Journal of Theoretical and Applied Information Technology
15th January 2023. Vol.101. No 1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

SENTIMENT ANALYSIS OF THE COVID-19 BOOSTER VACCINATION PROGRAM AS A REQUIREMENT FOR HOMECOMING DURING EID FITR IN INDONESIA

ANGGA PRATAMA[1], RAKSAKA INDRA ALHAQQ[2], YOVA RULDEVIYANI[3]
[1, 3]Magister of Information Technology, University of Indonesia, Indonesia
[2]Indonesian Agency for Meteorology, Climatology, and Geophysics (BMKG), Indonesia
E-mail: [1]angga.pratama11@ui.ac.id, [2]raksaka.indra@bmkg.go.id, [3]yova@cs.ui.ac.id

ABSTRACT

The COVID-19 vaccination program was carried out to overcome the pandemic. In addition to vaccination, there is a booster vaccine program which is also an obligation for the public to be followed but has not received a good response from the public. Indonesia's government is making booster vaccination a mandatory requirement for mass homecoming in the Islamic celebration day of Eid Fitr, known as mudik Lebaran. This study aims to find out public opinion and perception regarding the booster vaccination program for mudik Lebaran using sentiment analysis. This study uses eight classification modeling: Naïve Bayes, Support Vector Machine (SVM), Decision Tree, Logistic Regression, Random Forest, K-Nearest Neighbor, AdaBoost, and XGBoost. The best classification modeling is SVM with the best accuracy score 88% and the F1 score 88%. Then this SVM model is used to predict the sentiment of 30,582 tweet data from March 22 to May 02, 2022. The results are 11,507 giving negative sentiment (37.63%) and 19,075 giving positive sentiment (62.37%). This result shows that the government's strategy in accelerating the COVID-19 booster vaccination program was well accepted by making it a requirement for mudik Lebaran. Furthers analysis with visualization of time series, shows that the sentiment had eveloved. In the first week, negative sentiment prevailed due to reactions to this policy. The Indonesian people compare the policy of the MotoGP event in Mandalika which does not require a vaccine booster. After that, in the following weeks the positive sentiment prevailed because the community realized that boosters were important to maintain their health and that of their families back home. This research shows the importance of time series visualization because sentiment can change over time.
Keywords: *COVID-19, Vaccine Booster, Mudik Lebaran, Classification, Sentiment Analysis, Time Series*

1. INTRODUCTION

Coronavirus disease 19 (COVID-19) is a highly contagious pathogenic viral infection caused by the acute respiratory syndrome coronavirus 2 (SARS-Cov-2). This virus first appeared in Wuhan, China, and spread rapidly throughout the world [1]. The report issued by WHO until early March 2022, there have been 437 million cases of infection and 5.9 million deaths worldwide due to COVID-19 [2]. In addition to the health sector, the COVID-19 outbreak has caused a global economic depression in various countries [3].

In the early days of the pandemic, the spread of the virus and cases of death can be minimized by wearing masks, washing hands, and maintaining social distance. However, with the successful development, evaluation, and production of several vaccines, governments in various countries have begun to turn to vaccination programs as a solution to the COVID-19 pandemic [4]. By the end of February 2022, WHO recorded that 4.3 billion people had received the full 2-dose vaccination [2].

Since the start of the first COVID-19 vaccination program on December 13, 2020, a few controversies arose with the rejection of several people from various countries regarding the program [4]. It was recorded that until March 23, 2022, the vaccination program in Indonesia reached 93.29% for the first vaccine and 73.59% for the second vaccine. In addition to the first and second vaccines, which are referred to as primary vaccinations, the government also requires the public to participate in the booster vaccine program. However, the booster vaccine program is not very popular, the achievement is only up to 7.65% [5].

On March 23, 2022, the vice president of Indonesia discussed making booster vaccination a mandatory requirement for this year mass homecoming (mudik) [6]. Mudik is an annual activity that is awaited by the people of Indonesia.

Especially for those who go to the metropolitan area, back to visit their hometown, at the end of the Muslim fasting month (Lebaran), to regroup with their family in urban areas. Lebaran is Indonesian phrase to call Eid Fitr, which is a festival or celebration day after a month of Ramadhan fasting. This festival is occurred annually and in 2022 held on May 1st, causing mass movement from metropolitan to urban area. In this Covid situation, Government is very concerned that mudik Lebaran will cause virus transmission from metropolitan people to their family in the urban area [7].

Government will not forbid mudik Lebaran activity because last year is already forbidden. But pandemic is not over yet, so to prevent it get worse, booster vaccination is discussed will be mandatory requirement [6]. With this discourse, there are pros and cons related to booster vaccination as requirement for mudik. There are people said that the booster program is troublesome. They thought that 2 doses of vaccines are enough, some said that they afraid of getting fever after injection. But on another hand, some people are grateful because they could see their family in this year, another dose of vaccine is not a problem [8], [9]. To make it clear about public opinion on this discourse, we tend to conduct sentiment analysis of the covid-19 booster vaccination program as a requirement for homecoming during Eid Fitr in Indonesia with research questions:

- How public sentiment in Twitter towards covid-19 booster vaccination program as a requirement for homecoming during Eid Fitr?
- What drives public sentiment in Twitter towards covid-19 booster vaccination program as a requirement for homecoming during Eid Fitr?

Many studies have used sentiment analysis to find out public opinion and perception of a particular topic. Using sentiment analysis, we able to analyzes opinions, judgments, attitudes, and emotions towards an entity such as a particular issue or event. The importance of sentiment analysis is related to the rapid development of opinions on social media such as blog forums or Twitter [10].

Previous research related to sentiment analysis of the earlier COVID-19 vaccine program has been conducted. Research [11] aims to analyze public sentiment towards the COVID-19 vaccination program in Indonesia. Using 6,000 data taken from Twitter with the "vaccine covid-19" keywords. A Naïve Bayes algorithm is used to analyze the sentiment. The result of sentiment is over 3.4 thousand negatives (56%), over 2.4 thousand positive (39%), and the remaining 301 (1%) were neutral during January 15–22, 2021 period.

Research [12] also discusses sentiment analysis on the COVID-19 vaccination program in Iran. Researchers found differences in sentiment towards imported vaccines and domestically made vaccines within 6 months. This study uses deep learning algorithms; CNN and LSTM to classify Persian language datasets.

This study aims to find out public opinion and perception regarding the booster vaccination program for mudik Lebaran using sentiment analysis. To understand public opinion is the urgency of this research because public opinion should be considered by Government as one of materials to aligning policies. From previous studies that used only one or two models, this study is conducted by using eight classification models, it is expected to be able to obtain the best model to use in predicting the sentiment of future tweets. The best modeling is with the best accuracy value and F1-score value [11], [13].

We also conducted time series visualization and word cloud analysis for deeper discussion about public opinion because many studies of sentiment analysis are usually stop at the modelling result or prediction result. The time series visualization helps to analize change regarding of the public sentiment and word cloud analysis helps to identify frequent word to reflect the topic of discussion [14]. The results of this study are expected to be able to provide information to the government regarding the booster vaccination policy for mudik Lebaran, are get a positive or negative response, and why that sentiment is driven.

2. LITERATURE REVIEW

Literature review is carried out based on desk research by looking for studies that are related to this research. Related studies must be able to provide us with an understanding of the theory, methods, results, and discussions that we will convey later regarding this research [15]. The field of studies we are looking for is sentiment analysis, data preprocessing, text classification and methods of testing and validation. In the end, we were able to figure out the gap of the previous studies about sentiment analysis of the Covid-19 vaccination program and filled in the gap in this research.

2.1 Sentiment Analysis

Sentiment analysis or opinion mining is a study to analyze people's opinion. Sentiment analysis mainly focus on opinions which express or imply positive or negative sentiments [10]. In practice, sentiment analysis is carried out from a text data source, usually publicly available from digital media or social media. Social media that is often used as a data source is Twitter. This is because Twitter provides an API to retrieve tweets from users who post to the platform. Of course, this makes it easier for researchers to find data sources. In addition, until now Twitter remains a popular social media in Indonesia. Often when there is a government policy for a certain topic, Indonesian people then spill their opinions on Twitter [11], [13].

Many have done sentiment analysis with data sourced from Twitter. Research [16] took data from Twitter for sentiment analysis related to the response of the Saudi community to the Covid-19 vaccination program. Research [17] also uses Twitter as a data source for sentiment analysis regarding the Covid-19 vaccination in Africa. After data collection, the data obtained cannot be directly processed, this is because the data is unstructured and difficult to analyze quantitatively. Therefore, a preprocessing method is needed that aims to use the unstructured data into structured data so that it can then be analyzed [18], [19].

2.2 Data Preprocessing

Research [20] compared 15 preprocessing techniques on sentiment analysis from Twitter. To find out the best preprocessing technique, three different machine learning algorithms were used, namely Linear SVC, Bernoulli-Naïve Bayes, and Logistic Regression. This research produces preprocessing techniques with the best accuracy performance, namely stemming, number removal, and replacing elongated words.

Research [21], [22] adding a lemmatization technique to the preprocessing. Lemmatization is a more complicated technique than stemming. If stemming only removes affixes to a word, in lemmatization it also changes the words that have affixes into their standard structure. This preprocessing is only the initial stage of sentiment analysis. Furthermore, if the data is in a structured form, the data in the form needs to be classified into certain text classes according to the purpose of the analyst sentiment. This classification of data is usually positive and negative sentiments [10], [20]. In this study lemmatization is conducted as part of pre-processing techniques.

2.3 Text Classification

Text Classification is the most important stage in sentiment analysis research. The data obtained need to be classified into texts with positive and negative sentiments. This can be done manually by reading the text one by one, but if the data is very large, it will certainly drain energy and time [23], [24]. For this reason, usually in sentiment analysis, there are two methods that are often used for classification data, namely with the lexicon dictionary or with machine learning [25].

Classification text with lexicon is an easy way, by using an existing sentiment dictionary, then matching and scoring on the datasets that we have. The results of the classification can be directly analyzed. However, this lexicon technique is not without problems. It could be that this lexicon dictionary is not relevant to the data being tested. Different topics can have different results. Then with the development of the current language, the old lexicon dictionary is no longer relevant to the new dataset. This makes manual labeling techniques still used because they are more relevant and accurate [25]–[27].

Machine learning techniques use a certain algorithm to train machines to classify text predictively. To train it, of course, required training data. This training data comes from sample of the overall dataset taken at random, then labeling or classify the data manually. Many machine learning algorithms have been used to classify text which are also used in sentiment analysis [22], [28], [29]. In this study, 8 types of algorithms were used to obtain the most accurate classification modeling, namely Naïve Bayes, Support Vector Machine (SVM), Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbor, AdaBoost, and XGBoost.

2.4 Naïve Bayes (NB)

Naïve Bayes is a text classification technique that is often used because its theoretical approach is quite good in terms of consistency and calculation [11], [30]. In general, the calculation of Naïve Bayes uses the Bayes algorithm with the following equation:

$$P(C|U) = \frac{\mathrm{P(U|C)P(C)}}{\mathrm{P(U)}} \quad (1)$$

Equation (1) states that C is a class, U is data whose class has not been defined (undefined), while P(C|U) is a possible hypothesis depending on conditions. P(C) and P(U|C) are the previous probabilities of a class based on hypothetical conditions, while P(U) is the probability C. There are several variations of Nave Bayes used for text classification with this Bayes algorithm approach, namely Unigram Naïve Bayes [11], Maximum Entropy Classification, Multinomial Naïve Bayes, and Bernoulli Naïve Bayes. In this study, Bernoulli Naïve Bayes will be used because it is already available in the Python Scikit Learn Library [28].

2.5 Support Vector Machine (SVM)

SVM is used to find the best hyperplane by maximizing the distance between classes. Hyperplane is a function that may be used to differentiate between classes. Lines are used to categorize items in two dimensions, planes are used to categorize objects in three dimensions, and hyperplanes are used to categorize objects in higher-dimensional class spaces. The hyperplane discovered by SVM separates two classes, which means that it is located further from the data items than the outermost class. A support vector is the outermost data item in SVM that is closest to the hyperplane [31].

In general, data issues prevent linear input space separation; soft margin SVM is not accurate and does not generalize well since it cannot locate the separator in the hyperplane. As a result, a kernel is required to translate data into kernel space, a higher dimensional space that is beneficial for linearly dividing data. In general, the linear, polynomial, and radial basis function (RBF) kernels are the most often employed kernel functions [18], [31]. In this study, Linear SVC will be used as the SVM algorithm.

2.6 Logistic Regression (LR)

The link between a response variable (dependent variable) with two or more categories and one or more explanatory factors (independent variables) on a categorical or interval scale is described using the statistical analysis approach of logistic regression. Logistics Regression is a non-linear regression that is used to explain relationships between variables X and Y that are not linear, have irregular Y distributions, and have variable reaction times that cannot be described by conventional linear regression models. Some research sentiment analysis that uses this technique are found that the accuracy of the technique is not inferior to SVM [18], [19], [32]. In this study, the Logistic Regression algorithm from the Scikit Learn Library will be used.

2.7 Decision Tree (DT)

Decision Tree is a classifier algorithm that is smart and quite simple. DT can make predictions with an approach that represents a tree with branches, starting from observations for a thing to conclusions on the value that is targeted for it. DT is a combination of mathematical and computational techniques for categorizing, describing, and simplifying the tested dataset. DT is represented by leaves and branches, where the label class is like the leaf and the conjunctions between branches are the features that bring it to the label [18]. DT has a formula like the following:

$$(f, T) = (f1, f2, f3, \dots, fk, T) \quad (2)$$

Based on the equation (2), T is the dependent variable that becomes the target to be classified. While f is a vector composed of features f1, f2, f3, and so on [18]. In this study, DT algorithm from the Scikit Learn Library will be used.

2.8 K-Nearest Neighbor (KNN)

By designating class membership to each input, the KNN algorithm is used to categorize a group of inputs. An item is classified by KNN using the votes of many its neighbors and is then given the class that receives the most votes from the object's k closest neighbors. When k is set to 1, the object is only allocated to the one nearest neighbor class, which is often a positive small integer. KNN is a non-parametric method that solely employs the distance function to gauge how similar each item is to its neighbors [18]. In this study, the KNN algorithm from the Scikit Learn Library will be used.

2.9 Random Forest (RF)

RF is one of approach for supervised machine learning is a classification method built up of several decision trees that makes predictions that are more accurate than those provided by any one tree [19]. Leo Breiman first proposed the concept of RF (Random Forest) in 2001. Using the voting process, random forests are an amalgamation of bagging and random trees. It improves the bagged decision tree. Multiple decision trees are constructed using the

bagging method once the instances have been trained. Trees are created to fit the data set once data is randomly chosen. Following the training procedure, every created tree will cast a vote, and the class is predicted using the results of the majority vote. Large datasets benefit greatly from Random Forest's improved classification accuracy [33]. In this study, the Random Forest algorithm will be used in the Scikit Learn Library.

2.10 AdaBoost (AdaB)

AdaBoost is a boosting technique to improve predictive ability based on the Decision Tree algorithm. Another Boosting technique is Gradient Boosting. AdaBoost's 'ada' is taken from the word adaptive [34]. Introduced for the first time by Freund and Schapire which is the first feasible algorithm for boosting. AdaBoost works well on hard-to-train data. The training process initially predicts the original dataset and assigns equal weight to each training data. If the prediction is wrong using the first training, the weight of the training data will be increased, and so on until the sequence is complete. Since this process is an iterative process, the algorithm continues to add new trainers. That's why it's called adaptive [35]. In this study, the AdaBoost algorithm will be used in the Scikit Learn Library.

2.11 XGBoost (XGB)

XGBoost or Extreme Gradient Boosting is an algorithm that was first introduced by Chen and Guestrin. They improvised the Gradient Boosting algorithm by optimizing and improving the algorithm. This algorithm is done to increase the speed and performance of the classification model [35]. In this study, XGBoost will be used which is already available in the Scikit Learn Library.

2.12 Confusion Matrix and Cross Validation

The confusion matrix technique is used to evaluate and summarize the performance of machine learning classification modeling. The confusion matrix is a 2x2 matrix that summarizes the total correct and incorrect classification results with combinations [36] as shown in *Table 1*.

Table 1: Confusion Matrix 2 Classes

		Actual Classification	
	Label	Positive	Negative
Prediction classification	Positive	TP	FP
	Negative	FN	TN

TP means true positive, FP means false positive, TN means true negative, and FN means false negative. This combination will produce four measurement variables. The accuracy value is obtained through (3), precision is obtained through (4), recall is obtained through (5), and the F1-score is obtained through (6).

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (3)$$

$$Precission = \frac{TP}{TP + FP} \quad (4)$$

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

$$F1 - score = \frac{2\,.\,recall\,.\,precission}{recall + precission} \quad (6)$$

A dataset is initially arbitrarily split into k disjoint folds with about the same number of instances before the outcome is assessed. Then, each fold has a turn evaluating the model that the previous k-1 folds have created. The accuracy estimates variance for statistical inference may be high since the partition is random [36]. In this study, k=10 was used for cross-validation of the classification modeling. Then the parameter of classification modeling needs to be tuned. This study uses hyper tuning of parameters with GridSearchCV from Scikit-learn. Hyper tuning of parameters aims to obtain more effective results and the best performance from the modeling results [24].

2.13 Word Cloud

Basically, a word cloud is a visual representation of text. There are many uses for word clouds. Typically, this may be accomplished using simple text summarizing. The major application of word clouds is in text analytics. The fundamental source of word clouds is a body of literature. Word clouds allow you to see how comparable the information is for a given study project. They sum up single words without understanding their linguistic relationships or meaning. They have little or very limited interaction capabilities and are utilized statistically [37], [38].

2.14 The Gap

We examined study about sentiment analysis of Covid-19 Vaccine in Indonesia from year 2021. The previous study on sentiment analysis of the Covid-19 vaccine only took data over a short span of time, only 1 week. This study uses data over a longer period, 1 month. Previous study also stopped at the

results of the proportion of positive and negative sentiments without finding out the causes. Therefore, in this study visualization of time series is also carried out by searching for words that often appear so that the causes that encourage sentiment in public can be seen [11].

3. METHODS

This research, generally, consists of several stages, starting from data collection, data labeling, data pre-processing, feature extraction, classification modeling, evaluation and validation, sentiment analysis, and social network analysis. These stages are illustrated in *Figure 1*.

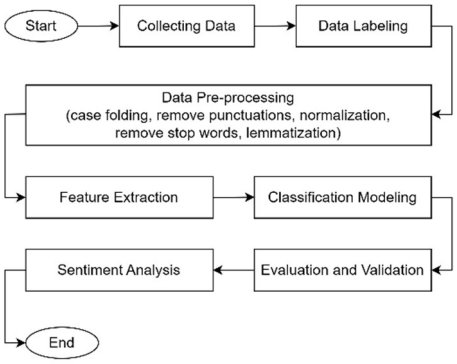

Figure 1: Research Stages

3.1 Data Collecting

At this stage, data is collected from Twitter. The keywords used are "vaksin", "booster", and "mudik". Data collection is done by crawling using the Twitter API [39] via the Tweepy Library [40]. The data that has been collected for this research purpose is 30,582 data, from March 22 to May 02, 2022.

3.2 Data Labelling

At this stage, tweet data labeling (annotation) using the manual method. Annotations are done by two annotators, the authors with a background in information technology studies. Tweet data labeled positive, and negative are then saved into a CSV (comma-separated values) document.

3.3 Data Pre-processing

Pre-processing of data is the first stage in classifying a text which has several stages, case folding, punctuation removal, normalization, stop word removal, and lemmatization. The method of case folding involves changing all text characters to lowercase. Punctuation removal is the process of eliminating any text characters, including numerals, URLs, ASCII codes, and emoji, that contain punctuation marks. Normalization The process of changing words in the form of abbreviations or slang into standard words [41]. In this study, normalization was carried out by standardizing words with facilities from Google Translate.

Next is Stop words removal is the process of eliminating words that often appear in the text that are considered meaningless by using 759 stop word databases in Indonesian [42] and the NLTK (Natural Language Toolkit) library. Lemmatization is the procedure used to standardize text and words based on the fundamental form, or lemma (the word from dictionary form) [43]. In this research, the lemmatization process uses the NLTK library [44].

3.4 Feature Extraction

The next step is feature extraction from a word using TF-IDF (Term Frequency-Inverse Document Frequency). TF-IDF is an algorithm that is commonly used in calculating word weights in a document. TF represents the number of words that appear frequently in a document. Often a word that appears frequently will interfere with the unique word search process. IDF's role is to reduce the weight of a word that often appears and can measure the importance of the meaning of a word in a document [23]. This research uses unigram feature extraction and TF-IDF vectorization.

3.5 Classification Modelling

After going through the word feature extraction stage, the next stage is applying classification modeling with the eight machine learning algorithms approach. There are Naïve Bayes, Support Vector Machine (SVM), Logistic Regression, Decision Tree, K-Nearest Neighbor, Random Forest, AdaBoost, and XGBoost using a Python programming library called Scikit-learn [28]. The classification modeling stage aims to get the best accuracy results among the eight classification models used. To perform classification modeling, training data is needed from the dataset. The training data used in this study is 80% of the total database [29].

3.6 Evaluation and Validation

The last stage in this research is to measure the evaluation of the performance of machine learning classification modeling that has been done in the previous stage. The purpose of using this evaluation measurement is to be able to compare the performance and effectiveness of machine learning classification models used [36]. The confusion matrix technique and k-fold cross validation is conducted. In this study, k=10 was used for cross-validation of the classification modeling. Then the parameter of classification modeling is tuned using GridSearchCV from Scikit-learn [28].

3.7 Sentiment Analysis

At this stage, all the data that has been collected will be predicted for the sentiment results automatically. Sentiment prediction using the best pre-stored modeling. After knowing the results of positive and negative sentiments, the results of the day-to-day sentiments are also displayed time series chart to see the trend. To find out the most used words for each sentiment used Word Cloud. Word cloud is a picture representing words used in a particular dataset. Words with bigger font sizes indicate more frequent appearances than those in smaller sizes [14].

4. RESULT AND DISCUSSION

4.1 Data Collecting

The dataset that was collected is 3,000 data. The dataset's time range is from March 22 to April 12, 2022. Only tweets that are considered the best can be used as a dataset for classification modeling purposes.

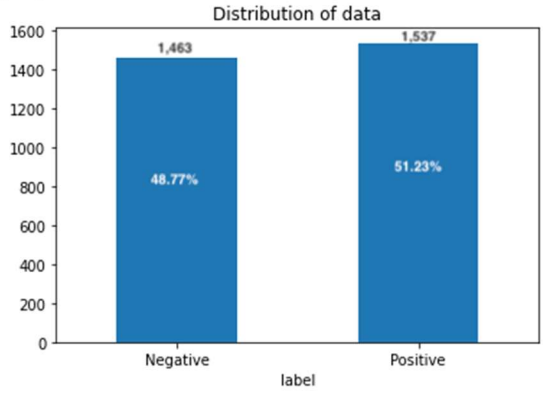

Figure 2: Distribution of Data

In creating the dataset, labeling (annotation) was done manually by two authors. The labeling dataset shows, that 1,463 have negative labels (48.77%) and 1,537 have positive labels (51.23%). *Figure 2* shows the distribution of dataset labels.

4.2 Data Pre-processing Result

After the tweet data is labeled, the data pre-processing stage is carried out in 5 steps, case folding, punctuation removal, normalization, stop words removal, and lemmatization. The results of each data pre-processing are shown in *Table 2*. The tweet has been transformed from raw text into structured text with the process from case folding until lemmatization.

Table 2: Data Preprocessing Result

Stages	Result
Tweet	Dengerin temen sekantor yg cerita gagal vaksin booster hr ini. Alasannya krn vaksin primernya sinopharm. Udah cb tnya ke dinkes, jwbannya tetep ga bs, hrus nunggu vaksin apa gt tp di jatim ga ada. Bingung krn kalo ga vaksin jd ga bs mudik nanti. 😂😂
Case folding	dengerin temen sekantor yg cerita gagal vaksin booster hr ini. alasannya krn vaksin primernya sinopharm. udah cb tnya ke dinkes, jwbannya tetep ga bs, hrus nunggu vaksin apa gt tp di jatim ga ada. bingung krn kalo ga vaksin jd ga bs mudik nanti. 😂😂
Remove punctuations	dengerin temen sekantor yg cerita gagal vaksin booster hr ini alasannya krn vaksin primernya sinopharm udah cb tnya ke dinkes jwbannya tetep ga bs hrus nunggu vaksin apa gt tp di jatim ga ada bingung krn kalo ga vaksin jd ga bs mudik nanti
Normalization	dengarkan rekan kerja yang menceritakan kisah bahwa vaksin booster gagal hari ini alasannya karena vaksin utamanya adalah sinopharm saya sudah coba tanya ke dinas kesehatan jawabannya masih belum berhasil saya harus menunggu vaksin apa tapi di jawa timur belum ada bingung karena kalau tidak ada vaksinnya nanti tidak bisa pulang
Stop words	dengarkan rekan kerja menceritakan kisah vaksin booster gagal alasannya vaksin utamanya sinopharm coba dinas kesehatan jawabannya berhasil menunggu vaksin jawa timur bingung vaksinnya pulang
Lemmatization	dengar rekan kerja cerita kisah vaksin booster gagal alasan vaksin utama sinopharm coba dinas kesehatan jawaban berhasil tunggu vaksin jawa timur bingung vaksin pulang

In the punctuation removal process, emoji had been removed because emoji created from combination of punctuations. In this step, there were word “cb” which is abbreviated from “coba” and

word "tnya" which is typo from "tanya". These kinds of word were normalized into standard word in the normalization step. In the stop word removal step, many words had been deleted because those words have no meaning. Last step is lemmatization, some words that still has suffix or prefix were changed into its base word.

4.3 Classification Modelling Result

Test data taken from the dataset were used to evaluate the prediction results from the classification model processes. In this research, the comparison of training and test data was 80:20, which split was obtained through the Scikit-learn. *Table 3* shows the distribution of training and test data from 3,000 datasets. The number of feature extraction using TF-IDF vectorization on the test data is 2,385 feature words. Then this word feature will be classified using machine learning algorithms modeling.

Table 3: Split of Training Data and Test Data

Label	Training Data	Test Data
Negative	1,238	301
Positive	1,162	299
Total	2,400	600

The 8 classifications modelling need to be evaluated using the confusion matrix to reap the excellent accuracy value, precision, recall, and F1-score. This study uses eight classifiers, NB, SVM, LR, DT, K-Nearest Neighbor KNN, RF, AdaB, and XGB. *Table 4* shows the confusion matrix from eight classifiers.

Table 4: Confusion Matrix Result

Classifier	Accuracy	Precision	Recall	F1-score
NB	0.8483	0.8040	0.9197	0.8580
SVM	0.8900	**0.8770**	0.9063	0.8914
LR	0.8716	0.8580	0.8896	0.8735
DT	0.8516	0.8387	0.8695	0.8538
KNN	0.8133	0.7841	0.8628	0.8216
RF	**0.8933**	0.8753	**0.9163**	**0.8954**
AdaB	0.7850	0.8102	0.7424	0.7748
XGB	0.8016	0.7922	0.8160	0.8039

The results of Table 4 show that the Random Forest classifier gets the best accuracy 0.893 (89.3%), recall 0.916 (91.6%), and F1-score 0.895 (89.5%). The best precision is obtained by SVM with 0.877 (87.7%). The result of Table 4 needs to be validated using k-fold cross-validation, with k value is 10. Table 5 shows the result of the confusion matrix between all classifiers after the cross-validation process. SVM classifier get the best value in all confusion matrix score; accuracy 0.88 (88%), precision 0.881 (88.1%), recall 0.88 (88%), dan F1-score 0.88 (88%).

Table 5: Confusion Matrix After Cross Validation

Classifier	Avg. Accuracy macro	Avg. Precision macro	Avg. Recall macro	Avg. F1-score macro
NB	0.8362	0.8398	0.8347	0.8352
SVM	**0.8804**	**0.8814**	**0.8803**	**0.8801**
LR	0.8429	0.8432	0.8427	0.8426
DT	0.8195	0.8205	0.8200	0.8194
KNN	0.7991	0.8016	0.7979	0.7982
RF	0.8750	0.8758	0.8752	0.8748
AdaB	0.7441	0.7453	0.7444	0.7438
XGB	0.7670	0.7679	0.7676	0.7669

After validation, the next step is hyper tunning of a parameter of the SVM classifier using GridSearchCV from Scikit-learn. The library used in the SVM classifier is LinearSVC. The tuned parameter is the value C. The value of the C parameter in LinearSVC tells the SVM optimization to avoid misclassifying each training example [45]. By assigning a range of C values from 1 to 10, hyper tunning of parameters is carried out with 10-fold cross-validation. The best value of parameter C is 1 with an accuracy score of 0.88 (88%). This accuracy value is similar to the results shown in Table 4 because the previous LinearSVC modeling used the default value, C=1.

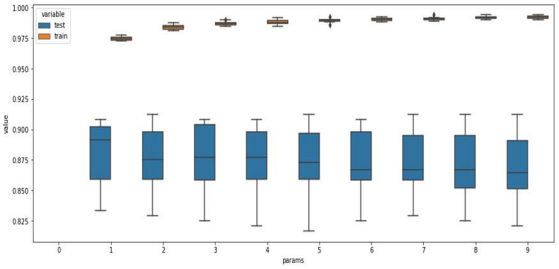

Figure 3: Hyper Tunning of a Parameter (C-Value) for SVM Classifier

Figure 3 shows if the value of C is getting bigger, the accuracy value will decrease. So, the best classifier model from this research is using SVM. Furthermore, the results of this modeling will be saved in pickle (pkl) format. In the future, we can predict the sentiment of the booster vaccine related to mudik Lebaran.

Journal of Theoretical and Applied Information Technology
15th January 2023. Vol.101. No 1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4.4 Sentiment Analysis Result

Total of 30,582 data were collected in the previous stage needs to be labeled with a sentiment. The data in the user tweet are related to booster vaccinations for mudik Lebaran. By using the SVM modeling that was saved before, sentiment labeling can be predicted automatically. The results of the prediction of sentiment towards the government's policy of the COVID-19 booster vaccine for mudik Lebaran were 11,507 giving negative sentiment (37.63%) and 19,075 giving positive sentiment (62.37%). *Figure 4* shows the distribution of sentiment prediction on tweet data related to the booster vaccine for mudik Lebaran.

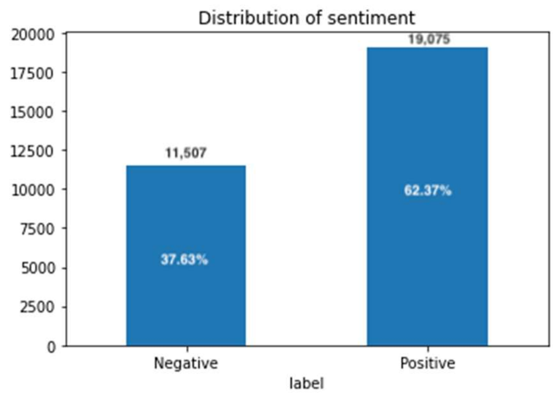

Figure 4: Distribution of Sentiment Prediction

To see the details of the distribution of sentiment every day, it is necessary to create a time series chart shown in *Figure 5*. Daily sentiment results show Twitter users in the first 6 days showed negative sentiment. This is due to the policy that was just issued by the Government regarding the requirements for mudik Lebaran to have a booster vaccine [6]. However, in the next few days, most Twitter users gave positive sentiments. This could be due to the high enthusiasm for mudik Lebaran. This is natural because the Indonesian government in 2020 [46] and 2021 [47] issued a policy prohibiting mudik Lebaran due to the covid-19 pandemic.

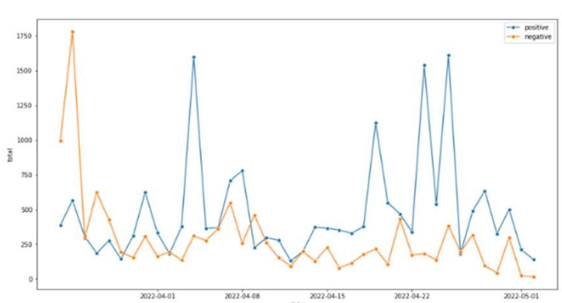

Figure 5: Daily Sentiment Time-Series

The enthusiasm of people who want to mudik Lebaran can be seen at the vaccination center which is always filled with people who want to get a booster vaccine [8].

The highest positive sentiment was achieved a few days before Eid. So, the government's strategy in accelerating the Covid-19 booster vaccination program was accepted by most people by making it a requirement for mudik Lebaran. If we look closely at earlier line from *Figure 5*, proportion of negative sentiment is higher than positive, but then the proportion is changed as following days. For further analysis, we want to see word cloud visualization first.

To find out what words often appear in negative sentiments, a word cloud is used as shown in *Figure 6*. There are the words "syarat" (requirement) and "pulang" (going home) that appear the most. This shows that people are burdened with the booster vaccine policy as a requirement for going home [9]. There are also the words "motogp", "presiden" (president), and "kerumuman" (crowd). Many people compare it to the MotoGP event in Mandalika, Lombok. The president was present at the event and there was a crowd, but the government did not require booster vaccinations for spectators [48].

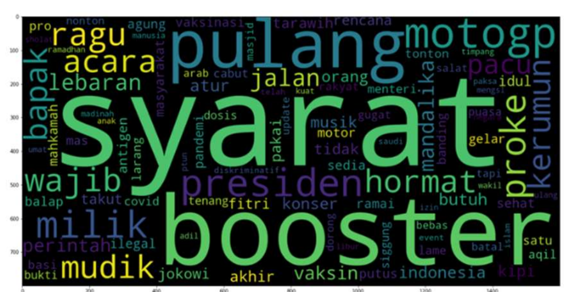

Figure 6: Word Cloud for Negative Sentiment

On the other hand, *Figure 7* shows the word cloud of positive sentiment. a lot of positive sentiment words, such as "sehat" (healthy), "aman" (safe), and "strong" (kuat). This is the public's optimism that the booster vaccination program is safe and makes it healthy and strong in warding off the COVID-19 pandemic [49]. This shows that many people are very enthusiastic and support this government policy. Other words that appear only relate to mudik Lebaran activities.

Figure 7: Word Cloud for Positive Sentiment

From our perspective when analyzing the words that occur frequently from the word cloud image, we realize that we also want to know how often the words occur frequently. Incidentally, the visualization of the word cloud in *Figure 6* and *Figure 7* is quite significant, so what if there are 2 or 3 words with a balanced frequency, it will certainly be difficult to distinguish which one is bigger. For that we also try to visualize it on a table form by taking the 5 words that appear the most for negative and positive sentiments.

Table 6: Word Appearance for Whole Data

No	Negative		Positive	
	Word	Count	Word	Count
1	syarat	5532	sehat	4883
2	booster	2783	lebaran	4624
3	pulang	2739	vaksinasi	5215
4	motogp	2573	mudik	3447
5	presiden	2201	aman	3170

From *Table 6* the visualization is clearer because it is quantified. With a table like that, we can analyze further on the visualization of the time series in *Figure 5* about the change in the proportion of negative sentiment. We extracted word frequency only on days where there was a significant peak, namely on 3/23, 4/4, 4/19, 4/23, and 4/25.

From *Table 7* we highlighted the words that occur most frequently on 5 different days at the peak. On 3/23 the word “syarat” (requirement) is the most frequently seen. This is because on this date the booster vaccination policy is a requirement for going home for Eid. The public responded reactively by comparing it to the MotoGP event in Mandalika where booster vaccination was not a requirement, so this created a negative sentiment [48]. However, over time, on 4/4 the word “sehat” (healthy) appeared most often, as well as on 4/19. This indicates that people are starting to realize the benefits of booster vaccination, namely for health to avoid Covid. On that date, the word "syarat" (requirement) which became a negative scourge on 3/25 was no longer visible.

Table 7: Word Appearance for Peak Point Dates

Date	No	Negative		Positive	
		Word	Count	Word	Count
3/25	1	**syarat**	**1631**	lebaran	348
	2	presiden	1347	syarat	310
	3	milik	1342	booster	249
	4	motogp	1311	izin	233
	5	prokes	1249	sedia	218
4/4	1	sehat	192	**sehat**	**1202**
	2	tarawih	103	kuat	678
	3	sholat	97	vaksinasi	322
	4	jokowi	96	cepat	219
	5	disiplin	96	lebaran	198
4/19	1	perintah	227	**sehat**	**814**
	2	lebaran	117	nyaman	679
	3	butuh	113	vaksinasi	349
	4	telah	112	disiplin	299
	5	saudi	112	prokes	204
4/23	1	papua	63	**lindung**	**1234**
	2	pulang	44	kuat	665
	3	martabat	38	papua	384
	4	lindung	33	covid	281
	5	pacu	32	lebaran	231
4/25	1	tenang	280	**tenang**	**1275**
	2	jokowi	173	kuat	437
	3	goreng	94	lebaran	404
	4	minyak	94	vaksinasi	395
	5	larang	93	papua	276

Continued to 4/23 where the word “lindung” (protect) became the most frequent, followed by the word “kuat” (strong). On that date, the public talked about the benefits of booster vaccination when going home, namely so that the body of the person who is going home will be strong so that it will protect the family in the village from Covid. On 4/25 the word "lindung" was replaced with the word "tenang" (calm), but both were followed by the word "kuat”. On that date, the public discussed that they should immediately carry out a booster vaccination so that when they go home, they will be calm and not be haunted by the side effects of the booster which can cause fever and headaches [9].

As Eid approaches on May 2, people are becoming more and more aware of the main reason for the booster vaccination policy as a condition for going home, which is to protect families in the village from being infected so that the transmission of covid can be suppressed and closer to herd immunity [49].

With the fact that booster vaccination, which is a requirement, was received positively by the public, we compared with the booster vaccine uptake data. On May 9 the booster vaccine had an uptake of 19.07% [50], an increase of 11.42% from the initial policy issued which is 7.65% [5]. Although the uptake is not as much as the first and second vaccines, this policy is sufficient to increase the booster vaccine absorption by more than 10%.

Even though this research has carried out an analysis of words that often appear to provide an overview of what is being discussed, it would be better if topic modeling was also carried out. This is so that the topics hidden in the document can be seen to give an idea of what was discussed, thereby encouraging sentiment. With a clearer picture, a recommended solution can also be formulated for the government to make its vaccine program successful.

For future study, we recommend that time series and analysis of frequently occurred word are always conducted in sentiment analysis study. For more advance analysis, using topic modelling will give more enrichment for the study instead of just using word frequency. This makes study of sentiment analysis should always visualize it as time series to represent public opinion whether it daily, weekly, or monthly.

5. CONCLUSION

We conducted sentiment analysis about Covid-19 booster vaccine from Twitter. Classification modeling using SVM has the best value in terms of all assessment components which include accuracy, precision, recall and F1-score. Using 8 text classification algorithms made us got the best choice to determine the model to be used for sentiment prediction. The distribution of sentiment is also quite different, with a negative sentiment value of 37.63% and a positive sentiment of 62.37%. From this, more people gave positive sentiment regarding booster vaccination as a requirement for mudik Lebaran.

Instead of just getting sentiment predictions from the entire dataset, we further examined what drives these sentiments with time series visualization and analysis of frequently occurring words. Even though overall positive sentiment prevailed, in the first week, negative sentiment prevailed. We also got an overview of the causes by analyzing the words that often appeared, namely due to the initial reaction from the public comparing the mandatory booster vaccine policy for going home compared to the implementation of MotoGP which is not required for booster vaccines. However, in the following weeks, positive sentiment gradually prevailed. The public's reaction to MotoGP was decreased and talk about the importance of booster vaccines was became a topic of conversation. People were become aware that booster vaccines could protect travelers and their families from the dangers of Covid-19.

This study shows the importance of time series visualization and word analysis which often appears in sentiment analysis studies. Research to find the right algorithm model is indeed a good thing to do, but it shouldn't stop there, but do further research about why that sentiment can arise. Time series visualization also shows that sentiment can change from time to time, therefore sentiment analysis research should not stop at sentiment predictions.

ACKNOWLEGDEMENT:
This research was sponsored by Ministry of Communication and Informatics (KOMINFO) Republic of Indonesia.

REFERENCES:

[1] M. A. Shereen, S. Khan, A. Kazmi, N. Bashir, and R. Siddique, "COVID-19 infection: Origin, transmission, and characteristics of human coronaviruses," *Journal of Advanced Research*, vol. 24. Elsevier B.V., pp. 91–98, Jul. 01, 2020. doi: 10.1016/j.jare.2020.03.005.

[2] WHO, "WHO Coronavirus (COVID-19) Dashboard," 2022. https://covid19.who.int/table (accessed Mar. 04, 2022).

[3] C. A. Tisdell, "Economic, social and political issues raised by the COVID-19 pandemic," *Econ Anal Policy*, vol. 68, pp. 17–28, 2020, doi: 10.1016/j.eap.2020.08.002.

[4] E. Mathieu *et al.*, "A global database of COVID-19 vaccinations," *Nat Hum Behav*, vol. 5, no. 7, pp. 947–953, Jul. 2021, doi: 10.1038/S41562-021-01122-8.

[5] Kemenkes, "Vaksin Dashboard," *Kemenkes*, 2022. https://vaksin.kemkes.go.id/#/vaccines (accessed Mar. 30, 2022).

[6] L. S. Rahayu, "Ma'ruf Amin Ungkap Kemungkinan Vaksin Booster Jadi Syarat untuk Mudik," *Detik.com*, 2022. https://news.detik.com/berita/d-5995581/maruf-amin-ungkap-kemungkinan-

vaksin-booster-jadi-syarat-untuk-mudik (accessed Mar. 30, 2022).

[7] V. I. Yulianto, "Is the Past Another Country? A Case Study of Rural Urban Affinity on Mudik Lebaran in Central Java," *Journal of Indonesian Social Sciences and Humanities*, vol. 4, pp. 49–66, Mar. 2019, doi: 10.14203/jissh.v4i0.118.

[8] M. I. Bustomi and I. A. Arbi, "Vaksinasi Booster Jadi Syarat Mudik, Antusiasme Warga Jaksel Meningkat Dua Kali Lipat," *Kompas.com*, 2022. https://megapolitan.kompas.com/read/2022/04/21/17150501/vaksinasi-booster-jadi-syarat-mudik-antusiasme-warga-jaksel-meningkat-dua (accessed Jun. 18, 2022).

[9] D. Kurnia and I. Tirta, "Syarat Vaksin Booster untuk Mudik Dirasa Memberatkan," *Repubika*, 2022. https://www.republika.co.id/berita/r98zh1485/syarat-vaksin-booster-untuk-mudik-dirasa-memberatkan (accessed Jun. 18, 2022).

[10] B. Liu, *Sentiment Analysis and Opinion Mining*. 2012. doi: 10.5120/ijca2016911545.

[11] Pristiyono, M. Ritonga, M. A. al Ihsan, A. Anjar, and F. H. Rambe, "Sentiment Analysis of COVID-19 Vaccine in Indonesia using Naïve Bayes Algorithm," *IOP Conf Ser Mater Sci Eng*, vol. 1088, no. 1, p. 012045, Feb. 2021, doi: 10.1088/1757-899x/1088/1/012045.

[12] Z. Bokaee Nezhad and M. A. Deihimi, "Twitter Sentiment Analysis from Iran About COVID 19 Vaccine," *Diabetes and Metabolic Syndrome: Clinical Research and Reviews*, vol. 16, no. 1, Jan. 2022, doi: 10.1016/j.dsx.2021.102367.

[13] I. Hasti, A. Nurmandi, I. Muallidin, D. Kurniawan, and Salahudin, "Pros and Cons of Vaccine Program in Indonesia (Social Media Analysis on Twitter)," in *Lecture Notes in Networks and Systems*, 2022, vol. 319, pp. 100–107. doi: 10.1007/978-3-030-85540-6_13.

[14] F. Heimerl, S. Lohmann, S. Lange, and T. Ertl, "Word Cloud Explorer: Text Analytics Based on Word Clouds," in *2014 47th Hawaii International Conference on System Sciences*, Mar. 2014, pp. 1833–1842. doi: 10.1109/HICSS.2014.231.

[15] A. W. Pradana, A. Y. Asmara, B. Triyono, R. Jayanthi, A. Dinaseviani, and Purwadi, "Analisis Desk Research Kebijakan Technology Transfer Office Sebagai Solusi Hambatan Teknologi Transfer di Lembaga Litbang Indonesia," *Matra Pembaruan: Jurnal Inovasi Kebijakan*, 2021.

[16] F. M. Alliheibi, A. Omar, and N. Al-Horais, "Opinion Mining of Saudi Responses to COVID-19 Vaccines on Twitter A Computational Linguistic Approach," 2021. [Online]. Available: www.ijacsa.thesai.org

[17] H. Budhwani, T. Maycock, W. Murrell, and T. Simpson, "COVID-19 Vaccine Sentiments Among African American or Black Adolescents in Rural Alabama," *Journal of Adolescent Health*, vol. 69, no. 6, pp. 1041–1043, Dec. 2021, doi: 10.1016/j.jadohealth.2021.09.010.

[18] M. A. Abdelaal, M. A. Fattah, and M. M. Arafa, "Predicting Sarcasm and Polarity in Arabic Text Automatically: Supervised Machine Learning Approach," *J Theor Appl Inf Technol*, vol. 100, no. 8, pp. 2550–2560, 2022, [Online]. Available: www.jatit.org

[19] E. al Maghayreh, "Using Machine Learning to Predict The Sentiment of Arabic Tweets Related to Covid-19," *J Theor Appl Inf Technol*, vol. 31, no. 14, pp. 5368–5375, 2022, [Online]. Available: www.jatit.org

[20] D. Effrosynidis, S. Symeonidis, and A. Arampatzis, "A Comparison of Pre-processing Techniquesfor Twitter Sentiment Analysis," in *TPDL 2017: Research and Advanced Technology for Digital Libraries*, 2017, pp. 394–406. doi: 10.1007/978-3-319-67008-9.

[21] C. Yan, M. Law, S. Nguyen, J. Cheung, and J. Kong, "Comparing public sentiment toward COVID-19 vaccines across Canadian cities: Analysis of comments on reddit," *J Med Internet Res*, vol. 23, no. 9, Sep. 2021, doi: 10.2196/32685.

[22] P. Sv, J. Tandon, Vikas, and H. Hinduja, "Indian citizen's perspective about side effects of COVID-19 vaccine – A machine learning study," *Diabetes and Metabolic Syndrome: Clinical Research and Reviews*, vol. 15, no. 4, Jul. 2021, doi: 10.1016/j.dsx.2021.06.009.

[23] A. S. Neogi, K. A. Garg, R. K. Mishra, and Y. K. Dwivedi, "Sentiment Analysis and Classification of Indian Farmers' Protest Using Twitter Data," *International Journal of Information Management Data Insights*, vol. 1, no. 100019, pp. 1–11, 2021, doi: 10.1016/j.jjimei.2021.100019.

[24] Pirjatullah, D. Kartini, D. T. Nugrahadi, Muliadi, and A. Farmadi, "Hyperparameter Tuning using GridsearchCV on the

Comparison of the Activation Function of the ELM Method to the Classification of Pneumonia in Toddlers," in *Proceedings - 2021 4th International Conference on Computer and Informatics Engineering: IT-Based Digital Industrial Innovation for the Welfare of Society, IC2IE 2021*, 2021, pp. 390–395. doi: 10.1109/IC2IE53219.2021.9649207.

[25] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093–1113, 2014, doi: 10.1016/j.asej.2014.04.011.

[26] C. A. Melton, O. A. Olusanya, N. Ammar, and A. Shaban-Nejad, "Public sentiment analysis and topic modeling regarding COVID-19 vaccines on the Reddit social media platform: A call to action for strengthening vaccine confidence," *J Infect Public Health*, vol. 14, no. 10, pp. 1505–1512, Oct. 2021, doi: 10.1016/j.jiph.2021.08.010.

[27] H. Lee *et al.*, "COvid-19 vaccine perception in south korea:web crawling approach," *JMIR Public Health Surveill*, vol. 7, no. 9, Sep. 2021, doi: 10.2196/31409.

[28] Scikit-learn, "Scikit-learn: Machine Learning in Python," *scikit-learn.org*, 2022. https://scikit-learn.org/stable/ (accessed Nov. 24, 2021).

[29] A. Alabrah, H. M. Alawadh, O. D. Okon, T. Meraj, and H. T. Rauf, "Gulf Countries' Citizens' Acceptance of COVID-19 Vaccines—A Machine Learning Approach," *Mathematics*, vol. 10, no. 3, p. 467, Jan. 2022, doi: 10.3390/math10030467.

[30] C. Villavicencio, J. J. Macrohon, X. A. Inbaraj, J. H. Jeng, and J. G. Hsieh, "Twitter sentiment analysis towards covid-19 vaccines in the Philippines using naïve bayes," *Information (Switzerland)*, vol. 12, no. 5, May 2021, doi: 10.3390/info12050204.

[31] S. F. N. Hadju and R. Jayadi, "Sentiment Analysis of Indonesian E-Commerce Product Reviews Using Support Vector Machine Based Term Frequency Inverse Document Frequency," *J Theor Appl Inf Technol*, vol. 99, no. 17, pp. 4316–4325, 2021, [Online]. Available: www.jatit.org

[32] M. Ahmad, S. Aftab, and I. Ali, "Sentiment Analysis of Tweets using SVM," *Int J Comput Appl*, vol. 177, no. 5, pp. 25–29, Nov. 2017, doi: 10.5120/ijca2017915758.

[33] S. Rani and N. Singh Gill, "Hybrid Model for Twitter Data Sentiment Analysis Based on Ensemble of Dictionary Based Classifier and Stacked Machine Learning Classifiers-SVM, KNN and C5.0," *J Theor Appl Inf Technol*, vol. 98, no. 04, pp. 624–635, 2020, [Online]. Available: www.jatit.org

[34] S. M. Alsubaie, K. M. Almutairi, N. A. Alnuaim, R. A. Almuqbil, N. Aslam, and I. Ullah, "Automatic Semantic Sentiment Analysis on Twitter Tweets Using Machine Learning: A Comparative Study," *J Theor Appl Inf Technol*, vol. 15, no. 23, 2019, [Online]. Available: www.jatit.org

[35] N. Weeraddana and S. Premaratne, "Unique Approach for Cricket Match Outcome Prediction Using Xgboost Algorithms," *J Theor Appl Inf Technol*, vol. 15, no. 9, 2021, [Online]. Available: www.jatit.org

[36] T. T. Wong and P. Y. Yeh, "Reliable Accuracy Estimates from k-Fold Cross Validation," *IEEE Trans Knowl Data Eng*, vol. 32, no. 8, pp. 1586–1594, Aug. 2020, doi: 10.1109/TKDE.2019.2912815.

[37] A. I. KABIR, K. AHMED, and R. KARIM, "Word Cloud and Sentiment Analysis of Amazon Earphones Reviews with R Programming Language," *Informatica Economica*, vol. 24, no. 4/2020, pp. 55–71, Dec. 2020, doi: 10.24818/issn14531305/24.4.2020.05.

[38] F. Heimerl, S. Lohmann, S. Lange, and T. Ertl, "Word cloud explorer: Text analytics based on word clouds," in *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2014, pp. 1833–1842. doi: 10.1109/HICSS.2014.231.

[39] Twitter, "Twitter API Documentation," *Twitter.com*. https://developer.twitter.com/en/docs/twitter-api (accessed Apr. 02, 2022).

[40] Tweepy, "Tweepy Documentation," *Tweepy.org*. https://docs.tweepy.org/en/stable/ (accessed Mar. 04, 2022).

[41] D. A. Nurdeni, I. Budi, and A. B. Santoso, "Sentiment Analysis on Covid19 Vaccines in Indonesia: From the Perspective of Sinovac and Pfizer," in *3rd 2021 East Indonesia Conference on Computer and Information Technology, EIConCIT 2021*, Apr. 2021, pp. 122–127. doi: 10.1109/EIConCIT50028.2021.9431852.

[42] F. Z. Tala, *A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia*.

Institute for Logic, Language, and Computation Universiteit van Amsterdam, 2003.

[43] M. Nutu, "Deep Learning Approach for Automatic Romanian Lemmatization," in *Procedia Computer Science*, 2021, vol. 192, pp. 49–58. doi: 10.1016/j.procs.2021.08.006.

[44] NLTK, "NLTK-Natural Language Toolkit," *nltk.org*. https://www.nltk.org/ (accessed Jun. 17, 2022).

[45] Scikit-learn, "sklearn.svm.LinearSVC," *scikit-learn.org*. https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html (accessed Jun. 18, 2022).

[46] N. R. Aida and V. R. Ratriani, "Larangan Mudik 2020, antara Cegah Virus Corona dan Dampak Ekonomi," *Kompas.com*, 2020. https://www.kompas.com/tren/read/2020/04/22/133325865/larangan-mudik-2020-antara-cegah-virus-corona-dan-dampak-ekonomi (accessed Jun. 18, 2022).

[47] P. Mutiara, "Pemerintah Larang Mudik Lebaran 2021," *Kemenko PMK*, 2021. https://www.kemenkopmk.go.id/pemerintah-larang-mudik-lebaran-2021 (accessed Jun. 18, 2022).

[48] N. R. Aida and I. E. Prastiwi, "Mengapa Vaksin Booster Tak Wajib di MotoGP tapi Jadi Syarat Mudik? Ini Jawaban Satgas," *Kompas.com*, 2022. https://www.kompas.com/tren/read/2022/03/25/133500465/mengapa-vaksin-booster-tak-wajib-di-motogp-tapi-jadi-syarat-mudik-ini (accessed Jun. 18, 2022).

[49] R. Sulaiman, "Kunci Mudik Sehat dan Aman, Jangan Lupa Vaksin Booster dan Disiplin Protokol Kesehatan," *Suara.com*, 2022. https://www.suara.com/health/2022/04/25/143902/kunci-mudik-sehat-dan-aman-jangan-lupa-vaksin-booster-dan-disiplin-protokol-kesehatan (accessed Jun. 18, 2022).

[50] WHO, "Indonesia: WHO Coronavirus Disease (COVID-19) Dashboard With Vaccination Data | WHO Coronavirus (COVID-19) Dashboard With Vaccination Data," 2022. https://covid19.who.int/data (accessed Oct. 26, 2022).