

Multilabel sentiment analysis for classification of the spread of COVID-19 in Indonesia using machine learning

by Fathoni Fathoni

Submission date: 12-Jun-2023 02:27PM (UTC+0700)

Submission ID: 2114317246

File name: 39_30578.pdf (518.24K)

Word count: 7191

Character count: 36614

Multilabel sentiment analysis for classification of the spread of COVID-19 in Indonesia using machine learning

Fathoni¹, Erwin², Abdiansah³

¹Department of Informatics System, University of Sriwijaya, Indralaya, Indonesia

²Department of Computer Engineering, Computer Science Faculty, Universitas Sriwijaya, Indralaya, Indonesia

³Department of Informatics Technique, Computer Science Faculty, Universitas Sriwijaya, Indralaya, Indonesia

Article Info

Article history:

Received Oct 14, 2022

Revised Mar 28, 2023

Accepted Apr 2, 2023

Keywords:

COVID-19

Decision tree

Indonesia

K-nearest neighbor

Naïve Bayes

Sentiment analysis

ABSTRACT

This study aims to use datasets on Twitter to find out public opinion on the spread of coronavirus in Indonesia by conducting sentiment analysis. The resulting sentiment analysis will benefit the community by helping the Indonesian government take various strategic measures to prevent and counter the spread of the COVID-19. This research was conducted through the data collection stage, namely crawling data tweet words in Bahasa Indonesia containing the meaning of the spread of COVID-19, the next stage of the process of creating labels manually. Next, the pre-process stage by removing the character, symbols and special features from Twitter. The last stage, classification using learning machine with 3(three) methods namely K-nearest neighbor (K-NN), Naïve Bayes and decision tree. The study analyzed sentiment of 1,119 valid Tweets data and found that K-NN algorithm had the highest accuracy value compared to Naïve Bayes and decision tree algorithms, which was 95.10%. However, the Twitter data analyzed obtained 78.19% of Tweets that fall into the negative category and only 13.85% of public opinion that is positive. This indicates that most of the Tweets of Indonesians in twitter do not mean the spread of COVID-19 disease somewhere.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Erwin

Department of Computer Engineering Computer Science Faculty, Sriwijaya University

Indralaya, South Sumatra, Indonesia

Email: erwin@unsri.ac.id

1. INTRODUCTION

A new disease called Coronavirus Disease (COVID-19) caused by the Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) virus [1]-[3] that appeared in Hubei province of Wuhan, China [4], has horrified the world. The spread of COVID-19 disease in Indonesia was first reported in early March 2020 [5], [6]. Within five months, 130,718 Indonesians tested positive for COVID-19 and 5,903 people died [7]. July 2021, the disease has spread very quickly to various regions and is transmitted to 2,228,938 Indonesians with 59,534 deaths [8]. The procedure to obtain data affected by COVID-19 disease is carried out through a manual data collection mechanism through health centers, government and private hospitals, health clinics spread throughout Indonesia as well as rapid tests conducted at specific times and locations. Surveillance systems like this require time, a lot of health workers and expensive costs. In addition, the geographical condition of Indonesia consisting of many large and small islands and the area of Indonesia requires another strategy to know and facilitate the complete data of people affected by COVID-19, such as the utilization of information technology.

The rapid use of information technology, especially mobile technology including in Indonesia, makes it very easy for people to obtain and share information quickly and widely. One of the most widely used mobile

technology trends of the world community [9], [10] and Indonesia's sharing of information is social media Twitter. Indonesians who regularly Tweets number 78 million people from 150 million active users of social media. The use of information from Twitter as a dataset to detect the spread of disease in regions or countries has been widely done by researchers [11], [12]. To obtain the dataset needed on Twitter, the researchers crawled the data by utilizing the functions provided by Twitter and developed their own new additional functions tailored to the characteristics and topics of their respective studies. Based on the results of the published study, Crawling Twitter developed by the researchers has a level of accuracy that can still be improved, especially for the results of crawling data Twitter spread COVID-19 disease in Indonesia.

The complexity of achieving a good level of Twitter crawling accuracy is influenced by several factors [13], such as: i) a person's psychological condition (such as being angry and happy); ii) characteristics and behaviors of Twitter users; iii) synonyms and vocabulary are very many and mean the same; iv) the total number of downloadable Tweets is limited. These factors cause the catch data in Twitter crawling to contain bias and may lead to errors in the process of drawing a conclusion as well as further action. The conclusions made will be even more dangerous if the public opinion displayed in the Tweetser turns out to contain elements of hoaxes.

Crawling information from Twitter has been widely used by researchers to analyze the influence and development of pandemic COVID-19 [14], [15], including researchers in Indonesia [16], [17]. This provides opportunities for interesting new research topics such as to know the opinions of the Indonesian people on the spread of COVID-19 in Indonesia. The sentiment of Indonesian people's analysis of the COVID-19 pandemic in Tweetsters is not an easy thing to do manually. This difficulty is caused by the large number of Tweets as well as other complexity factors. Therefore, a special method is needed that can make it easier to analyze public opinion in Tweetsters automatically, and can identify the existence of the COVID-19 pandemic in Indonesia through sentiment analysis. Sentiment analysis is an investigation of an event or event that informs about one's behavior, emotions and opinions [15]. The results of sentiment analysis of Indonesian people on Twitter will be processed and evaluated using machine learning approach. Machine learning is necessary because the resulting sentiment analysis has not produced much information needed because there is still a total recapitulation of each level of sentiment.

Research on the use of Machine learning to build models that can identify and identify Tweets indicating the presence of disease in Indonesia has been conducted by several previous researchers [11], [18], [19]. The use of machine learning in this study is intended to build a classification model and group the results of the identification of COVID-19 disease data Tweets in Indonesia using naïve bayes algorithm, algorithm decision tree and K-nearest neighbor (K-NN) algorithm. The use of these three algorithms aims to know the classification model that best suits the form of datasets processed so that it will produce the best level of accuracy.

The novelty found in this research is that the sentiment analysis of Twitter users in Indonesia is mostly in a negative category, which indicates that most of the Tweets of the Indonesian people in Twitter do not mean the spread of the disease COVID-19. The truth of the information conveyed on Twitter cannot be ascertained [10], this is the latest research offer that must be carried out, so that information on the spread of COVID-19 conveyed on social media can be trusted. The most contribution obtained in this research is beneficial to the people of Indonesia, namely helping the government to take various strategic actions to prevent and control the spread of the COVID-19 virus. The structure of this paper consists of section 2 presents research methods, data collection, category identification, pre-processing, classification using machine learning, section 3 discusses the results and analysis, and section 4 presents the conclusions.

2. METHOD

The study has several steps, namely crawling, identifying COVID-19 categories, pre-processing, classification using machine learning, and evaluation in Figure 1. The research stages starting from the early stages of research focused on building the twitter data corpus and vocabulary corpus. The first process that will be carried out is to collect natural language datasets from Indonesian-language Tweets on Twitter. Next, carry out the process of extracting incident sentence information with steps, namely making standard sentence formats, simplifying sentences, identifying important words in sentences, and determining input and target words in sentences.

2.1. Data collection

Twitter is one of the social media that is growing rapidly with an increasing number of users throughout the world, including in Indonesia. Basically, tweets on Twitter are text. Text is a sentence of information and knowledge that is disseminated at a certain time and medium [20]. In the current information era, there is a lot of information available on the internet in the form of text from various types of documents such as research documents, magazines, electronic books and articles that are unstructured (unstructured data),

including emails, pdf files, social media (Twitter, Facebook), and others, video, audio, image and business content in bulk [21]. The volume of text documents is growing rapidly and experts predict a growth of 80% by 2025. Search technologies states that 80% of data in organizations is unstructured data. This stage of collecting data through Tweets using keywords related to the spread of COVID-19 in Indonesia by utilizing the Twitter search API operator function contained in the RapidMiner v9.9 application. The results of Tweets data collection will be processed manually to determine information relevant to the spread of COVID-19 in Indonesia through identical categories of sentences containing the meaning of contracting COVID-19 disease.

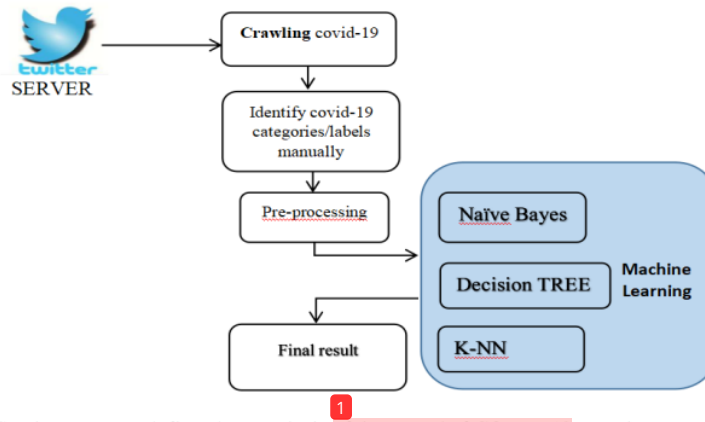


Figure 1. Sentiment research flowchart analysis of the spread of COVID-19 on twitter users in Indonesia

2.2. Pre-processing

Pre-processing is a stage to get a good dataset and reduce data inconsistencies and improve quality results. This step will be done to remove characters and symbols and eliminate unnecessary Twitter-specific features. The data from the category identification process still contains foreign characters, symbols and words that are not common to the Indonesian people. Pre-processing process is done with the following stages:

- Case folding: This first step is the process to change the characters contained in a sentence in a document (crawling result) to be converted to a standard sentence. Standardization of these characters is very important in order to facilitate the process of searching for the desired word in a sentence.
- Tokenizing (Parsing): Parsing is the process of cutting sentences in a document into several parts of a word as well as performing the process of removing certain special characters or symbols that are not needed.
- Stop word removal: Stop word removal is the process of removing words that are considered unimportant and do not affect the meaning of the sentences contained in the document. This process is very important to facilitate and speed up the process of selecting words to be searched for at a later stage.
- Feature selection: This step will perform the process of selecting the words contained in the document that correspond to the search keyword.

2.3. Classification with machine learning

Classification in machine learning is a grouping of data where the data used has a label or target class. So that the algorithms for solving classification problems are categorized into supervised learning or supervised learning. To measure the accuracy of sentiment analysis will be used algorithm Naïve Bayes and algorithm decision tree and algorithm K-NN. The use of these three different algorithms is intended to know and obtain the algorithm that best matches the resulting dataset and analysis that has the highest accuracy level value.

2.3.1. K-nearest neighbor

Many approaches are for estimating the conditional distribution of Y given X, and then classifying the observations given to the class with the highest probability estimate. One such method is the K-NN classifier. Given a positive K-nearest integer K and a test observation of x_0 , the K-NN classifier first identifies its neighbor K points in the training data that is closest to x_0 , represented by N_0 . To determine the distance of data points, the K-NN Euclidean distance algorithm [22] is used (1). Then estimate the conditional probability

for class j as a fraction of pointing to N_0 whose response value equals j in (2). Finally, the K-NN applies Bayes' rule and classifies the x_0 test observations into the class with the greatest probability.

$$d(x, x') = \sqrt{(x_1 - x'_1)^2 + \dots + (x_n - x'_n)^2} \quad (1)$$

$$Pr(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in N_0} I(y_i = j) \quad (2)$$

The K-NN or K-Nearest Neighbor Algorithm 1 is one of the algorithms that is widely used in machine learning for classification cases. The K-NN algorithm is a classification algorithm that works by taking a number of K nearest data (neighbors) as a reference for determining the class of new data. This algorithm classifies data based on similarity or similarity or proximity to other data. In general, the way the K-NN algorithm works is to determine the number of neighbors (K) that will be used for class determination considerations. Calculate the distance from the new data to each data point in the dataset. Take a number of K data with the shortest distance, then determine the class of the new data.

Algorithm 1. K-NN

Input: x, S, d, k

Output: class of x

Set k value

for $(x', l') \in S$:

Compute the Distance $d(x', x)$

(Euclidean Distance, Jaccard, ect)

end for

Sort the $|S|$ distances in descending order

Take the top- k distance

Count the number of occurrences of each class l_j

Assign label to x based on the most frequent class

2.3.2. Naïve Bayes

Naïve Bayes is a classification method based on probability and is designed to be used with the assumption that one class is independent of each other. In the Naïve Bayes classification, the learning process is more focused on estimating probabilities, and the determination of the probability of an event based on the distribution of the probability that the event occurs from other specified events [23], [24]. The advantage of this approach is that the classification will get a smaller error value when the data set is large. The Naïve Bayes formulation for the (3).

$$P(Y|X) = \frac{P(Y) \prod_{i=1}^q P(X_i|Y)}{P(X)} \quad (3)$$

Where:

$P(Y|X)$ is the probability of data with vector X in class Y

$P(Y)$ is the initial probability of class Y (prior probability)

$\prod_{i=1}^q P(X_i|Y)$ is the Y class independent probability of all features in vector X

$P(X)$ is the probability of X .

The probability of $P(X)$ is always constant so that in the calculation of predictions later it can be ignored and only calculate the part of $P(Y) \prod_{i=1}^q P(X_i|Y)$ only by choosing the largest value as the class of prediction results or commonly known as Maximum A Posteriori (MAP) where this MAP can be denoted by (4).

$$MAP = arg(max (P(Y) \prod_{i=1}^q P(X_i|Y))) \quad (4)$$

While the probability of independence $\prod_{i=1}^q P(X_i|Y)$ is the effect of all features of the data on each class Y , which is denoted by (5),

$$P(X|Y = y) = \prod_{i=1}^q P(X_i|Y = y) \quad (5)$$

each feature set $X = [X_1, X_2, \dots, X_q]$ consists of q attributes.

Naive Bayes algorithm which is a type of supervised learning Algorithm 2, where the algorithm cannot learn on its own but must be given an example first by labeling the dataset that we have. Labeling here means that our dataset has been given a truth value which will be used as a target value or reference value. Naive

Bayes is the most popular classification method used with a good level of accuracy. Naive Bayes is a simple probability-based classification method and is designed to be used with the assumption that the explanatory variables are independent. In this algorithm learning is more emphasized on probability estimation. The advantage of the Naive Bayes algorithm is that the error rate is lower when the dataset is large, besides that the accuracy of Naive Bayes and the speed is higher when applied to a larger dataset.

Algorithm 2. Naïve Bayes

```

1. for q = 1 ... s //loop for each mining model's element
2.     μ[q] = 0; // initialization of mining model's //elements
3. for j = 1 ... μ // loop for each vector
4.     μ[d[j,p]]++ //increment count of vectors for value // xj,k of vector xj
5.     for k = 1 ... //loop for each attribute
6.         p-1
           μ[φ(k-
           1) + (d[j,k - 1] · φ(0) +
           d[j,p]]++) //increment count of vectors for value // xj,k of vector xj,p
7.     end for
8. end for
    
```

2.3.3. Decision tree

A decision tree is a structure that contains nodes and edges and is built from a dataset (table of columns representing features/attributes and rows corresponds to records). Each node is either used to make a decision (known as decision node) or represent an outcome (known as leaf node). The model given the training uses the training data to estimate the results of the test data in the prediction phase [25], [26].

ID3 stands for Iterative Dichotomiser 3 and is named such because the algorithm iteratively (repeatedly) dichotomizes (divides) features into two or more groups at each step. ID3 uses a top-down greedy approach to build a decision tree [27]. In simple words, the top-down approach means that we start building the tree from the top and the greedy approach means that at each iteration we select the best feature at the present moment to create a node.

In order to perform ID3, we first select the predictor X_j and the cut point s such that splitting the predictor space into the regions {X|X_j < s} and {X|X_j ≥ s} leads to the greatest possible reduction in residual sum of squares (RSS). The notation {X|X_j < s} means the region of predictor space in which X_j takes on a value less than s. That is, we consider all predictors X₁, . . . , X_p, and all possible values of the cutpoint s for each of the predictors, and then choose the predictor and cutpoint such that the resulting tree has the lowest RSS. In greater detail, for any j and s, we define the pair of half-planes:

$$R1(j, s) = \{X|X_j < s\} \text{ and } R2(j, s) = \{X|X_j \geq s\},$$

and we seek the value of j and s that minimize in (6).

$$\sum_{i: x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i: x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2 \tag{6}$$

where \hat{y}_{R_1} is the mean response for the training observations in $R_1(j, s)$ and \hat{y}_{R_2} is the mean response for the training observations in $R_2(j, s)$. Finding the values of j and s that minimize in (6) can be done quite quickly, especially when the number of features p is not too large.

Next, we repeat the process, looking for the best predictor and best cut point in order to split the data further so as to minimize the RSS within each of the resulting regions. However, this time, instead of splitting the entire predictor space, we split one of the two previously identified regions. We now have three regions. Again, we look to split one of these three regions further, so as to minimize the RSS. The process continues until a stopping criterion is reached; for instance, we may continue until no region contains more than five observations.

Decision tree Algorithm 3 is a type of classification algorithm whose structure is similar to a tree that has roots, branches and leaves. The root node (internal node) represents the features in the dataset, the branch node represents the decision rule, and each leaf node represents the output. The purpose of using a decision tree is to create a training model that can be used to predict the class or value of the target variable by studying simple decision rules deduced from previous data (training data). To predict the class of a given dataset, the decision tree algorithm starts from the root node of the tree. This algorithm compares the value of the root attribute with the record attribute. Based on this comparison, the algorithm traces the branch and goes to the next node. For the next node, the algorithm again compares the attribute values with other sub-nodes and moves towards a deeper node.

Algorithm 3. Decision tree (ID3)

```

ID3(Samples_V, Attributes_A, ClassLabel C):
Create RootNode
if all members of samples are in the same class C
    then RootNode = single-node tree with label = C
else if Attributes is empty
    then below Branch add Leaf with the most occurred label in sample
else
    A = element in Attributes which maximizes InformationGain(Sample, A)
A is decision attribute for RootNode
for each possible value v of A
    add a Branch below RootNode, testing for A = v
    samples_v = subset of Samples with A = v
    if samples_c is empty
        then below Branch add Leaf with the most occurred label in sample
    else
        below Branch add Subtree ID3(samples_v, Attributes A,
ClassLabel C)

```

3. RESULTS AND DISCUSSION

The study obtained sentiment analysis datasets of 1,119 valid Indonesian-language Tweets from Tweets collection against the spread of COVID-19 disease in March 2021 using keywords #kena COVID and #kena corona. The query keyword is based on the consideration of medical words and words commonly used in Indonesian society that are close to synonyms that reflect common symptoms or symptoms of COVID-19 disease. Some examples of sentences that match search keywords are shown in Table 1. Generally, the categories used in classification techniques use at most 3 labels, in this study proposed adding the number of labels to 4 classes namely P, N, Neu and None. Category P represents the class with the predicted opinion of the true event occurring, category N with opinion prediction events do not occur, Neu category with neutral predictions and None category with opinion predictions that cannot be known according to the core of the required sentence.

Table 1. Sentiment results of Indonesian Tweets sentences that match search keywords

Sentiment (label)	Examples of Tweets sentences
P	Finally, I updated my progress again because when I got COVID I was lazy to study. I'm starting to study again, but I'm just working on tomorrow's questions, I'll swipe again, hopefully the results will be negative.
P	Initially got COVID on January 1, mild symptoms a week later after being recovered. Now I have just taken a test. Why does it fluctuate
P	Did you know that according to research blood type A is said to be more susceptible to COVID? What causes it
N	Self quarantine or no COVID, I ask forgiveness from God, and we praise God
N	It must be hard to keep your distance on the show because you feel close emotionally and the symptoms of getting Corona in the early stages are not too visible
N	So earlier at the office suddenly a father came into his room and suddenly sneezed and then casually said I think I have Corona
NEU	I thought I was going to get out of this COVID because there was already a vaccine, uhh instead a corona appeared
NEU	don't let it get you corona, just stick to the health protocol
NEU	Initially got COVID on January 1, mild symptoms a week later after being recovered. Now I have just taken a test. Why does it fluctuate
NONE	Did you know that according to research blood type A is said to be more susceptible to COVID? What causes it
NONE	Later, if you get COVID, I will laugh
NONE	It must be hard to keep your distance on the show because you feel close emotionally and the symptoms of getting Corona in the early stages are not too visible

Further, sentiment analysis that will be focused and discussed in this study is analyzing the results of document classification values obtained using a predetermined algorithm consisting of several derived classes, consisting of:

- True positive (TP): the results of the data processing value in this class will inform that the Tweet data obtained is predicted to be true of the occurrence of the spread of COVID-19 disease and the results of the sentence checking show the true occurrence of COVID-19 disease somewhere.
- True negative (TN): the results of the data processing value in the TN class will inform that the Tweet data obtained is predicted to be incorrect occurrence of the spread of COVID-19 disease and the results of the sentence checking show the incorrect occurrence of COVID-19 disease somewhere.

- c) False positive (FP): the results of the data processing value in the FP class will inform that the Tweet data obtained is predicted to be true of the occurrence of the spread of COVID-19 disease and the results of the sentence checking show the incorrect occurrence of COVID-19 disease somewhere.
- d) False negative (FN): the results of the data processing value in the FN class will inform that the Tweet data obtained is predicted to be incorrect occurrence of the spread of COVID-19 disease and the results of the sentence check show the true occurrence of COVID-19 disease somewhere.

3.1. Classification using Naïve Bayes algorithm

The results of sentiment analysis calculation using algorithm naïve bayes are calculated accuracy rate value of 78.50%, with Recall value of 84.58% and Precision value of 76.03%. The value is obtained based on the data spread and calculation of matrix confusion as shown in Tables 2 and 3. Figure 2 informs the spread of positive (True) opinions to all opinion prediction values stating that predictions that cannot be known (none) and get the correctness score get the highest value of 38.10% with a precision rate of 95.34%, followed by positive predictions and neutral predictions (Neu) that both get a value of 28.04%, but the precision value of positive predictions is greater than the precision value of NEU, that is; 98.15% and 91.67%.

Table 2. Matrix confusion results using algorithm Naïve Bayes

Parameter	True(N)	True(P)	True(N+)	True(None)	True(Neu)	True(P+)	True(Polarity)
Pred.N	123	0	0	0	0	0	0
Pred.P	18	71	0	0	0	0	0
Pred.N+	48	28	55	0	0	0	0
Pred.None	31	89	0	449	0	0	0
Pred.Neu	0	70	0	0	41	0	0
Pred.P+	76	0	0	0	0	20	0
Pred.Polarity	0	0	0	0	0	0	3
Class Recall	50%	42.06%	100%	100%	100%	100%	100%

Table 3. Precision class using algorithm Naïve Bayes

Parameter	Precision
Pred.N	100%
Pred.P	98.15%
Pred.N+	46.15%
Pred.None	95.34%
Pred.Neu	91.67%
Pred.P+	0.88%
Pred.Polarity	100%

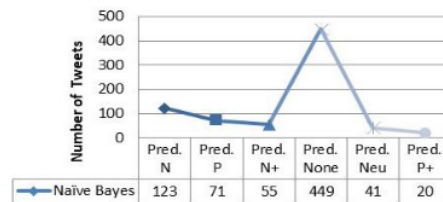


Figure 2. Tweet polarity using Naïve Bayes classification

Table 4 is a table that informs the results of sentiment processing analysis using algorithm Naïve Bayes which states that out of 119 tweet data there are 71 tweets that actually contain the meaning of the sentence has occurred the spread of COVID-19 disease somewhere. While 141 tweets stated that there is no spread of COVID-19. Although the information has spread COVID-19 only amounted to 7% of the total tweets, with a calculation accuracy rate of 78.50%, the Indonesian government still has to take various strategic measures to prevent and reduce the spread of the disease.

Table 4. The result of the Tweet value by class is different from the predicted value and the actual value (true) using the Naive Bayes algorithm

Sentiment Analysis	True Negative (0)	True Positive (1)
Prediction Negative (0)	TN=123	FN=0
Prediction Positive (1)	FP=18	TP=71
Total	141	71

3.2. Classification using algorithm decision tree

Tables 5 and 6 are matrix confusion tables resulting from sentiment analysis calculations using algorithm decision tree classification that informs accuracy value obtained by 50.53% with Recall value of 14.29% and Precision value of only 7.22%. The low accuracy and Precision value of decision tree calculation results is further clarified through Figure 3 which informs the spread of sentiment that is positive (True) to all predicted events that are only worth none prediction (unknown) with an absolute value of 100%. While the predicted value of other opinions is worth 0%.

Table 5. Matrix confusion results using algorithm decision tree

Parameter	True N	True P	True N+	True None	True Neu	True P+	True Polarity
Pred.N	0	0	0	0	0	0	0
Pred.P	0	0	0	0	0	0	0
Pred.N+	0	0	0	0	0	0	0
Pred.None	308	155	567	53	34	2	3
Pred.Neu	0	0	0	0	0	0	0
Pred.P+	0	0	0	0	0	0	0
Pred.Polarity	0	0	0	0	0	0	0
Class Recall	0	0	0	100%	0	0	0

Table 6. Precision class using algorithm decision tree

Parameter	Precision
Pred.N	0
Pred.P	0
Pred.N+	0
Pred.None	50.53%
Pred.Neu	0
Pred.P+	0
Pred.Polarity	0

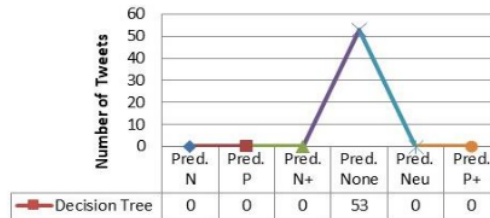


Figure 3. Tweet polarity using decision tree classification

Based on the information shown in Table 7, it appears that algorithm decision tree cannot get the results of the number of tweets containing the meaning of the sentence has occurred or not the occurrence of a disease spread somewhere. The result of this calculation is supported by a low level of accuracy of calculations that is only 50.53%. This indicates that the decision tree algorithm used in the calculation is not in accordance with the characteristics of the dataset obtained at the previous stage.

Table 7. The result of the Tweet value by class is different from the predicted value and the actual value (true) using the decision tree algorithm

Sentiment analysis	True Negative (0)	True Positive (1)
Prediction Negative (0)	TN=0	FN=0
Prediction Positive (1)	FP=0	TP=0
Total	0	0

3.3.3. Classification using K-NN algorithm

The result of sentiment analysis calculation using algorithm K-NN with A value is 5 get accuracy result of 95.10% with Recall value of 68.72% and Precision value of 81.97%. The breakdown spread of

calculation result values using the K-NN algorithm is displayed in the Matrix confusion table in Tables 8 and 9. This high accuracy and precision value is also shown in Figure 4 which informs the spread of positive opinion (True) against all event predictions that are only worth 100% positive prediction, and all other sentiment prediction values are worth 0%.

Table 8. Matrix confusion results using algorithm K-NN (K=5)

Parameter	True N	True P	True N+	True None	True Neu	True P+	True Polarity
Pred.N	308	0	0	0	0	0	0
Pred.P	0	155	0	35	18	2	0
Pred.N+	0	0	567	0	0	0	0
Pred.None	0	0	0	18	0	0	0
Pred.Neu	0	0	0	0	16	0	0
Pred.P+	0	0	0	0	0	0	0
Pred.Polarity	0	0	0	0	0	0	3
Class Recall	100%	100%	100%	33.96	47.06	0%	100%

Table 9. Precision class using algorithm K-NN (K=5)

Parameter	Precision
Pred.N	100%
Pred.P	73.81%
Pred.N+	100%
Pred.None	100%
Pred.Neu	100%
Pred.P+	0%
Pred.Polarity	100%

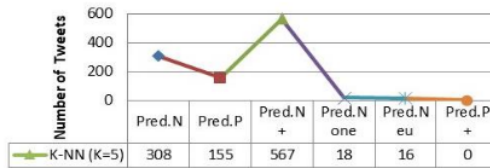


Figure 4. Tweet polarity using K-NN classification (K=5)

Table 10 informs that there are 14% or 155 tweets from 1,119 tweets analyzed stating the truth that there has been a spread of COVID-19 disease somewhere, while 28% or 308 tweets indicate that the information submitted on Twitter does not occur the spread of the disease. Although the correctness of information has occurred the spread of COVID-19 disease is only 14%, with the accuracy of the calculation of algorithm K-NN with a value of k is 5 as big as 95.10%, then the Government of Indonesia must immediately take various strategic measures to overcome the spread of COVID-19 disease in the place.

Table 10. The result of the Tweet value by class is different from the predicted value and the actual value (true) using the K-NN algorithm

Sentiment Analysis	True Negative (0)	True Positive (1)
Prediction Negative (0)	TN = 308	FN = 0
Prediction Positive (1)	FP = 0	TP = 155
Total	308	155

3.4. Sentiment results comparison analysis

Figure 5 inform the graph of sentiment analysis comparison results using algorithm Naïve Bayes (N.B) classification with decision tree algorithm (D.T). And K-NN algorithm that informs that the algorithm that has the highest accuracy value is algorithm K-NN with a value of K is 5, and algorithm that has the lowest accuracy value is decision tree. As for the highest precision and Recall values produced by calculations using Naïve Bayes algorithm and the lowest is algorithm decision tree.

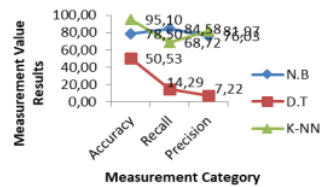


Figure 5. Comparison of sentiment analysis results

4. CONCLUSION

This study resulted in the classification of Tweets based on opinions and categories with keywords #kena COVID and #kena corona that get the K-NN algorithm as a classification algorithm that has the highest accuracy value of 95.10% and has a precision value of 81.97%. Based on the results of confusion matrix algorithm K-NN known that the positive sentiment of Indonesian people who inform the spread of COVID-19 by 13.85% of the 1,119 valid Tweets analyzed, while the negative opinion value of the Tweets is 78.19%.

The short dataset collection period (March 2021) and the small amount of data analyzed tend not to be constant, causing the results of research and conclusions made cannot be used to justify the opinion of the Indonesian people on the spread of the COVID-19 pandemic in other and longer periods of time. This offers great opportunities for future research improvement and development. In addition, the researchers were still able to improve the accuracy value of the calculation results by utilizing other classification algorithms such as fuzzy clustering including using a deep learning approach.

ACKNOWLEDGEMENTS

This article is partly supported by the Directorate of Research and Community Service, the Directorate General of Strengthening Research and Development and the Rector of the University of Sriwijaya.




REFERENCES

- [1] S. Das and A. K. Kolya, "Predicting the pandemic: sentiment evaluation and predictive analysis from large-scale tweets on COVID-19 by deep convolutional neural network," *Evolutionary Intelligence*, Mar. 2021, doi: 10.1007/s12065-021-00598-7.
- [2] B. Ganesh *et al.*, "Epidemiology and pathobiology of SARS-CoV-2 (COVID-19) in comparison with SARS, MERS: An updated overview of current knowledge and future perspectives," *Clinical Epidemiology and Global Health*, vol. 10, 2021, doi: 10.1016/j.cegh.2020.100694.
- [3] WHO, "Naming the coronavirus disease (COVID-19) and the virus that causes it," World Health Organization, 2020, Accessed: Feb. 8, 2020. [Online]. Available: [https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-\(COVID-2019\)-and-the-virus-that-causes-it](https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(COVID-2019)-and-the-virus-that-causes-it).
- [4] H. A. Rothan and S. N. Byraredd, "The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak," *Journal of Autoimmunity*, vol. 109, p. 102433, 2020, doi: 10.1016/j.jaut.2020.102433.
- [5] D. K. Sari, R. Amelia, R. Dharmajaya, L. M. Sari, and N. K. Fitri, "Positive correlation between general public knowledge and attitudes regarding COVID-19 outbreak 1 month after first cases reported in Indonesia," *Journal Community Health*, vol. 46, no. 1, 2021, doi: 10.1007/s10900-020-00866-0.
- [6] D. N. Aisyah, C. A. Mayadewi, H. Diva, Z. Kozlakidis, Siswanto, and W. Adisasmito, "A spatial-temporal description of the SARSCoV-2 infections in Indonesia during the first six months of outbreak," *PLoS One*, vol. 15, no. 12 December, 2020, doi: 10.1371/journal.pone.0243703.
- [7] C. Suratnoaji, Nurhadi, and I. D. Arianto, "Public opinion on lockdown (PSBB) policy in overcoming COVID-19 pandemic in Indonesia: Analysis based on big data twitter," *Asian Journal for Public Opinion Research*, vol. 8, no. 3, pp. 393-406, 2020, doi: 10.15206/ajpor.2020.8.3.393.
- [8] L. Moradi, "COVID-19 in Southeast Asia," *Journal of Archives in Military Medicine*, vol. 9, no. 3, 2021, doi: 10.5812/jamm.117787.
- [9] R. Nagar *et al.*, "A case study of the New York City 2012-2013 influenza season with daily geocoded Twitter data from temporal and spatiotemporal perspectives," *Journal of Medical Internet Research*, vol. 16, no. 10, p. e236, 2014, doi: 10.2196/jmir.3416.
- [10] K. Spurlock and H. Elgazzar, "Predicting COVID-19 infection groups using social networks and machine learning algorithms," in *2020 11th IEEE Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON)*, Oct. 2020, pp. 0245-0251, doi: 10.1109/UEMCON51285.2020.9298093.
- [11] M. Adriani, F. Azzahro, and A. N. Hidayanto, "Disease surveillance in Indonesia through Twitter posts," *Journal of Applied Research and Technology*, vol. 18, no. 3, Jun. 2020, doi: 10.22201/icat.24486736e.2020.18.3.1091.
- [12] S. Lim, C. S. Tucker, and S. Kumara, "An unsupervised machine learning model for discovering latent infectious diseases using social media data," *Journal of Biomedical Informatics*, vol. 66, pp. 82-94, 2017, doi: 10.1016/j.jbi.2016.12.007.
- [13] S. Joshi and D. Deshpande, "Twitter sentiment analysis system," *International Journal of Computer Applications (IJCA)*, vol. 180, no. 47, pp. 35-39, 2018, doi: 10.5120/ijca2018917319.
- [14] G. K. Shahi, A. Dirkson, and T. A. Majchrzak, "An exploratory study of COVID-19 misinformation on Twitter," *Online Social Networks and Media*, vol. 22, 2021, doi: 10.1016/j.osnem.2020.100104.
- [15] A. R. Rahmanti, D. N. A. Ningrum, L. Lazuardi, H. C. Yang, and Y. C. Li, "Social media data analytics for outbreak risk communication: public attention on the 'new normal' during the COVID-19 pandemic in Indonesia," *Computer Methods Programs Biomed*, vol. 205, 2021, doi: 10.1016/j.cmpb.2021.106083.




- [16] S. W. Wijaya and I. Handoko, "Examining a COVID-19 Twitter hashtag conversation in Indonesia: a social network analysis approach," *In 2021 15th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, 2021, doi: 10.1109/IMCOM51814.2021.9377382.
- [17] M. Machmud, B. Irawan, K. Karinda, J. Susilo, and Salahudin, "Analysis of the intensity of communication and coordination of government officials on twitter social media during the COVID-19 handling in Indonesia," *Academic Journal of Interdisciplinary Studies*, vol. 10, no. 3, 2021, doi: 10.36941/AJIS-2021-0087.
- [18] M. Abduh, M. Hamka, T. Taniredja, A. Zainuddin, and W. N. Habiby, "Indonesian perceptions on online learning amidst COVID-19: a Twitter sentiment analysis," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 30, no. 1, pp. 567-576, Apr. 2023, doi: 10.11591/ijeecs.v30.i1.pp567-576.
- [19] R. Ranovan, A. Doewes, and R. Saptano, "Twitter data classification using multinomial naive bayes for tropical diseases mapping in Indonesia," *Journal Telecommunication Electronic Computer Enggenering*, vol. 10, no. 2-4, pp. 155-159, 2018.
- [20] Fathoni, Erwin, and Abdiansah, "Extraction of event sentence information in the COVID-19 distribution location detection system based on the Indonesian language corpus," *International Conferences Electronic Engenering Computer Science Informatics*, vol. 2022-October, no. October, pp. 383-388, 2022, doi: 10.23919/EECSIS6542.2022.9946530.
- [21] D. Baviskar, S. Ahirrao, and K. Kotecha, "Multi-layout unstructured invoice documents dataset: a dataset for template-free invoice processing and its evaluation using AI approaches," *IEEE Access*, vol. 9, pp. 101494-101512, 2021, doi: 10.1109/ACCESS.2021.3096739.
- [22] F. M. J. M. Shamrat *et al.*, "Sentiment analysis on twitter tweets about COVID-19 vaccines using NLP and supervised KNN classification algorithm," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 23, no. 1, pp. 463-470, Jul. 2021, doi: 10.11591/ijeecs.v23.i1.pp463-470.
- [23] M. B. Rissan and R. F. Hassan, "Naïve-Bayes family for sentiment analysis during COVID-19 pandemic and classification tweets," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 28, no. 1, pp. 375-383, Oct. 2022, doi: 10.11591/ijeecs.v28.i1.pp375-383.
- [24] H. R. Arabia, K. Daimi, R. Stahlbock, C. Soviany, L. Heilig, and K. Brüssau, "Correction to: principles of data science," *Principles of Data Science*, 2020, doi: 10.1007/978-3-030-43981-1.
- [25] N. M. Abdulkareem, A. M. Abdulazez, D. Q. Zeebaree, and D. A. Hasan, "COVID-19 world vaccination progress using machine learning classification algorithms," *Qubahan Academi Journal*, vol. 1, no. 2, 2021, doi: 10.48161/qaj.v1n2a53.
- [26] B. Charbuty and A. Abdulazez, "Classification based on decision tree algorithm for machine learning," *Journal Application Science Technology Trends*, vol. 2, no. 01, 2021, doi: 10.38094/jast20165.
- [27] W. W. Wei, M. H. W. Wei, B. Z. M. Hui, R. S. B. Zhang, and R. D. R. Scherer, "Research on decision tree based on rough set," *Journal Internet Technology*, vol. 22, no. 6, pp. 1385-1394, Nov. 2021, doi: 10.53106/160792642021112206015.

BIOGRAPHIES OF AUTHORS






Fathoni    was born in Palembang, Indonesia, in 1972. In 2020. He is currently working on a project for his Doctorate in Engineering, Faculty of Engineering, Sriwijaya University, Indonesia. He received his bachelors' degree in Informatics Management And Computer Engineering from the Institut Sains dan Teknologi Akprind in 1998 and Magister of Information System from Universitas Gunadarma in 2001. In 2008, he joined as a lecturer at Information System Department in Universitas Sriwijaya where is working until now. His current research interests include the field of data mining, pattern recognition, and artificial intelligence. He can be contacted at email: fathoni@unsri.ac.id.



Prof. Dr. Erwin    was born in Palembang, Indonesia, in 1971. He received his Bachelor of Mathematics from Universitas Sriwijaya, Indonesia, in 1994, and an M.Sc. degree in Actuarial from the Bandung Institute of Technology (ITB), Bandung, Indonesia, in 2002. In 1994, he joined Universitas Sriwijaya, as a Lecturer. Since December 2006, he has been with the Department of Informatics, Universitas Sriwijaya, where he is an Professor in 2023. Since 2012, he has been with the Department of Computer Engineering, Universitas Sriwijaya. Then, in 2019, he received his Doctorate in Engineering, Faculty of Engineering, Universitas Sriwijaya. His current research interests include image processing, and computer vision. Prof. Dr. Erwin, S.Si., M.Si. is a member of IAENG and IEEE He has authored or coauthored more than 54 publications: 30 proceedings and 18 journals, with 9 H-index and more than 200 citations. He can be contacted at email: erwin@unsri.ac.id.



Abdiansah    was born in Pagar Alam, Indonesia, in 1984. He received his Bachelor of Informatic Engineering from IST Akprind, Yogyakarta, Indonesia, in 2006. He also received his Master and Doctoral degree in Computer Science at Universitas Gadjah Mada in 2008 and 2019 respectively. In 2008 he joined as lecturer at Department of Informatics, Faculty of Computer Science, Universitas Sriwijaya. His current research interests include natural language processing, question answering system, chatbot system, text mining, big data, and computer vision. He has authored or co-authored some publications. He can be contacted at email: abdiansah@unsri.ac.id.

Multilabel sentiment analysis for classification of the spread of COVID-19 in Indonesia ... (Fathoni)

Multilabel sentiment analysis for classification of the spread of COVID-19 in Indonesia using machine learning

ORIGINALITY REPORT

14%

SIMILARITY INDEX

5%

INTERNET SOURCES

15%

PUBLICATIONS

5%

STUDENT PAPERS

PRIMARY SOURCES

1 Fathoni, Erwin, Abdiansah. "Extraction of Event Sentence Information in the Covid-19 Distribution Location Detection System based on the Indonesian Language Corpus", 2022 9th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), 2022 **10%**
Publication

2 Submitted to Ain Shams University **2%**
Student Paper

3 issuu.com **2%**
Internet Source

Exclude quotes On

Exclude matches < 2%

Exclude bibliography On