ISSN 2085-4579

Sentiment Analysis of Indonesian Presidential Candidate Before and After the Election

Muhammad Hilmy Rasyad Sofyan[1], Akmal Zulkifli[2], Rasim[3*]
[1]Department of Computer Science, Universitas Pendidikan Indonesia, Bandung, Indonesia
hilmyrs123@upi.edu
[2]Department of Computer Science, Universitas Pendidikan Indonesia, Bandung, Indonesia
akmalzulkifli29@upi.edu
[3]Department of Computer Science, Universitas Pendidikan Indonesia, Bandung, Indonesia
rasim@upi.edu*

Accepted on June 25th, 2024
Approved on December 6th, 2024

***Abstract*— As one of the world's democratic countries, Indonesia has just held a general election to choose its next president. The development of the times encourages presidential candidates to make a new breakthrough, such as the use of social media as campaign media. Currently, X is a popular social media used as a campaign medium. On X, users are given the freedom to share their opinions. Various opinions related to one of the presidential candidates were used by the researchers to collect data using the data crawling method. The results of the data crawling are first processed with different methods in the pre-process to make the data ready for use. Some of the steps that need to be taken in the pre-process are such as cleaning, normalisation, stopword, tokenisation, stemming and translation processes. All the processes carried out in the pre-process stage will produce mature or usable data. The mature data is then classified into positive, negative and neutral using the Naïve Bayes classification method. Once the classification is complete, the results are evaluated in terms of sentiment towards one of the presidential candidates. The results of a total of 2117 data collected from 01 February 2024 to 20 May 2024, there are 390 data used for the pre-presidential election sentiment analysis and 1618 data used for the post-presidential election sentiment analysis. Both before and after the presidential election was held, this presidential candidate had more positive sentiments than the negative and neutral sentiments he received from the public.**

***Index Terms*—Data Crawling; Data Preprocessing; Naïve Bayes Classification Method; Sentiment Analysis.**

I. INTRODUCTION

Indonesia is one of the countries that implement a democratic political system. In a democratic system, citizens are given the freedom to participate in national development both in political development and other fields [1]. According to Alhamid & Hanim and also Kelibay [2] one form of citizen participation in a democratic country is by participating in general elections, where in general elections citizens are given the freedom to choose prospective leaders who deserve to take office next. Then on February 14, 2024, Indonesia held a general election to choose the next presidential candidate.

In this election, there were three people running as Indonesian presidential candidates, namely Anies Baswedan, Prabowo Subianto, and Ganjar Pranowo. The rapid development of technology has led these three presidential candidates to campaign on social media during the campaign period. As a result of the social media campaigns, the public became more free to express their opinions, which made this election quite interesting [3]. One of the social media that is often used to convey good ideas and opinions by the public is X [4].

The popularity of X to be used as a place to convey opinions cannot be separated from its social media model in the form of microblogging which allows users to send text messages of up to 280 characters so that they can convey their intentions and objectives briefly, concisely, and clearly [5]. The popularity of X is supported by data in 2023 which shows the number of X users in Indonesia is around 25.25 million users, making Indonesia the 4th country with the most X users. With so many X users in Indonesia, it will certainly make a lot of tweets circulating discussing various things related to presidential candidates in the current election.

The number of tweets circulating on X, especially those discussing presidential candidates in this election, makes researchers interested in looking at sentiment analysis of tweets on X. According to Septiani et al [6] the purpose of sentiment analysis is to classify the text in a sentence so that it can be determined whether the opinion contained in a sentence contains positive, negative, or neutral sentiment. There are various methods for performing sentiment analysis such as Naive Bayes Classifier, Decision Tree, and Support Vector Machine [7].

There are many studies on sentiment analysis that have been done before. First, research conducted by Septiani [8] made a sentiment analysis system with the Naive Bayes Classifier method to see how public

sentiment towards moving the National Capital. The results obtained are that the system does not work optimally because there is a mismatch of the dataset used, which affects the data classification process and there are obstacles in understanding sentiment classification such as in the system a sentence is considered positive, but when manually classified the sentence should be negative. Second, research conducted by Gunawan [9] made a sentiment analysis system with the Naive Bayes Classifier method to analyze public sentiment on product reviews, from this study getting accuracy results above 50% of two types of tests with different numbers of classes. Where in testing 5 classes show a lower level of accuracy with a value of 52.66% - 59.33% compared to testing 3 classes which have a value of 73.89% - 77.78%. Third, research conducted by Mahfud [10] used the Naive Bayes Classifier method to determine the trend of Indonesian presidential candidates in 2024. The results obtained from this study are the high accuracy value of the Naive Bayes method used which strengthens the studies stating that Naive Bayes is a fairly reliable classification method.

II. METHODS

In this research, there are several steps taken to complete this research and get good results. These steps are data crawling, data preprocessing, Naive Bayes classification, and result evaluation. Figure 1 is the stages of the research conducted:

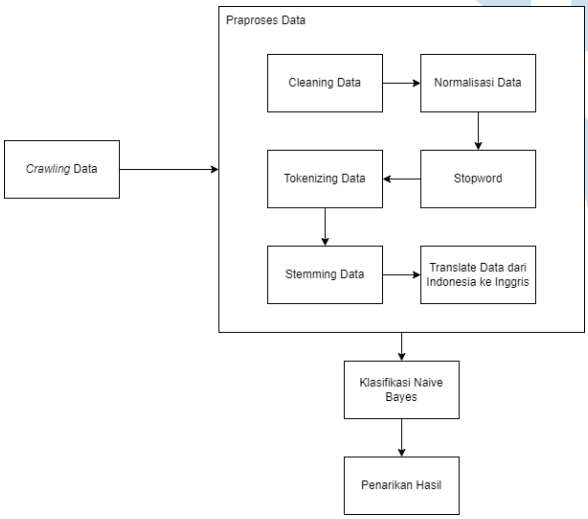

Fig. 1. Research Stages

A. Data Crawling

Data crawling is the process of collecting data from a database. In this study, data was retrieved from X using tweet-harvest. Tweet-harvest is a tool for crawling data from X by utilizing the API from X itself. The keywords used for data retrieval are "pilpres", "pemilu", and the name of one of the presidential candidates in Indonesian with a period of time from February 1, 2024 to May 20, 2024.

B. Data Preprocessing

Data preprocessing is the process of preparing raw data into data that is ready for use by eliminating and converting data into a form that can be processed by the system [11]. In pre-processing, there are several stages that need to be passed, namely:

1. Data cleaning: this stage is used to clean data from unused characters or symbols [12]. Cleaning is needed because the data provided by X contains a lot of dirty data such as URLs, retweet text, hashtags, and other things that are not used in research. In addition, a process is carried out to change the text to lowercase to make the data the same.
2. Data normalization: this stage is used to correct incorrect spelling so that it can be converted into standard words.
3. Stopword: the stage used to eliminate conjunctions.
4. Data tokenizing: this stage is used to break sentences into words.
5. Data stemming: this stage is used to eliminate connecting words, so that words become basic words only.
6. Data translation: at this stage all words will be translated into English to make the data can be processed in the next stage.

C. Naïve Bayes Classification

After preprocessing the data, the data is classified using Naïve Bayes method. Naïve Bayes method is used because it has fast modeling capabilities, has the ability to predict, and also provides new methods in exploring and understanding data [13]. At this stage we use a library in Python called TextBlob, the reason why the data needs to be translated into English is because this library can only be used for English, German, and French. In this classification, the data will be classified by the NaiveBayesClassifier function from TextBlob into three types of sentiment, namely positive, negative, or neutral sentiment [14].

D. Result Recall

After classifying with Naïve Bayes, results will be drawn to see the distribution of sentiment on the presidential candidate who is the target of the research. After the distribution of sentiment is obtained, a comparison will be made between the results of the presidential candidate sentiment before the presidential election and after the presidential election is carried out.

III. RESULT

The results of this research are carried out in accordance with the research flow arrangement. The structure of this research flow starts from crawling

data, data preprocessing, naïve bayes classification, and results recall.

A. *Data Crawling*

At this stage, researchers retrieved data from the X server using the available API. This stage resulted in a total of 2117 tweet data, each tweet from February 2024 to May 2024 and derived from the keywords “pemilu”, “pilpres”, and the name of one of the presidential candidates selected for this study. Table I is an example of tweet data obtained from crawling data.

TABLE I. DATA CRAWLING RESULTS

Conversation_id_str	Created_at	Full_text	Username
1759911802991939748	Tue Feb 20 12:04:06 +0000 2024	Pemilu dah selesai tapi test print standee baru dateng But its okay! CAKEP BANGET GW TADI YG LEMES PULANG KERJA JADI SEMANGAT LAGI https://t.co/ygbsaYfLgf	jejenjreng
1790537038883905791	Wed May 15 00:17:51 +0000 2024	Negeri ini memang Terlalu kejam bagi manusia2 Kelas Bawah Padahal di setiap PEMILU Lima Tahunan Derajat orang Kaya orang Miskin bahkan ORANG GILA Sekalipun itu sama2 Sama2 Punya Satu hak Suara Alehhh22 https://t.co/ZyMRdFTsxG	Heraloebss
1757786015115420086	Wed Feb 14 15:16:58 +0000 2024	Minta tolong drop bukti-bukti kecurangan pemilu di kolom reply dong. Thx ya. #KawalSampaiFinal https://t.co/UADUN0rVRT	timpenguinnas

The crawling data results have 15 columns columns that are relevant and will be used for this research are conversation_id_str, created_at, full_text, and username. Those columns are required for the next stage of the research. On a later stage, data will be preprocessed to be a fully prepared data.

B. *Data Preprocessing*

Before classifying data, the dataset obtained from data crawling needs to be preprocessed first until the dataset is clean and ready to be used in the next stage. At this stage the data must go through a cleaning process, normalization, stopwords, tokenizing, stemming, and translating data from Indonesian to English.

In the cleaning stage, each tweet is cleaned from URLs, punctuation marks, emoticons, hashtags, retweet text, and others. It also creates a dataframe that uses only three columns, namely 'full_text', 'username', and 'created_at'. In addition, cleaning up duplicate data, changing strings to lower case and changing the date format in the 'created_at' column to year-month-date hour:minute:second. Table II is one of the results of the cleaning stage.

TABLE II. DATA CLEANING RESULTS

Created_at	Full_text (after cleaning)	Username
2024-02-20 12:04:06	pemilu dah selesai tapi test print standee baru dateng but its okay! cakep banget gw tadi yg lemes pulang kerja jadi semangat lagi	jejenjreng

In the normalization stage, some words will be changed to become standard words. Table III shows the word forms before and after being changed with a total of 14 normalized words.

TABLE III. NORMALIZATION WORDS LIST

Before	After
Gt	Gitu
Cth	Contoh
Tp	Tapi
Tpi	Tapi
Yg	Yang
Bgt	Banget
Srkg	Sekarang
Prabowo	“null”
Prabowogibran	“null”
Gibran	“null”
Ganjar	“null”
Ganjarmahfud	“null”
Mahfud	“null”
Jgn	Jangan

From the table, there are some words that changed to null, because the other presidential candidates name are not relevant for this research, as we only focused to one of them. Also, not every slangs are covered in this normalization list because Indonesian have so many slangs to begin with.

After normalization, stopwords are performed to remove tokens or unimportant words such as conjunctions. To do stopwords using the Sastrawi library in Python. Sastrawi is used because it can identify stopwords in Indonesian.

At the tokenizing stage, existing sentences are converted into a collection of tokens or made into words. After the sentences converted into tokens, then the next stage data will be stemmed which means the affix words are removed to become basic words. Sastrawi is also used for stemming the data, because Sastrawi allows the user to do language processing for

words or text in Indonesian and finally, the data is translated from Indonesian into English. The translation process is carried out with the help of the unlimited machine translator library which is a free translator library without word limits for translation in Python. This translation process is done to facilitate the data classification process. Table IV shows the data that has been preprocessed.

TABLE IV. PREPROCESSED DATA RESULTS

Full_text	Full_text_en
milu dah selesai test print standee baru dateng but its okay cakep banget gw tadi lemes pulang kerja jadi semangat	milu dah selesai test print standee baru dateng but its okay cakep banget gw tadi lemes pulang kerja jadi semangat
negeri memang terlalu kejam manusia2 kelas bawah padahal tiap milu lima tahun derajat orang kaya orang miskin bahkan orang gila sekalipun sama2 sama2 punya satu hak suara alehhh22	The country is indeed too cruel to lower class people, even though every five years the rich, poor and even crazy people all have the same right to vote alehhh22
kabar baik pasang calon bukti laku curang milu diskualifikasi posisi dua naik menang banyak doa zikir seluruh dukung amin kun faya kun	good news put up candidates proof of fraudulent behavior milu disqualification second place up win lots of prayers of remembrance all support amen kun faya kun

From the preprocessing results, the data shown is not perfectly processed. There is some words that not covered in normalization stage that will affect the translation and error in removing affix for some words. Word like pemilu has an affix removal error, the function identifies “pe” in “pemilu” as a prefix and removed it when it shouldn’t get removed.

C. *Naïve Bayes Classification and Result*

Before classifying the data that has gone through the preprocessing, 50% of them will be used as train data. Both for pre-election and post-election datasets will use the same treatment. Figure 2 shows the code for making training data.

```
import random

set_positif = []
set_negatif = []
set_netral = []

for n in datasetpasca:
  if (n[1] == 'Positif'):
    set_positif.append(n)
  elif (n[1] == 'Negatif'):
    set_negatif.append(n)
  else:
    set_netral.append(n)

set_positif = random.sample(set_positif, k=int(len(set_positif)/2))
set_negatif = random.sample(set_negatif, k=int(len(set_negatif)/2))
set_netral = random.sample(set_netral, k=int(len(set_netral)/2))

train = set_positif + set_negatif + set_netral

train_set = []

for n in train:
  train_set.append(n)
```

Fig. 2. Code for Splitting Training Data

After dividing the training data, data classification began using the taining data with NaiveBayesClassifier function that provided by Textblob library. Besides classification, labelling is also done to see which data is belong to which sentiment category. Figure 3 shows the code for classifying and labeling sentiment data.

```
from textblob.classifiers import NaiveBayesClassifier

cl = NaiveBayesClassifier(train_set)
print("Akurasi Test : ", cl.accuracy(datasetpasca))

Akurasi Test :  0.7719406674907293

#labeling

data_tweetpasca = list(datapasca['full_text_en'])
polaritas = 0

status = []
total_positif = total_negatif = total_netral = total = 0

for i, tweet in enumerate(data_tweetpasca):
  analysis = TextBlob(tweet, classifier = cl)

  if analysis.classify() == 'Positif':
    total_positif += 1
  elif analysis.classify() == 'Netral':
    total_netral += 1
  else:
    total_negatif += 1

  status.append(analysis.classify())
  total+=1
```

Fig. 3. Code for NaiveBayesClassifier & labelling

After classifying and labeling the data using naïve bayes classification, the accuracy rate for pre-election is 72,8% and 77,2% for post-election. The results of the pre-presidential election data classification have a total of 390 data and post-presidential election data have a total of 1618 data. Then Table V shows the results of data classification.

TABLE V. DATA CLASSIFICATION RESULTS

Data Classification Results			
	Positive	***Negative***	***Neutral***
Pre-election	191	51	148
Post-election	963	523	132

From Table V, the amount of positive, negative, and neutral data can be converted into sentiment distribution results. Sentiment distribution results are shown in Figure 4 & 5.

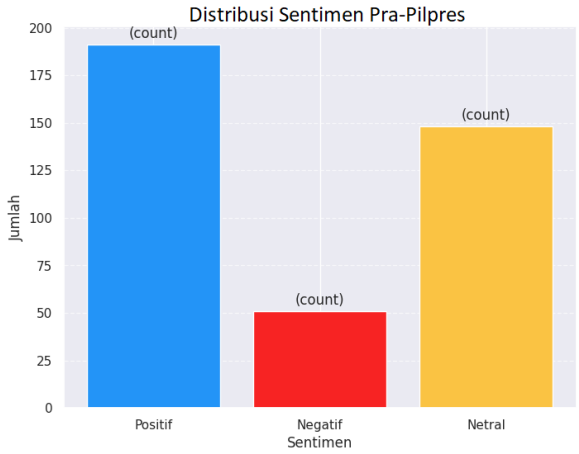

Fig. 4. Sentiment Distribution Pre-Presidential Election

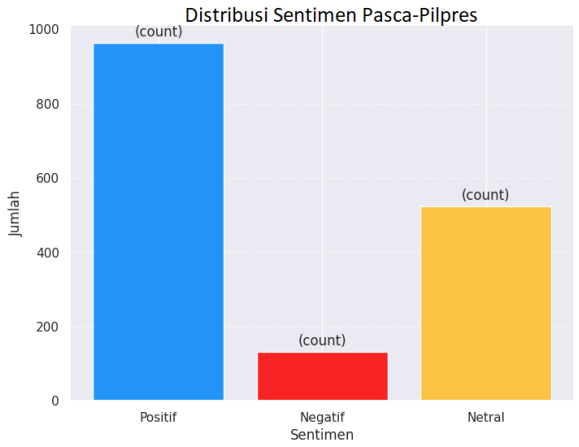

Fig. 5. Sentiment Distribution Post-Presidential Election

From the results obtained both before and after the election was held, positive sentiment is still owned by one of the presidential candidates, and with quite high results when compared to the negative sentiment he got. Looking at the results before the election was held, from a total of 390 data, there were 49% positive sentiments, 38% neutral sentiments, and 13% negative sentiments. And after the election, starting from before the announcement of the election results until after they were announced, there was a total of 1618 data, consisting of 59% positive sentiment, 32% negative sentiment, and 8% neutral sentiment. From these results there is an increase in the number of positive sentiments when the election has been held, but with a note that there is a considerable difference in the amount of data and can make this increase only influenced by the amount of data and not because of the quality of the existing data. In addition to sentiment distribution, word cloud results are also obtained as shown in Figure 6.

Fig. 6. Word Cloud from The Data Used

When viewed from the word cloud obtained, the keyword that appears most often is one of the names of the presidential election candidate which is the focus of the discussion. Followed by several keywords that lead to the context of the general election, such as the keywords “presiden”, “politik”, and “rakyat”. In addition, some words also indicate that many tweets discuss hopes and expectations related to candidates or elections in general, shown by words such as “mau”, “buat”, and “bisa”. Then, when looking at the sentiment dimension, there are more potentially positive and neutral words than negative ones. These results can strengthen the results of sentiment analysis, which shows a positive trend both before and after the election, and the consistency between the results of the sentiment distribution and the word cloud shows the reliability of the method used.

IV. Conclusion and Recommendation

From this research, it can be concluded that one of the presidential candidates used as the object of research has more positive sentiments both before and after the presidential election. From a total of 390 data for before the presidential election, 191 data were obtained with positive sentiments, then from a total of 1618 data for after the presidential election, 963 data were obtained with positive sentiments. This shows that people on X still have sentiments that tend to be positive towards this one of the presidential candidates.

Future research that wants to conduct research on the comparison of sentiment analysis before and after the implementation of the 2024 or subsequent elections, it can be done by examining sentiment analysis of all existing candidates and comparing between each candidate, then it can also use other methods in classifying sentiment to see the difference in results from different methods. Also, could add more words into normalization list to normalize slangs and acronym that Indonesian mainly used on X to achieve more accurate result in this type of research.

REFERENCES

[1] F. Rahma Bachtiar, "PEMILU INDONESIA: KIBLAT NEGARA DEMOKRASI DARI BERBAGAI REFRESENTASI 1." [Online]. Available: http://isites.harvard.edu/fs/docs/icb.topic925740.files/Week%206/Mainwaring_Latin.pdf

[2] S. Pemilu, B. B. Pemilih Pemula Di, P. Sitti Rahmah, and I. Negeri Sultan Syarif Kasim Riau, "Socialisation Election 2024 For First-Time Voters In BT8 Pekanbaru," 2024. [Online]. Available: http://journal.al-matani.com/index.php/arsy,Online

[3] R. Vindua and A. U. Zailani, "Analisis Sentimen Pemilu Indonesia Tahun 2024 Dari Media Sosial Twitter Menggunakan Python," *JURIKOM (Jurnal Riset Komputer)*, vol. 10, no. 2, p. 479, Apr. 2023, doi: 10.30865/jurikom.v10i2.5945.

[4] S. W. Ritonga, . Y., M. Fikry, and E. P. Cynthia, "Klasifikasi Sentimen Masyarakat di Twitter terhadap Ganjar Pranowo dengan Metode Naïve Bayes Classifier," *Building of Informatics, Technology and Science (BITS)*, vol. 5, no. 1, Jun. 2023, doi: 10.47065/bits.v5i1.3535.

[5] S. Nurul, J. Fitriyyah, N. Safriadi, E. Esyudha, and P. #3, "JEPIN (Jurnal Edukasi dan Penelitian Informatika) Analisis Sentimen Calon Presiden Indonesia 2019 dari Media Sosial Twitter Menggunakan Metode Naive Bayes", [Online]. Available: http://dev.twitter.com.

[6] Syahril Dwi Prasetyo, Shofa Shofiah Hilabi, and Fitri Nurapriani, "Analisis Sentimen Relokasi Ibukota Nusantara Menggunakan Algoritma Naïve Bayes dan KNN," *Jurnal KomtekInfo*, pp. 1–7, Jan. 2023, doi: 10.35134/komtekinfo.v10i1.330.

[7] F. Nurhuda, S. W. Sihwi, and A. Doewes, "Analisis Sentimen Masyarakat terhadap Calon Presiden Indonesia 2014 berdasarkan Opini dari Twitter Menggunakan Metode Naive Bayes Classifier," vol. 2, no. 2, 2013.

[8] J. Teknika, R. K. Septiani, S. Anggraeni, and S. D. Saraswati, "Teknika 16 (02): 245-254 Klasifikasi Sentimen Terhadap Ibu Kota Nusantara (IKN) pada Media Sosial Menggunakan Naive Bayes," *IJCCS*, vol. x, No.x, pp. 1–5.

[9] B. Gunawan, H. Sasty, P. #2, E. Esyudha, and P. #3, "JEPIN (Jurnal Edukasi dan Penelitian Informatika) Sistem Analisis Sentimen pada Ulasan Produk Menggunakan Metode Naive Bayes," vol. 4, no. 2, pp. 17–29, 2018, [Online]. Available: www.femaledaily.com

[10] F. Khusnu Reza Mahfud, W. Hariyanto, G. Chandra Puspitadewi, and P. Perpustakaan dan Ilmu, "ANALISIS TREN CALON PRESIDEN INDONESIA 2024," 2024.

[11] L. Ellyanti, Yova Ruldeviyani, Lelianto Eko Pradana, and Andro Harjanto, "Sentiment Analysis of Twitter Users to the PeduliLindungi Using Naïve Bayes Algorithm," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 7, no. 2, pp. 414–421, Mar. 2023, doi: 10.29207/resti.v7i2.4684.

[12] D. Kusnanda and A. A. Permana, "Implementation Of Naive Bayes Classifier (NBC) For Sentiment Analysis On Twitter In Mobile Legends." [Online]. Available: http://ijstm.inarah.co.id1132

[13] A. Erfina and M. R. N. R. Alamsyah, "Implementation of Naive Bayes classification algorithm for Twitter user sentiment analysis on ChatGPT using Python programming language," *Data and Metadata*, vol. 2, Jan. 2023, doi: 10.56294/dm202345.

[14] I. G. S. Mas Diyasa, N. M. I. Marini Mandenni, M. I. Fachrurrozi, S. I. Pradika, K. R. Nur Manab, and N. R. Sasmita, "Twitter Sentiment Analysis as an Evaluation and Service Base On Python Textblob," *IOP Conf Ser Mater Sci Eng*, vol. 1125, no. 1, p. 012034, May 2021, doi: 10.1088/1757-899x/1125/1/012034.