no. 2, pp. 171–176, 2019.

[9] N. Sinha, P. Singh, M. Gupta, and P. Singh, "Robotics at workplace: An integrated Twitter analytics – SEM based approach for behavioral intention to accept," *Int. J. Inf. Manage.*, vol. 55, p. 102210, 2020, doi: https://doi.org/10.1016/j.ijinfomgt.2020.102210.

[10] M. Vinodkumar Sadhuram and A. Soni, "Natural Language Processing based New Approach to Design Factoid Question Answering System," in *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, Jul. 2020, pp. 276–281, doi: 10.1109/ICIRCA48905.2020.9182972.

[11] Statista, "Countries with the most Twitter users 2021," 2021. https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/.

[12] J. Brownlee, "Supervised and Unsupervised Machine Learning Algorithms," in *Discover How Machine Learning Algorithms Work*, 2021.

[13] I. Steinwart and A. Christmann, *Support Vector Machines*. Springer Publishing Company, Incorporated, 2008.

[14] A. U. Khasanah, "Educational Data Mining Techniques Approach to Predict Student ' s Performance," vol. 9, no. 2, pp. 115–118, 2019, doi: 10.18178/ijiet.2019.9.2.1184.

[15] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, vol. 49, no. 06. The Morgan Kaufmann Series in Data Management Systems Morgan Kaufmann, 2011.

[16] vaksin.kemkes.go.id. (2021, 28 Oktober). Vaksinasi COVID-19 Nasional. Accessed on 28 Oktober 2021. from https://vaksin.kemkes.go.id/#/vaccines

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/ISRITI54043.2021.9702774

THE COMMITTEE

STEERING COMMITTEE

Chuan-Ming Liu	(National Taipei University of Technology, Taiwan)
Totok Suprawoto	(STMIK AKAKOM Yogyakarta, Indonesia)
Setyawan Widyarto	(Universiti Selangor, Malaysia)

ORGANIZING COMMITTEE

General Chair	
Bambang Purnomosidi Dwi Putranto	(STMIK AKAKOM Yogyakarta, Indonesia)
Co-Chair	
Maria Mediatrix	(STMIK AKAKOM Yogyakarta, Indonesia)
Secretary	
Sumiyatun	(STMIK AKAKOM Yogyakarta, Indonesia)
Treasury	
Muhammad Agung Nugroho	(STMIK AKAKOM Yogyakarta, Indonesia)
Program Chair	
Widyastuti Andriyani	(STMIK AKAKOM Yogyakarta, Indonesia)

TECHNICAL COMMITTEE

Domy Kristomo	(STMIK AKAKOM Yogyakarta, Indonesia)
Robby Cokro Buwono	(STMIK AKAKOM Yogyakarta, Indonesia)
Danny Kriestanto	(STMIK AKAKOM Yogyakarta, Indonesia)
Luthfan Hadi Pramono	(STMIK AKAKOM Yogyakarta, Indonesia)
Cosmas Haryawan	(STMIK AKAKOM Yogyakarta, Indonesia)

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/ISRITI54043.2021.9702846

AUTHOR INDEX

Sentiment Analysis on Social Media (Twitter) about Vaccine-19 Using Support Vector Machine Algorithm

Agus Sulistyono
Computer Science Department, BINUS Online Learning
Bina Nusantara University
Jakarta, Indonesia, 11480
Agus.sulistyono@binus.ac.id

Sri Mulyani
Computer Science Department, BINUS Online Learning
Bina Nusantara University
Jakarta, Indonesia, 11480
Sri.mulyani@binus.ac.id

Emny Harna Yossy*
Computer Science Department, BINUS Online Learning
Bina Nusantara University
Jakarta, Indonesia, 11480
emny.yossy@binus.ac.id

Rakhmi Khalida
Computer Science Department
Gunadarma University
Depok, Indonesia, 16424
Rakhmikhalida7@gmail.com

***Abstract*—Currently the world is experiencing a Corona Virus Disease (Covid-19) pandemic which attacks the respiratory tract and spreads very quickly to various countries including Indonesia, so the World Health Organization (WHO) has declared Covid-19 as a pandemic. To overcome this pandemic, experts in the medical field also intervened by making vaccinations to strengthen human immunity against the Covid virus. This sentiment analysis was carried out to see opinions on the object, namely the existence of a Covid-19 vaccine. Data collection by crawling data with the keyword 'Covid Vaccine'. The method that will be used is the Support Vector Machine (SVM). The analysis was carried out by comparing the classification accuracy values of the two SVM kernel functions, namely linear and Radial Basic Function (RBF). The results of the study obtained positive sentiment of 43.5%, negative of 19.1%, and neutral of 37.4%. Then the evaluation of the system using the confusion matrix obtained an accuracy value for the linear kernel of 79.15%, a precision value of 77.31%, and a recall value of 78.09%. While the RBF kernel has an accuracy of 84.25%, a precision value of 83.67%, and a recall value of 81.99%. While the cross validation obtained the optimum value at k = 1 with an accuracy value of 80.18% for the linear kernel and 85.88% for the RBF kernel. So the RBF kernel has a higher accuracy than the linear kernel.**

***Keywords*—Covid-19, Vaccine, Support Vector Machine, Linear, Radial Basis Function.**

I. INTRODUCTION

Advances in technology, one of which is the internet, makes it easy for us to exchange information or express thoughts, opinions, and respond to events through online media [1]. Online media to exchange information with others can be called social media. One type of social media that is widely used is Twitter. Twitter is a microblogging service where the flow of interaction is faster than blogs [2]. Microblogging can be optimized as a channel for fast interaction, so that concise and important information can be known by other users. This makes Twitter often used as a medium to provide experiences, share opinions, and respond to events. This response or experience can be classified to determine sentiment on a topic or can be called sentiment analysis [3]. Therefore, sentiment analysis is one solution to the problem of classifying opinions or reviews into positive or negative opinions. The opinion that will be reviewed in this study is the Covid Vaccine in Bahasa.

Currently the world is experiencing a Covid-19 pandemic (Corona Virus Disease-19) where this virus attacks the respiratory tract and spreads very quickly. In addition to the implementation of strict health protocols, other effective interventions are needed to break the chain of disease transmission, namely through vaccination efforts. Covid-19 vaccination aims to reduce the transmission / transmission of Covid-19, reduce morbidity and mortality due to Covid-19, achieve group immunity in the community (herd immunity) and protect the community from Covid-19 in order to remain socially and economically productive [4].

Research conducted is the application of the SVM algorithm for sentiment analysis on the twitter data of the Corruption Eradication Commission of the Republic of Indonesia [5]. Classification is divided into negative, neutral, or positive responses. The results of the testing and evaluation of the research are the accuracy of the test results by 82% and precision testing by 90%, as well as recall by 88% and f1-score by 89%. Another sesearch conducted is applying the SVM method in the classification of public figure tweet sentiment [6]. Classification is done by using the RBF kernel and polynomials on the SVM method. In this study to see the level of accuracy produced. From these results, it was found that the RBF kernel provides a better accuracy rate than the polynomial kernel with the accuracy for the RBF kernel on unigram features of 72.5% and the accuracy of the polynomial kernel only 72%.

In addition, the SVM algorithm for sentiment analysis reviews ruang guru applications [7]. Classification is done to see the positive or negative response. The SVM kernel used in this study are linear, RBF, and polynomial. From the results of this study, it was found that a high accuracy value was in the range of 90% with a linear kernel giving a better accuracy value than RBF and polynomials. A Study on the Implementation of Support Vector Machines for Sentiment Analysis of Twitter Users towards Telkom and Biznet Services [8]. With the aim of analyzing the sentiments of twitter users towards Telkom and Biznet services. Tests using the Confusion Matrix and K-Fold Cross validation are intended to share training information and testing information. K-Fold Cross validation and Confusion Matrix share the results of an accuracy value of 79.6%, precision 76.5%, recall 72.8%, and F1-score 74.6% for Telkom, and

*Corresponding Author

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/ISRITI54043.2021.9702775

accuracy 83.2%, precision 78, 8% recall 71, 6%, and F1-score 75% for biznet.

Based on the description above, this research is about sentiment analysis on social media (Twitter) to classify Twitter user responses to the keyword in Bahasa is "Vaksin Covid "into positive, negative or neutral responses. In this study the classification method used is the Support Vector Machine (SVM).

II. Literature Review

A. Sentiment Analysis

Sentiment analysis is a branch of text mining research. Specifically, the purpose of text mining can be divided into two, namely text data categorization and text clustering. In categorization, text mining is used as a tool to find categories that match the specified class (supervised learning), while grouping in text mining functions as a tool to group text data based on similar characteristics, and clustering can be used to label unknown classes [9].

Text mining is a process of mining data in the form of text where the data source is usually obtained from documents and the goal is to find words that can represent the contents of the document so that an analysis of the connectivity between documents can be carried out. Data mining (Pattern Discovery) is the process of seeking knowledge or patterns that are interesting/valuable. Evaluation is the interpretation of patterns found [10]. The purpose of text mining is to extract useful information from data sources. So, the data source used in text mining is a collection of documents that have an unstructured format through the identification and exploration of interesting patterns. The stages of text mining are tokenization, lower case, removing tad abaca, stemming, and filtering.

B. Twitter

Twitter is a microblogging service that was officially released on July 13, 2006. Twitter's main activity is posting short things (tweets) via the web or mobile. The maximum length of a tweet is 140 characters, about the character length of a newspaper title. Twitter being an almost unlimited source used in text classification, there are many characteristics of twitter tweets. Meanwhile, in Indonesia, the number of Twitter users reached 14.05 million as of January 2021 [11].

C. Support Vector Machine

Supervised Learning Method is a learning method to find the relationship between input attributes and target/class attributes from the training data to be used as a model and can be used to predict the value of the target attribute. In the supervised learning method, the attribute already has a label and is then used as a model. The model is used for classification at the next test stage. In sentiment analysis, the supervised learning method is useful in determining the opinion of a product that is more likely to be positive or negative [12].

The Support Vector Machine (SVM) algorithm is a type of supervised learning method [13]. The general characteristics of SVM are summarized as follows: SVM is a linear classifier, pattern recognition is done by transforming the data in the input space to a higher-dimensional space (feature space) and optimization is carried out on the new vector space, implementing a Structural Risk Minimization strategy (SRM), basically only able to handle the classification of two classes, but has been developed for the classification of more than two classes with pattern recognition.

SVM technique aims to find the optimal hyperplane. Hyperplane that can divide the two classes with the farthest margin between classes. Margin is the distance between the hyperplame and the closest pattern of each class. The instance closest to this is called the support vector. In real-world problems, data sets are generally non-linearly separated. To solve this problem, you can use the kernel trick on space. In general, the kernel function is to convert the power set in the input space into a feature space with a higher dimension. By using the kernel trick, you only need to know the kernel function used to determine the support vector and you don't need to know the shape of the nonlinear function. In general, there are 3 types of kernel functions, namely kernel linear, kernel Gaussian / Radial Basis Function (RBF), dan kernel polynomial.

D. K-Fold Cross Validation

K-fold cross validation is a technique for validating datasets to find good accuracy [14]. This technique divides the dataset into k subsets. One of these subsets will be used as test data and the remaining k-1 subsets will be used to process training data. This process is carried out k times so that each subset will be the test data of the model. This process will get k performance scores from the learning process. All these performance values will be averaged and the value with the highest average will be selected as the model. K-fold cross validation has the advantage of being able to classify datasets more efficiently, but this method has a weakness in that the computational process used will be larger because it processes k times.

E. Confussion Matrix

Confusion matrix is a matrix that displays a visualization of the performance of the data classification algorithm in the matrix [15]. It compares the predicted classification to the actual classification. The confusion matrix table can be seen in table 2.1.

TABLE 1. CONFUSION MATRIX

Actual	Prediction	
	Positive	Negative
Positive	True Positive	False Negative
Negative	False Positive	True Negative

From the confusion matrix table, the accuracy, precision, and recall values can be calculated. Accuracy value is a value that describes how accurate the method used in classifying is correctly from the entire existing data. The precision value describes the number of correctly classified positive category data divided by the total data classified as positive. The recall value shows what percentage of the positive category data is classified correctly. It can be seen in the following formula:

$$\text{Acuracy} = \frac{TP+TN}{TP+TN+FN+FP} \; x \; 100\%$$

*Corresponding Author.

$$\text{Precision} = \frac{TP}{TP+FP}\ x\ 100\%$$

$$\text{Recall} = \frac{TP}{TP+FN}\ x\ 100\%$$

III. Research Methods

In this study, a sentiment analysis system will be built on social media, namely Twitter. The use of Twitter as a data source is because Twitter uses more text in it and Twitter is widely used to exchange opinions or express opinions related to a topic. So that sentiment can be analyzed based on opinions or opinions from Twitter users. Data retrieval or data crawling uses the Python language with a workspace, namely Jupyter Notebook.

Data collection based on opinions related to a topic on Twitter. The topic that will be taken to discuss the sentiment analysis is "Covid Vaccine". This topic had become a hot discussion and controversy in the community because the vaccine was not yet clinically tested and there were side effects. After specifying keywords, in order to access data and collect data or crawling requires a developer twitter account. So a twitter developer account needs to be created in order to get the twitter API which will be used to collect data. After that, the data will be processed at the preprocessing and labeling stages. The processed data will then be analyzed using a support vector machine (SVM) algorithm with a linear kernel and a radial basis function (RBF). The flow of the research method can be seen in Figure below.

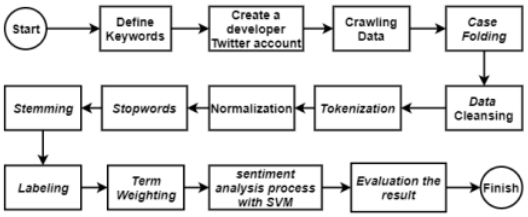

Figure 1. Research Methods.

IV. Result and Discussion

A. Data Collection

The retrieval of data from tweets is known as the crawling process. The crawling process is carried out in the python programming language using the API provided by twitter with the keyword in Bahasa is "Vaksin Covid-19". Before the program is run, it is necessary to register with the Twitter developer first to get the API token. After registering, the tokens used are consumer key, consumer secret, access token and access token secret, which are APIs for access to Twitter data. In the crawling process, only use attributes as needed that will be used in the processing of sentiment analysis, namely id and text. Id is the identity of each tweet, and the text contains the tweet itself. The crawled data is shown in table 2.

Table 2. Crawling Data Result.

Id	*Tweet*
1411539632014060000	4 July - Number of COVID-19 vaccines administered to 3 July 'The one that's valid on RTM News' #RTM #RTM75 #TerusUnggul #BeritaRTM #COVID19Malaysia #PKP #LindungDiriLindungAll #PEMULIH https://t.co/0Yo1jbApRO

B. Pre-Processing

Before carrying out the sentiment analysis process on tweets, it is necessary to clean the data first, which aims to reduce words that have no effect on the results of data processing, so that the system can process accurately. The stages of pre-processing are as follows:

1. Case Folding

   The crawled text contains a variety of uppercase and lowercase letters. In case folding will change all letters to lowercase. The results of case folding are shown in table 3.

Table 3. Case Folding Result.

Before case folding	**After case folding**
4 July - Number of COVID-19 vaccines administered to 3 July 'The one that's valid on RTM News' #RTM #RTM75 #TerusUnggul #BeritaRTM #COVID19Malaysia #PKP #LindungDiriLindungAll #PEMULIH https://t.co/0Yo1jbApRO	4 july - the number of covid-19 vaccines given to 3 july 'valid on rtm news #rtm #rtm75 #terusunggul #beritartm #covid19malaysia #pkp #lindungdiridindungall #pemulih https://t.co/0yo1jbapro

2. Cleansing Data

   In cleansing the data will be deleted such as hashtags, numbers, tags, users, and other characters other than letters will be removed. This process aims to reduce random errors (noise) in the tweet data to be classified. The results of data cleansing are shown in Table 4.

Table 4. Cleansing Data Result.

After Case folding	Data Cleansing
4 july - the number of covid-19 vaccines given to 3 july 'valid on rtm news #rtm #rtm75 #terusunggul #beritartm #covid19malaysia #pkp #lindungdiridindungall #pemulih https://t.co/0yo1jbapro	july the number of valid covid vaccines so far on the news rtm

3. Tokenization

   At this stage it will break the paragraph or sentence in the tweet into smaller parts, namely words that stand alone. Tokenization results are shown in table 5.

Table 5. Tokenization Result

Data Cleansing	Tokenization
july the number of valid covid vaccines so far on the news rtm	['july', 'amount', 'administration', 'vaccine', 'covid', 'so far', 'july', 'which', 'valid', 'at', 'news', 'rtm']

4. Normalization

   In the normalization process, words originating from tweets will be corrected for spelling to conform to the KBBI. The results of normalization are shown in table 6.

Table 6. Normalization Result.

Tokenization	Normalization
['july', 'amount', 'administration', 'vaccine', 'covid', 'so far', 'july', 'which', 'valid', 'at', 'news', 'rtm']	['july', 'amount', 'administration', 'vaccine', 'covid', 'agree', 'july', 'which', 'valid', 'at', 'news', 'rtm']

5. Stopwords

This step aims to eliminate words that are considered meaningless such as conjunctions, prepositions, and others. The results of the stopwords are shown in Table 7.

TABLE 7. STOPWORDS RESULT.

Normalization	Stopwords
['july', 'amount', 'administration', 'vaccine', 'covid', 'agree', 'july', 'which', 'valid', 'at', 'news', 'rtm']	['july', 'gift', 'vaccine', 'covid', 'agreed', 'july', 'authentic', 'news']

6. Stemming

The stemming process is looking for basic words from the stopwords results. Each word will be changed to its root word, this change includes the removal of affixes and derivatives of the word. The stemming process uses the Sastrawi library. The results of stemming can be seen in table 8.

TABLE 8. STEMMING RESULT.

Stopwords	Stemming
['july', 'gift', 'vaccine', 'covid', 'agreed', 'july', 'authentic', 'news']	['july', 'berry', 'vaccine', 'covid', agreed, 'july', 'authentic', 'news"]

7. Merge Word and Translation into English

After the data goes through the preprocessing stage, it needs to be recombined. After that, the words were translated into English. This change was made because the data still used foreign words. So to make the word uniform, it needs to be changed into English, in addition to making the labeling process easier. The process of converting to English is carried out with the function of the google spreadsheet, namely "=GOOGLETRANSLATE("text","id","en")" where the word text is the cell that will be changed from Indonesian to English. Examples of translated data can be seen in Table 9 below.

TABLE 9. TRANSLATION RESULT.

Text Bahasa	Text English
july gave the covid vaccine agreed july is valid news	July give valid takat covid vaccine news

C. Labelling and Sentiment algorithm SVM

The labeling classification process is carried out using python with the TextBlob library to determine the polarization of each tweet data so that the tweet data will automatically get the polarization value. After getting the polarization value, the tweet data will be categorized into 3, namely positive, negative, and neutral. If a tweet with a polarization greater than 0 (>0) will be labeled positive, a tweet with a polarization less than 0 (< 0) will be labeled negative and a tweet with a polarization equal to 0 (= 0) will be labeled neutral. Examples of polarization results are shown in table 10.

TABLE 10. LABELLING RESULT.

Text English	Polarization	Label
July give valid takat covid vaccine news	0.05	positive
covid drug dock drink let sick vaccine vaccine vaccine	-0.714285714	negative
Siti Fadilah said Nidom prof tire advocacy team leader covid vaccine vaccination appointment	0	neutral

From the results of the labeling process, there were 4087 positive classifications, 1808 negative classifications, and 3314 neutral classifications. The results can be seen in the following figure.

```
In [9]: print("Netral :", (sum(df['Score']=='neutral')))
        print("positif :", (sum(df['Score']=='positive')))
        print("Negatif :", (sum(df['Score']=='negative')))

        Netral : 3314
        positif : 4087
        Negatif : 1808
```

Figure 3. Labeling Process Results.

The results of the labeling obtained will be visualized in the form of a word cloud. In making WordCloud using the WordCloud library and matplotlib.pyplot. The results of the visualization of all data which is a combination of positive, negative, and neutral labeling are shown in Figure 2.

Figure 2. WordCloud of All Tweet.

From the labeling classification process that has been carried out, the percentage of positive sentiment is 44.38%, the percentage of negative sentiment is 19.64%, and the percentage of neutral sentiment is 35.98%. Based on the percentage results obtained, it is known that the positive response to the covid vaccine is good because the number is greater than the negative response. However, there are still some people who are neutral about the existence of a covid vaccine.

Reported from the official website of the Ministry of Health's vaccine on 28 Oktober 2021 that 56 per 100 Indonesian population targets had received one dose (the total target of the vaccination target until the final stage). So people tend to vaccinate.

D. Evaluation Classification

Evaluation of the classification algorithm uses two methods, namely the confusion matrix and k-fold cross validation.

1. Evaluation Using Confusion matrix

The evaluation is done by knowing the accuracy value on the SVM linear kernel and RBF. The method used is the confusion matrix. After classifying the SVM algorithm with a linear kernel, the accuracy value is 79.15% with the confusion matrix table shown in table 11.

TABLE 11. CONFUSION MATRIX LINEAR KERNEL.

		Actual		
		Negative	Neutral	Positive
Predict	Negative	258	49	55
	Neutral	49	568	46
	Positive	74	111	632

To prove the calculation of the accuracy value obtained from the confusion matrix table with the formula:

$$\text{Accuracy} = \frac{Tpos + TNet + TN}{Tpos + FPos + TNet + FNet + TN + FN}$$

$$\text{Accuracy} = \frac{632 + 568 + 258}{632 + (55+46) + 568 + (49 + 111) + 258 + (49+74)}$$

$$\text{Accuracy} = \frac{1458}{1842} = 0{,}7915309 = 79{,}15\%$$

Also obtained a precision value of 77.3198% or 77.32%. Furthermore, the recall value obtained in the linear SVM classification is 78.09%. The evaluation of the classification of the SVM algorithm with the RBF kernel obtained an accuracy value of 84.25% with the confusion matrix table shown in table 12.

TABLE 12. CONFUSION MATRIX RBF KERNEL.

		Actual		
		Negative	Neutral	Positive
Predict	Negative	248	54	60
	Neutral	16	629	18
	Positive	42	100	675

To prove the calculation of the accuracy value obtained from the confusion matrix table with the formula:

$$\text{Accuracy} = \frac{Tpos + TNet + TNeg}{Tpos + FPos + TNet + FNet + TNeg + FNeg}$$

$$\text{Accuracy} = \frac{675 + 629 + 248}{675 + 142 + 629 + 34 + 248 + 308}$$

$$\text{Accuracy} = \frac{1552}{1842} = 0{,}84256 = 84{,}25\%$$

The results of the evaluation carried out on the classification of the SVM kernel RBF algorithm can be shown in the following figure.

Figure 4. SVM RBF Classification Evaluation.

The precision value of the SVM RBF classification was also obtained at 83.67%. Furthermore, the recall value obtained in the SVM RBF classification is 81.99%. The comparison of the evaluation results of linear SVM classification with SVM RBF can be seen in table 13.

TABLE 13. COMPARISON OF EVALUATION RESULT.

Kernel	Accuracy	Precision	Recall
Linear	79,15%	77,31%	78,09%
RBF	84,25%	83,67%	81,99%

The results of this study obtained less than optimal results because the amount of labeling classification data from the results of positive, negative, and neutral sentiments was not balanced or not the same amount of data. In addition, there are still foreign words from regions in Indonesia. So that in the preprocessing process these foreign words are also processed. Then in the process of translating words into English these words cannot be translated. Thus affecting the results of the accuracy value.

2. Evaluation Using K-Fold Cross Validation

At this stage, an evaluation of the sentiment classification of the SVM kernel linear algorithm and RBF is carried out using the k-fold cross validation method. The fold size used is ten (k=10) because 10-fold cross validation is one of the recommended k-fold cross validations for selecting the best model because it tends to provide less biased estimates. This is to determine the composition of the kernel that has the most optimum performance. The results of the 10-fold cross validation test can be seen in table 14.

TABLE 14. CROSS VALIDATION SVM.

Fold	SVM (%)	
	Linear	RBF
1	80.18	85.88
2	76.25	83.17
3	77.74	83.98
4	79.64	84.80
5	78.42	83.58
6	77.06	84.39
7	77.20	83.03
8	78.26	84.23
9	78.53	85.86
10	77.98	84.51

From table 15 the form of the graph is shown in Figure 15.

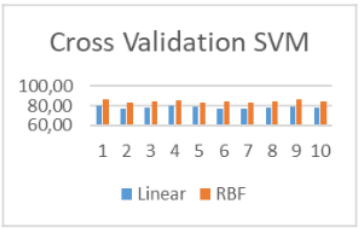

Figure 5. Cross Validation Chart

From the results of the cross-validation test, it is known that the SVM with the linear kernel and the RBF kernel has the optimum result or score at k=1, with an accuracy of 80.18% for the linear kernel and 85.88% for the RBF kernel.

The results of the performance comparison and validation test with cross validation show that the SVM method with the RBF kernel has the best results compared to the linear kernel.

V. CONCLUSIONS AND SUGGESTION

A. *Conclusion*

The conclusions that can be drawn from the results of sentiment analysis research with the SVM algorithm are as follows:

1. Sentiment analysis is carried out using the crawling method for data collection, then preprocessing or data processing is carried out. After the data is processed, it is then labeled based on positive, negative, and neutral sentiments. Furthermore, classification is carried out using the SVM algorithm with linear kernel and RBF. After classification, positive sentiment is 43.5%, negative sentiment is 19.1%, and neutral sentiment is 37.4%. It can be concluded that with the existence of this COVID-19 vaccine program, there is a greater positive response in the community.
2. Testing the performance of sentiment classification about Vaccine-19 with the SVM algorithm, namely the confusion matrix and cross validation. In the confusion matrix, the accuracy value for the linear kernel is 79.15%, the precision value is 77.31%, and the recall value is 78.09%. While the RBF kernel has an accuracy of 84.25%, a precision value of 83.67%, and a recall value of 81.99%. In cross validation, the optimum value is obtained at k=1 with an accuracy value of 80.18% for the linear kernel and 85.88% for the RBF kernel. So that the RBF kernel has a higher accuracy than the linear kernel. The results of the classification evaluation are less than optimal because there are still Indonesian regional languages that are also processed. The results of the classification evaluation are not optimal because in the labeling classification process the results for positive, negative, and neutral labels have different results that affect the accuracy of the algorithm.

B. *Suggestion*

The suggestions that can be given by the author are:

1. For preprocessing at the normalization stage, a more complete regional language dictionary is needed because there are still tweet users about Vaccine-19 who use abbreviations or regional languages in Indonesia.
2. To maximize the accuracy of each kernel, the amount of data from the labeling classification process for each sentiment for Vaccine-19 should be balanced.
3. Further sentiment analysis about Vaccine-19 can be done by comparing other SVM kernels or with other algorithms such as Naïve Bayes, or Random Forest.

ACKNOWLEDGEMENT

Thanks to Binus Online Learning Computer Science for guiding and facilitating the author in completing the research.

REFERENCES

[1] J. Serrano-Guerrero, J. A. Olivas, F. P. Romero, and E. Herrera-Viedma, "Sentiment analysis: A review and comparative analysis of web services," *Inf. Sci. (Ny).*, vol. 311, pp. 18–38, 2015, doi: https://doi.org/10.1016/j.ins.2015.03.040.

[2] K. Rudra, A. Sharma, N. Ganguly, and M. Imran, "Classifying Information from Microblogs during Epidemics," in *Proceedings of the 2017 International Conference on Digital Health*, 2017, pp. 104–108, doi: 10.1145/3079452.3079491.

[3] S. M. Mohammad, "9 - Sentiment Analysis: Detecting Valence, Emotions, and Other Affectual States from Text," in *Emotion Measurement*, H. L. Meiselman, Ed. Woodhead Publishing, 2016, pp. 201–237.

[4] P. Pemerintah, "KEPUTUSAN MENTERI KESEHATAN REPUBLIK INDONESIA NOMOR HK.01.07/MENKES/4638/2021," 2021.

[5] D. Darwis, E. S. Pratiwi, and A. F. O. Pasaribu, "Penerapan Algoritma SVM untuk Analisis Sentimen Pada Data Twitter Komisi Pemberantasan KorupsiRepublik indonesia," *J. Ilm. Edutic*, vol. 7, no. 1, pp. 1–11, 2020, [Online]. Available: https://journal.trunojoyo.ac.id/edutic/article/view/8779.

[6] M. A. Rizaty, "Siapa Tokoh Terpopuler di Twitter pada 2021?," *databoks*, 2021. https://databoks.katadata.co.id/datapublish/2021/07/09/siapa-tokoh-terpopuler-di-twitter-pada-2021.

[7] F. F. Irfani, "Analisis Sentimen Review Aplikasi Ruangguru Menggunakan Algoritma Support Vector Machine," *J. Bisnis,Manajemen dan Inform.*, pp. 258–266, 2020.

[8] B. W. Sari and F. F. Haranto, "Implementasi Support Vector Machine Untuk Analisis Sentimen Pengguna Twitter Terhadap Pelayanan Telkom dan Biznet," *J. PILAR Nusa Mandiri*, vol. 15, no. 2, pp. 171–176, 2019.

[9] N. Sinha, P. Singh, M. Gupta, and P. Singh, "Robotics at workplace: An integrated Twitter analytics – SEM based approach for behavioral intention to accept," *Int. J. Inf. Manage.*, vol. 55, p. 102210, 2020, doi: https://doi.org/10.1016/j.ijinfomgt.2020.102210.

[10] M. Vinodkumar Sadhuram and A. Soni, "Natural Language Processing based New Approach to Design Factoid Question Answering System," in *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, Jul. 2020, pp. 276–281, doi: 10.1109/ICIRCA48905.2020.9182972.

[11] Statista, "Countries with the most Twitter users 2021," 2021. https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/.

[12] J. Brownlee, "Supervised and Unsupervised Machine Learning Algorithms," in *Discover How Machine Learning Algorithms Work*, 2021.

[13] I. Steinwart and A. Christmann, *Support Vector Machines*. Springer Publishing Company, Incorporated, 2008.

[14] A. U. Khasanah, "Educational Data Mining Techniques Approach to Predict Student ' s Performance," vol. 9, no. 2, pp. 115–118, 2019, doi: 10.18178/ijiet.2019.9.2.1184.

[15] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, vol. 49, no. 06. The Morgan Kaufmann Series in Data Management Systems Morgan Kaufmann, 2011.

[16] vaksin.kemkes.go.id. (2021, 28 Oktober). Vaksinasi COVID-19 Nasional. Accessed on 28 Oktober 2021. from https://vaksin.kemkes.go.id/#/vaccines

Cek Plagiasi 13

ORIGINALITY REPORT

21%	16%	13%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

MATCH ALL SOURCES (ONLY SELECTED SOURCE PRINTED)

9%

★ edas.info

Internet Source

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		