Sentiment Analysis of Student on Online Lectured During Covid-19 Pandemic Using K-Means and Naïve Bayes Classifier

Yusuf Affandi[1*], Endang Sugiharti[1]

[1] Department of Computer Science, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Semarang, Indonesia
*Corresponding author: ysufaffandi@students.unnes.ac.id

ARTICLE INFO

Article history
Received 21 March 2023
Revised 31 March 2023
Accepted 1 April 2023

Keywords
Covid-19 Pandemic
K-Means
Naïve Bayes Classifier
Sentiment Analysis
Twitter

ABSTRACT

The Covid-19 pandemic that occurred at the end of 2019 caused life changes, one of which was the learning process in universities. in accordance with the instructions issued by the Minister of Education as an effort to prevent the spread of Covid-19 by conducting online learning. Learning that is carried out online with a long period of time there are many obstacles such as networks and learning processes that are not optimal. Thus, students have mixed opinions on online lectures. Twiter is one of the social media used by students in expressing opinions on online lectures. The sentiment that users write on Twitter has not been determined in a more positive or negative direction. Sentiment analysis is needed to determine the tendency of student opinions towards online lectures. In this study, a sentiment analysis of online lectures was carried out using the K-Means and Naïve Bayes Classifier methods. The K-Means method is used to perform labeling or clustering and the Naïve Bayes Classifier is used as the classification. Based on research conducted with testing the Naïve Bayes Classifier model with a 70% division of training data and 30% test data using matrix confussion resulted in an accuracy of 95.67%.

1 Introduction

One of the sectors affected by the spread of Covid-19 is the education sector. The establishment of Covid-19 as a pandemic by WHO, the Minister of Education issued a prevention by making a Circular Letter (SE) on March 17 2020, the Minister of Education and Culture of the Republic of Indonesia Number 3 of 2020 concerning the Implementation of Education Policies in the Emergency Period of the Spread of Covid-19. In the circular letter, it is explained that in order to prevent the spread of Covid-19, all activities or learning processes carried out face-to-face are replaced with online learning / distance using video conferencing or the like until an undetermined time limit (Kementrian Pendidikan dan Kebudayaan, 2020).

However, the implementation of long-term distance learning due to Covid-19 did not just go smoothly. Online learning also has several disadvantages, namely requiring an adequate internet network, requiring more costs, there are various obstacles/slow in communicating (Waryanto, 2006). This causes students to have sentiments or opinions towards online lectures that take place, due to obstacles faced in the learning process.

In expressing opinions or sentiments, one of the social media used is Twitter. Twitter is a social media that works in real-time, allowing users to express opinions or sentiments about issues or problems that occur (Pravina et al., 2019). On Twitter social media, not all users write academic meaningful messages, so it takes a classification of documents to be able to find out messages that are academic and non-academic in which they contain sentiments from the author (Susilo & Rochimah, 2013).

Sentiment analysis is a field of study that analyzes the opinions, attitudes, and emotions of a person related to products, organizations, services, individuals and other topics (Sasikala & Marry Immaculate Sheela, 2020). The technique that develops in extracting documents in the form of text today is text mining (Budi, 2017). Text mining is a concept of data mining techniques to find a pattern in text, which is useful for finding information with a specific purpose. Text mining can be processed in various purposes including summarization, text document search, and sentiment analysis (Hasan & Wahyudi, 2018). Student opinions on online lectures during the Covid-19 pandemic need to be studied in the processing of sentiment analysis texts. Student sentiment analysis is a process to filter student opinions and classified into positive and negative classes, so that classification results are expected to help universities and students as a basis for consideration in running online lectures in the long term.

The Naïve Bayes algorithm is one of the classification algorithms based on Bayes' probabilisitic theorem. As in the research conducted by Andika et al. (2019) The use of the Naïve Bayes Classifier algorithm in sentiment analysis of the quick count results of the presidential election obtained an accuracy rate of 82.90%. Another study by Kurniawan & Susanto (2019) found that the use of two methods, namely K-Means and Naïve Bayes Clasifier in the analysis of the 2019 presidential election, resulted in better accuracy compared to only one method. The classification method using K-Means and the Naïve Bayes Classifier model resulted in an average accuracy of 93.35%. Furthermore, the research conducted by Lestari et al. (2017) entitled "Sentiment Analysis of DKI 2017 Regional Election Opinions on Indonesian Twitter Documents Using Naïve Bayes and Emoji Weighting" used Twitter data as many as 900 tweets. The Naïve Bayes and emoji methods are used as non-textual data weighting. The results of the accuracy of this study on textual weighting by 68.52%, non-textual weighting by 75.93%, and on the merger of the two weightings by 74.81%, with the conclusion that the merger of the two weightings improved the results of system accuracy.

Another study was conducted by Faesal et al. (2020) entitled "Sentiment Analysis on Twitter User Tweet Data Against Online Store Sales Products Using the K-Means method". In this study, 1,130 used Twitter data with Tokopedia keywords and obtained data as much as 1,130. Then the preprocessing and anti-word stages are carried out to limit the number of words that often appear, namely as many as 14 words for the clustering stage to be carried out. After that, it is divided into 3 clusters based on the frequency of occurrence of the word, namely often used, medium and rarely used. The K-Means method gives good results with an accuracy rate of 92.86%.

Analysis of student sentiment on online lectures is considered important (Musfiroh et al., 2021). Online learning or online lectures still have several obstacles including obstacles in the field of internet networks, limited features of online learning applications, and obstacles in terms of learning services (Hatauruk, 2020). Student opinions on online lectures during the Covid-19 pandemic need to be studied in the processing of sentiment analysis texts. Student sentiment analysis is a process to filter student opinions and classified into positive and negative classes, so that classification results are expected to help universities and students as a basis for consideration in running online lectures in the long term. Based on the description above, this research focuses on analyzing student sentiment using the K-Means and Naïve Bayes Classifier methods, this research is entitled "Sentiment Analysis on Online Lectured During Covid-19 Pandemic Using K-Means and Naïve Bayes Classifier Methods".

2 The Proposed Algorithm

2.1 Term Frequency – Invers Document Frequency (TF-IDF)

TF-IDF (Term Frequency–Inverse Document Frequency) is a method used to convert textual data into numerical data through weighting used in information retrival. TF-IDF is an algorithm used to calculate the value or weighting of terms or words in a document (Baeza-Yates & Ribeiro-Neto, 1999).

TF-IDF is a statistical weighting method that measures how important a word is to a document (Nugraha & Sebastian, 2018). TF is defined as the frequency with which words appear in a document. The IDF is a value used to indicate how important a word is to a document. The IDF value will get smaller as the frequency of occurrence of more words in the document. The value of the IDF will be even greater if the frequency with which words appear is only a small amount in the document. The IDF is defined the more frequency a word appears on a document, the less influence that word has on the document. The smaller the frequency with which words appear in the document, the greater the influence of the word on the document (Feldman & Sanger, 2007). According to Nugraha & Sebastian (2015) TF is defined as the frequency of words that appear in a document. The TF formula can be seen in Equation 1.

$$\boldsymbol{tf(t,d) = number\ of\ occurrences\ of\ t\ in\ d} \qquad (1)$$

Where:

$tf(t,d)$= The number of times the word t appears in document d

t = the word to be weighted

d = a document that has the word t in it

While the IDF is a value used to indicate how important a word is in a document. The IDF value will get smaller as the frequency of occurrence of more words in the document. Here is the Equation of the Inverse Document Frequency in Equation 2.

$$\boldsymbol{idf_t = log\ log\ \frac{N}{n_t}} \qquad (2)$$

Where :

idf_t = *idf* value for the word *t*

N = total documents in a class

n_t = number of documents containing the word *t*

After obtaining the TF value and the IDF value, the next stage is to calculate the value from the TF-IDF using Equation 3.

$$\boldsymbol{tfidf(t,d,D) = tf(t,d) * idf(t,d)} \qquad (3)$$

Where :

$tf(t,d)$ = term frequency of the word *t* in document *d*

$idf(t,d)$ = inverse dokument frequency of the word *t* in document *d*

2.2 K-Means

K-Means is one of the simplest unsupervised learning algorithms and has good performance to solve data clustering problems (Velmurugan & Santhanam, 2010). K-Means is a non-hierarchical data clustering method that classifies data into the form of one or more clusters. Data that has the same characteristics is grouped in one cluster and different data is grouped with other clusters. So that the data in the cluster has a small variation (Yudi Agusta, 2007).

The purpose of the K-means algorithm is to divide data into several clusters without training and labeling by randomly determining centroid points and determining clusters according to user needs. This centroid has a high degree of sensitivity to clustering results (Cui et al., 2005). Centroid determination and cluster count determination result in different cluster quality.

K-Means is a method that has several methods to calculate the nearest centroid, one of which is Euclidean distance, the stages of clustering data using the K-Means method are (Sulastri & Diartono, 2019):

1) Determine the number of k-clusters.
2) Grouping data up to obtaining k-cluster fruit with centroid points of each cluster which is a predetermined centroid point.

3) Calculates the centroid center of the existing data of each group. The centroid of each group is taken from the average (mean) of all data values on each feature.

4) Classify each of the data on the nearest centroid. To measure the distance of the data to the centroid center by calculating using the Euclidean distance formula in Equation 4.

$$w = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)} \tag{4}$$

The data is calculated and reclassified into groups that have centroids closest to the data. The process of calculating the reclassification of this data uses Equation 5.

$$\boldsymbol{d_{ij}} = \{\mathbf{1}\boldsymbol{w} = \{\boldsymbol{B}(\boldsymbol{x_i}, \boldsymbol{c_i})\} \quad \mathbf{0}\, other \tag{5}$$

d_{ij} is the value of the Xi point group to the C_1 centroid, w is the smallest distance from the Xi data to the k group after comparison, C_1 is the 1st centroid.

5) Repeat steps 2-4 until the cluster center and cluster members have not changed or remained the same.

2.3 Gaussian Naïve Bayes Classifier

This method is based on Thomas Bayes, an English scientist who introduced the Bayes Theorem, which is to predict future opportunities based on past experiences (Lee et al., 2001). The Naïve Bayes Classifier is one of several methods that can be used to classify data. Bayesian classification is a statistical classification that can be used to predict the probability of membership of a class.

In the implementation of sentiment analysis, past experiences are likened to training data and the future as test data. The Naïve Bayes Classifier method goes through two stages in the text classification process, namely the training stage and the classification stage. At the training stage, a process is carried out on a sample of data that can be represented by the data wherever possible. Next is the determination of the prior probability for each category based on sample data. At the classification stage, the category value of a data is determined based on the terms that appear in the classified data. (Nurhuda et al., 2016).

This classification method is suitable for large amounts of data. The Naïve Bayes Classifier classification is also preferred for its speed and simplicity (Buntoro, 2016). The Naïve Bayes Classifier algorithm aims to perform data classification on certain classes. The result of classifying work is measured by the value of predictive accuracy (Kusumadewi, 2009). The advantage of the Naïve Bayes Classifier algorithm is that Naïve Bayes only requires a small amount of training data to determine the estimated parameters required in the classification process (Kao & Poteet, 2007). According to (Karaca & Cattani, 2018) the Naïve Bayes Classifier has a general form as in Equation 6.

$$\boldsymbol{P}\,(\boldsymbol{X}) = \frac{\boldsymbol{P}(\boldsymbol{H})\boldsymbol{P}(\boldsymbol{H})}{\boldsymbol{P}(\boldsymbol{x})} \tag{6}$$

Where:

X = Data hypothesis x is a special class

H = data with unknown classes

P(H|X) = hypotestic probability H is based on condition X

P(H) = probability hypothesis H

P(X|H) = probability hypothesis X is based on condition H

P(X) = probability X

3 Method

In this study, student sentiment analysis in online lectures by applying TF-IDF feature extraction, the K-Means method as a data grouping method or automatic labeling and Naïve Bayes Classifier as a classification on the online lecture dataset. The flow of data analysis diagrams for student sentiment analysis at online lectures during the Covid-19 pandemic using the K-Means and Naïve Bayes Classifier methods can be seen in Figure 1.

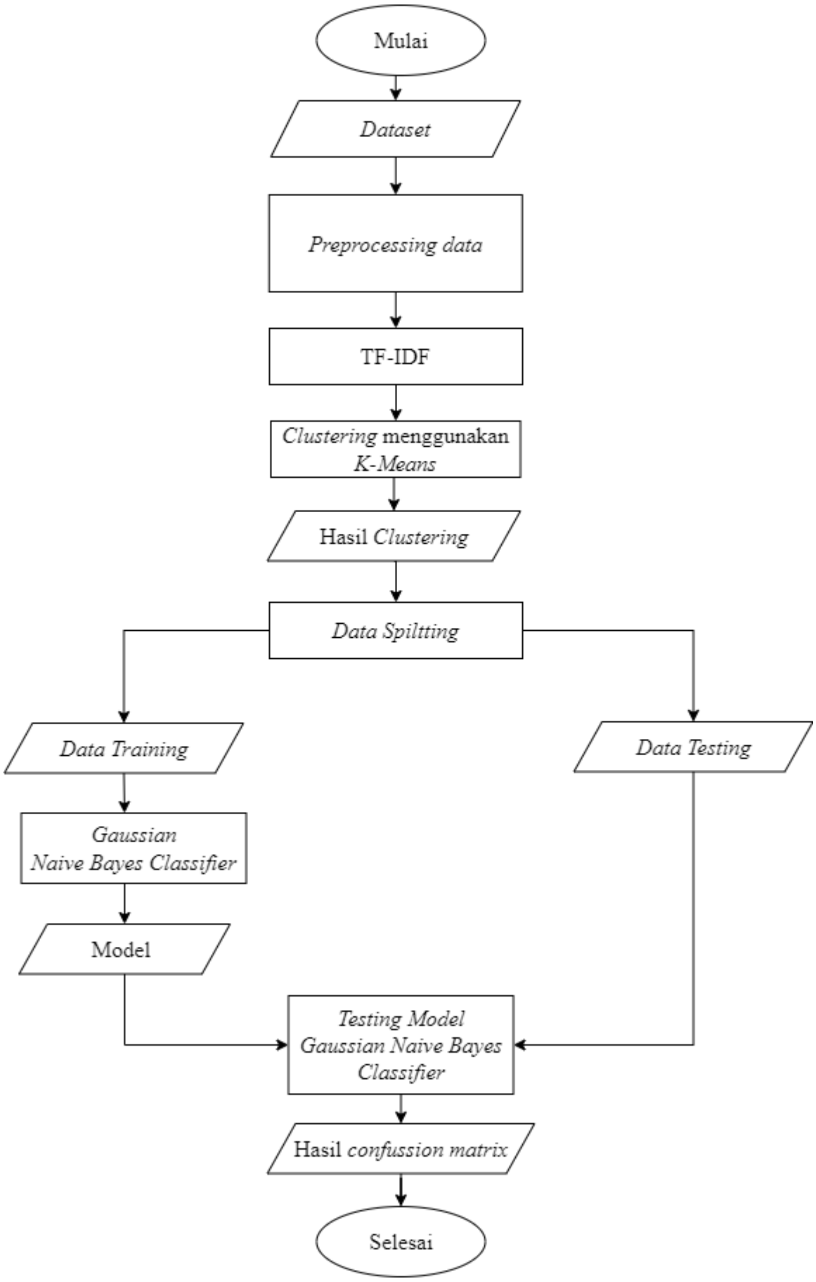

Figure 1. Figure of research steps

In the research of student sentiment analysis in online lectures using the K-Means and Naïve Bayes Classifier methods, it was carried out in several stages. The first stage is crawling Twitter data with kuliah online keywords. Then data preprocessing is carried out with stages of cleansing, case folding, normalization, stemming, stopword removal, tokenizing. The last stage performs classification using the Naïve Bayes Classifier method and accuracy calculations using matrix confusion.

4 Results and Discussion

This section is divided into 2 parts, results and discussion. Results are a description of the data and findings obtained using the methods and procedures described in the data collection method. The discussion is a review of the results that answer the research questions more comprehensively.

4.1 Result

In this study, the analysis of student sentiment in online lectures by applying the extraction of TF-IDF features, the K-Means method as an automatic labeling method and the Naïve Bayes Classifier as a classification on the kuliah online dataset on Twitter, had four research results. The four results are the results of data collection, data processing results, K-Means labeling results and classification results. The research results are as follows.

4.1.1 *Data Collection Result*

The dataset results were carried out by crawling with kuliah online keywords on Twitter with a period of February 23, 2022 to February 25, 2022 and obtained as many as 696 data in Indonesian.

Table 1. Data crawling results

Created-At	From-User	Text	Id
23/02/2022 07:39	Wicaksono ??	Kuliah tatap muka ditunda, tapi kafe-kafe penuh mahasiswa belajar online... ?	1,5E+18
23/02/2022 12:00	Ilham Kurniawan Sucipto????	@rianovskaya iya kan kak kuliah itu entah kenapa kek capek banget padahal sehari cuman 2-3 kelas doang, tambah lagi online malah jadi hmmmmmmm	1,5E+18
20/02/2022 21:26	BAYMAX	beli parfum mahal mahal buat kuliah malah daring terus asw.... udah mau masuk minggu ke 4 daring...	1,5E+18
20/02/2022 00:07	Nadin	@collegemenfess Belum pernah luring tapi aku suka kuliah daring. Sebenernya sih suka soalnya ga perlu ke kampus terus juga hemat gaperlu ngekost dan cari makan sendiri	

The data obtained from crawling Twitter still has noise so that preprocessing is carried out to remove words and characters that have no meaning so that they can produce a good dataset for the next process. The preprocessing stage that will be carried out in this study is case folding, cleansing, normalization, stemming, stopword removal, tokenizing.

4.1.2 *Case folding*

The principle of case folding is to change the character of the uppercase letter to lowercase. The results of the case folding can be seen in Table 2.

Table 2. Case folding results

Text	Case folding
Kuliah tatap muka ditunda, tapi kafe-kafe penuh mahasiswa belajar online... ?	kuliah tatap muka ditunda, tapi kafe-kafe penuh mahasiswa belajar online... ?
@rianovskaya iya kan kak kuliah itu entah kenapa kek capek banget padahal sehari cuman 2-3 kelas doang, tambah lagi online malah jadi hmmmmmmm	@rianovskaya iya kan kak kuliah itu entah kenapa kek capek banget padahal sehari cuman 2-3 kelas doang, tambah lagi online malah jadi hmmmmmmm

beli parfum mahal mahal buat kuliah malah daring terus asw.... udah mau masuk minggu ke 4 daring...	beli parfum mahal mahal buat kuliah malah daring terus asw.... udah mau masuk minggu ke 4 daring...
@collegemenfess Belum pernah luring tapi aku suka kuliah daring. Sebenernya sih suka soalnya ga perlu ke kampus terus juga hemat gaperlu ngekost dan cari makan sendiri	@collegemenfess belum pernah luring tapi aku suka kuliah daring. sebenernya sih suka soalnya ga perlu ke kampus terus juga hemat gaperlu ngekost dan cari makan sendiri

4.1.3 *Cleansing*

At this stage, all less important characters will be removed. The data cleaning process will remove punctuation, tags, URLs, emoticons. The cleansing results can be seen in Table 3.

Table 3. Cleansing Result

Text	Cleansing
kuliah tatap muka ditunda, tapi kafe-kafe penuh mahasiswa belajar online... ?	kuliah tatap muka ditunda tapi kafe kafe penuh mahasiswa belajar online
@rianovskaya iya kan kak kuliah itu entah kenapa kek capek banget padahal sehari cuman 2-3 kelas doang, tambah lagi online malah jadi hmmmmmmm	iya kan kak kuliah itu entah kenapa kek capek banget padahal sehari cuman kelas doang tambah lagi online malah jadi hmmmmmmm
beli parfum mahal mahal buat kuliah malah daring terus asw.... udah mau masuk minggu ke 4 daring...	beli parfum mahal mahal buat kuliah malah daring terus asw udah mau masuk minggu ke daring
@collegemenfess belum pernah luring tapi aku suka kuliah daring. sebenernya sih suka soalnya ga perlu ke kampus terus juga hemat gaperlu ngekost dan cari makan sendiri	belum pernah luring tapi aku suka kuliah daring sebenernya sih suka soalnya ga perlu ke kampus terus juga hemat gaperlu ngekost dan cari makan sendiri

4.1.4 *Normalization*

At this stage, the process of converting non-standard words into standard words is carried out. Normalization results can be seen in Table 4.

Table 4. Normalization result

Text	Normalization
kuliah tatap muka ditunda tapi kafe kafe penuh mahasiswa belajar online	kuliah tatap muka ditunda tapi kafe kafe penuh mahasiswa belajar online
iya kan kak kuliah itu entah kenapa kek capek banget padahal sehari cuman kelas doang tambah lagi online malah jadi hmmmmmmm	iya kan kak kuliah itu entah kenapa kek capai sekali padahal sehari hanya kelas saja tambah lagi online malah jadi hmmmmmmm
beli parfum mahal mahal buat kuliah malah beli parfum mahal mahal buat kuliah malah daring terus asw udah mau masuk minggu ke daring	beli parfum mahal mahal buat kuliah malah daring terus asw sudah mau masuk minggu ke daring
belum pernah luring tapi aku suka kuliah daring sebenernya sih suka soalnya ga perlu ke kampus terus juga hemat gaperlu ngekost dan cari makan sendiri	belum pernah luring tapi saya suka kuliah daring sebenarnya sih suka soalnya tidak perlu ke kampus terus juga hemat gaperlu ngekost dan cari makan sendiri

4.1.5 *Stemming*

The stemming process is the process of converting a word form into a basic word form. Stemming results can be seen in Table 5.

Table 5. Stemming result

Text	Stemming
kuliah tatap muka ditunda tapi kafe kafe penuh mahasiswa belajar online	kuliah tatap muka tunda tapi kafe kafe penuh mahasiswa ajar online
iya kan kak kuliah itu entah kenapa kek capek banget padahal sehari cuman kelas doang tambah lagi online malah jadi hmmmmmmm	iya kan kak kuliah itu entah kenapa kek capai sekali padahal hari hanya kelas saja tambah lagi online malah jadi hmmmmmmm
beli parfum mahal mahal buat kuliah malah beli parfum mahal mahal buat kuliah malah daring terus asw udah mau masuk minggu ke daring	beli parfum mahal mahal buat kuliah malah daring terus asw sudah mau masuk minggu ke daring
belum pernah luring tapi aku suka kuliah daring sebenernya sih suka soalnya ga perlu ke kampus terus juga hemat gaperlu ngekost dan cari makan sendiri	belum pernah luring tapi saya suka kuliah daring benar sih suka soal tidak perlu ke kampus terus juga hemat gaperlu ngekost dan cari makan sendiri

4.1.6 *Stopword Removal*

The stopword removal process aims to eliminate words that have a high frequency of occurrence and have no meaning or influence in the classification process. Thus, resulting in the selection of important words that represent a sentence. The results of stopword removal can be seen in Table 6.

Table 6. Stopword removal result

Text	Stopword removal
kuliah tatap muka tunda tapi kafe kafe penuh mahasiswa ajar online	kuliah tatap muka tunda kafe kafe penuh mahasiswa ajar online
iya kan kak kuliah itu entah kenapa kek capai sekali padahal hari hanya kelas saja tambah lagi online malah jadi hmmmmmmm	iya kan kak kuliah entah kek capai sekali padahal hari kelas tambah online malah jadi hmmmmmmm
beli parfum mahal mahal buat kuliah malah daring terus asw sudah mau masuk minggu ke daring	beli parfum mahal mahal buat kuliah malah daring terus asw mau masuk minggu daring
belum pernah luring tapi saya suka kuliah daring benar sih suka soal tidak perlu ke kampus terus juga hemat gaperlu ngekost dan cari makan sendiri	pernah luring saya suka kuliah daring benar sih suka soal perlu kampus terus hemat gaperlu ngekost cari makan sendiri

4.1.7 *Tokenizing*

Tokenizing aims to turn words into tokens or break sentences into word by word. Tokenizing results can be seen in Table 7.

Table 7. Tokenizing result

Text	Tokenizing
kuliah tatap muka tunda kafe kafe penuh mahasiswa ajar online	['kuliah', 'tatap', 'muka', 'tunda', 'kafe', 'kafe', 'penuh', 'mahasiswa', 'ajar', 'online']

iya kan kak kuliah entah kek capai sekali padahal hari kelas tambah online malah jadi hmmmmmmm	['iya', 'kan', 'kak', 'kuliah', 'entah', 'kek', 'capai', 'sekali', 'padahal', 'hari', 'kelas', 'tambah', 'online', 'malah', 'jadi', 'hmmmmmmm']
beli parfum mahal mahal buat kuliah malah beli parfum mahal mahal buat kuliah malah daring terus asw mau masuk minggu daring	['beli', 'parfum', 'mahal', 'mahal', 'buat', 'kuliah', 'malah', 'daring', 'terus', 'asw', 'mau', 'masuk', 'minggu', 'daring']
pernah luring saya suka kuliah daring benar sih suka soal perlu kampus terus hemat gaperlu ngekost cari makan sendiri	['pernah', 'luring', 'saya', 'suka', 'kuliah', 'daring', 'benar', 'sih', 'suka', 'soal', 'perlu', 'kampus', 'terus', 'hemat', 'gaperlu', 'ngekost', 'cari', 'makan', 'sendiri']

4.1.8 *Term Frequency – Invers Document Frequency*

Datasets that have passed the preprocessing stage are still in the form of text, in sentiment analysis in order for the algorithm to read the dataset used, it is necessary to convert text data into numerical data. The data must first be converted to a number form using the TF-IDF extraction feature. The results of the TF-IDF can be seen in Table 8.

Table 8. TF-IDF result

Preprocessing result	Word	TF-IDF result
['kuliah', 'tatap', 'muka', 'tunda', 'kafe', 'penuh', 'mahasiswa', 'ajar', 'online']	kuliah	0,05890163
	tatap	0,31150183
	muka	0,31150183
	tunda	0,37547338
	kafe	0,63573170
	penuh	0,37547338
	mahasiswa	0,23077871
	ajar	0,23948586
	online	0,07573776

4.1.9 *K-Means*

Data that has been carried out the process of preprocessing and word weighting using TF-IDF then the data is labeled using the K-Means clustering method. At this stage the researcher determines k = 2, namely cluster 1 as a positive cluster and cluster 2 as a negative cluster. Cumulative results of labeling based on two categories namely 504 positive and 192 negative from 696 data. The results of the K-Means can be seen in Table 9.

Table 9. K-Means result

Text	Sentiment
kuliah tatap muka tunda kafe kafe penuh mahasiswa ajar online	Positif
iya kan kak kuliah entah kek capai sekali padahal hari kelas tambah online malah jadi hmmmmmmm	Positif
beli parfum mahal mahal buat kuliah malah beli parfum mahal mahal buat kuliah malah daring terus asw mau masuk minggu daring	Negatif

pernah luring saya suka kuliah daring benar sih suka soal perlu kampus terus hemat gaperlu ngekost cari makan sendiri	Negatif

4.1.10 *Classification Naïve Bayes Classifier result*

The classification produced in this study using the K-Means clustering method as labeling obtained tweets with a positive sentiment label of 504 and tweets with a negative sentiment label of 192. From the results of classification testing on the Naïve Bayes Classifier, it is known that the accuracy performance produced by the Naïve Bayes Classifier method. Based on the test results, the Naïve Bayes Classifier method produced an accuracy of 95.67%. The accuracy results using the Naïve Bayes Classifier confusion matrix can be seen in Table 10.

Table 10. Confussion matrix result

		Prediction	
		Positive	Negative
Actual	Positive	156	2
	Negative	7	44
Accuracy		95,69378%	

4.2 Discussion

This study used Indonesian tweet data containing online lecture keywords on Twiiter social media. This study used the K-Means method in the sentiment analysis process. The Naïve Bayes Classifier method is used as a model to classify the sentiment analysis process on online lecture tweet data.

In this study, the data retrieval process was carried out using crawling techniques using RapidMiner studio. The crawling process is carried out by determining the data to be taken as many as 1000, provided that the tweet data taken contains online lecture keywords. Determining data crawling is based on restrictions on the Twitter API, the Twittter API also restricts users from retrieving data only within one week of the data crawl. To obtain data that is longer than a week's time span Twitter APIs require users to crawl using a paid account. From the crawling process, 696 tweet data were obtained in the period from February 23 to February 25, 2022. Then the data obtained is carried out a preprocessing process to clean the data from feature features that are not needed in the classification process.

The labeling process in this study uses the K-Means method to cluster datasets that have been carried out by the preprocessing and word weighting process using TF-IDF. The clusetring process is carried out by specifying k = 2. C1 as a positive cluster and C2 as a negative cluster. Then a random determination of the initial centroid is carried out to carry out the clustering process. Determination of clusters is carried out by calculating the distance of the closest object to the centroid. This process is carried out until the object does not undergo changes. The process on the K-Means method is carried out with the help of the sklearn.cluster library.

From the clustering process of the K-Means online lecture Twitter dataset will be categorized into two sentiments, namely positive sentiment and negative sentiment based on the results of the clustering that has been carried out. From the K-Means clustering process, results were obtained with the number of positive sentiment labels (C1) as many as 504 and negative sentiment labels (C2) as many as 192 from a total data of 696. This K-Means clustering result data will be used for the testing process on model creation for sentiment analysis. At the model creation stage, data training is carried out by the Naïve Bayes Classifier method.

In the study, 70% of training data and 30% of testing data were distributed. The results of the accuracy test from making a model using the Naïve Bayes Classifier method showed an accuracy of 95.67% based on the calculation of the confussion matrix with false positive predictions 7, true positive prediction 156, true negative predictions 44, false negative predictions 2.

The advantage of this study is that the results of testing the model obtained good accuracy in processing the dataset obtained from crawling results using online lecture keywords on Twitter social media by clustering K-Means in the labeling process through the preprocessing process, TF-IDF word weighting and classification with the Naïve Bayes classifier method. As well as a simple web-based system that is easy to use by users.

However, this study also has a drawback, namely that this study only tested the Twitter dataset about online lectures in Indonesian and has not been fully maximized in the process of eliminating words that are not up to Indonesian standards and eliminating words that do not have important weight in a document. As well as the labeling process using the K-Means method which does not have a database of positive words and negative words. So that the results of the labeling results are unbalanced and do not have labeling accuracy. Therefore, further development is needed in the labeling process so that the labeling data can produce better labeling data based on the positive and negative words contained in the document.

5 Conclusion

The implementation in analyzing student sentiment towards online lectures on Twitter social media using the K-Means and Naïve Bayes Classifier methods was carried out in several stages. The initial stage carried out in this study was to crawl Twitter with online lecture keywords on Twitter social media. The second stage is to preprocess data with the stages of case folding, cleansing, normalization, stemming, stopword removal and tokenizing. The third stage is to do word weighting using the TF-IDF extraction feature. The next stage is clustering using the K-Means method as automatic labeling in this study. The last stage is to classify using the Naïve Bayes Classifier method. This study applies the K-Means method in clustering or automatic labeling and tests the accuracy of the Naïve Bayes Classifier method in analyzing student sentiment at online lectures during the Covid-19 pandemic with a dataset obtained by crawling Twitter social media. The accuracy testing process of the Naïve Bayes Classifier method was carried out using a confussion matrix with a division ratio of 70% of training data and 30% of testing data and obtained an accuracy of 95.67%. Based on the classification results, it can be concluded that K-Means can do clustering well and the accuracy results of the Naïve Bayes Classifier method go well in analyzing student sentiments about online lectures.

6 Reference

Andika, L. A., Azizah, P. A. N., & Respatiwulan, R. (2019). Analisis sentimen masyarakat terhadap hasil quick count pemilihan presiden indonesia 2019 pada media sosial twitter menggunakan metode Naive Bayes Classifier. *Indonesian Journal of Applied Statistics*, *2*(1), 34–41.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval* (Vol. 463). ACM press New York.

Budi, S. (2017). Text mining untuk analisis sentimen review film menggunakan algoritma K-means. *Techno.Com*, *16*(1), 1–8.

Buntoro, G. A. (2016). Analisis sentimen hatespeech pada twitter dengan metode Naïve Bayes Classifier dan support vector machine. *Jurnal Dinamika Informatika*, *5*(2), 1–13.

Cui, X., Potok, T. E., & Palathingal, P. (2005). Document clustering using particle swarm optimization. *Proceedings - 2005 IEEE Swarm Intelligence Symposium, SIS 2005*, *2005*, 191–197.

Feldman, R., & Sanger, J. (2007). Advanced approaches in analyzing unstructured data. *The Text Mining Handbook, Cambrige University*.

Hasan, F. N., & Wahyudi, M. (2018). Analisis sentimen artikel berita tokoh sepak bola dunia menggunakan algoritma support vector machine dan Naive Bayes berbasis particle swarm optimization. *Jurnal Akrab Juara*, *3*(4), 42–55.

Kao, A., & Poteet, S. R. (2007). *Natural language processing and text mining*. Springer Science & Business Media.

Kementerian Pendidikan dan Kebudayaan. (2020). Surat edaran menteri pendidikan dan kebudayaan republik Indonesia tentang pencegahan corona virus disease (covid-19) pada satuan pendidikan. *Surat Edaran Menteri Pendidikan Dan Kebudayaan Republik Indonesia Nomor 3 Tahun 2020*, *3*(1), 2.

Kurniawan, I., & Susanto, A. (2019). Implementasi metode K-means dan Naïve Bayes classifier untuk analisis sentimen pemilihan presiden (pilpres) 2019. *Eksplora Informatika*, *9*(1), 1–10.

Kusumadewi, S. (2009). Klasifikasi status gizi menggunakan naive bayesian classification. *CommIT (Communication and Information Technology) Journal*, *3*(1), 6.

Lee, E. P. F., Lee, E. P. F., Lozeille, J., Soldán, P., Daire, S. E., Dyke, J. M., & Wright, T. G. (2001). An ab initio study of RbO, CsO and FrO (X2∑+; A2∏) and their cations (X3∑-; A3∏). Physical Chemistry Chemical Physics, 3(22), 4863–4869.

Nugraha, K. A., & Sebastian, D. (2018). Analisis trend akun media sosial twitter menggunakan TF-IDF dan Cosine similarity. *Prosiding Seminar Nasional ReTII*, *0*(0), 103–110

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*, *2*(2), 35.

Pravina, A. M., Cholissodin, I., & Adikara, P. P. (2019). Analisis sentimen tentang opini maskapai penerbangan pada dokumen twitter menggunakan algoritme support vector machine (svm). *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer E-ISSN*, *2548*, 964X.

Sulastri, S., & Diartono, D. A. (2019). Analisa jejaring sosial twitter menggunakan klastering K-means dan hirarki agglomeratif. *Prosiding Sendi*, 261-271.

Susilo, T. H., & Rochimah, S. (2013). Pengklasifikasian topik dan analisis sentimen dalam media sosial. *Snasti*, 1–9.

Velmurugan, T., & Santhanam, T. (2010). Computational complexity between K-means and K-medoids clustering algorithms for normal and uniform distributions of data points. *Journal of Computer Science*, *6*(3), 363.

Waryanto, N. H. (2006). Online learning sebagai salah satu inovasi pembelajaran. *Pythagoras*, *2*(1), 10–23.

Yudi Agusta. (2007). K-Means – Penerapan, permasalahan dan metode terkait. *jurnal sistem dan informatika*, *3*(Februari), 47–60.