Comparison of Support Vector Machine and Random Forest Algorithms in Sentiment Analysis on Covid-19 Vaccination on Twitter using Vader and Textblob Labelling

Berliana Putri Meliani, Oktariani Nurul Pratiwi, Rachmadita Andreswari
Information System Department
Telkom University
Bandung, Indonesia
berlianaputrii@student.telkomuniversity.ac.id, onurulp@telkomuniversity.ac.id, andreswari@telkomuniversity.ac.id

***Abstract*—Corona Virus Disease 2019 is a world outbreak that was first reported in Wuhan in December 2019. The first case of Covid-19 in Indonesia was confirmed on March 2, 2020. Covid-19 is caused by infection with virus named SARS-Cov-2. Currently, social media is widely used to find out public opinion. Generally, on Twitter social media, issues that are currently hot and much discussed by the public will become Trending Conversations. To find out and filter the opinions on social media, whether they include positive or negative opinions, you can use Sentiment Analysis. In this study, the sentiment analysis about covid-19 vaccination will use the Support Vector Machine (SVM) and Random Forest algorithms. The dataset will be labeled using the VaderSentiment and Textblob libraries found in Python. This comparison results that the SVM algorithm with textblob labeling produces an accuracy of 0.8940. Meanwhile, the sentiment results show that people tend to have negative opinions. Therefore, the best modeling for sentiment analysis is to use the Support Vector Machine with Textblob labeling.**

Keywords—support vector machine; random forest; vadersentiment; textblob; confusion matrix

I. INTRODUCTION

Corona Virus Disease 2019 is a world outbreak that was first reported in Wuhan in December 2019. The first case of Covid-19 in Indonesia was reported on March 2, 2020. Covid-19 is caused by infection with the SARS-Cov-2 virus [1]. This virus can be transmitted through droplets from an infected person to another person. The vaccination program in Indonesia itself has only started to be implemented in January 2021. Until October 2021, there are several types of vaccines used in Indonesia [2]. Initially, the implementation of vaccination reaped many pros and cons from the community regarding the side effects caused by the vaccine. Lack of education to the public about vaccines is one of the causes. Currently, social media is widely used to find out public opinion. Generally, on Twitter social media, issues that are currently hot and much discussed by the public will become Trending Conversations. To find out and filter the opinions on social media, whether they include positive or negative opinions, you can use Sentiment Analysis. Sentiment analysis is very helpful in the decision-making process. Most of the articles and research publications prove that the Support Vector Machine method is superior to other data classification techniques [3][4][5]. As a comparison, this study also applies Random Forest to measure higher accuracy. Several studies have been carried out for text classification before, for example for BBC Random Forest news text classification, song classification based on song lyrics using a Support Vector Machine [6]. The research resulted in a fairly good classification in classifying high-dimensional data in the case of news and song lyric. In a comparative study of algorithms that can be used for sentiment stated that classification with the Random Forest algorithm clearly has the advantage of high accuracy and performance, simplicity in analysis, and improvement in results over a period of time [7]. Based on the literature study in previous research, the authors compare the Support Vector Machine and Random Forest algorithms. Therefore, it can be seen which algorithm is more suitable for sentiment analysis on public opinion on the covid-19 vaccination program and produces a better accuracy value.

With this, it is hoped that it can provide valid information regarding the analysis of public sentiment towards the Covid-19 vaccination. The objectives of this research are: how the implementation of the Support Vector Machine and Random Forest algorithms for analyzing public sentiment regarding the Covid-19 Vaccination on Twitter social media; knowing the level of accuracy of the Support Vector Machine and Random Forest algorithms in analyzing public sentiment on Covid-19 vaccination on Twitter social media.; how the comparison of positive and negative public sentiment towards the Covid-19 Vaccination on Twitter social media.

With this research, it is hoped that it can be useful to help determine strategies for implementing vaccinations in Indonesia so that they can be carried out smoothly and measure the level of public awareness regarding the importance of covid-19 vaccination during the pandemic. And can provide information and references that can be used to make decisions.

II. RELATED WORKS

A. Text Mining

Text mining is the process of filtering new information from different sources. Text mining can be used to identify practical social media posts that meet your organization's needs. It can also be used by human resources for a variety of purposes, including understanding a candidate's employer perception or comparing job descriptions to resumes. There are several steps that need to be done in Text Mining, including [8]: Identifying problems and specific objectives; identifying text to collect; organizing text and combining everything easily in CSV; identifying to clean up reviews and analyzing text using the extraction feature; scanning keywords to analyze text; and Gaining insight and recommendations

A. Azizah and E. D. Ariyani (Eds.): iCAST-SS 2022, ASSEHR 719, pp. 620–626, 2022.
https://doi.org/10.2991/978-2-494069-83-1_108

In this study, text mining was used to filter information about vaccination programs obtained from social media Twitter. The information will then be identified to determine the public's perception of the vaccination program. Before text mining, the dataset containing the required information will be carried out with data cleansing and also feature extraction.

B. Sentiment Analysis

Sentiment analysis is the process of understanding, extracting, and processing textual data to obtain emotional information contained in opinion or sentiment statements. Sentiment analysis is used to identify the negative and positive attitudes and tendencies of people who think about a problem or object. Opinion mining is a combination of text mining and natural language processing [9]. In this case, sentiment analysis was used to determine public opinion on the COVID-19 vaccination program.

C. Covid-19

SARS-COV2, a member of the vast family of Coronaviruses, a group of viruses that infect the respiratory system, is the cause of Coronavirus illness (COVID-19). The virus typically only has the potential to cause minor respiratory infections like the flu. However, this virus can also result in fatal respiratory illnesses such Middle East respiratory syndrome (MERS), severe acute respiratory syndrome (SARS), and lung infections (pneumonia) [10].

When coughing or sneezing, COVID-19 infection can be transferred through the nose and mouth by tiny droplets of water. After that, the drops land on neighboring items. Then, if a different individual comes into close contact with an object that has been contaminated with these droplets and touches their mouth, nose, or eyes (the triangle of the face), they may contract COVID-19. Alternately, a person can contract COVID-19 if they unintentionally breathe droplets from an infected person. Therefore, it's crucial to keep a minimum of 1 meter between you and somebody who doesn't seem to be in excellent health [11].

D. Vaccination

Vaccination is a bodily process whereby a person becomes immune or protected from disease, and at some point when exposed to a disease, usually someone who has been vaccinated will not feel sick or only get a mild illness. Vaccines are biological products that contain microorganisms or some form of antigen, or substances that have been safely processed, and when administered to humans, actively develop specific immunity against certain diseases. Vaccines will stimulate the formation of immunity against certain diseases of the human body, and the body knows how to remember and recognize viruses and bacteria that carry diseases and fight them [12]. The vaccination program itself has pros and cons, especially for the side effects of the vaccine itself.

E. Data Preprocessing

Data preprocessing is a tool for analyzing and processing complex data but can take a lot of time to process. It covers a wide range of areas, including data preparation and data reduction technology. The first includes data transformation, integration, cleaning, and normalization. The latter aims to reduce data complexity through feature selection, sample selection, or discretization. After successfully implementing the Data preprocessing, the final data set obtained can be considered a usable source and suitable for the next data mining algorithm [13].

F. Support Vector Machine

The basic concept of Support Vector Machine (SVM) is a combination of several decades of existing computational theory, such as hyperplane margin, kernel, and other supporting concepts. However, until 1992, no attempt was made to assemble this component. The basic principle of SVM is the linear classifier, which has been further developed to allow you to solve non-linear problems using the concept of kernel tricks in higher-dimensional workspaces. This has generated interest in research in the area of pattern recognition and determined the potential theoretical and application-related features of SVM. Currently, SVM is well applied to real-world problems and generally offers better solutions than traditional methods such as Artificial Neural Networks [14].

G. Random Forest

Random Forest algorithm supervised learning. The basic concept of this algorithm is to form a forest and make it randomly. Random Forest is created to combine several decision trees. Similar to bagging, each decision tree is constructed using a different bootstrap pattern. However, unlike bagging, instead of selecting a separate rule of all predictive attributes on each node, only a predefined number of attributes that are randomly selected at each node are used [15].

H. VaderSentiment

VADER (Valence Aware Dictionary and Sentiment Reasoner) is a rule-based and lexicon sentiment analysis system designed specifically for social media sentiments. Vader employs a stingy rule-based model to assess tweet sentiment. According to research, Vader enhances the capabilities of traditional sentiment lexicons such as Linguistic Inquiry and WordCount. The valence score of each word in the lexicon is added up, adjusted according to the rules, and then normalized to be between -1 (extremely negative) and +1. (extremely positive) [16].

I. Textblob

Textblob is a text data processing library in Python. This library contains simple APIs for experimenting with Natural Language Processing (NLP) tools such as noun phrase extraction, sentiment analysis, classification, and others. The sentiment property returns a named tuple containing the sentiment form such as polarity and subjectivity. A float in the range between -1.0 until 1.0 represents the polarity score, with a higher polarity value indicating a more positive sentiment. The float value of the range is subjectivity where 0.0 represents extreme objectivity and 1.0 represents extreme subjectivity [17].

J. Confusion Matrix

The confusion matrix shows how frequently a specific activity is properly identified and how frequently it is

mistaken for another action. Typically, performance metrics such as accuracy, sensitivity, and specificity are used to summarize classification accuracy [18]. With the evaluation method using a confusion matrix, it can be seen the success rate of the algorithm used is based on the True Positive, True Negative, False Positive, and False Negative values.

K. Receiver Operating Characteristic (ROC)

The Receiver Operating Characteristic (ROC) or ROC curve is a two-dimensional plot that depicts how well performance of the classifier system as the discriminant threshold value varies across the predictor range. The x-axis represents the predictive test's false positive rate. The y-axis represents the prediction test's true positive rate.

Area Under Curve (AUC) or commonly known as c-statistic can be used as an evaluation method to distinguish the actual status. In general, the rules for interpreting the AUC value are as follows [19]: AUC = 0.5 : None; 0.6 > AUC > 0.5 : Poor; 0.7 > AUC > 0.6 : Acceptable; 0.8 > AUC > 0.7 : Very Good; AUC > 0.9 : Excellent.

III. METHODOLOGY

A. System Overview

Systematic problem-solving in this research is to use Knowledge Discovery in Database [20] where the dataset used will be analyzed so that the patterns contained in the dataset can be identified. The systematic solution is divided into several stages as shown below.

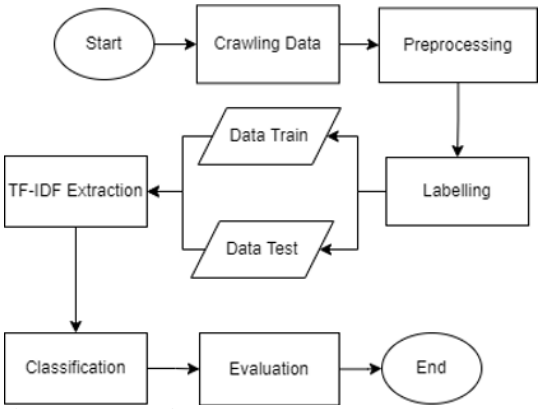

Fig. 1 System Overview

Based on Fig. 1 This research will collect tweet data from Twitter social media using the data crawling method, the data is then stored as raw data. Raw data will go through the preprocessing step to become data that is ready to be analyzed. The data will be labeled automatically by using VaderSentiment and Textblob. Datasets that have passed the labeling stage will be subjected to data splitting to divide the data into training set and testing set. The training data and data testing will be feature extracted using the TF-IDF Vectorizer where considering a word's overall document weightage as a measure of how frequently it appears in the documents using TfidfTransformer in sklearn module in python. If the data has been extracted using a vectorizer, the data will be classified using the Support Vector Machine and Random Forest. The results of the classification will be evaluated using a confusion matrix.

B. Dataset

The dataset used by the author in this study is a dataset in the form of tweets on social media Twitter containing the keywords COVID-19 vaccine and related words. The data taken is in the form of tweets with several conditions, the conditions for data retrieval are: Tweets taken are tweets with a period from December 2021 to May 2022; Tweets taken are only tweets that use Indonesian. Based on these provisions, the tweets collected using the Tweepy library on python were 17539. The following are some of the tweets obtained from the data crawling stage, which can be seen in Table 1

TABLE I. EXAMPLE DATASET

Text	Timestamp
COVAD-COVID!!!!! Muak saya dengan COVID-19 yg katanya MEMATIKAN!!!!! mematikan apa!!!!?????? Mematikan perekonomian masyarakat!!!!!! Mau sampai kapan masyarakat mau dibodohi dengan PANDEMI yg tak berujung!!!!!! saya 3x terpapar covid!! Tidak vaksin,,nyatanya masih hidup!!	2021-12-09 09:22:57+00:00
Perlindungan terhadap jangkitan COVID-19 berkurangan 3-5 bulan selepas imunisasi, menurut kajian RECoVaM. Pengambilan dos penggalak vaksin COVID-19 adalah untuk meningkatkan tahap immuniti. #vaksin #dospenggalak #COVID19 https://t.co/YO06d76IqA	2021-12-09 09:14:16+00:00
Data terkini menunjukkan dos penggalak vaksin COVID-19 diperlukan untuk melawan varian Omicron. https://t.co/oJRSMYCIli	2021-12-09 09:12:53+00:00
Kemenkes Jepang telah memperingatkan kenaikan risiko dan kasus Myocarditis pada usia muda akibat vaksin covid Pfizer dan Moderna. Kalau @KemenkesRI bagaimana? Health ministry warns of vaccine's side effects | NHK WORLD-JAPAN News https://t.co/Z0vJeqWkv4	2021-12-09 09:12:04+00:00
Data terkini menunjukkan dos penggalak vaksin COVID-19 diperlukan untuk melawan varian Omicron. https://t.co/oJRSMYCIli	2021-12-09 09:09:48+00:00

C. Preprocessing

After the dataset is collected using the data crawling method using the Twitter API, the raw data must go through a data preprocessing process. The data collected is still redundant, so preprocessing needs to be done to eliminate duplicate data, hyperlinks, mentions, and also unnecessary words.. This data preprocessing stage is needed so that the existing dataset becomes more structured so that it can facilitate the data analysis process later. The stages of the preprocessing process are:

- Data cleansing is the process of cleaning data redundancy such as deleting duplicate data and also data containing hyperlinks, retweets, numbers, emojis, and also user mentions that can influence the data analysis process.
- Tokenization is the process of breaking down data that is used to form a sentence word for word.
- Stopwords removal is the process of removing words that have no important meaning but often appear in a sentence such as conjunctions.

- Stemming is the process of changing words that have affixes into the basic word form.

The following table is a comparison of the data before and after preprocessing.

TABLE II. PREPROCESSING DATA

Before Preprocessing	After Preprocessing
@junebluemoon Aamiin... pulang vaksinnya nanti langsung istirahat kak. Soalnya sepengalaman aku bulan lalu vaksin booster aku cuma lemes dikit trs pulangnya dibawa istirahat, bangun" Alhamdulillah udh segar lg. tp balik lg ya imun org beda"	aamiin pulang vaksin langsung istirahat kak alam vaksin booster lemes dikit trs pulang bawa istirahat bangun alhamdulillah udh segar lg tp lg ya imun org beda
@AbdulLatif_BE @TedHilbert @ohorella_b @JodohBisaDiatur @__Sridiana_3va @dr_koko28 Abdull Latif, Coba tunjukkan data dan fakta ilmiah tentang: 1. Keamanan vaksin covid 2. Validitas test PCR 3. hasil penelitian covid itu lebih bahaya dr TBC	abdull latif coba tunjuk data fakta ilmiah aman vaksin covid validitas test pcr hasil teliti covid bahaya dr tbc
@convomf Saya. Hari ini abias vaksin booster. Efeknya sedikit demam + ngantuk	abias vaksin booster efek demam ngantuk
yang abis booster bekas suntikannya lebih ngilu dari vaksin 1 sama 2 ga si?	abis booster bekas suntik ngilu vaksin ga si
abis booster vaksin ke 3 seharian masih biasa aja, pas kerja juga masih biasa, pas malem baru kerasa efeknya	abis booster vaksin hari aja pas kerja pas malem rasa efek

To facilitate the data labeling process, datasets that have gone through the preprocessing stage is then translated into English.

TABLE III. TRANSLATED DATASET

After Preprocessing	After Translated
aamiin pulang vaksin langsung istirahat kak alam vaksin booster lemes dikit trs pulang bawa istirahat bangun alhamdulillah udh segar lg tp lg ya imun org beda	amen, go home, take a break immediately, sis, when the booster vaccine is weak, go home, take a break, wake up, thank god, it's fresh again, but again, people's immunity is different
abdull latif coba tunjuk data fakta ilmiah aman vaksin covid validitas test pcr hasil teliti covid bahaya dr tbc	abdull latif try to show scientific fact data is safe covid vaccine validity pcr test accurate results covid is dangerous from TB
abis vaksin booster efek demam ngantuk	after booster vaccine effect of sleepy fever
abis booster bekas suntik ngilu vaksin ga si	after the booster, the former injection hurts, does the vaccine hurt?
abis booster vaksin hari aja pas kerja pas malem rasa efek	after the vaccine booster, just on the day of work, at night, feel the effect

D. Labeling

Class Labeling used is positive and negative. The labeling process is carried out using Vader Sentiment Intensity Analyzer and Texblob as a comparison. However, the implementation of Vader Sentiment Intensity Analyzer and Textblob currently does not support the labeling process in Indonesian, so tweet data using Indonesian needs to be translated into English first. Based on [21] by using VaderSentiment if the compound value is more than or equal to 0.05 then the data line is considered to have a positive sentiment. Otherwise, if the compound value is less than 0.05 then the data line is considered to have a negative sentiment. The same value is also applied using Textblob, data that produces a sentiment polarity of more than or equal to 0.05 then the data will be labeled positive. While the data that produces a sentiment polarity of less than 0.05, the data will be labeled negative. Table IV displays data labels using VaderSentiment and Textblob after translation.

TABLE IV. DATA LABELLING

Translated Tweet	VaderSentiment	Textblob
i-i haven't had the booster vaccine yet	Negative	negative
amen, allahumma, amen, for the prayers for the family. honestly, think about the condition of the family, the booster vaccine, and friends at work.	Positive	positive
amen, tomorrow is a booster vaccine even though	Negative	negative
amen, go home, take a break immediately, sis, when the booster vaccine is weak, go home, take a break, wake up, thank god, it's fresh again, but again, the immune system is different.	Positive	negative

E. Data Splitting

Data Splitting is done with 3 ratios of sharing data testing and training data. The comparison of the accuracy results produced by the three data ratios shown in Table V below.

TABLE V. RATIO COMPARISON

	SVM			Random Forest		
	70:30	80:20	90:10	70:30	80:20	90:10
Vader	0.8218	0.8362	0.8522	0.7947	0.8017	0.8325
Textblob	0.8374	0.8657	0.8940	0.8136	0.8263	0.8596

Dataset will divided into training set and testing set with ratios 90:10. The ratio was chosen because it is based on Table V that the ratio produces the highest accuracy value. So, using a ratio of 90:10, the amount of training data is 3654 tweets and testing data is 406 tweets.

IV. RESULT AND ANALYSIS

A. Sentiment Analysis

The dataset used is data in the form of tweets taken from November 2021 to May 2022. The dataset is then determined for positive and negative labeling. The labeling is done to find out whether tweets that contain positive or negative sentiments. The author performs labeling in two ways, namely by using Vader Sentiment Analyzer and Textblob for automatic labeling. The total dataset used for sentiment labeling is 4060 tweets.

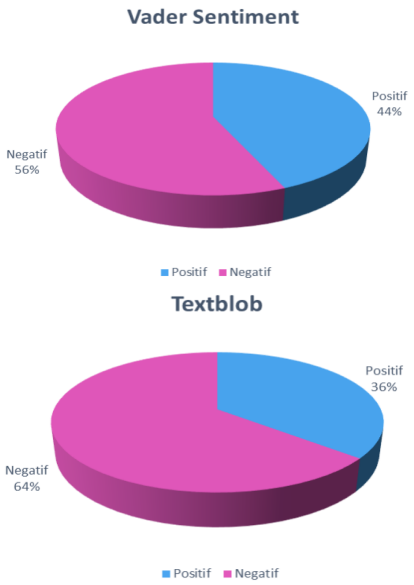

Fig. 2 Sentiment Labeling Comparison

Based on Fig. 2 Based on these data, labeling using Textblob negative sentiment generated more when compared to labeling results using VaderSentiment. The labeling results generated using textblob labeling are 64% for negative sentiment labels and 36% for positive sentiment labels. When compared, labeling using Vader sentiment only produces 56% negative sentiment and 44% positive sentiment.

B. *Classification Result*

With a ratio of 90:10 with 90% as training data and 10% as testing data and labeling using Textblob, the Support Vector Machine Algorithm produces an accuracy value of 0.8940 for testing data. If the labeling is done using Vader Sentiment Analyzer, the SVM algorithm produces an accuracy of 0.8522.

Using the Random Forest algorithm with an estimator of 100 and a data ratio of 90:10 produces the highest test data accuracy value with labeling using Textblob, which is 0.8596. Meanwhile, with the same data ratio in labeling using Vader, the accuracy value is 0.8325.

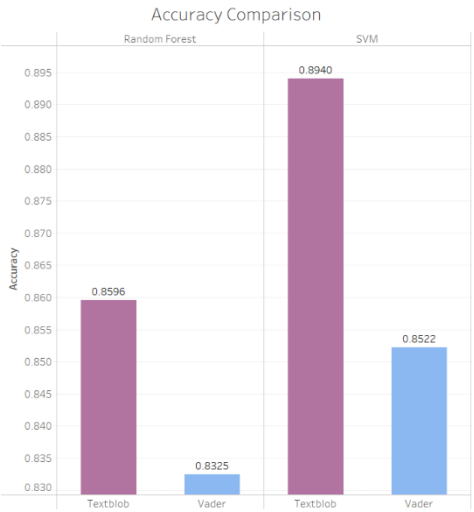

Fig. 3 Accuracy Comparison

Based on Fig. 3 The classification result that produces the highest accuracy value is the Support Vector Machine algorithm with automatic labeling using Textblob. The Support Vector Machine algorithm will be more efficient if it is implemented in fewer classes [22]. In this study, the number of classes used was 2 labels, namely positive and negative. Therefore, the SVM algorithm produces higher accuracy when compared to Random Forest.

C. *Performance of Support Vector Machine and Random Forest.*

The performance results measured based on the accuracy value show that the Support Vector Machine algorithm with labeling using Textblob produces the highest accuracy, which is 0.8940. While the Random Forest algorithm with Textblob labeling produces an accuracy value of 0.8596. Both algorithms produce higher accuracy with labeling using Textblob compared to using Vader Sentiment Analyzer.

D. *Performance Evaluation*

After the classification stage, the next stage needs to be a performance evaluation. In this study, confusion matrix and Receiver Operating Characteristic (ROC) will be used for evaluation method. Performance evaluation using the confusion matrix is carried out using the skit-learn library. The table generated by the confusion matrix contains True Positive, True Negative, False Positive, and False Negative values.

TABLE VI. CONFUSION MATRIX

ALGORITHM – AUTOMATIC LABELLING	Confusion Matrix			
	True Positive (TP)	False Positive (FP)	True Negative (TN)	False Negative (FN)
SVM – Vader	116	21	230	39
SVM – Textblob	97	9	266	34
Random Forest – Vader	111	24	227	44
Random Forest - Textblob	85	11	264	46

Based on the Confusion Matrix Table, the Support Vector Machine algorithm with labeling using Textblob produces a False Negative value of 34 and a True Negative value of 266. If you look at Fig. 3 above, the algorithm with the Textblob labeling method produces the highest accuracy value when compared to other algorithms and labeling.

Receiver Operating Characteristics (ROC) is a curve that describes the confusion matrix. The ROC curve is depicted as a two-dimensional graph with the horizontal line as the value of the False Positive Rate and the vertical line as the value of the True Positive Rate. The ROC curve will describe the area value that is in the area under the curve or called the Area Under the Curve (AUC) value. The greater the AUC value, the stronger the classification model used.

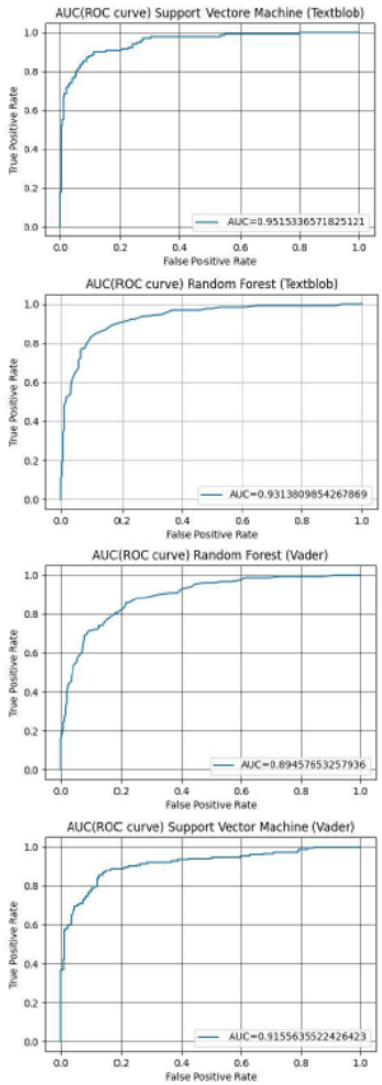

Fig. 4. ROC Comparison

Based on Fig. 4, Support Vector Machine algorithm modeling using Textblob labeling produces the highest AUC value when compared to other modeling algorithms with AUC value 0.9515.

V. CONCLUSION

Based on the results of the analysis and evaluation, the conclusions drawn from this study are the implementation results show that Support Vector Machine algorithm with labeling using Textblob and a ratio of 90:10 produces the highest level of accuracy, that is equal to 0.8940. While the accuracy generated by the Random Forest algorithm with labeling using Textblob is 0.8522. Textblob labeling produces higher accuracy than using VaderSentiment with a comparison of positive and negative sentiments for the Covid-19 vaccination program is 36% for positive sentiment and 64% for negative sentiment.

REFERENCES

[1] S. Sardjoko, *Proyeksi COVID-19 di Indonesia*. 2021.

[2] "Vaksinasi COVID-19 Nasional," 2021. https://vaksin.kemkes.go.id/#/vaccines (accessed Oct. 20, 2021).

[3] S. Rani and S. Bhatt, "Sentiment Analysis on twitter data using Machine Learning," *J. Xidian Univ.*, vol. 14, no. 12, pp. 1–4, 2020, doi: 10.37896/jxu14.12/039.

[4] 2018) (Al Amrani, Lazaar, El Kadiri., "Random Forest and Support Vector Machine based Hybrid Approach to SA -- RF.pdf," *The First International Conference on Intelligent Computing in Data Sciences*. pp. 511–520, 2018.

[5] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," *Neurocomputing*, no. xxxx, 2020, doi: 10.1016/j.neucom.2019.10.118.

[6] W. Willy, D. P. Rini, and S. Samsuryadi, "Perbandingan Algoritma Random Forest Classifier, Support Vector Machine dan Logistic Regression Clasifier Pada Masalah High Dimension (Studi Kasus: Klasifikasi Fake News)," *J. Media Inform. Budidarma*, vol. 5, no. 4, p. 1720, 2021, doi: 10.30865/mib.v5i4.3177.

[7] A. Gupte, S. Joshi, P. Gadgul, and A. Kadam, "Comparative Study of Classification Algorithms used in Sentiment Analysis," *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 5, pp. 6261–6264, 2014.

[8] T. Kwarrler, *Text Mining in Practice with R*. Wiley, 2017.

[9] I. Rozi, S. Pramono, and E. Dahlan, "Implementasi Opinion Mining (Analisis Sentimen) Untuk Ekstraksi Data Opini Publik Pada Perguruan Tinggi," *J. EECCIS*, vol. 6, no. 1, pp. 37–43, 2012.

[10] Bio Farma, "Kenali Virus Covid-19," 2021. https://www.biofarma.co.id/id/berita-terbaru/detail/kenali-virus-covid19 (accessed Nov. 10, 2021).

[11] Kementrian Kesehatan RI, "Pertanyaan dan Jawaban Terkait Covid-19," 2020. https://www.kemkes.go.id/folder/view/full-content/structure-faq.html (accessed Nov. 10, 2021).

[12] Kementerian Kesehatan RI, "Question (Faq) Pelaksanaan Vaksinasi Covid-," *2020*, vol. 2, no. 1, pp. 1–16, 2021, [Online]. Available: https://kesmas.kemkes.go.id/assets/uploads/contents/others/FAQ_VAKSINASI_COVID__call_center.pdf.

[13] S. García, S. Ramírez-Gallego, J. Luengo, J. M. Benítez, and F. Herrera, "Big data preprocessing: methods and prospects," *Big Data Anal.*, vol. 1, no. 1, pp. 1–22, 2016, doi: 10.1186/s41044-016-0014-0.

[14] S. van Plaosan, "Support Vector Machine (SVM)." http://learningbox.coffeecup.com/SVM.html (accessed Nov. 10, 2021).

[15] K. M. Bottenberg, *A General Introduction to Data Analytics*. 2017.

[16] C. J. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," *Proc. 8th Int. Conf. Weblogs Soc. Media, ICWSM 2014*, pp. 216–225, 2014.

[17] S. Lorla, "TextBlob Documentation," *TextBlob*, p. 69, 2020.

[18] S. Ruuska, W. Hämäläinen, S. Kajava, M. Mughal, P. Matilainen, and J. Mononen, "Evaluation of the confusion matrix method in the validation of an automated system for measuring feeding behaviour of cattle," *Behav. Processes*, vol. 148, pp. 56–62, 2018, doi: 10.1016/j.beproc.2018.01.004.

[19] S. Yang and G. Berdine, "The receiver operating characteristic (ROC) curve," *Southwest Respir. Crit. Care Chronicles*, vol. 5, no. 19, p. 34, 2017, doi: 10.12746/swrccc.v5i19.391.

[20] A. Azevedo, "Data Mining and Knowledge Discovery in Databases," pp. 502–514, 2019, doi: 10.4018/978-1-5225-7598-6.CH037:

[21] I. Irawaty, R. Andreswari, and D. Pramesti, "Vectorizer Comparison for Sentiment Analysis on Social Media Youtube: A Case Study," *2020 3rd Int. Conf. Comput. Informatics Eng. IC2IE 2020*, pp. 69–74, 2020, doi: 10.1109/IC2IE50715.2020.9274650.

[22] M. Sheykhmousa, M. Mahdianpari, H. Ghanbari, F. Mohammadimanesh, P. Ghamisi, and S. Homayouni, "Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 13, pp. 6308–6325, 2020, doi: 10.1109/JSTARS.2020.3026724.