Sentiment Analysis of the Covid-19 Virus Infection in Indonesian Public Transportation on Twitter Data: A Case Study of Commuter Line Passengers

Intania Cahya Sari
Magister of Technology Information
Universitas Indonesia
Jakarta,Indonesia
intania.cahya@ui.ac.id

Yova Ruldeviyani
Magister of Technology Information
Universitas Indonesia
Jakarta, Indonesia
yova@cs.ui.ac.id

Abstract— **The appearance of the Covid-19 virus in early 2020 became a frightening pandemic for the world, including Indonesia. The infection of the Covid-19 virus was rapid because of its transmission can be through human contact. This condition causes worrying in society. Besides, these worrying also occurs in the passenger of public transportation, especially the commuter line. Passengers in large numbers and push each other will cause worry if commuter line passengers will transmit the Covid-19 virus to the commuter line. Many passengers write their opinions about the transmission of the Covid-19 pandemic on social media Twitter. This causes various opinions that can be positive, negative, or even neutral. Therefore, to see the opinions on commuter line passengers, a research was made to analyze the sentiment of the Covid-19 transmission to commuter line passengers. This research was implemented using a comparison of 2 methods, Naïve Bayes outperformed the Decision Tree with an accuracy of 73.59%. Furthermore, the result of sentiment analysis was a positive classification compared to the other 2 classes.**

Keywords— Sentiment Analysis, Covid-19, Naïve Bayes, Decision Tree

I. Introduction

The new year 2020 was surprised by the pandemic of a new virus, which was Novel Coronavirus or currently known as Covid-19. This virus has a suspect started in China, which began with the finding of a case of pneumonia whose cause is unknown. That finding was announced on January 5, 2020. Furthermore, on January 12, 2020, has been known that cause symptoms such as pneumonia are called Novel Coronavirus. This virus is transmited rapidly in China with 81,116 cases (within 3 months) with a death rate of 3,231. Besides in China, this virus has attacked various parts of the world. In total cases recorded by the World Health Organization (WHO), 173,344 cases occurred in 152 countries [1].

Indonesia is one of the countries that be a place of transmitting the Covid-19 virus. On March 2, 2020, was the first time it has been announced that there have been 2 positive cases of Covid-19 in Indonesia [2]. Every day there are additional cases that are positive infection to the Covid-19 virus. This is because the transmission of this virus can be through contact between humans, so that the transmission is rapid. This situation makes people worrying because of all their activities every day have always related to each other. The worrying also occurs in passengers of public transportation.

Research in China conducted from December 2019 to March 2020 tried to measure the risk of transmission on various public transportation, such as flights, trains, and buses.[3]. The train is one of the most common and important forms of transportation in many countries, as it can carry 3 times as many passengers as buses and planes. With far more passengers, an investigation was carried out to measure the risk of the individual level of Covid-19 transmission to the train. As a result, the train has a high risk of transmission [4].

In Indonesia, trains are also important transportation. There are types of trains that carry more passengers each day. The route is around Jabodetabek, namely the commuter line. Every day, the commuter line will never be free from passengers. Especially during rush hour such as the morning when people go to work and after work hours, passengers are often seen in large numbers [5]. This situation makes the space for passengers on the commuter line to be limited and may occur between the passengers push each other. This has caused worry of commuter line passengers about the transmission of the Covid-19 virus.

Many commuter line passengers who felt worried and scared write their opinion through social media. Because basically everyone is free to give an opinion, moreover an opinion on social media. The development of social media is also growing rapidly, as on Twitter. Choosing Twitter as a source of opinion mining because of its popularity [6]. Twitter supports a short description of a person's ideas and opinions which can be up to 140 characters. This can be used as a source of data to be analyzed [7]. Therefore, what people say on Twitter can be classified as positive, negative, or neutral opinions.

Sentiment analysis can be done to see opinions written by passengers on Twitter. Sentiment analysis in a observe is carried out to research people's opinions, evaluations, attitudes, and feelings primarily based totally on what they write. Sentiment analysis is applied because it is related to perception and reality and the choices to be made. Just as when making decisions, it is not rare for someone to look for other people's opinions [8].

In this study, sentiment analysis is done to see the opinions of commuter line passengers who were written on Twitter. The aim is to know the reaction of commuter line passengers to the transmission of the Covid-19 virus. The data analyzed is the commuter line passengers tweet data. Furthermore, the data collected were classified into 3 classes, namely Positive, Negative, and Neutral. Next, sentiment analysis can be used to obtain the real influence of people [7].

The method to be used in this study is the Naïve Bayes method. This method was chosen based on previous research which has a title as Sentiment Evaluation of Public Transport in Social Media using the Naïve Bayes Method [9]. This study applied sentiment analysis for evaluation and received responses and feedback from public transportation users. The responses and feedback from users are very diverse and each has a perspective. Classification is done using Naïve Bayes because the algorithm is simple and effective for text classification, and Naïve Bayes is selected to reduce the complexity in which the information on social media may be very large [9]. Furthermore, to compare the performance of Naïve Bayes, this study also uses the Decision Tree method [10].

Naïve Bayes was also selected based on a survey that had been conducted, which is about the use of an algorithm that will be used in sentiment analysis. The results show that the Naïve Bayes is one of the two models that are often used to solve problems in sentiment analysis [11]. Furthermore, the Decision Tree method was chosen because it's provided better accuracy compared to the other [12]. The decision tree also is simpler and easier to use. Both methods are easy methods with high accuracy results. Therefore, this research will apply the naïve Bayes method and the decision tree to see which method will produce better accuracy.

The systematics of writing in this study consists of an introduction that explains the background of the study and its objectives. The literature study discusses the theories derived from the literature that supports the preparation of this research. The research methodology describes the stages of research activities. Then, results and discussion show the results of research conducted based on the stages of the study by analysis of results and discussion. Conclusions explain the conclusions of the study and suggestions for future research. References contain a list of references used in this study.

II. Literature Study

Sentiment analysis is a observe that analyzes people's opinions, sentiments, evaluations, attitudes, and feelings primarily based totally on what they write [8]. Other opinions also say that sentiment analysis is done to find out what other people think based on information, such as written opinions [13]. From the two opinions said that the analysis was carried out on the written opinion. People often express and write their opinions on social media because of the development of the digital era that makes us unable to get away from social media.

Discussing about social media, Twitter is one of the platforms that is usually used to write people's opinions. This can be used as a source of data to be analyzed [7]. Choosing Twitter as a source of opinion mining because of its popularity [6]. Twitter supports a short description of a person's ideas and opinions which can be up to 140 characters. This can be used as data in sentiment analysis to produce information and obtain the real influence of people [7].

In this sentiment analysis research is carried out utilizing text mining, which is the method of extricating valuable data in a content [14]. Text mining is part of data mining and involves the preparation of documents [15]. Preprocessing on documents makes a text from documents that were not structured initially into structured data. The structured data is then classified using a classification method in data mining.

The method implemented in this research is Naïve Bayes and Decision Tree. The reason for applying the Naïve Bayes is due to strong assumptions and high accuracy [16][17]. While the Decision Tree algorithm was chosen because it is simpler and easier to use [10]. The two methods are explained as follows:

A. Naïve Bayes

Naïve Bayes is a grouping model which calculates probabilities in each class based on the division of words in a document [11]. This method has good accuracy used for sentiment analysis [17] and has been tested with several other algorithms [14][16][11]. The theory will be used to predict probabilities are as follows [11]:

$$P(label\,|features) = \frac{P(label) * P(features\,|label)}{P(features)} \tag{1}$$

$P(label)$ is the previous probability of the label. $P(features\,|label)$ is the previous probability that is classified as a label. $P(features)$ is the previous probability that occurred. Naïve was given a presumption which states that all features are free, the formula might be rewritten as:

$$P(label\,|features) = \frac{P(label) * P(f1\,|label) * \ldots * P(fn\,|label)}{P(features)} \tag{2}$$

B. Decision Tree

The decision tree method is a representative model of a tree or tree to predict test results [14]. Attributes are like a node and as a tree branch. The highest node of a decision tree is named the root [18]. The classification process begin with the root node of the tree. The decision tree algorithm stages start from:

1. Prepare training data.
2. Select attributes as root using Information Gain (ID3).
3. Create a branch for each value.
4. Repeat the process for each branch until all cases in the branch have the same class

Furthermore, to obtain optimal accuracy and compare the performance of several methods, a cross-validation process is needed. The cross-validation process is dividing all documents into 2 (two) parts, namely training data and testing data. An important factor in the implementation of cross-validation is the partition of ratios for training data and testing data. This partition of the ratio is called validation size. Determination of validation size uses *k-fold*, i.e. the data is divided by as many as *k* parts.

Furthermore, training data and testing data iterate as much as *k*. It can be analogized in the first iteration of the testing data is 1 part of *k* while the training data is all data minus the testing data [(*k*-1) part]. This process continues until the k-iteration with randomized data again [19].

At the research stage, the process was carried out utilizing CRISP-DM (CRoss-Industry Standard Process for Data Mining). It was proposed within the mid-1990s by a European consortium of companies to service as a nonproprietary basic methodology for data mining. CRISP-DM is a form of data mining model that is used to standard process models in various data mining methods [20]. This model contains 6 steps there are business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The following six stages are used in research methodology, which can be seen in figure 1.

Some studies have discussed sentiment analysis with case studies on social media. The following are some of the previous studies related to sentiment analysis and research on the use of classification methods. The first research was the result of a survey of 54 articles which were then concluded. All articles that examined are then categorized into 6 categories to see how the articles use algorithms in solving problems in sentiment analysis. The results show that the Naïve Bayes algorithm and the Support Vector Machine algorithm are models that are often used to solve problems in sentiment analysis [11].

The second research was about sentiment analysis in social media using Naïve Bayes, Decision Tree, and Random Forest. Sentiments on data taken from Twitter are grouped into positive, negative, or neutral. The result is that Indonesian Twitter users give more neutral comments. The highest accuracy of the three algorithms obtained from testing data in rapid miner tools was compared. The accuracy of Naïve Bayes is higher accuracy than the other algorithm with 86.43%. The accuracy of the Decision Tree and Random Forest 82.91% [17].

Furthermore, the third research still about sentiment analysis from tweets using Decision Tree, K-Nearest Neighbor, and Naïve Bayes. The case study in this research is a tweet from e-commerce Tokopedia and Bukalapak. The techniques are used in this research such as text mining, preprocessing text, classification, etc. These techniques are used to create classification and analysis of sentiment analysis. Rapid miner is moreover used to help in making analysis sentiments for comparison by using three different classifications within the dataset. The results of this research are the highest accuracy is the Naïve Bayes algorithm of 77% [21].

Next is the fourth research was about classification using Naïve Bayes Classifier and Decision Tree Algorithms. For this research, they are suggesting a more accurate and effective prediction in the assumption kind of brain tumor is Naïve Bayes classification and decision tree algorithm. The purpose of this study is to prove that the Decision Tree algorithm is simpler and easier than the Naïve Bayes algorithm. Utilizing these two algorithms, the kind of tumor has been found and it allows analysis of historical information from data sets which makes neurologists assume the kind of tumor. The result is true that the Decision Tree algorithm is faster and more accurate than the Naïve Bayes algorithm [10].

The fifth study, about Naïve Bayes vs. Decision Trees vs. Neural Networks in the Classification. This research categorizes web pages by using a comparison of the three algorithms. This research shows that they enhanced Naïve Bayes classifier not only outperforms the traditional Naïve Bayes but also perform similarly as good. The result that can be taken is that the Naïve Bayes algorithm is the best choice for training web pages [16].

The difference between 5 previous research and author research can be seen in table I.

TABLE I. THE DIFFERENCE BETWEEN PREVIOUS RESEARCH

Research	Previous Research	Author's research
1	Use many data mining methods to find out which method is suitable for sentiment analysis. The result is naïve bayes and svm are model that are often to use.	Use naïve bayes based on the result of this prevoius research.
2	Compare the performance of 3 methods (Naïve Bayes, Decision Tree, Random Forest) to find out the best accuracy result.	Use naïve bayes based on the result of this prevoius research.
3	Compare the performance of 3 methods (Decision Tree, K-NN, Naïve Bayes) to find out the best accuracy result.	Use naïve bayes based on the result of this prevoius research.
4	Compare naïve bayes and decision tree. Decision tree has faster and more accurate than naïve bayes.	Based on this previous research, author choose decision tree to compare with naïve bayes.
5	Compare the performance of 3 methods (Naïve Bayes, Decision tree, Neural Netwirk) to find out the best accuracy result	Use naïve bayes based on the result of this prevoius research.

III. RESEARCH METHODOLOGY

In this sentiment analysis, several steps must be done to obtain the best results. The following are the six stages of developing data mining used in the research methodology can be seen in figure 1.

A. Business Understanding

This study was done to analyze public sentiment, especially commuter line public transport passengers at the beginning of its virus infection, which is March 2020. The aim is to see "*How are the sentiments of commuter line passengers about the transmission of the Covid-19 virus in Indonesia?*" and another aim is to find out "*How are sentiments analysis using the Naïve Bayes and Decision Tree?*".

B. Data Understanding

The process of collecting data starts with crawling Twitter data using the Rapid Miner 9.6.0 tool. The data collected are tweets in Indonesian containing the words of

KRL, commuter line, corona, korona, and COVID-19. At least one tweet must contain the KRL / commuter line and one of the words corona/corona / COVID-19. At this stage, the tweet data is eliminated by duplicating data, due to the retweet feature. Then the data is selected and only the tweet text will be taken.

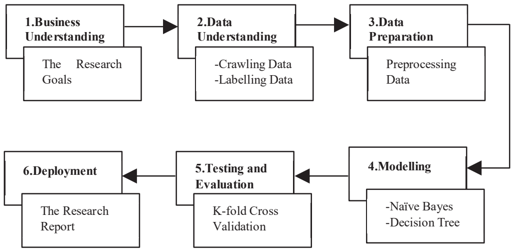

Fig. 1. The step-in research methodology using CRISP-DM

The next stage is the data labeling process in twitter data. The tweet data that has been collected is then labeled manually before they are classified. The purpose of manual labeling is the annotators manually categorized every tweet as positive, negative, neutral [22]. The categorized data might be used to train classifiers. Labeling is done with the category of emotional approaches, such as joy, sadness, anger, fear, surprise, disgust, disbelief, anticipation [9][11]. The guideline to labeling the tweet is if the tweet contains positive or negative emotion are categorized as positive or negative, and tweets with common statements are categorize as neutral.

C. Data Preparation

The next process after collecting data is preparing data to be analyzed using data mining. In rapid miner, the name of this step is process document, as in figure 3. There are the following processes:

1) *Data Cleaning*
The tweet data that has been successfully crawled will enter the stage of removing URLs, mentions, hashtags, emoticons and removing unneeded characters such as punctuation (such as: .;,; @; #, etc.).

2) *Data Transformation*
In this stage, there were case folding and filtering processes. The case folding process is the stage of changing all the characters in a tweet to be converted to lowercase letters, no longer capital letters in the tweet. The filtering stage is to discard meaningless words such as conjunctions.

3) *Data Reduction*
The tweet data enters the stemming process, which replaces the words to the basic words. Stemming used is a stemming porter. Before entering the data stemming process, it is first made a token (tokenizing process). Then each token is converted into a basic word form.

D. Modeling

Naïve Bayes and Decision Tree methods are used for data classification and model. Naïve Bayes was chosen because of its high accuracy [16][17] and Decision was used because of the ease of use of its algorithm [10]. K-Fold Cross Validation is used as a modeling technique, by testing training data and testing data *k* times testing.

E. Testing and Evaluation

After getting the classification model, then the training data testing and the data testing is done using K-Fold Cross-Validation. In this study, the k-fold value to be used is 5,10,15 and 20 to get the best accuracy.

F. Deployment

The following Deployment Phase is creating a simple report or implementation process of data mining. This is used to develop conclusions. Figure 2 is containing stages of data analysis using classification algorithms.

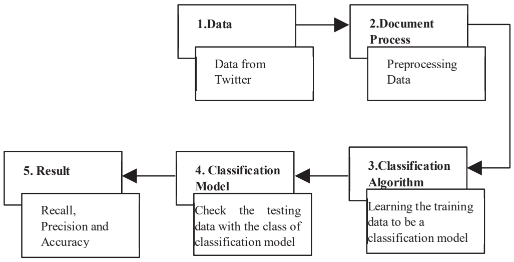
1.Data
Data from Twitter
2.Document Process
Preprocessing Data
3.Classification Algorithm
Learning the training data to be a classification model
4. Classification Model
Check the testing data with the class of classification model
5. Result
Recall, Precision and Accuracy

Fig. 2. Data Analysis Stage

IV. Result and Discussion

A. Result

The following are the results of data processing using the Rapid Miner tool.

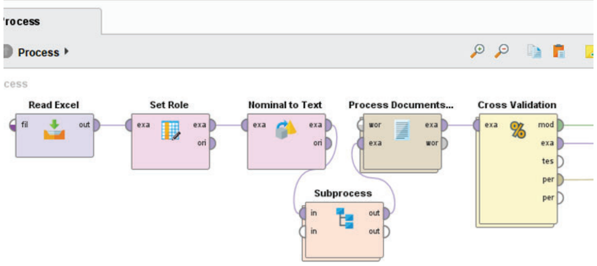

Fig. 3. Data processing on Rapid miner

The process starts with reading the data in the form of .xlsx to the cross-validation process as in figure 3.

1) *Data Understanding*

The total data that was successfully collected using Rapid Miner 9.6.0 tools are 340 data. This data is free from duplication or free from retweets. It can be seen in Table II.

TABLE II. The Number of Twitter Data

Sentiment	Number
Positive	135
Negative	152
Neutral	53
Total	**340**

Next is the data labeling stage, which can be seen in Table III as follows. Positive label categories such as good, like, great, cool, thank you. Meanwhile, negative labels such as bad, hate, disappointed and angry.

TABLE III. LABELING DATA

Text	Sentiment
Adanya Upaya antisipasi sudah dilakukan secara masif oleh KRL sejak bulan lalu. Seperti pemberian edukasi kepada para penumpang untuk mencegah penyebaran virus Korona #KAILawanCorona	Positive
PT Kereta Commuter Indonesia menyatakan bakal mengikuti rapat penanggulangan virus corona bersama Pemerintah Provinsi DKI hari ini, 12 Maret 2020.	Neutral
Coba jangan dibilang krl jalur jakk-boo aja yang risikonya gede penularan korona, jalur lain pun kalo mau dibilang pasti sama juga potensinya gede hhhhhhh dipikir jalur bogor aja kali yang desek2an tiap hari	Negative

2) *Data Preraration*

Furthermore, the data collected enters the Data Preparation stage as shown in Table IV. In the table, a tweet with positive sentiment will be presented.

TABLE IV. DATA PREPARATION

Process	Data
Real Data	Setelah korona dinyatakan pandemi, Indonesia harus mulai berpikir untuk membatasi pertemuan di atas 25 orang. Naik bus/KRL setiap hari ke kantor memudahkan transmisi. Baiknya mulai dipikirkan kerja dari rumah.
Cleansing	Setelah korona dinyatakan pandemi Indonesia harus mulai berpikir untuk membatasi pertemuan di atas orang Naik bus KRL setiap hari ke kantor memudahkan transmisi Baiknya mulai dipikirkan kerja dari rumah
Case Folding	setelah korona dinyatakan pandemi indonesia harus mulai berpikir untuk membatasi pertemuan di atas orang naik bus krl setiap hari ke kantor memudahkan transmisi baiknya mulai dipikirkan kerja dari rumah
Filtering	korona pandemi indonesia berpikir membatasi pertemuan orang bus krl hari kantor memudahkan transmisi dipikirkan kerja rumah
Stemming	korona pandemi indonesia pikir batas temu 25 orang bus krl hari kantor mudah transmisi pikir kerja rumah
Tokenizing	‘korona’, ‘pandemi’, ‘indonesia’, ‘pikir’, ‘batas’, ‘temu’, ‘orang’, ‘bus’, ‘krl’, ‘hari’, ‘kantor’, ‘mudah’, ‘transmisi’, ‘pikir’, ‘kerja’, ‘rumah’

All data that has been processed is entered into the cross-validation process to share training and testing data.

3) *Testing and Evaluation*

Here are the results of tests using the Naïve Bayes classification model and the Decision tree on each k-fold value of 5, 10, 15, and 20, which can see in Table V.

TABLE V. THE ACCURACY IN EACH K-VALUE

k	*Decision Tree*	*Naïve Bayes*
5	56.76 %	73.24 %
10	58.24 %	73.82 %
15	56.15 %	74.37 %
20	56.18 %	72.94 %
Average	56.83 %	73.59 %

B. *Discussion*

The testing was conducted on 340 Twitter sentiment data with various k-fold values can be seen in Table IV. The values are used in the k-fold cross-validation test affect the accuracy of each algorithm. The effect of a *k* value that is too small will produce a small accuracy too. This is because the small *k* value causes the analysis to be affected by errors in the data[23]. Whereas a large *k* value will cause high accuracy results.

The determination of the optimal *k* value by looking at Table IV. The optimal k value is chosen which is the value of *k*=10 for Decision Tree and *k*=15 for Naive Bayes. This value was chosen because its accuracy is stable, in the sense, that it is not too small and too large. This is also in compliance with previous research which states that the value of *k*=10 is the optimal value for testing the k-fold cross-validation [23].

The testing also produces classifications with positive, negative, and neutral classes, as in Table V. The results displayed in table VI are the results of testing the entire data. A total of 340 data were used as test data and training data by dividing the ratio using k-fold cross-validation.

TABLE VI. THE RESULT OF DATA TESTING USING K=10

	Sentiment	The Initial Data	The Test Result Data
Decision Tree	Positive	135	104
	Negative	152	82
	Neutral	53	12
	Total	**340**	**340**
Naïve Bayes	Positive	135	117
	Negative	152	101
	Neutral	53	33
	Total	**340**	**340**

The confusion matrix can be seen in Table VII.

TABLE VII. THE CONFUSION MATRIX

Naïve Bayes (Accuracy: 73.82%)					
		True			
		Positive	**Neutral**	**Negative**	**Precision**
Prediction	**Positive**	117	10	8	86.67%
	Neutral	10	33	10	62,26%
	Negative	21	30	101	64.65%
	Recall	79.5%	45.21%	84.87%	
Decision Tree (Accuracy: 58.24%)					
		True			
		Positive	**Neutral**	**Negative**	**Precision**
Prediction	**Positive**	104	30	25	65.41%
	Neutral	15	12	12	30.77%
	Negative	29	31	82	57.75%
	Recall	70.27%	16.44%	68.91%	

Decision Tree was processes data with positive sentiment totaling 135 processed resulting in a positive class of 104 data. Negative sentiments from 152 data, the results of negative classification are 82. Furthermore, the data with neutral sentiment are 53 data, the true neutral data classification is 12 data. Naïve Bayes successfully processed data with positive sentiment totaling 135 processed resulting in positive classification totaling 117 data. Negative sentiments from 152 data, negative classification results are 101. Furthermore, data with neutral sentiments are 53 data, true neutral data classification is 33 data. So, in Table V, it shows that the result of sentiment analysis was a positive classification compared to the other 2 classes.

To see the result of the two algorithms, it can be seen with the recall precision and accuracy. Furthermore, for the comparison, the average accuracy of Naïve Bayes is 73.59%, while the average accuracy of the Decision Tree algorithm is 56.83%. The result is that Naïve Bayes is superior to the Decision Tree. This significant difference in accuracy is due to the noise factor in the data. As we know that Naïve Bayes are not interdependent between their attributes. Meanwhile, the Decision Tree in each attribute affects each node. A comparison of the average of the two algorithms also proves the theory that the Naïve Bayes algorithm produces better accuracy than the other classification algorithms [16].

V. Conclusion

The results of this research show that the Naïve Bayes algorithm outperforming the Decision Tree with an accuracy of 73.59%. It concluded that the value of *k* on k-fold cross validation affects the recall, precision, and accuracy results. Therefore, a test with several *k* values is obtained and the optimal *k* value is *k*=10 for Decision Tree and *k*=15 for Naive Bayes. All accuracy values in the test are then calculated on average and are compared between the Decision Tree and Naïve Bayes. The conclusion is that the sentiment that exists in the community is more to the positive sentiment that contains an appeal and a call for prevention and control of the Covid-19 outbreak. For future research, the development on this research can be done with other method that better than this method. And in the next research, it is expected that the data can be labeled automatically. Because manual labeling can lead to different points of view in determining sentiment.

VI. References

[1] "WHO | China." [Online]. Available: https://www.who.int/countries/chn/en/. [Accessed: 17-Mar-2020].

[2] "Kasus Covid-19 Pertama, Masyarakat Jangan Panik | Indonesia.go.id." [Online]. Available: https://indonesia.go.id/narasi/indonesia-dalam-angka/ekonomi/kasus-covid-19-pertama-masyarakat-jangan-panik. [Accessed: 07-Sep-2020].

[3] S. L. Maogui Hu, Hui Lin, Jinfeng Wang, Chengdong Xu, Andrew J Tatem, Bin Meng, Xin Zhang, Yifeng Liu, Pengda Wang, Guizhen Wu, Haiyong Xie, "The Risk of COVID-19 Transmission in Train Passengers: An Epidemiological and Modelling Study," vol. 306, pp. 0–19, 2020.

[4] S. Zhao *et al.*, "The association between domestic train transportation and novel coronavirus (2019-nCoV) outbreak in China from 2019 to 2020: A data-driven correlational report," *Travel Med. Infect. Dis.*, vol. 33, no. January, pp. 2019–2021, 2020.

[5] "Membludak, Pengguna KRL Diimbau Hindari Jam Sibuk - Info Publik |." [Online]. Available: https://rri.co.id/humaniora/info-publik/863530/membludak-pengguna-krl-diimbau-hindari-jam-sibuk. [Accessed: 11-Sep-2020].

[6] H. Parveen and S. Pandey, "Sentiment analysis on Twitter Data-set using Naive Bayes algorithm," *Proc. 2016 2nd Int. Conf. Appl. Theor. Comput. Commun. Technol. iCATccT 2016*, pp. 416–419, 2017.

[7] M. Vadivukarassi, N. Puviarasan, and P. Aruna, "Sentimental Analysis of Tweets Using Naive Bayes Algorithm," *World Appl. Sci. J.*, vol. 35, no. 1, pp. 54–59, 2017.

[8] B. Liu, *Sentiment Analysis and Opinion Mining*. Chicago: University of Illinois: Morgan & Claypool Publisher, 2012.

[9] N. Othman, M. Hussin, and R. A. R. Mahmood, "Sentiment Evaluation of Public Transport in Social Media using Naïve Bayes Method," *Int. J. Eng. Adv. Technol.*, vol. 9, no. 1, pp. 2305–2308, 2019.

[10] D. S. Reddy, C. N. Harshitha, and C. M. Belinda, "Brain tumor prediction using naïve Bayes' classifier and decision tree algorithms," *Int. J. Eng. Technol.*, vol. 7, no. 1.7 Special Issue 7, pp. 137–141, 2018.

[11] W. Medhat, A. Hassan, and H. Korashy, "Sentiment Analysis Algorithms and Applications: A Survey," *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, 2014.

[12] V. M., J. Vala, and P. Balani, "A Survey on Sentiment Analysis Algorithms for Opinion Mining," *Int. J. Comput. Appl.*, vol. 133, no. 9, pp. 7–11, 2016.

[13] B. Pang and L. Lee, *Opinion Mining and Sentiment Analysis*. computer Science Department, Cornell University, USA, 2008.

[14] M. Guia, R. R. Silva, and J. Bernardino, "Comparison of Naive Bayes, support vector machine, decision trees and random forest on sentiment analysis," *IC3K 2019 - Proc. 11th Int. Jt. Conf. Knowl. Discov. Knowl. Eng. Knowl. Manag.*, vol. 1, pp. 525–531, 2019.

[15] R. Feldman and J. Sanger, *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, 2007.

[16] D. Xhemali, C. J. Hinde, and R. G. Stone, "Naive Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages," *Int. J. Comput. Sci.*, vol. 4, no. 1, pp. 16–23, 2009.

[17] V. A. Fitri, R. Andreswari, and M. A. Hasibuan, "Sentiment analysis of social media Twitter with a case of Anti-LGBT campaign in Indonesia using Naïve Bayes, decision tree, and random forest algorithm," *Procedia Comput. Sci.*, vol. 161, pp. 765–772, 2019.

[18] O. Somantri and D. Dairoh, "Analisis Sentimen Penilaian Tempat Tujuan Wisata Kota Tegal Berbasis Text Mining," *J. Edukasi dan Penelit. Inform.*, vol. 5, no. 2, p. 191, 2019.

[19] S. Yadav and S. Shukla, "Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification," *Proc. - 6th Int. Adv. Comput. Conf. IACC 2016*, no. Cv, pp. 78–83, 2016.

[20] P. Chapman *et al.*, *CRISP-DM 1.0 Step-by-step data mining guide*. USA: SPSS Inc., 2000.

[21] A. Bayhaqy, S. Sfenrianto, K. Nainggolan, and E. R. Kaburuan, "Sentiment Analysis about E-Commerce from Tweets Using Decision Tree, K-Nearest Neighbor, and Naïve Bayes," *2018 Int. Conf. Orange Technol. ICOT 2018*, no. October 2018.

[22] S. Anastasia and I. Budi, "Twitter sentiment analysis of online transportation service providers," *2016 Int. Conf. Adv. Comput. Sci. Inf. Syst. ICACSIS 2016*, pp. 359–365, 2017.

[23] S. K. Lidya, O. S. Sitompul, and S. Efendi, "Sentiment Analysis Pada Teks Bahasa Indonesia Menggunakan Support Vector Machine (SVM)," *Semin. Nas. Teknol. dan Komun. 2015*, vol. 2015, no. Sentika, pp. 1–8, 2015.