Analyzing Public Sentiment on COVID-19 Using TF-IDF and K-Nearest Neighbors (K-NN) on Twitter Data

Arip[1)],Anna Dina Kalifia[2)]

[1-2)] Fakultas Sains Dan Teknologi, Universits Teknologi Yogyakarta

Correspondence Author: aarip4333@gmail.com

Article Info :

Article History :

Received :
11 January 2024
Revised :
04 February 2024
Accepted :
03 July 2024
Available Online :
28 August 2024

Keyword :
COVID-19, Sentiment Analysis, Twitter, K-Nearest Neighber, Text Mining, Public Opinion

ABSTRACT

The coronavirus outbreak that occurred in almost all countries in the world has had an impact not only on the health sector, but also on other sectors such as tourism, finance, transportation, etc. This has given rise to various kinds of sentiments from the public with the emergence of the corona virus as a trending topic on social media Twitter. Twitter was chosen by the public because it can disseminate information in real time and can see the market's reaction quickly. In this study, "tweet" data or public tweets related to the "Corona Virus" were used to see how the polarity of sentiment emerged. Text mining techniques and K-Nearest Neighbor (K-NN) machine learning classification algorithms were used to build a tweet classification model on sentiment whether it has a positive, negative, or neutral polarity. The test results were produced by the algorithm with an average result for a precision value of 57.93% and for an average recall niali of 55.21% with an accuracy value of 64.52%

1. INTRODUCTION

Based on Groot, et al. in the "Ninth Report of the International Committee on Taxonomy of Viruses", Coronavirus 2019 (COVID-19) is an infectious disease caused by an acute respiratory syndrome, namely Coronavirus 2 (SARS-CoV-2). SARS-CoV-2 is known as the subfamily Orthocoronavirinae, the family Coronaviridae, the order Nidovirales, and the realm Riboviria. Corona viruses are a group of related viruses that cause disease in mammals and birds. In humans, the corona virus causes respiratory tract infections that can range from mild to deadly. Mild illness includes some cases of the flu, while more deadly varieties can cause SARS, MERS, and COVID-19. Coronavirus Disease 2019 or COVID-19 is a new disease that can cause respiratory distress and pneumonia. This disease is caused by Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) infection. Clinical symptoms that appear are diverse, ranging from common cold symptoms (cough, runny nose, sore throat, muscle pain, headache) to severe complications (pneumonia or sepsis).

ISSN Print : 1979-7141
ISSN Online : 2541-1942

COVID-19 is a new disease and researchers are still studying how it is transmitted. From various studies, the main method of spreading this disease is suspected to be through respiratory droplets and close contact with sufferers. Droplets are small particles from the mouth of patients that contain the disease virus, which are produced when coughing, sneezing, or talking. Droplets can pass up to a certain distance (usually 1 meter). Droplets can stick to clothes or objects around the sufferer when coughing or sneezing. However, droplet particles are large enough that they will not survive or settle in the air for a long time. Therefore, people who are sick, are required to wear masks to prevent the spread of droplets. For transmission through food, until now there is no scientific evidence.

The Covid-19 pandemic has indeed had a very wide impact, not only limited to the health sector, but also attacking other sectors such as tourism, finance, business, and transportation. This has raised various sentiments from various parties from all over the world because this pandemic is taking place in almost all countries in the world.

Twitter or X is a social media created by Jack Dorsey in 2006. In 2019, based on Twitter's press-release, there were 500 million tweets or tweets by Twitter users per day. As many as 500 million tweets are used to post things about users and share information, the content of tweets can also express feelings. This opinion through tweets can be used to see how sentiment is raised towards events that are of interest to the public, one of which is related to the Corona Virus which is trending on Twitter's social media line. Sentiment analysis or opinion mining is the computational study of internet users' opinions, sentiments, and emotions through entities and attributes that are owned and expressed in the form of text. Sentiment analysis will group (classify) the polarity of the text into sentences or documents to find out whether the opinions expressed in the form of sentences or documents are positive, negative, or neutral. In this study, sentiment analysis was carried out to see the opinion or tendency of opinion towards a problem or object, in this case a tweet or tweet "Covid19" whether it contains a polarity of negative, positive, or neutral sentiment.

According to (Liu, 2012) sentiment analysis or opinion mining refers to a broad field of natural language processing, linguistic computation and text mining which has the purpose of analyzing a person's opinions, sentiments, evaluations, attitudes, judgments and emotions whether the speaker or writer is satisfied with a particular topic, product, service, organization, individual, or activity. The task of sentiment analysis is to group the text into sentences or documents and then determine the opinions expressed in the analyzed sentence or document whether they are positive, negative, or neutral.

2. RESEARCH METHODS

Sentiment analysis is the process of analyzing data obtained from various social media platforms and the internet. This analysis process requires understanding and processing text data to obtain sentiment information from a text or opinion sentence on social media platforms. This sentiment analysis is carried out with the aim of finding out whether the opinion text leads to a positive or negative view.

There is also in the research I did here I used tweet data that I took from Kabgle, a website that provides a collection of datasets, then the data that I use here I took tweet data in 2020 with the amount of data I took was 2000 data which is still raw data that has not been processed cleaned. Here are the stages that I did in this research

2.1 Preprocessing Data / Text Preprocessing

Preprocessing is the initial stage of unstructured data processing. The goal is to simplify the process of searching queries in documents, speed up the processing of documents, and simplify the process of sorting the retrieved data. This preprocessing has stages in processing a text data, these stages include

ISSN Print : 1979-7141

ISSN Online : 2541-1942

a. Cleansing Data

Cleanup refers to the identification, correction and removal of information that is inaccurate, incomplete, irrelevant or irregular from a data set. The goal of cleaning is to reduce interference or noise and missing in the dataset.

b. Case Folding

Case folding is a process in text preprocessing that is carried out to standardize the characters in the data. The case folding process is the process of converting all letters into lowercase letters. In this process the characters 'A'-'Z' contained in the data were changed to the characters 'a'-'z', in this study we changed the letters to pseudo-lowercase letters

c. Filtering

Filtering is the stage of extracting key words from the token results. You can use a stop list algorithm (remove less important words) or a word list (save important words).

d. Tokenizing

Tokenizing is the process of breaking text or paragraphs into small units called tokens. In the context of natural language processing, tokens can be words, phrases, or punctuation. The tokenizing process helps transform text that is in the form of a continuous into a sequence of discrete tokens.

e. Voting

Stemming is a very important process for finding the root word of a derivative word. The essence of the stemming process is to eliminate the suffix in a word.

2.2 Labeling

Sentiment labeling is the process of assigning a specific label or category to a text or document based on the sentiment it contains. Sentiment labels generally include categories such as "Very positive", "Positive", "Very negative", "Negative", or "Neutral". However, in this case, only "Positive" and "Negative" sentiments are used.

2.3 TF-IDF(Term Frequency-Inverse Document Frequency)

TF-IDF is an algorithm that digs between Term frequency and Inverse Document Frequency. Term frequency is the number of occurrences of a term in a document. Inverse Document Frequency is the reduction of the dominance of terms that often appear in various documents, by taking into account the inverse frequency of documents that contain a word. Techniques used in natural language processing and information retrieval to evaluate how important a word is in a document or corpus of text

2.4 Data Visualization

At this stage it is useful to make it easier to understand the information contained in the data and to make it easier to understand the results of the processing stages that have been passed. There are several types of plots or graphics used are as follows, namely wordcloud. WordCloud is a graph that can visually represent a collection of words in text based on the greatest frequency, arranged in such a way that the frequency with which the words appear is reflected in their size or intensity of color. WordCloud is typically used to provide a visual representation of the words that appear most frequently in a text or text collection.

ISSN Print : 1979-7141

ISSN Online : 2541-1942

2.5 Modeling

The K-Nearest Neighbor (K-NN) algorithm is a method for classifying objects based on learning data that is closest to the object. Learning data is projected into a multi-dimensional space, where each dimension represents a feature of the data.

3. RESULTS AND ANALYSIS

Row No.	conversatio...	date	time	user_id	username	tweet	mentions	replies_count
1	1258320972...	2020-05-07	23:57:30	1179769476	its_dul	Klo kata gw Pemerint...	['mas__piyuu...	0
2	1258356644...	2020-05-07	23:53:20	1012156669...	meonkbaong	Saat ini yang bisa sa...	['oiivert']	0
3	Sumber Dina...	Infografis ini ...	tetap waspad...	ikuti himbaua...	?	?	?	?
4	Menteri Keua...	keterbatasan ...	?	?	?	?	?	?
5	Jokowi: Kita ...	?	?	?	?	?	?	?
6	1258420607...	2020-05-07	23:37:07	1249264214...	nadyathara	Yang berbahaya dari ...	[]	0
7	Bebas sih jrx ...	tapi ga usah ...	?	?	?	?	?	?
8	Misalnya ini a...	tapi PENYAKI...	?	?	?	?	?	?
9	1258419017...	2020-05-07	23:30:48	410572714	zyghozy	salah satu keluarga ...	[]	6
10	1258418951...	2020-05-07	23:30:32	507298967	infonyamks	2 Nenek Ini Tolak Ba...	[]	0
11	1258418825...	2020-05-07	23:30:03	2340155395	suaradotcom	Pemerintah Jepang ...	?	?
12	1258389904...	2020-05-07	23:20:27	1245359512...	anthonraharu...	Dengan diberlakukan...	['tvonenews']	0
13	1258038791...	2020-05-07	23:19:52	7038677016...	nano_daru	Bgmn bisa turun kala...	['dr_koko28']	0
14	1258415773...	2020-05-07	23:17:55	1236935724...	pwanea	Bhabinkamtibmas P...	[]	0

Figure 1. Raw data

The first stage I did was to do labeling in order to define a class from the dataset. The data set that has been manually classified produces three category outputs, positive, negative sentiment and.

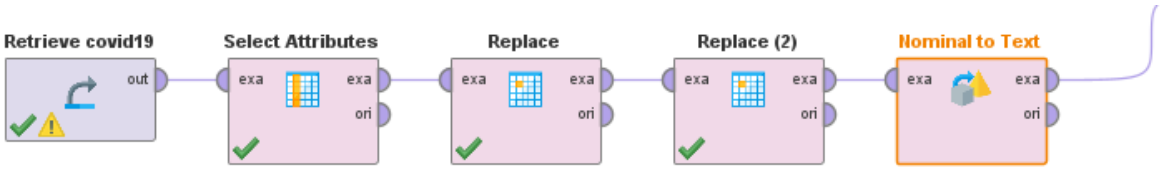

Figure 2. Data processing process

At this stage, the process of selecting attributes to be used is carried out using the "select attribute" operator. Then in the next stage, a cleaning process is carried out using the "Replace" operator, at this stage aims to remove symbols and characters that are considered unimportant, then after that the conversion process is carried out using the "Nominal to Text" operator which changes the initial nominal to the form of text

ISSN Print : 1979-7141

ISSN Online : 2541-1942

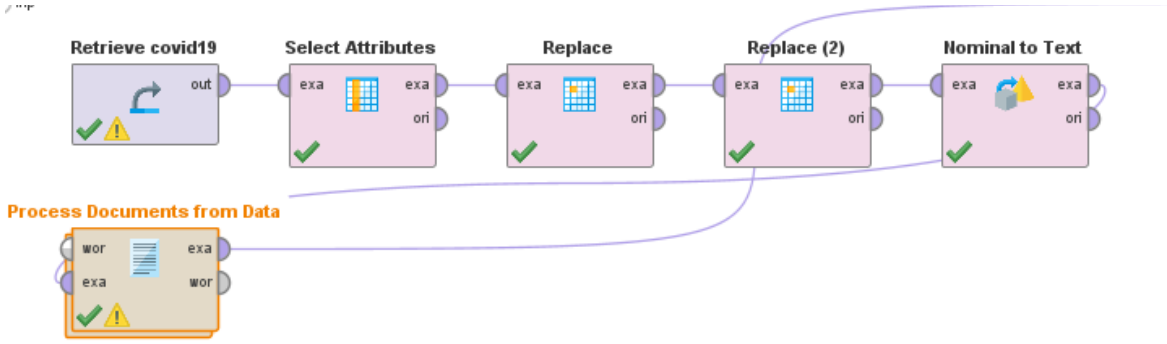

Figure 3. Process Documents From Data Stage

The next stage is the preprocessing process stage using the process documents from data operator. Once the data is collected, start to the data processing stage. This pre-existing data cannot be directly used in processing for sentiment analysis. So it is continued to the data preparation stage.

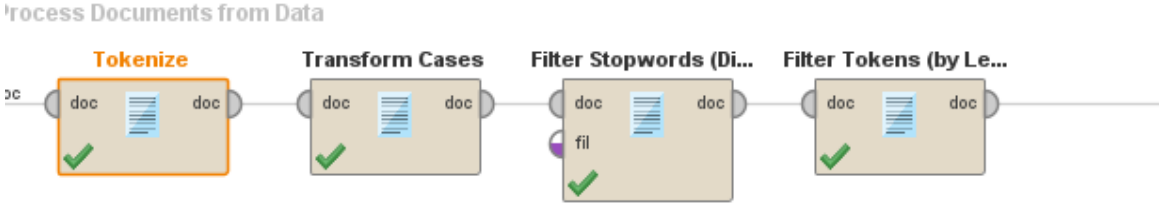

Figure 4. Pre-Processing Process

The next step is the Pre-Processing of data stage which in this stage includes 4 operators used including Tokenize, Transform Caces, Filter Stopwords, Filter Tokens

- *Tokenize*

The tokenization process here is the okenization step of separating sentences into a collection of words. Where each sentence is broken down into word by word. As well as removing symbols and numbers.

- *Transform Case*

In this process, the transform case operator is used to convert capital letters to lowercase letters in text.

- *Filter Stopwords (Dictionary)*

Stopword, which is the process of removing words that do not contain meanings such as conjunctions. The stopword process uses a stopword dictionary obtained from the results of the research and a stopword dictionary that has been created based on the findings of the word in the data set that will be combined into one file

ISSN Print : 1979-7141
ISSN Online : 2541-1942

- *Filter Token (by length)*

In the operator, this token filter is used in omitting or deleting a predetermined number of words. In this study, the researcher used a minimum character length of four characters and a maximum character length of 25 characters. In other words, sentence lengths less than four and more than 25 will be removed. After the preprocessing process of the existing dataset

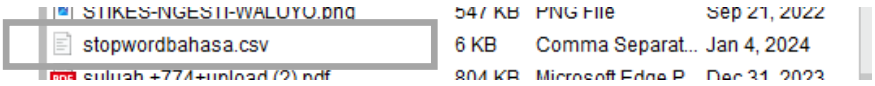

Figure 5. Files used to perform stopwords

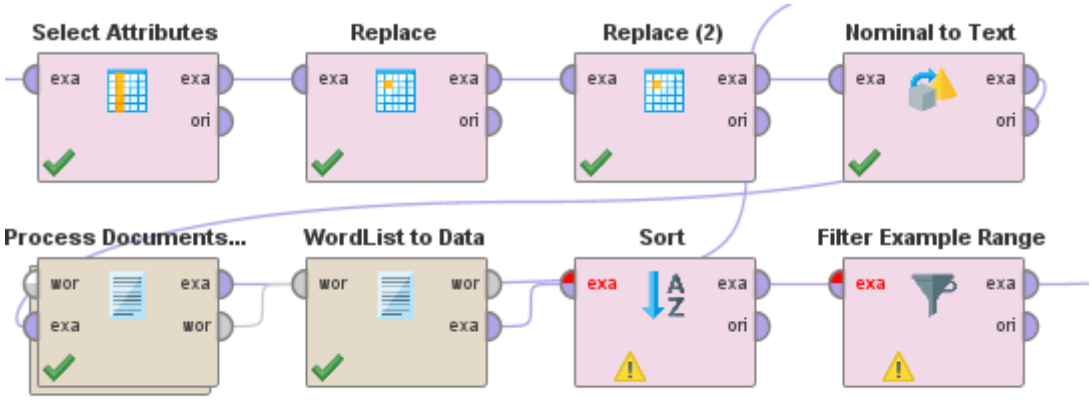

Figure 6. The Process of Selecting the Word Used

After completing the Pre-Processing stage, then the stage of selecting words to be used in the next stage is carried out. At this stage, there are 3 operators used, the first is wordList to Data which is used to create a dataset from a word list. The dataset contains rows for each word and attributes for the word itself, the number of documents that give rise to it, the number of documents labeled where it appears, and for each class number of its occurrence in that class document. The second is the "Sort" operator used to filter the most words and then it will be sorted based on the largest or smallest that will be displayed later. The Example Range filter is used to determine or filter how many words will be raised, as well as later with the visualization process regarding how many words appear can be set on this operator.

ISSN Print : 1979-7141

ISSN Online : 2541-1942

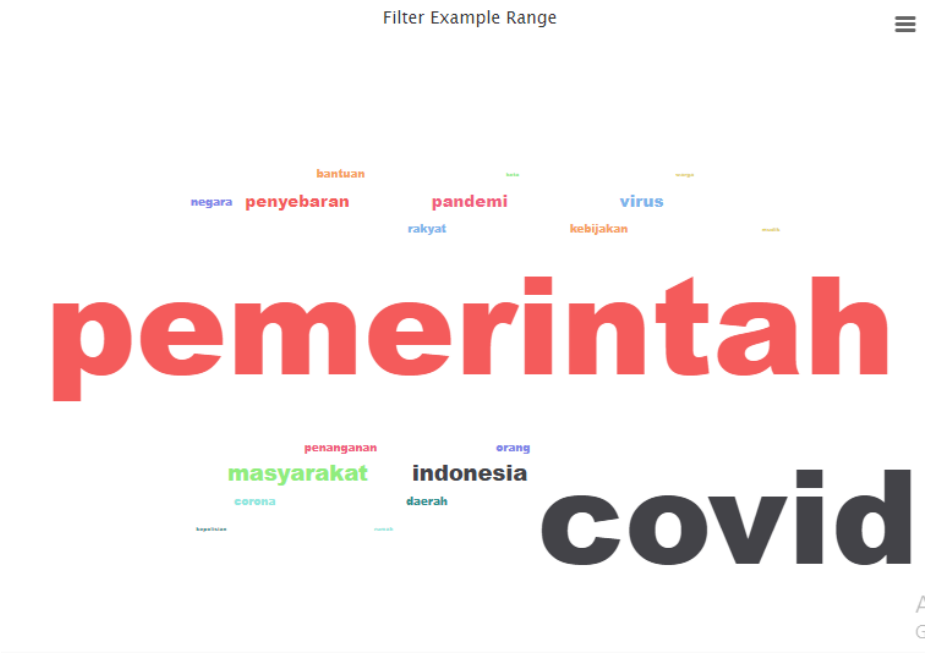

Figure 7. Wordcloud Negative Sentiment

From the image above, it can be seen that words that have a large size are words that often appear or have a large number of terms. There are words; government, covid, indonesia, society, pandemic, spread, etc. These words show the negative sentiments of the people's tweets,

Figure 8. Wordcloud Positive Sentiment

ISSN Print : 1979-7141
ISSN Online : 2541-1942

From the picture above, it can be seen that there is a difference where the word covid is larger than before, namely sentime, the words found above include; Covid, government, Indonesia, society, pandemic, spread, etc. These words show the negative sentiments of the people's tweets.

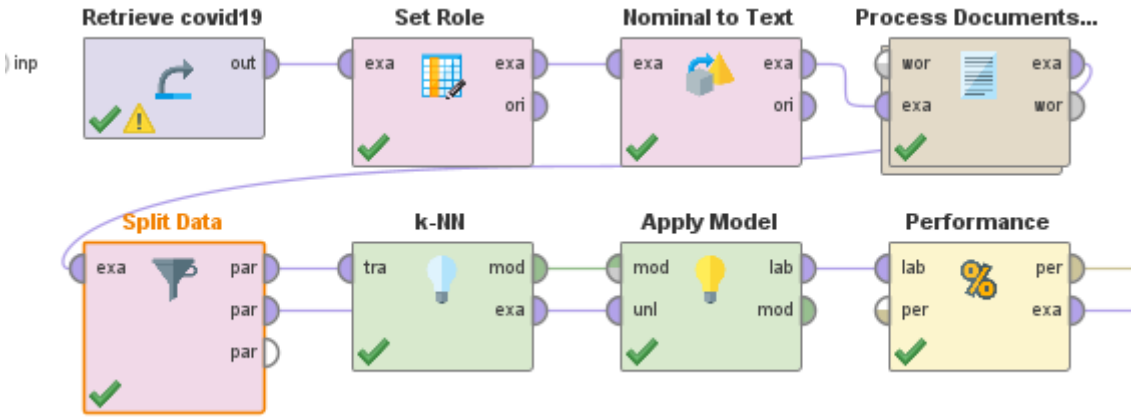

Figure 9. Modeling and K-NN Processors

At this stage I use the K-Nearest Neighbors (KNN) Algorithm, then here I also use several operators, one of which is the Set Role operator which is used to set an attribute to be selected, then there is Splite Data used to divide a data into training data and also training data. The Apply Model operator functions to apply the model that has been learned or trained to the training data sample. The Performance operator is used to evaluate the statistical performance of a binominal classification task, i.e. a classification task where the label attribute has a binominal type.

accuracy: 64.52%

	true POSITIF	true NEGATIF	class precision
pred. POSITIF	252	117	68.29%
pred. NEGATIF	43	39	47.56%
class recall	85.42%	25.00%	

Gambar 10. confussion matrix

Based on the results of the classification of training data with training data with the application of the KNN algorithm. Where in the results of the confusion matrix the accuracy obtained was 64.52%. TP or positive labels that were predicted to be true were 252 while FP or negative labels that were predicted to be wrong were 117 and TN or negative labels that were predicted to be true were 39 while FN or positive labels that were predicted to be wrong were 43.The average for the precision value was 57.93% and for the average recall niali was 55.21%

4. CONCLUSION

Based on the results of the Sentiment Analysis of Public Opinion Related to Covid-19 by Applying the k-nearest neighbors (K-NN) algorithm based on TF-IDF on Tweets Data, the average

ISSN Print : 1979-7141

ISSN Online : 2541-1942

results for the precision value were 57.93% and for the average recall value was 55.21% with an accuracy value of 64.52% using the K-NN algorithm and the TF-IDF method. The conclusions of the results show that although the model has a fairly good level of accuracy, the current precision and recall values indicate some limitations in the model's ability to consistently classify sentiment from Covid-19-related Tweets data. Hopefully this research can be a reference for research in the field of text mining and in the future this research can be developed by choosing more appropriate keywords and can be visualized with a network graph or with a system

5. DECLARATION OF COMPETING INTEREST

We declare that we have no conflict of interest.

6. REFERENCES

Taufan, R., Rivanie, T., Rahayu, S., & Gata, W. (2020). Sentimen analisis twitter terhadap isolasi diri masyarakat Indonesia akibat dampak COVID-19. *MATICS: Jurnal Ilmu Komputer dan Teknologi Informasi (Journal of Computer Science and Information Technology)*, *12*(2), 99-103.

Samsir, S., Ambiyar, A., Verawardina, U., Edi, F., & Watrianthos, R. (2021). Analisis Sentimen Pembelajaran Daring Pada Twitter di Masa Pandemi COVID-19 Menggunakan Metode Naïve Bayes. *Jurnal Media Informatika Budidarma*, *5*(1), 157-163.

Fathonah, F., & Herliana, A. (2021). Penerapan Text Mining Analisis Sentimen Mengenai Vaksin Covid-19 Menggunakan Metode Naïve Bayes. *Jurnal Sains dan Informatika*, *7*(2), 155-164.

Rosari, M. A., Wasino, W., & Tony, T. (2022). Analisis Sentimen Tanggapan Masyarakat Terhadapbantuan Sosialpemerintah Di Masa Pandemi Covid-19 Pada Platform Twitter. *Jurnal Ilmu Komputer dan Sistem Informasi*, *10*(1).

Halim, A., & Safuwan, A. (2023). Analisis Sentimen Opini Warganet Twitter Terhadap Tes Screening Genose Pendeteksi Virus Covid-19 Menggunakan Metode Naïve Bayes Berbasis Particle Swarm Optimization. *Jurnal Informatika Teknologi dan Sains (Jinteks)*, *5*(1), 170-178.

Risnantoyo, R., Nugroho, A., & Mandara, K. (2020). Sentiment analysis on corona virus pandemic using machine learning algorithm. *Journal of Informatics and Telecommunication Engineering*, *4*(1), 86-96.

Habibi, H. A. N. S., Nugroho, A., & Firliana, R. (2023). Perbandingan Algoritma Naïve Bayes Classifier Dan K-Nearest Neighbors Untuk Analisis Sentimen Covid-19 Di Twitter. *JURNAL ILMIAH INFORMATIKA*, *11*(01), 54-62.

Riza, F. (2022). ANALISA SENTIMEN VAKSINASI COVID-19 DENGAN METODE SUPPORT VECTOR MACHINE DAN NAÏVE BAYES BERBASIS TEKNIK SMOTE.

Kaparang, S., Kaparang, D. R., & Rantung, V. P. (2021). Analisis Sentimen New Normal Pada Masa Covid-19 Menggunakan Algoritma Naive Bayes Classifier. *JOINTER: Journal of Informatics Engineering*, *2*(01), 16-23.

ISSN Print : 1979-7141

ISSN Online : 2541-1942