Journal of Harbin Engineering University Vol 44 No. 7
ISSN: 1006-7043 July 2023

Sentiment Analysis of *Komisi Pemberantasan Korupsi* (KPK) on Twitter Social Media by applying the Algorithm Naïve Bayes Classifier

Gigih Forda Nama
Department of Informatics, Faculty of Engineering, University of Lampung, Indonesia
Doctoral Program of Environmental Science, University of Lampung, Indonesia

Hendy Syuhada
Department of Informatics, Faculty of Engineering, University of Lampung, Indonesia

Titin Yulianti
Department of Informatics, Faculty of Engineering, University of Lampung, Indonesia

*** Coresponding Author:-** Gigih Forda Nama
Email: gigih@eng.unila.ac.id

Abstract-Komisi Pemberantasan Korupsi (KPK) is an official institution in Indonesia specifically assigned to handle corruption cases. Along with the rise of corruption cases in Indonesia, the public also expressed opinions on the performance of the KPK, which was conveyed through Twitter. However, this opinion is still vaguely positive or negative. Therefore, in this study, sentiment analysis was carried out on this opinion using the machine learning-based naïve bayes classifier algorithm.
Data comes from Twitter which is taken by crawling technique through API (Application Programming Interface). The data is processed through several stages, namely preprocessing which includes removing punctuation marks, removing repetitive words and words that often appear but do not really have meaning in sentences. The next stage is data labeling which is done manually by assigning a label or class to the data. Next is the modeling process, which is the process of building an appropriate model to predict the probability of incoming data and classifying them according to the previous probability calculations. The data used in the modeling process is 2055 tweet data which is divided into training sets and testing sets with a ratio of 80:20. Next, a system deployment with the chosen model was carried out to analyze sentiment towards the KPK on Twitter.
The results of this study indicate that using the multinomial Naïve Bayes Classifier model, the precision value is 0.69, the recall is 0.89, the F-1 Score is 0.74, and the accuracy is 64%. In this study, a website was also developed to retrieve new data which then automatically classified it into positive, or neutral labels. This website also displays the results in the form of tables and graphs.

Keywords: Sentiment Analysis, KPK, Multinomial model, Naïve Bayes Classifier, Twitter

I. INTRODUCTION

Corruption is categorized as one of the extraordinary crimes because corruption causes damage to the democratic process and the social and economic rights of the wider community. Therefore, extraordinary prevention and treatment efforts are needed as well. The institution authorized to do that is Komisi Pemberantasan Korupsi or KPK.
Komisi Pemberantasan Korupsi is an institution engaged in eradicating corruption. However, based on Article 6 of Law no. 30 of 2002 concerning the Komisi Pemberantasan Korupsi, the task of the KPK is not only in terms of eradicating, but also coordinating with agencies authorized to eradicate corruption, conducting supervision, investigation, investigation, and prosecution of criminal acts of corruption, and monitoring to the administration of the State government [1].
News about KPK many displayed in online news portals such as kompas and detikcom, but these news portals do not provide an API so that they can access more about the news contained in them. Therefore, other media are needed that accommodate a lot of news about the KPK and have an API to access the news. One of the social media that is widely used by the public and provides an API is Twitter.
Twitter is media which used to facilitate users in communicating and obtaining information on various topics. Many people give their opinion on the KPK through Twitter, but the opinion is still unclear whether it is positive or not negative. Based on these problems, it is necessary to conduct a sentiment analysis of public opinion so that it is known that the opinion is positive or negative so that in the end the performance of the KPK can be judged as good or bad.

II. MATERIALS AND METHODS

2.1Related Research

There are several related studies that serve as comparisons and references regarding the methods used in this study.

Research by Alec Go et al. from Stanford University aims to classify the sentiment of Twitter messages as positive messages and negative messages. This study uses three algorithms from machine learning, that is naïve bayes, maximum entropy and support vector machine. The results of this study have an accuracy above 80% when trained usingdata emoticonsfor the three algorithms [2].

In the 2018 study, a public opinion survey was conducted on an event, object, or location. Then these sentiments can be used by other tourists to decide whether to go to that place or not [3].

In 2009 a study was conducted to find a comparison of algorithms in data mining by Daniela Xhemali et al. from Loughborough University. This study focuses on the comparison naïve bayes, decision tree, and neural network to automatically analyze and classify data on web training course. The results of this study prove that the Algorithm naïve bayes more outperform in system classifying data [4].

In other studies, modeling and forecasting are carried outtime series with Python and Jupyter Notebook as software open source to run the system on the shares of PT. Bank Negara Indonesia in 2020. In the forecasting results, it was concluded that the stock price of PT BNI in the next 3 years has an upward trend. In this forecasting there is a possibility that stock prices are not only influenced by time but can also be influenced by other factors [5].

In 2019, a sentiment analysis was also carried out on the performance of the House of Representatives (DPR) which was expressed by the public through Twitter social media. This research uses the method of naïve bayes classifier and using as many as 1546 data tweets. The results of this study found that the DPR got 95 tweets positive with 0.75 polarity or 75% positive sentiment, 693 tweets neutral with a polarity of 0.79 or 79% neutral sentiment and 758 tweets negative with a polarity of 0.82 or 82% negative sentiment with an accuracy score of 0.8 or 80% based on data testing as much as 20% [6].

Analysis text mining from Twitter about infrastructure in Indonesia with the classification method naïve bayes in 2019 the results showed that the proportion of negative sentiment was greater than positive sentiment. In addition, the results of the classification using the method naïve bayes the best model is obtained on the airport model with an accuracy of 82%, a precision of 0.84 and a recall of 0.48 [7].

Algorithm naïve bayes also used in research on film opinion sentiment analysis on Twitter using about 500 test data divided into 400 data training and 100 data testing by using the evaluation method K-Fold Cross Validation. Based on the experimental results, the sentiment analysis that can be done by the system with accuracy obtained is 90% with details of 92% precision value, 90% recall and 90% f-measure [8].

In another study using the KNN algorithm (K-Nearest Neighbor) in the sentiment analysis of Twitter users on the topic of the 2017 DKI Pilkada. Data tweets used is as much as 2000 data tweets Indonesian language that was collected during January 2017 using a Python package called Twitter scraper. Using the KNN algorithm with TF-IDF word weighting and the Cosine Similarity function, and classifying sentiment values into two classes: positive and negative. The test results show that the highest accuracy value is 67.2% when k=5, the highest precision is 56.94% when k=5, and recall is 78.24% with k=15 [9].

2.2 Twitter

Twitter is a social networking media that allows its users to send and read text-based messages of up to 280 characters which are known as tweets or tweets. One of the advantages of Twitter social media is that it provides an API (Application Programming Interface) which is very good, making it easier for everyone to retrieve data from Twitter [10].

2.3 Sentiment Analysis

Sentiment analysis is a process in data mining that is used to identify and extract information from a text that aims to understand social sentiment on the text. Sentiment analysis also used to obtain information about the attitudes, opinions and emotions contained in the information text. Task from sentiment analysis is to classify the polarity of the text in the text, classified as positive text or negative text [11].

2.4 Python

Python is an object-based programming language and can be run on all platforms. Programs in Python are run through an interpreter so that programs in Python can be tested directly. Python programming language supports object-oriented programming concepts so that in Python there are various library and framework used to analyze the data [12]. Here are some library found in Python:

1. Tweepy.
2. NLTK *(Natural Language Toolkit).*
3. Sastrawi.
4. *Scikit-learn.*

2.5 Naïve Bayes

Naïve bayes is one of the popular classification algorithms and has competitive performance in the classification process proposed by Thomas Bayes. Algorithm naïve bayes predicting future opportunities based on past experience so it is known as Bayes theorem [13]. Algorithm method naïve bayes classified into several types based on their function, namely:

1. Multinomial Naïve Bayes.
2. Bernoulli Naïve Bayes.
3. Gaussian Naïve Bayes [14].

In some practical forms, the parameters for the calculation of the model naïve bayes using method maximum likelihood or the highest similarity. For the realm of classification, the calculated P(H|X) is the

probability that the hypothesis is true for the observed sample X data, which can be applied to equation (1).

$$P(H|X)=\frac{P(H|X)P(H)}{P(X)} \quad (1)$$

Keterangan:

X = Sample data with labels not yet known
H = Hypothesis that X is data with label C
P(H|X) = Probability that hypothesisi is true for the observed sample X data
P(X|H) = Probability of sample data X, if it is assumed that the hypothesis is true.
P(H) = Probability of the hypothesis H
P(X) = Probability of observed sample data [8]

2.6 Jupyter Notebook

Jupyter Notebook makes it possible to integrate code with output in a single document interactively. Jupyter supports several computing products that can be used, including:

1. Jupiter Notebook.
2. JupyterHub.
3. JupyterLab [15].

2.7 PostgreSQL

PostgreSQL is a SQL database system (Structured Query Language) object-relational which is robust and highly extensible and popular for its reliability, robustness of features and high performance. PostgreSQL using models client-server where the client and server can reside on different hosts in a networked environment. PostgreSQL supports several data types including primitives (such as string, integer, numeric, and boolean), structured (such as time, array, range, and UUID), documents (JSON, JSONB, XML), and geometries (points, lines, circles, and polygons) [16].

2.8 Framework Django

Django is a popular python based framework, and is used for web development. Django follows its own convention of the Model-View-Controller (MVC) architecture called the Model View Template (MVT). MVT is a software design pattern consisting mainly of 3 components Model, View, and Template. Each user requests some resource, then Django acts as a controller and looks for the resource in the urls.py file. If a URL maps, the view associated with that URL is called. After that, the view interacts with the model and template. In the end, Django responds to the user and returns a template in response [17].

2.9 CRISP-DM

Cross Industry Standard Process for Data Mining (CRISP-DM) was developed in 1996 by analysis from several industries such as Daimler Chrysler standardization, SPSS, NCR. CRISP-DM provides a standard data mining process as a general problem solving strategy of a business or research unit [18]. Some related work to processing data using open source application was found on [19][20] [21] [22][23]

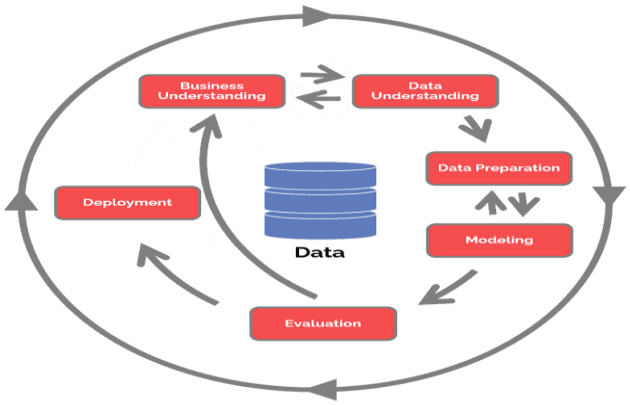


Figure 1 Stages of the CRISP-DM research method

The process in CRISP-DM consists of 6 phases of activities, namely:

1. Business Understanding.
2. Data Understanding.
3. Data Preparation.
4. Modeling.
5. Evaluation.
6. Deployment.

III. Results and Discussions

The method used in this research is Cross Industry Standard Process for Data Mining (CRISP-DM). The following is a flow chart in the system development process carried out in this study.

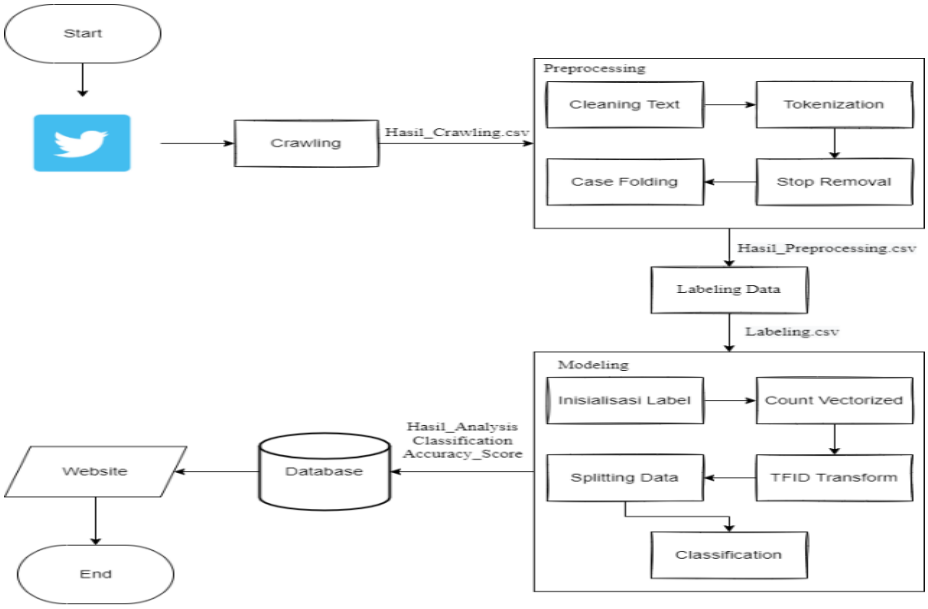


Figure 2 *Flowchart* research stages

Business Understanding

Business understanding generally focuses on understanding goals and needs from a business point of view, which is then translated into problems in data mining. Then a specific plan is made to achieve that goal.

The purpose of this research is to create a model or sentiment analysis system based on tweets the community on Twitter social media about the performance of the Komisi Pemberantasan Korupsi, then analyze the performance of the model naïve bayes in classifying tweets that can be obtained, and then make website for data visualization.

3.1 Data Understanding

At the data understanding stage, an understanding of the data needs related to the previous business objectives will be carried out and collect the dataset obtained. In this study, the dataset was obtained through Twitter social media using the API (Application Programming Interface) in the form of data tweets on Twitter. In this process, data is collected on Twitter which contains attributes, including date tweets made, username users, as well as the contents of tweets with query KPK. The dataset obtained in this process is saved with the extension .csv. Then the data that has been obtained will be analyzed to obtain information.

3.2 Data Preparation

The first step in data preparation is data crawl. Next is the stage data pre-processing, on data pre-processing carried out several steps such as cleaning text, tokenization, stopwords removal, and case folding. Then the next stage is data labelling the process of labeling or class on tweets manually stored in the dataset.

3.3.1 Crawling Data

Crawl is technique taking data use API (Application Programming Interface) on Twitter with Python programming language and using tweepy library. In this research, crawling on Twitter with the keyword KPK to get tweets from Twitter social media users who discussed the KPK.

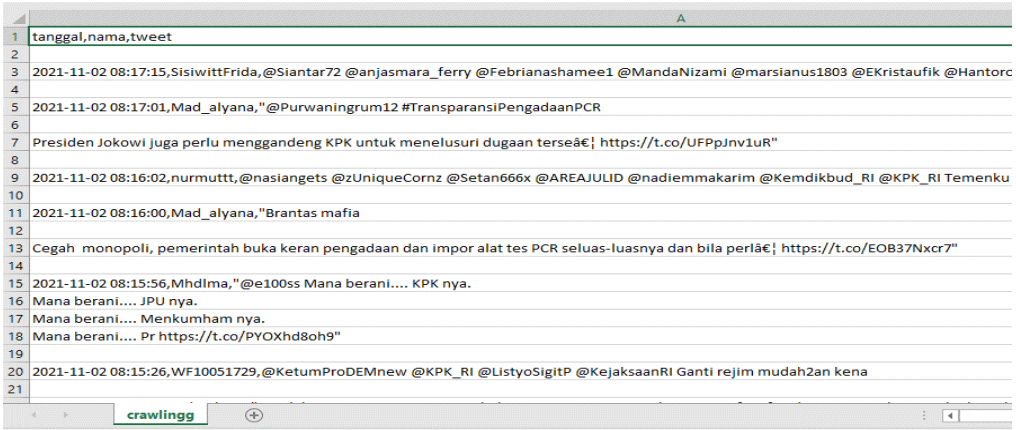


Figure 3 Display of crawl on files CSV

Based on figure 3.2, it can be seen that in this process the data taken contains attributes, including date tweets made, user name users, as well as the contents of tweets. In this study, the data taken were 2055 data tweets.

3.3.2 Preprocessing Data

At stage preprocessing the process of mapping the raw data into data that is more suitable for the final data modeling, such as removing punctuation marks, removing repeated words, and words that often appear but do not really have meaning in sentences.

a. Cleaning text

```
[7]: #Cleaning Text
def remove_punct(text):
    #text  = "".join([char for char in text if char not in string.punctuation])
    text = re.sub(r'[^a-zA-Z0-9]', ' ', str(text))
    text = re.sub(r'\b\w{1,2}\b', '', text)  # menghilangkan 2 kata
    text = re.sub(r'\s\s+', ' ', text)
    return text

df['ISI'] = df['tweet'].apply(lambda x: remove_punct(x))
df.head(10)
```

Figure 4 Source code cleaning text

tweet	ISI
@Siantar72 @anjasmara_ferry @Febrianashamee1 @...	Siantar72 anjasmara ferry Febrianashamee1 Man...
@Purwaningrum12 #TransparansiPengadaanPCR\r\n\...	Purwaningrum12 TransparansiPengadaanPCR Presi...
@nasiangets @zUniqueCornz @Setan666x @AREAJULI...	nasiangets zUniqueCornz Setan666x AREAJULID n...
Brantas mafia\r\n\r\nCegah monopoli, pemerint...	Brantas mafia Cegah monopoli pemerintah buka k...
@e100ss Mana berani.... KPK nya.\r\nMana beran...	e100ss Mana berani KPK nya Mana berani JPU ny...
@KetumProDEMnew @KPK_RI @ListyoSigitP @Kejaksa...	KetumProDEMnew KPK ListyoSigitP KejaksaanRI G...
@Balak_tiga @azwarsiregar @MbahKun04402064 @_D...	Balak tiga azwarsiregar MbahKun04402064 Dauph...
@Yulia_4ja @KPK_RI Ada yg mau dilindungi ? ð□¤...	Yulia 4ja KPK Ada mau dilindungi
@nurmuttt @zUniqueCornz @Setan666x @AREAJULID ...	nurmuttt zUniqueCornz Setan666x AREAJULID nad...
@NoviR007 @Andiarief__ @SBYudhoyono @PDemokrat...	NoviR007 Andiarief SBYudhoyono PDemokrat dari...

Figure 5 Display of results from cleaning text on Jupyter Notebook

Cleaning text is a process to remove punctuation marks or characters and delete words or sentences that are repeated. As can be seen in figure 3.4, all characters other than the letters A to Z and the digits 0 to 9 as well as words consisting of less than 2 characters will be deleted as shown in the "ISI" column.

b. Tokenization

```
[9]: #Tokenization

def tokenization(text):
    text = re.split('\W+', text)
    return text

df['TOKENIZATION'] = df['ISI'].apply(lambda x: tokenization(x.lower()))
df.head(10)
```

Figure 6 Source code tokenization

ISI	TOKENIZATION
Siantar72 anjasmara ferry Febrianashamee1 Man...	[, siantar72, anjasmara, ferry, febrianashamee...
Purwaningrum12 TransparansiPengadaanPCR Presi...	[, purwaningrum12, transparansipengadaanpcr, p...
nasiangets zUniqueCornz Setan666x AREAJULID n...	[, nasiangets, zuniquecornz, setan666x, areaju...
Brantas mafia Cegah monopoli pemerintah buka k...	[brantas, mafia, cegah, monopoli, pemerintah, ...
e100ss Mana berani KPK nya Mana berani JPU ny...	[, e100ss, mana, berani, kpk, nya, mana, beran...
KetumProDEMnew KPK ListyoSigitP KejaksaanRI G...	[, ketumprodemnew, kpk, listyosigitp, kejaksaa...

tweet	ISI
@Siantar72 @anjasmara_ferry @Febrianashamee1 @...	Siantar72 anjasmara ferry Febrianashamee1 Man...
@Purwaningrum12 #TransparansiPengadaanPCR\r\n\...	Purwaningrum12 TransparansiPengadaanPCR Presi...
@nasiangets @zUniqueCornz @Setan666x @AREAJULI...	nasiangets zUniqueCornz Setan666x AREAJULID n...
Brantas mafia\r\n\r\nCegah monopoli, pemerint...	Brantas mafia Cegah monopoli pemerintah buka k...
@e100ss Mana berani.... KPK nya.\r\nMana beran...	e100ss Mana berani KPK nya Mana berani JPU ny...
@KetumProDEMnew @KPK_RI @ListyoSigitP @Kejaksa...	KetumProDEMnew KPK ListyoSigitP KejaksaanRI G...
@Balak_tiga @azwarsiregar @MbahKun04402064 @_D...	Balak tiga azwarsiregar MbahKun04402064 Dauph...
@Yulia_4ja @KPK_RI Ada yg mau dilindungi ? ð□¤...	Yulia 4ja KPK Ada mau dilindungi
@nurmuttt @zUniqueCornz @Setan666x @AREAJULID ...	nurmuttt zUniqueCornz Setan666x AREAJULID nad...
@NoviR007 @Andiarief__ @SBYudhoyono @PDemokrat...	NoviR007 Andiarief SBYudhoyono PDemokrat dari...

Figure 7 Display of results from tokenization on Jupyter Notebook

Tokenization is the process of separating the obtained sentences into several words. It can be seen in figure 3.6, the sentence in the content tweets will be split word by word as shown in the "TOKENIZATION" column.

Stopword Removal

```
[10]: #Stop_Removal

nltk.download('stopwords')
from nltk.corpus import stopwords

stopword = nltk.corpus.stopwords.words('indonesian')

def remove_stopwords(text):
    text = [word for word in text if word not in stopword]
    return text

------
df['STOP_REMOVAL'] = df['TOKENIZATION'].apply(lambda x: remove_stopwords(x))
df.head(5)
```

Figure 8 *Source code stopword removal*

TOKENIZATION	STOP_REMOVAL
[, e100ss, mana, berani, kpk, nya, mana, beran...	[, e100ss, berani, kpk, nya, berani, jpu, nya,...
[, ketumprodemnew, kpk, listyosigitp, kejaksaa...	[, ketumprodemnew, kpk, listyosigitp, kejaksaa...
[, balak, tiga, azwarsiregar, mbahkun04402064,...	[, balak, azwarsiregar, mbahkun04402064, dauph...

Figure 9 Display of results from stopword removal on Jupyter Notebook

Stopwords Removal namely the process of eliminating words that often appear but do not really have meaning in sentences. It can be seen in figure 3.8, the sentence in the content tweets which has been separated in the process tokenization Previously, some words that often appear but do not have much meaning will be removed as shown in the "STOP_REMOVAL" column.

Case Folding

```
[11]: #Case Folding
      df['isi'] = df['ISI'].str.lower()
      df['user_name'] = df['nama'].str.lower()
      df.head(5)
      # len(df.index)
```

Figure 10 *Source code case folding*

isi	user_name
siantar72 anjasmara ferry febrianashamee1 man...	sisiwittfrida
purwaningrum12 transparansipengadaanpcr presi...	mad_alyana
nasiangets zuniquecornz setan666x areajulid n...	nurmuttt
brantas mafia cegah monopoli pemerintah buka k...	mad_alyana
e100ss mana berani kpk nya mana berani jpu ny...	mhdlma

[11]:

	tanggal	nama	tweet	ISI
0	2021-11-02 08:17:15	SisiwittFrida	@Siantar72 @anjasmara_ferry @Febrianashamee1 @...	Siantar72 anjasmara ferry Febrianashamee1 Man...
1	2021-11-02 08:17:01	Mad_alyana	@Purwaningrum12 #TransparansiPengadaanPCR\r\n\...	Purwaningrum12 TransparansiPengadaanPCR Presi...
2	2021-11-02 08:16:02	nurmuttt	@nasiangets @zUniqueCornz @Setan666x @AREAJULI...	nasiangets zUniqueCornz Setan666x AREAJULID n...
3	2021-11-02 08:16:00	Mad_alyana	Brantas mafia\r\n\r\nCegah monopoli, pemerint...	Brantas mafia Cegah monopoli pemerintah buka k...
4	2021-11-02 08:15:56	Mhdlma	@e100ss Mana berani.... KPK nya.\r\nMana beran...	e100ss Mana berani KPK nya Mana berani JPU ny...

Figure 11 Display of results from case folding on Jupyter Notebook

Case Folding is the process used to change each word to be the same. It can be seen in figure 3.10, the entire contents of the sentence in the content tweets will be changed to the same, in this case it is changed to lowercase as shown in the "content" and "user_name" columns.

3.3.3 Labeling Data

At this stage, labeling the contents of the sentence tweets carried out to determine classifier in order to know the sentiment of the sentence. This process is carried out manually with the aim of being used as data training. In this study there are 3 labels, namely positive, neutral, and negative.

Table 1 Process *Labeling* data

User_name	ISI	label
Mad_alyana	purwaningrum12 transparansipenga daanpcr presi…	neutral
nurmuttt	nasiangets zuniquecornz setan666x areajulid n…	neutral
Mad_alyana	brantas mafia cegah monopoli pemerintah buka…	positive
Mhdlma	e100ss mana berani kpk nya mana berani jpu…	negative
wf10051729	KetumProDEMne w KPK ListyoSigitP KejaksaanRI Ganti rejim muda..	positive
indrashaza	Balak tiga azwarsiregar	negative

Modeling

At this modeling stage, data analysis is carried out, in the analysis process, initialization of the label value or is carried out class to a polarity, then divides the dataset into two parts (data training and data testing), then perform a classification to predict the probability of a data that will enter and group it accordingly with the previous probability calculations. The data classification process is carried out using an algorithm naïve bayes classifier. In this study, a test was carried out to determine the algorithm naïve bayes which type is better for the classification process.

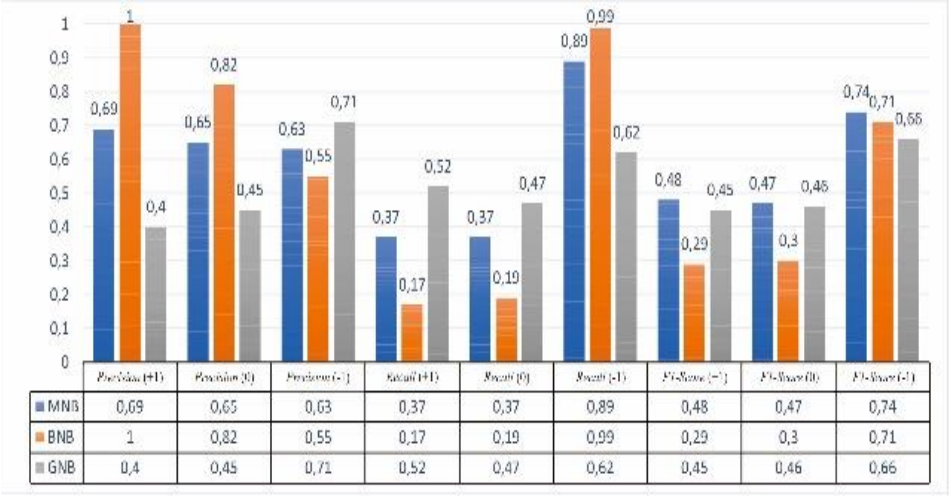


Figure 12 Comparison of values of precision, recalls, and F1-score of three types of algorithms naïve bayes

On figure 12 shows value comparison precision, recall, F1-score, and accuracy of the three types of algorithms naïve bayes. From this comparison, the classification method Multinomial NB have a superior average on the value of recall and F1. So the algorithm type naïve bayes used in this study for the classification process is multinomial naïve bayes.

3.3 Evaluation

Evaluation stage conducted for provide an assessment of the model that has been built previously. By using the algorithm naïve bayes, then the prediction results will be obtained from the data that will enter the system. Evaluation is needed to find out whether the model built is good enough or not.

Table 2 *Classification Report*

Label	*Precision*	*Recall*	*F1-score*	*Accuracy*
Positive	0,69	0,37	0,48	0,64
Neutral	0,65	0,37	0,47	
Negative	0,63	0,89	0,74	

Based on table 2, it can be seen the results of the evaluation of the modeling in this study use value *precision, recall,* and *F1-score*. Score *precision* the largest dataset is obtained with a positive label, which is 0.69, which means the system can accurately predict positive labels for 69% of the total predicted data. Score *recall* the largest dataset was obtained with a negative label, which was 0.89, which means that the system could accurately predict negative labels by 89% of the total negative labeled data on the data *testing*. Score *F1-score* the largest is obtained from the dataset with a negative label, which is 0.74, which means the system can predict the data accurately by 74% of the total data with a negative label.

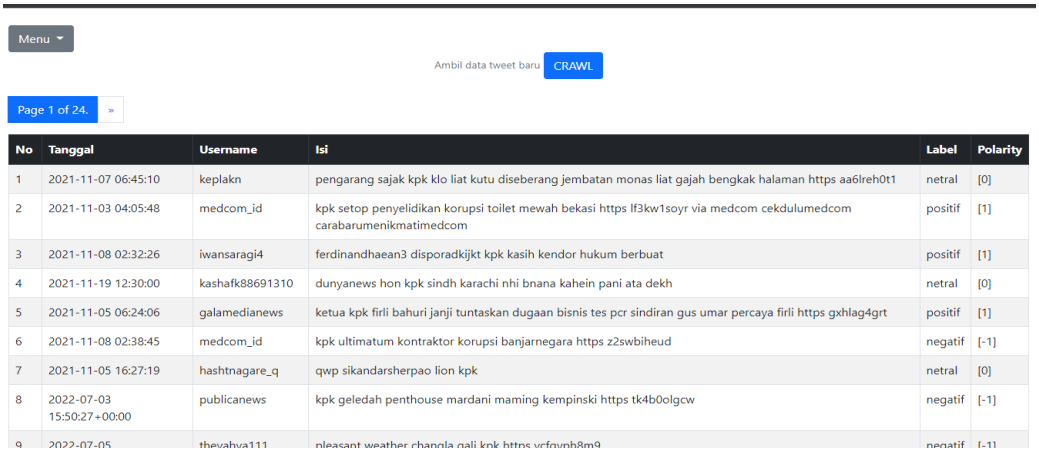
Besides that,*accuracy score* than 20% of all datasets used as data*testing*in this study get a value of 0.6447688564476886 or about 64%. So, it can be concluded that the model used in this study is good enough to predict sentiment*tweets*about the KPK on Twitter social media.

3.4 Deployment

At this deployment stage, the completed model will be presented to the customer as well as the display website for data visualization, so that customers can also assess the results.

3.6.1 Data Visualization

In this study, which is visualized on website not only datasets that have been carried out in previous processes, but also new data taken after the system has learned. Here is a view of website data visualization:



No	Tanggal	Username	Isi	Label	Polarity
1	2021-11-07 06:45:10	keplakn	pengarang sajak kpk klo liat kutu diseberang jembatan monas liat gajah bengkak halaman https aa6lreh0t1	netral	[0]
2	2021-11-03 04:05:48	medcom_id	kpk setop penyelidikan korupsi toilet mewah bekasi https lf3kw1soyr via medcom cekdulumedcom carabarumenikmatimedcom	positif	[1]
3	2021-11-08 02:32:26	iwansaragi4	ferdinandhaean3 disporadkijkt kpk kasih kendor hukum berbuat	positif	[1]
4	2021-11-19 12:30:00	kashafk88691310	dunyanews hon kpk sindh karachi nhi bnana kahein pani ata dekh	netral	[0]
5	2021-11-05 06:24:06	galamedianews	ketua kpk firli bahuri janji tuntaskan dugaan bisnis tes pcr sindiran gus umar percaya firli https gxhlag4grt	positif	[1]
6	2021-11-08 02:38:45	medcom_id	kpk ultimatum kontraktor korupsi banjarnegara https z2swbiheud	negatif	[-1]
7	2021-11-05 16:27:19	hashtnagare_q	qwp sikandarsherpao lion kpk	netral	[0]
8	2022-07-03 15:50:27+00:00	publicanews	kpk geledah penthouse mardani maming kempinski https tk4b0olgcw	negatif	[-1]
9	2022-07-05	thevahva111	pleasant weather chandla gali kpk https vcfgvnh8m9	negatif	[-1]

Figure 13 Table data page

On the table data page displays a list tweets in tabular form with several columns as shown in figure 3.12. On this page new data retrieval can also be done by pressing the blue “CRAWL” button. After the new data is successfully retrieved, the system will process the data to predict the right sentiment label based on the previously trained model.

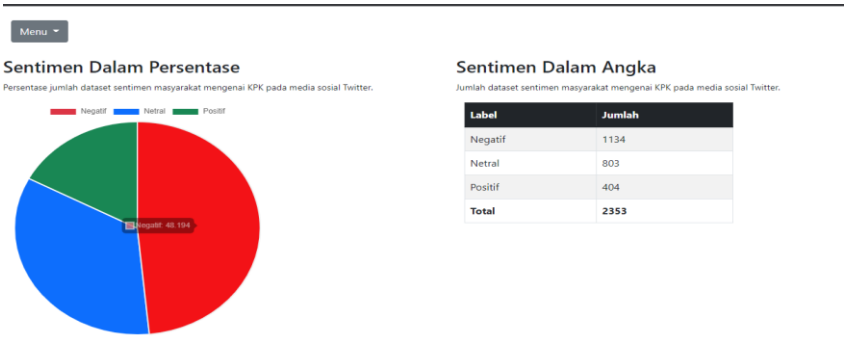
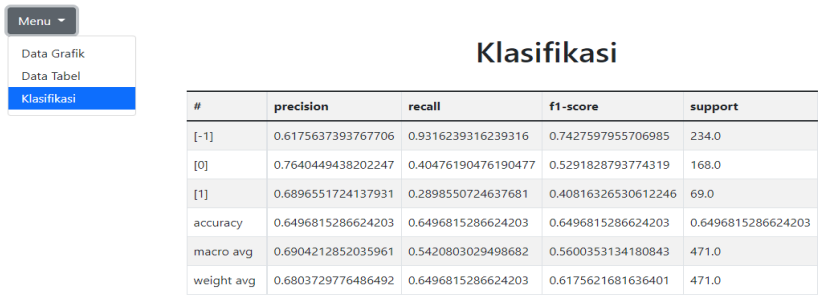


Figure 14 Graphic data page

Furthermore, on the data page the graph displays pie chart which contains information on the percentage of the total public sentiment dataset regarding the KPK on Twitter social media that has been analyzed, namely positive sentiment 17.17%, negative sentiment 48.20%, and neutral sentiment 34.13% as shown in the figure 3.13. In addition, there is also a number of datasets on public sentiment regarding the KPK on Twitter social media that have been analyzed, namely 404 positive sentiments tweets, negative sentiment 1134 tweets, and neutral sentiment 803 tweets of the total 2353 tweets which exists.



#	precision	recall	f1-score	support
[-1]	0.6175637393767706	0.9316239316239316	0.7427597955706985	234.0
[0]	0.7640449438202247	0.40476190476190477	0.5291828793774319	168.0
[1]	0.6896551724137931	0.2898550724637681	0.40816326530612246	69.0
accuracy	0.6496815286624203	0.6496815286624203	0.6496815286624203	0.6496815286624203
macro avg	0.6904212852035961	0.5420803029498682	0.5600353134180843	471.0
weight avg	0.6803729776486492	0.6496815286624203	0.6175621681636401	471.0

Figure 15 Classification page

On the classification page, as shown in figure 3.14, what is displayed is the value of precision, recall and F1-score of each sentiment, namely positive, negative, and neutral sentiment. There is also an

accuracy value of the system that has been made, which is 64%.

3.6.2 Implementation testing data

Testing data is done to see the performance of the model that has been built. Testing data in this study took 3 samples of data and at each retrieval, 100 data were taken. At each data collection, validation is carried out before being used as a model for further data collection.

Table 3 Comparison accuracy score

Accuracy score			
Preliminary data	New data I	New data II	New data III (Not validated yet)
0,64	0,64	0,65	0,66

Based on table 3.3 can be seen accuracy score tends to increase every time data is withdrawn. Because in this study only 3 times of data were withdrawn, this still needs to be proven again whether the more data taken will be directly proportional to the value of the data.accuracy obtained.

IV. CONCLUSIONS

1. An analysis system to determine sentiment regarding the performance of the Komisi Pemberantasan Korupsi based on public tweets on Twitter social media has been successfully created using Python.
2. The classification process using the multinomial naïve bayes method with a dataset comparison of 80:20, obtained an accuracy value of 0.64 or about 64%. The highest precision value is obtained from a dataset with a positive label, which is 0.69. The largest recall value was obtained from a dataset with a negative label, which was 0.89 of the total data with a negative label. The largest F1-score value was obtained in a dataset with a negative label, which was 0.74. So, it can be interpreted that the system used in this study is good enough to predict tweet sentiment about the KPK on Twitter social media.
3. Websites for data visualization that have been developed in this research, can directly retrieve new data which then the data will be used classified and visualized in the form of tables and graphs.

REFERENCES

1. U. M. Sosiawan, 'Peran Komisi Pemberantasan Korupsi (KPK) Dalam Pencegahan dan Pemberantasan Korupsi', *Jurnal Penelitian Hukum De Jure*, vol. 19, no. 4, p. 517, 2019, doi: 10.30641/ dejure. 2019.v19.517-538.
2. A. Go, R. Bhayani, and L. Huang, 'Twitter Sentiment Classification using Distant Supervision', *Processing*, vol., pp. 1–6, 2009.
3. S. S. Arote and R. L. Paikrao, 'A Modified Approach Towards Personalized Travel Recommendation System Using Sentiment Analysis', *2018 International Conference On Advances in Communication and Computing Technology (ICACCT)*, pp. 203–207, 2018, doi: 10.1109/ICACCT.2018.8529327.
4. D. Xhemali, C. J. Hinde, and R. G. Stone, 'Naïve Bayes vs . Decision Trees vs . Neural Networks in the Classification of Training Web Pages', *IJCSI International Journal of Computer Science Issues*, vol. 4, no. 1, pp. 16–23, 2009.
5. B. D. Prasetya, F. S. Pamungkas, and I. Kharisudin, 'Pemodelan dan Peramalan Data Saham dengan Analisis Time Series menggunakan Python', *PRISMA, Prosiding Seminar Nasional Matematika*, vol. 3, pp. 714–718, 2020.
6. D. D. Putri, G. F. Nama, and W. E. Sulistiono, 'Analisis Sentimen Kinerja Dewan Perwakilan Rakyat (DPR) Pada Twitter Menggunakan Metode Naive Bayes Classifier', *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 10, no. 1, pp. 34–40, Jan. 2022, doi: 10. 23960/ jitet.v10i1.2262.
7. W. Bimananda, I. Riski, K. Dwi, R. Nooraeni, T. Siahaan, and Y. Dhea, 'Analisis Text Mining dari Cuitan Twitter Mengenai Infrastruktur di Indonesia dengan Metode Klasifikasi Naïve Bayes', *Eigen Mathematics Journal*, vol. 2, no. 2, pp. 92–101, 2019, doi: 10.29303/ emj.v1i2.36.
8. F. Ratnawati, 'Implementasi Algoritma Naive Bayes Terhadap Analisis Sentimen Opini Film Pada Twitter', *Jurnal Inovtek Polbeng - Seri Informatika*, vol. 3, no. 1, pp. 50–59, 2018.
9. A. Deviyanto and M. D. R. Wahyudi, 'Penerapan Analisis Sentimen Pada Pengguna Twitter Menggunakan Metode K-Nearest Neighbor', *JISKA (Jurnal Informatika Sunan Kalijaga)*, vol. 3, no. 1, pp. 1–13, 2018, doi: 10.14421/jiska.2018.31-01.
10. B. M. Pintoko and K. Muslim, 'Analisis Sentimen Jasa Transportasi Online pada Twitter Menggunakan Metode Naive Bayes Classifier', *e-Proceeding of Engineering*, vol. 5, no. 3, pp. 8121–8130, 2018.
11. P. Sai and B. Balachander, 'Sentimental analysis of twitter data using tweepy and textblob', *International Journal of Advanced Science and Technology*, vol. 29, no. 3, pp. 6537–6544, 2020.
12. M. C. Kirana, N. P. Perkasa, M. Z. Lubis, and M. Fani, 'Visualisasi Kualitas Penyebaran Informasi Gempa Bumi di Indonesia Menggunakan Twitter', *Journal of Applied Informatics and Computing (JAIC)*, vol. 3, no. 1, pp. 23–32, 2019, doi: 10.30871/ jaic.v 0i0.1246.

13. D. Heksaputra, Y. Azani, Z. Naimah, and L. Iswari, 'Penentuan Pengaruh Iklim Terhadap Pertumbuhan Tanaman dengan Naïve Bayes', *Seminar Nasional Aplikasi Teknologi Informasi (SNATI)*, pp. 34–39, 2013, [Online].
14. H. K. C. A. Pratama, W. Suharso, and Q. A'yun, 'Pengklasifikasian Kanker Payudara Dan Kanker Paru-Paru Dengan Metode Gaussian Naïve Bayes , Multinomial Naïve Bayes , Dan Bernoulli Naïve Bayes', *Jurnal Smart Teknologi*, vol. 3, no. 4, pp. 350–355, 2022.
15. A. Ingargiola, 'What is the Jupyter Notebook?', *https://jupyter-notebook-beginner-guide. readthedocs.io/ en/latest/ what_is_jupyter.html*, 2015. https://jupyter-notebook-beginner-guide. readthedocs. io/ en/latest/what_is_jupyter.html (accessed Sep. 20, 2022).
16. golfsql, 'Apa Itu PostgreSQL? Bagaimana Cara Kerja PostgreSQL?', *https://www.pgsql. com/apa-itu-postgresql-bagaimana-cara-kerja-postgresql/*, Nov. 22, 2021. https:// www.pgsql.com/apa-itu-postgresql-bagaimana-cara-kerja-postgresql/ (accessed Apr. 04, 2022).
17. B. Kumar, 'What is Python Django and used for', *https://pythonguides.com/what-is-python-django/*, Aug. 04, 2021. https:// pythonguides.com/what-is-python-django/ (accessed Sep. 20, 2022).
18. N. Hotz, 'What is CRISP DM?', *https:// www.datascience-pm.com/crisp-dm-2/*, Aug. 08, 2022. https:// www. datascience-pm. com/crisp-dm-2/
19. Putri, D. D., Nama, G. F., & Sulistiono, W. E. (2022). Analisis Sentimen Kinerja Dewan Perwakilan Rakyat (DPR) Pada Twitter Menggunakan Metode Naive Bayes Classifier. Jurnal Informatika dan Teknik Elektro Terapan, 10(1).
20. Rohmalia, N., Nama, G. F., & Purwasih, N., 2021. Dashboard Monitoring Atmospheric Corrosion Sensor in Material Metal Using Laravel Framework. Journal of Engineering and Scientific Research, 3(1), 1-6.
21. Nama, G.F. and Kurniawan, D., 2017. An enterprise architecture planning for higher education using the open group architecture framework (togaf): Case study University of Lampung. In 2017 Second International Conference on Informatics and Computing (ICIC) (pp. 1-6). IEEE.
22. Nama, G.F. and Muludi, K., 2018. Implementation of two-factor authentication (2FA) to enhance the security of academic information system. Journal of Engineering and Applied Sciences, 13(8), pp.2209-2220.
23. Nama, G.F. and Despa, D., 2016, October. Real-time monitoring system of electrical quantities on ICT Centre building University of Lampung based on Embedded Single Board Computer BCM2835. In 2016 International Conference on Informatics and Computing (ICIC) (pp. 394-399). IEEE.