A Covid-19 Sentiment Analysis on Twitter Using K-Nearest Neighbours

Castaka Agus Sugianto*[1], Shandy Tresnawati[2]
[1]Teknik Informatika, [2]Teknik Komputer, Politeknik TEDC Bandung
Pesantren No. KM 2, Cibabat, Kec. Cimahi Utara, Cimahi City, Jawa Barat, 40513
E-mail : castaka@poltektedc.ac.id**1,* shandy.tresnawati@poltektedc.ac.id*2*
**Corresponding author*

Abstract - In December 2019, an outbreak named Corona Virus (SARS-CoV-2) occurred in the city of Wuhan, China which was later known as COVID-19. News of the development of the virus spread through various media, one of which was through the well-known platform Twitter. Twitter is one of the widely used media platforms to communicate about Covid-19. Information related to Covid-19 circulating in the community can be in the form of news or opinions or opinions. Then, the circulating information will be classified into three classes, namely positive, negative or neutral. The method used to calculate the prediction of text classification on Twitter is K-nearest neighbors (KNN). The dataset used in grouping on twitter by using the account name Covid19. Firstly, the dataset by crawling data or information on twitter. Secondly, the text mining stage to determine the class distance value and calculate the Euclidean distance formula based on all the training data to be tested. After the training process is complete, the evaluation model used will be used, the Euclidean results are taken based on the value of the closest distance. The accuracy of the model will be calculated using the previous Euclidean method. The results of this study he obtained with the highest value, one of which was 78% using a 50:50 sample comparison with k-5 and k-9 values.

Keywords – Twitter, Covid-19, KNN, Sentiment Analysis

1. INTRODUCTION

Since December 2019, the world has been shaken by the Covid-19 pandemic. Many people are looking for information related to Covid-19 which is currently shocking the world through print and electronic media. Social media support is one of the things that people need to increase trust and obtain information on various social media platforms, one of which is Twitter. Currently, Twitter is one of the media that has provided various kinds of information regarding Covid-19. Currently, information related to Covid-19 is obtained from the #kawalcovid19 website or Twitter media @KawalCOVID19, which is a source of information from pro-data Indonesian citizen volunteers, consisting of health practitioners, academics and professionals. In addition to the sources above, the public can obtain information on Covid-19 by using the account name, @KawalCOVID19. There is a lot of information circulating in the community regarding COVID-19, whether it is positive, negative or neutral information. The circulation of this information in the era of advanced technology is very fast. To deal with unclear information, which can lead to hoaxes or the like, it is necessary to have a system that can inform that the information is positive or negative. Information as a text management system needed is a text mining method that has a high commercial value potential [1]. According to Ronen Feldman and James Sanger (2007), defining text mining is an intensive knowledge process where users relate to documents from time to time using a set of analytical tools [2]. Research that discusses the use of text mining which is applied in further research to improve performance in classifying document data containing text and numbers [3].

Based on research [4] it can be concluded that the determination of positive or negative contradictions of an opinion is done manually. However, the increasing number of sources of opinion becomes an expansion, namely the time and effort required to classify the contradictions of these opinions will be more and more embedded. This research was carried out using a platform, especially twitter. This sentiment is given to the public for information on the platform so that the public can receive the information. This sentiment is not made into a fabrication but can be classified properly. Twitter is a microblogging service that has become an increasingly popular platform for web users to communicate with one another. In a study he did for his writing, which compared the median of twitter with traditional news media. As a result, twitter is a source of information that contains short news. However, in the future Twitter will provide information quickly to the whole world [5]. Research conducted by [6], this experiment was carried out using the MyPersonality dataset owned by Twitter users with the aim of text classification to predict personality through an online questionnaire. In a respondent test carried out with 3 (three) combined methods, namely Multi Nave Bayes, KNN, and SVM, the results obtained the best accuracy value is the combined method of 65% compared to other methods Multi Nave Bayes of 63%, KNN of 60%, and SVM by 61%.

In addition, the SVM and General Inquirer methods are also used to classify tweet sentiment as positive, negative, and neutral based on the query given. So the purpose of this context approach is to include tweets that make up the classification [7]. Where the context of the number of documents for positive classification will refer to sentences containing knowledge of information, not just denial of such information, for example "good", "ali won the ranking class champion", and so on. While the negative classification which contains denial of the incident of the information that he does not want includes "corruption", "corona crisis", and so on. For the neutral classification, the sentence contains positive and negative intermediaries in which it is expressed by the annotator. For example: "he is so moving, lying down!!!", "wae condition" and so on. The following review studies are used as the theoretical basis for this research. Some of them are: “Twitter Sentiment Classification using Naïve Bayes on Trainer Perception” by Ibrahim & Yusoff, in 2016 discussed the problem of 50 tweets of the keywords 'Malaysia' and 'Maybank' being used as perception training to classify the sentiment of 25 tweets from each keyword in order to be validly measurable. To examine the experiment, the Twython, MongoDB, and Nave Bayes methods were used which later showed that the test in the research conducted was that the tweet keyword classified by the Naïve Bayes method had an accuracy of 90% with a standard deviation of 14%. “Prediction of Indian Election Using Sentiment Analysis on Hindi Twitter” by Sharma & Moh, in 2016 discusses the problem of a system that is difficult to predict the Indian dictionary approach to classified tweets respectively. The Library Language, Hindi Indian, Naïve Bayes, and SVM methods are used. It was found that the BJP group as the best win in the prediction was valued at 78.4%, which was SVM compared to Naïve Bayes, which got 62.1%. “Sentiment Analysis on Twitter Posts: An analysis of Positive or Negative Opinion on Gojek” by Windasari et al, in 2017 discussed the problem of analyzing the grouping of tweet sentiment related to the Go-Jek service, which is a social media capture twitter. The N-gram and SVM methods are used. It was found that to test the dataset as much as 1000 positive tweets and 1000 negative tweets as a training data sentiment, the predicted accuracy of 86% is SVM. "Sentiment Analysis Approach Based on N-grams and KNN Classifier" by Kaur et al, in 2018 discussed the problem of rarely using feature selection to extract word features of a classified analysis sentiment. The N-gram method is used as a feature extract and also posts sentence tags, and KNN. The results showed that the results of the research analysis carried out were that SVM was divided into a comparison of the performance of the work system including the existing system and the proposed system that was used, which was better accurate with 86% accuracy, which was the proposed system. For the existing system obtained 81%. "Twitter Sentiment Analysis for Arabic Tweets" by Abuelenin et al., in 2018 discussed the problem of the sentiment analysis system not yet supporting the use of features including the corpus, namely Arabic. Cosine Similarity method is used, ISRI Arabic Stemmer, Machine Learning is used hybrid test. The results showed that the research carried out in identifying the use of Arabic language combination features was processed experimentally using Machine Learning with a large amount of data, the best value of a method used in feature selection, namely countVectorizer, TFIDFVectorixer and hashingVectorizer, the best feature selection value in the countVectorizer position was 92 ,98%.

Based on the basis of the theories above, it can be concluded that if the accuracy results obtained are 60% above, the Naïve Bayes algorithm and SVM are applied and obtained are quite good. For this reason, this research will implement the use of the algorithm used is K-Nearest Neighbors which is expected to get maximum and accurate results. In addition, the effect of the accuracy results is also analyzed based on the number of k neighbor values in order to determine the level of accuracy of the results.

2. RESEARCH METHOD

2.1. Text Mining

Text mining is the discovery and extraction of significant interesting knowledge from free or unstructured text. With knowledge that comes from patterns and relationships and is used to reveal facts, trends or ideas [8]. In addition, text mining is also needed to convert text into data then other data mining text is used for analysis. With the largest number of data sources where to analyze it manually, it requires text mining to handle the data. However, it is necessary to identify and separate any specific type of information from the text given the need for text mining techniques or methods to assist in grouping data into different groups according to certain requirements [9].

2.2. Text Processing

The use of txt processing is used to carry out the cleaning stage in order to get the results of tweet classification that work well. The steps that can be carried out in this research to perform text processing are :

1. Removing. This stage of the process can be done by cleaning letters that are not normalized or special twitter characters such as retweets need to be removed a certain document. Examples of using removing include @indrakusuma, #ujianTA, /, apleh, and so on.
2. Case folding. This stage is where to change sentences when uppercase letters into lowercase letters of a particular document.
3. Tokenisasi. Tokenization is used to break comments into words. Where this process is carried out by checking something in the existing comments, it is necessary to be a process of breaking or dividing words [10].
4. Filtering. This stage of the process is carried out where words that occur frequently are used, so it is necessary to eliminate every word that must stop without significant meaning in a document. Examples of the use of words in filtering include: "the", "if", and so on [11].

5. Stemming. This stage as the last process is used to find the root or in the formation of the basic words of each word of a document. By doing this the process of returning various word formations to the same representation [12].

2.3. K-Nearest Network

K-Nearest neighbors (KNN) is a simple learning method for all training documents that is used to predict test document labels and has a very large text similarity calculation [13]. According to [14] it can be concluded that K-nearest Neighbors is one of the simplest classification techniques. The performance of the KNN classifier is determined by the choice of k as well as the distance metric applied. Determining the value of k is difficult when the points are uniformly distributed. This KNN classification can be done in two steps :

1) Find the k nearest neighbors in the training dataset.
2) Give the label used based on the majority choosing among the neighbors determined by calculating the Euclidean distance between the new observation and the example in the training dataset.

2.4. Feature Extraction

Where in this research, the tf-idf is used for calculating the word weighting of a document. Term frequency is the occurrence of a term in a document where document frequency is the number of documents a term appears [12]. Where Wi,j is the term weight (tj) of a document (di). For tfi,j is the number of occurrences of the term (tj) of a document (di). The letter n is the total of all documents in a database and df i is also the number of documents containing the term (tj). So log (n/df) calls the IDF value with n being the number of data.

$$W_{i,j} = tf_{i,j}\ x\ log\ log\ \left(\frac{n}{df_i}\right) \qquad (1)$$

2.5. Euclidean Distance

In equation 2.1, it states that where the Euclidean distance(u,v) is the scalar distance of the two vector data u, it is used to classify the connectedness data between the test data and the total number of training data in a document. For data v is also used as training data [15]. Where $\sum \quad d\ i_{=1}$ is the sum of the distances of the i-th variable where i is the value 1, 2, 3.. so on until –n. u_i is i-th training datav_i is i-th test data.

$$Euclidean\ distance(u,v) = \sqrt{\sum_{i=1}^{d} \quad (u_i - v_i)^2} \qquad (2)$$

2.6. Text Classification

Text classification is the classification of text documents into a set of predetermined classes. This is a learning approach when a set of training documents {D1, D2,…Dn} is labeled with class {C1, C2, … Cn} for the process of finding a classification model to be built and also predicting the class of the representative document based on the training data model [17].]. If the dataset that represents the document is used to identify the classification in the prediction, when is the largest number of k closest, then take the greatest opportunity based on the Euclidean distance. Examples of text classification forms are Rocchio Classification, Super Vector Machine, Naïve Bayes, Reccurent Neural Network, K-Means, K-Nearest Neighbors and so on.

2.7. Evaluation of Experimental Models and Calculation Methods

The dataset that has been trained and it is necessary to carry out an evaluation test as a result of modeling in the confusion matrix which is used to determine the level of accuracy of the model that has been trained (training) with the dataset has been tested (testing). For this reason, this study chose the confusion matrix method of the K-Nearest Neighbors (KNN) algorithm to determine the results of the accuracy test. In this study, the KNN model is classified as a binary dataset, which is a positive (true) and negative (false) class. If the positive class is 1, then it is declared true. While the negative class is 0, then it is declared false. With this, the confusion matrix used has 2 binary classifications of output classes. To find out the performance measurement of the confusion matrix, there are 4 types as classifications. It is known that the 4 types are True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

Table 1. Binary Confusion Matrix Classification

Prediction	True Value	
Class	Classified Positive	Classified Negative
Positive	TP (True Positive)	FP (False Positive)
Negative	FN (False Negative)	TN (True Negative)

Table 2. Confusion Matrix Equation

Precision	$\frac{TP}{TP+FP} \times 100\%$
Recall	$\frac{TP}{TP+FN} \times 100\%$
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN} \times 100\%$

In Table 1, it can be explained that TP: True Positive, namely the number of positive text data that is classified correctly by the system. TN : True Negative, namely the number of negative text data that is classified correctly by the system. FP: False Positive, ie the number of positive text data but classified incorrectly by the system. FN: False Negative, which is the number of negative text data but is classified incorrectly by the system. Precision is the number of positive categorized text data that is classified correctly divided by the total data that is classified as positive with the aim of getting the system for users to be given answers to the information value of the level of accuracy. Recall is the number of percentages related to positive category data that can be classified correctly by the system with the aim of giving the system's success rate in retrieving information. While accuracy is a definition where the level of closeness between the predicted value and the actual or actual value with the aim of comparison between the data that is classified correctly and the whole data.

2.8. Data Collection Procedure

In this study, the data used for experimentation is secondary data. Secondary data is data obtained indirectly, such as documentation, journals, literature books, and various other information related to the problems studied [18]. The secondary data that this research collects is crawling data from the twitter.com/KawalCOVID19 site with an account name using the R programming language. This secondary data as a support is used as research material related to sentiment analysis information. The data used is a class sourced from social media, one of which is twitter data named @KawalCOVID19 after crawling the data using the R programming language starting from January 1, 2020 to December 31, 2020 format so that after crawling the data will be received randomly starting on June 24 until October 16, 2020 totaling 2065 datasets were grouped manually into positive, negative and neutral (label) classes.

2.9. Data Analysis Technique

Research that can be carried out after all the necessary data has been collected, there are several stages that must be carried out in analyzing the data as follows, where crawling data that has been manually labeled will be entered into the database and processed using the PHP programming language.

1. Selecting data that is in accordance with the research which is then used as secondary data, namely data containing case study information with Covid-19.
2. Collecting words that are in the pre-processing stage from the packagist site.
3. The secondary data that has been obtained is then carried out in the pre-processing stage to eliminate other characters.

2.10. Data Analysis Technique

In Figure 2. this training data is carried out after crawling the data and then doing sentiment labels manually with Microsoft Excel and then imported into the database using the PHP programming language which is used as sample data material for training data. In addition, it is necessary to do a manual labeling process as a form of class identification used as a result of the classification of the K-Nearest Neighbors algorithm. In detail in Figure 2, the training data flow can be explained as follows.

1. The first stage, the beginning of the process, the dataset is taken from Twitter by crawling it with the R programming language in the form of a dataset in CSV format.
2. The second stage, the dataset will be entered into Microsoft excel in order to label the data manually.
3. The third stage, after labeling the dataset, it will be entered into the database via PHPMyAdmin locally.

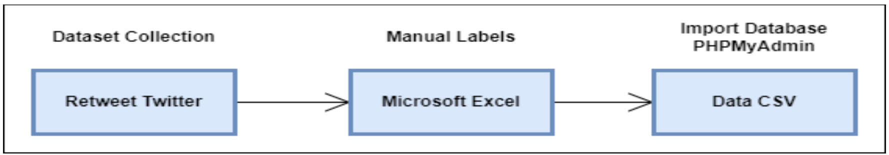


Figure 1. Training Data

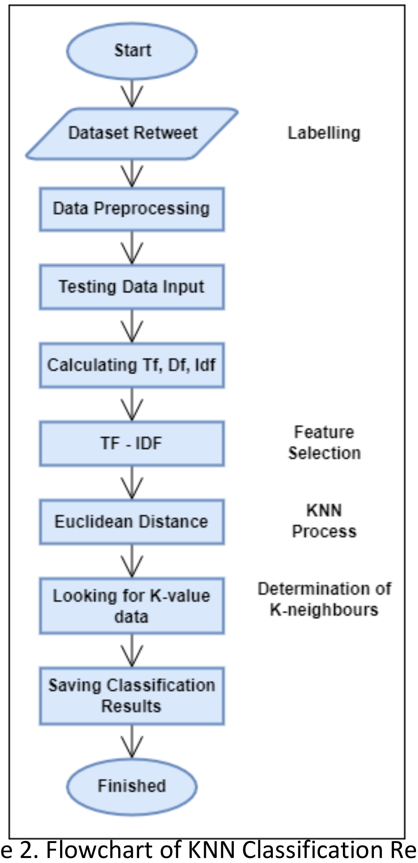


Figure 2. Flowchart of KNN Classification Results

2.11. DataTesting

In the picture above, the flowchart is a data test to determine the results of the classification process for the K-Nearest Neighbors method, which can be explained as follows.

1. After importing the dataset into the database via PHPMyAdmin, the training data is used.
2. Then, perform data processing to collect data for a document.
3. Enter the test data input used to calculate the Euclidean distance.
4. Counting tf, df, and idf used for word weighting.
5. Then calculate the Euclidean distance between the test data and the total number of document data containing tf-idf.
6. After that, the calculation results obtained will be sorted based on the minimum size of the smallest to the largest distance.

Then, it will determine the classification results based on the Euclidean distance and then stored.

3. RESULTS AND DISCUSSION

3.1. Data Preprocessing

After doing the class (label) the dataset will then carry out the data preprocessing stage, namely removing, case folding, tokenization, filtering and stemming where this process is contextual. So that it is easy to process as a text mining classification process. The various stages of data preprocessing can be seen below as follows.

3.1.1. Data Training

Training data is the part of the dataset that we train to make predictions or perform functions of an algorithm. The research data, the training data used comes from Twitter with @(at)covid19 which is then saved in CSV format. Furthermore, the training data is stored in the SQL database system.

Table 3. Example of Training Data

No	Data training	Class
1	RT @KawalCOVID19: Indonesia mengumumkan 1.113 kasus baru #COVID19 pd tgl 24 Juni 2020. Total: 49.009 Kasus aktif: 26.778 (+658) Sembuh: 19?	Positif
2	@mikachandra_ @KawalCOVID19 @BNPB_Indonesia Tujuan utama digalakkannya New Normal adalah sektor ekonomi. Pusat pere? https://t.co/NJjhRJc5vg	Negatif
3	@firdzaradiany @KawalCOVID19 @kamilmoon @mutiaranissa "Kalau pemerintah bilang pandemi sudah membaik, ya membaik!" -Presiden Burkina Faso	Positif
4	RT @septian: Dan, sepanjang sejarah pencatatan @KawalCovid19, Jawa Tengah konsisten menjadi provinsi dengan selisih yang terbesar antara da?	Positif

3.1.2. Removing

The removing stage is the first step that must be done, it is important to delete certain characters. For numeric data, changing with text means that in this study it is only neutral.

Table 4. Example of Removing Sample Data

No	Data training	Removing
1	RT @KawalCOVID19: Indonesia mengumumkan 1.113 kasus baru #COVID19 pd tgl 24 Juni 2020. Total: 49.009 Kasus aktif: 26.778 (+658) Sembuh: 19?	RT KawalCOVID Indonesia mengumumkan kasus baru COVID pd tgl Juni Total Kasus aktif Sembuh
2	@mikachandra_ @KawalCOVID19 @BNPB_Indonesia Tujuan utama digalakkannya New Normal adalah sektor ekonomi. Pusat pere? https://t.co/NJjhRJc5vg	mikachandra KawalCOVID BNPB Indonesia Tujuan utama digalakkannya New Normal adalah sektor ekonomi Pusat pere https t co NJjhRJc vg
3	@firdzaradiany @KawalCOVID19 @kamilmoon @mutiaranissa "Kalau pemerintah bilang pandemi sudah membaik, ya membaik!" -Presiden Burkina Faso	firdzaradiany KawalCOVID kamilmoon mutiaranissa Kalau pemerintah bilang pandemi sudah membaik ya membaik -Presiden Burkina Faso
4	RT @septian: Dan, sepanjang sejarah pencatatan @KawalCovid19, Jawa Tengah konsisten menjadi provinsi dengan selisih yang terbesar antara da?	RT septian Dan sepanjang sejarah pencatatan KawalCovid Jawa Tengah konsisten menjadi provinsi dengan selisih yang terbesar antara da

3.1.3. Case Folding

The second step is casefolding, this is a sentence or document that must be changed from uppercase or capital letters to lowercase letters.

Table 5. Example of Case Folding Sample Data

No	Removing	Case folding
1	RT KawalCOVID Indonesia mengumumkan kasus baru COVID pd tgl Juni Total Kasus aktif Sembuh	rt kawalcovid indonesia mengumumkan kasus baru covid pd tgl juni total kasus aktif sembuh
2	mikachandra KawalCOVID BNPB Indonesia Tujuan utama digalakkannya New Normal adalah sektor ekonomi Pusat pere https t co NJjhRJc vg	mikachandra kawalcovid bnpb indonesia tujuan utama digalakkannya new normal adalah sektor ekonomi pusat pere https t co njjhrjc vg
3	firdzaradiany KawalCOVID kamilmoon mutiaranissa Kalau pemerintah bilang pandemi sudah membaik ya membaik -Presiden Burkina Faso	firdzaradiany kawalcovid kamilmoon mutiaranissa kalau pemerintah bilang pandemi sudah membaik ya membaik -presiden burkina faso
4	RT septian Dan sepanjang sejarah pencatatan KawalCovid Jawa Tengah konsisten menjadi provinsi dengan selisih yang terbesar antara da	rt septian dan sepanjang sejarah pencatatan kawalcovid jawa tengah konsisten menjadi provinsi dengan selisih yang terbesar antara da

3.1.4. Tokenization

The third stage is tokenization, which is a process of cutting or separating the input string based on each constituent word. For the case folding process to tokenization, it is marked with a sign (-) as a list of word compilers.

Table 6. Example of Tokenization Sample Data

No	Case folding	Tokenization
1	rt kawalcovid indonesia mengumumkan kasus baru covid pd tgl juni total kasus aktif sembuh	rt-kawalcovid-indonesia-mengumumkan-kasus-baru-covid-pd-tgl-juni-total-kasus-aktif-sembuh
2	mikachandra kawalcovid bnpb indonesia tujuan utama digalakkannya new normal adalah sektor ekonomi pusat pere https t co njjhrjc vg	mikachandra-kawalcovid-bnpb-indonesia-tujuan-utama-digalakkannya-new-normal-adalah-sektor-ekonomi-pusat-pere-https-t-co-njjhrjc-vg
3	firdzaradiany kawalcovid kamilmoon mutiaranissa kalau pemerintah bilang pandemi sudah membaik ya membaik -presiden burkina faso	firdzaradiany-kawalcovid-kamilmoon-mutiaranissa-kalau-pemerintah-bilang-pandemi-sudah-membaik-ya-membaik--presiden-burkina-faso
4	rt septian dan sepanjang sejarah pencatatan kawalcovid jawa tengah konsisten menjadi provinsi dengan selisih yang terbesar antara da	rt-septian-dan-sepanjang-sejarah-pencatatan-kawalcovid-jawa-tengah-konsisten-menjadi-provinsi-dengan-selisih-yang-terbesar-antara da

3.1.5. *Filtering*

The fourth stage is filtering, which is the next stage of tokenization results where every word that composes a document must be deleted if the words are less relevant or important. In this study in the fourth stage using the stoplist method.

Table 7. Example of Filtering Sample Data

No	Tokenization	Filtering
1	rt-kawalcovid-indonesia-mengumumkan-kasus-baru-covid-pd-tgl-juni-total-kasus-aktif-sembuh	kawalcovid - indonesia - mengumumkan - juni - total - aktif - sembuh
2	mikachandra-kawalcovid-bnpb-indonesia-tujuan-utama-digalakkannya-new-normal-adalah-sektor-ekonomi-pusat-pere-https-t-co-njjhrjc-vg	mikachandra - kawalcovid - bnpb - indonesia - tujuan - utama - digalakkannya - new - normal - sektor - ekonomi - pusat
3	firdzaradiany-kawalcovid-kamilmoon-mutiaranissa-kalau-pemerintah-bilang-pandemi-sudah-membaik-ya-membaik--presiden-burkina-faso	irdzaradiany - kawalcovid - kamilmoon - mutiaranissa - pemerintah - pandemi - membaik - membaik - - presiden - burkina - faso
4	rt-septian-dan-sepanjang-sejarah-pencatatan-kawalcovid-jawa-tengah-konsisten-menjadi-provinsi-dengan-selisih-yang-terbesar-antara da	septian - sejarah - pencatatan - kawalcovid - jawa - konsisten - provinsi - selisih - terbesar

3.1.6. *Stemming*

The last stage is stemming, which is the next stage where a filtered document becomes the basic word.

Table 8. Example of Stemming Sample Data

No	Filtering	Stemming
1	kawalcovid - indonesia - mengumumkan - juni - total - aktif - sembuh	indonesia - umum - juni - total - aktif - sembuh
2	mikachandra - kawalcovid - bnpb - indonesia - tujuan - utama - digalakkannya - new - normal - sektor - ekonomi - pusat	indonesia - tuju - utama - galak - normal - sektor - ekonomi - pusat
3	irdzaradiany - kawalcovid - kamilmoon - mutiaranissa - pemerintah - pandemi - membaik - membaik - -presiden - burkina - faso	perintah - pandemi - baik - baik
4	septian - sejarah - pencatatan - kawalcovid - jawa - konsisten - provinsi - selisih - terbesar	sejarah - catat - konsisten - provinsi - selisih - besar

3.2. *Result using The Default Word*

In the use of words in a sentence where there are things that are not recognized by the rules or enhanced spelling dictionaries that apply. The application of the default set of words can be seen in the table below as follows.

Table 9. Example of Stemming Sample Data

No	Default Word	Improvement	Basic word	Nature of Words
1	Ongkosny	Ongkosnya	Ongkos	Noun
2	Ngerepotin	Merepotkan	Repot	Adjective
3	Kadrun	-	-	-

4	Tunjukin	Tunjukkan	Tunjuk	Verb
5	Ikutin	Ikutan	Ikut	Verb
6	Brjudul	Berjudul	Judul	Noun
7	Ditotalin	Ditotalkan	Total	Noun
8	Diterapin	Diterapkan	Terap	Noun or Verb
9	Ngrasa	Merasakan	Rasa	Noun
10	Ngikuti	Mengikuti	Ikut	Verb

3.3. *Sample Document Selection*

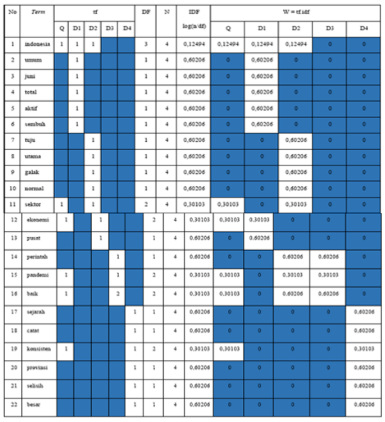
It is known that there are four sample documents in this study that are used are the results of the stemming stage. The equation formula for word weighting used in this study is the K-Nearest Neighbors method, namely tf-idf with Euclidean distance can be formulated as follows.

Table 10. Sample documents

No	Stemming	Class
1	indonesia umum juni total aktif sembuh	Positif
2	indonesia tuju utama galak normal sektor ekonomi pusat	Negatif
3	perintah pandemi baik baik	Positif
4	sejarah catat konsisten provinsi selisih besar	Positif

3.4. *Word Weighting Calculation Results (TF-IDF)*

The sample data used in this study above is calculated using the tf-idf (Term frequency-Inverse document frequency) formula which aims to classify the text of a document. TF (Term frequency) is the number of times the term appears in a document. For the use of this tf-idf by using the output on binary tf-idf, namely whether or not a term or word is present in the document, it can be said if it exists then it is given a value of 1 (one), while if it is not said, it is given a value of 0 (zero) " . To find out how to calculate tf-idf, the keyword (query) is prioritized because this query is randomly obtained from the data collection list from the entire stemming data above. As for how to calculate tf-idf can be seen below as follows. A search was conducted on the four sample documents of this study which were used with the information as in query is ”Indonesia's economic sector is consistent with the pandemic” and class as foloow in Figure 3.



No	*Term*	tf					DF	N	IDF	W = tf.idf				
		Q	D1	D2	D3	D4			log(n/df)	Q	D1	D2	D3	D4
1	indonesia	1	1	1			3	4	0,12494	0,12494	0,12494	0,12494	0	0
2	umum		1				1	4	0,60206	0	0,60206	0	0	0
3	juni		1				1	4	0,60206	0	0,60206	0	0	0
4	total		1				1	4	0,60206	0	0,60206	0	0	0
5	aktif		1				1	4	0,60206	0	0,60206	0	0	0
6	sembuh		1				1	4	0,60206	0	0,60206	0	0	0
7	tuju			1			1	4	0,60206	0	0	0,60206	0	0
8	utama			1			1	4	0,60206	0	0	0,60206	0	0
9	galak			1			1	4	0,60206	0	0	0,60206	0	0
10	normal			1			1	4	0,60206	0	0	0,60206	0	0
11	sektor	1		1			2	4	0,30103	0,30103	0	0,30103	0	0
12	ekonomi	1		1			2	4	0.30103	0.30103	0.30103	0	0	0
13	pusat			1			1	4	0.60206	0	0.60206	0	0	0
14	perintah				1		1	4	0.60206	0	0	0.60206	0.60206	0
15	pandemi	1			1		2	4	0.30103	0.30103	0	0.30103	0.30103	0
16	baik	1			2		2	4	0.30103	0.30103	0	0.60206	0.60206	0
17	sejarah					1	1	4	0,60206	0	0	0	0	0,60206
18	catat					1	1	4	0,60206	0	0	0	0	0,60206
19	konsisten	1				1	2	4	0.30103	0.30103	0	0	0	0.30103
20	provinsi					1	1	4	0,60206	0	0	0	0	0,60206
21	selisih					1	1	4	0,60206	0	0	0	0	0,60206
22	besar					1	1	4	0,60206	0	0	0	0	0,60206

Figure 3. TF-IDF Calculation Result

3.5. *Test Evaluation*

To find out the results of the calculation of the use of data splitting, the percentage values of 50:50, 60:40, 70:30 and 80:20 can be seen in the table below as follows :

Table 11. The results of calculating the best split dataset ratio

Split data 50:50%		
Class	Positif	Negatif
Positif	41	37
Negatif	0	97

Table 12. Results of confusion matrix

Precision	$\frac{41}{41+0} \times 100\%$	100%
Recall	$\frac{41}{41+37} \times 100\%$	0,525641 ≈ 0,52
Accuracy	$\frac{41+97}{41+37+0+97} \times 100\%$	0,788571 ≈ 0,78

Various uses of data splitting will evaluate the overall test results from the testing data above, the results obtained are as follows.

Table 13. Overall experimental evaluation of splitting the dataset

Sample Comparison (350)	K=5			K=7			K=9		
(Training/Testing)	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall
(50:50)	78%	100%	52%	77%	100%	50%	78%	100%	51%
(60:40)	75%	100%	50%	72%	100%	47%	72%	100%	47%
(70:30)	71%	100%	55%	69%	100%	51%	67%	100%	49%
(80:20)	77%	100%	65%	73%	100%	60%	71%	100%	57%

Based on Table 13, the research results obtained from the table above where a comparison of the distribution of the dataset between them, the best accuracy value is 78% with K-5 and K-9 values.

4. CONCLUSION

The results of the research analysis using the K-Nearest Neighbors method has been good percentage. The highest dataset splitting accuracy value is 50:50 data splitting with an accuracy of 78% with k-5 and k-9 values and the lowest value is 70:30 data splitting with an accuracy of 67 % with a value of k-9. The calculation of the K-Nearest Neighbors method to calculate the distance of the Euclidean distance value contained in this study there are 4 (four) sample documents that have positive and negative class labels and are combined with keywords (queries). Testing is done by k=3, from the 4 sample documents tested and yeilded the Euclidean distance values in the best accuracy is 78% using a 50:50 sample comparison with k-5 and k-9 values.

REFERENCES

[1] Chandra, D. N., Indrawan, G., & Sukajaya, I. N. (2016). Klas ifikasi Berita Lokal Radar Malang Menggunakan Metode Naïve Bayes Dengan Fitur N-Gram. *Jurnal Ilmiah Teknologi Dan Informasi ASIA (JITIKA)*.

[2] Kalokasari, D. H., Shofi, I. M., & Setyaningrum, A. H. (2017). IMPLEMENTASI ALGORITMA MULTINOMIAL NAIVE BAYES CLASSIFIER PADA SISTEM KLASIFIKASI SURAT KELUAR (Studi Kasus : DISKOMINFO Kabupaten Tangerang). *JURNAL TEKNIK INFORMATIKA*. https://doi.org/10.15408/jti.v10i2.6199

[3] Junianto, E., & Riana, D. (2017). Penerapan PSO Untuk Seleksi Fitur Pada Klasifikasi Dokumen Berita Menggunakan NBC. *Jurnal Informatika*.

[4] Nurhuda, F., Widya Sihwi, S., & Doewes, A. (2016). Analisis Sentimen Masyarakat terhadap Calon Presiden Indonesia 2014 berdasarkan Opini dari Twitter Menggunakan Metode Naive Bayes Classifier. *Jurnal Teknologi & Informasi ITSmart*. https://doi.org/10.20961/its.v2i2.630

[5] Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E. P., Yan, H., & Li, X. (2011). Comparing twitter and traditional media using topic models. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*.

[6] Pratama, B. Y., & Sarno, R. (2016). Personality classification based on Twitter text using Naive Bayes, KNN and SVM. *Proceedings of 2015 International Conference on Data and Software Engineering, ICODSE 2015*. https://doi.org/10.1109/ICODSE.2015.7436992

[7] Montejo-Ráez, A., Martínez-Cámara, E., Martín-Valdivia, M. T., & Ureña-López, L. A. (2014). Ranked WordNet graph for Sentiment Polarity Classification in Twitter. *Computer Speech and Language*. https://doi.org/10.1016/j.csl.2013.04.001

[8] Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., & Den Hartog, D. N. (2018). Text Mining in Organizational Research. *Organizational Research Methods*. https://doi.org/10.1177/1094428117722619

[9] Hashimi, H., Hafez, A., & Mathkour, H. (2015). Selection criteria for text mining approaches. *Computers in Human Behavior*. https://doi.org/10.1016/j.chb.2014.10.062

[10] Zulfa, I., & Winarko, E. (2017). Sentimen Analisis Tweet Berbahasa Indonesia Dengan Deep Belief Network. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*. https://doi.org/10.22146/ijccs.24716

[11] Tong, Z., & Zhang, H. (2016). *A Text Mining Research Based on LDA Topic Modelling*. https://doi.org/10.5121/csit.2016.60616

[12] Deviyanto, A., & Wahyudi, M. D. R. (2018). Penerapan Analisis Sentimen Pada Pengguna Twitter Menggunakan Metode K-Nearest Neighbor. Jiska. *JISKA (Jurnal Informatika Sunan Kalijaga)*. https://doi.org/10.14421/jiska.2018.31-01

[13] Jiang, S., Pang, G., Wu, M., & Kuang, L. (2012). An improved K-nearest-neighbor algorithm for text categorization. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2011.08.040

[14] Trishna, T. I., Emon, S. U., Ema, R. R., Sajal, G. I. H., Kundu, S., & Islam, T. (2019). Detection of Hepatitis (A, B, C and E) Viruses Based on Random Forest, K-nearest and Naïve Bayes Classifier. *2019 10th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2019*. https://doi.org/10.1109/ICCCNT45670.2019.8944455

[15] Istia, S. S., & Purnomo, H. D. (2018). Sentiment analysis of law enforcement performance using support vector machine and K-nearest neighbor. *Proceedings - 2018 3rd International Conference on Information Technology, Information Systems and Electrical Engineering, ICITISEE 2018*. https://doi.org/10.1109/ICITISEE.2018.8720969

[16] Sriwanna, K. (2018). Text classification for subjective scoring using K-nearest neighbors. *3rd International Conference on Digital Arts, Media and Technology, ICDAMT 2018*.

[17] Vijayan, V. K., Bindu, K. R., & Parameswaran, L. (2017). A comprehensive study of text classification algorithms. *2017 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2017*. https://doi.org/10.1109/ICACCI.2017.8125990

[18] Ary, M. (2016). Pengklasifikasian Karakteristik Mahasiswa Baru Dalam Memilih Program Studi Menggunakan Analisis Cluster. *Jurnal Informatika*. https://doi.org/10.31311/ji.v2i1.58