Facebook Sentiment Analysis for the 2024 Presidential Candidate Election using the Naive Bayes Method

Destri Komalasari
Information System Study Program, Nusa Putra University
Jl. Raya Cibatu Cisaat No.21, Kec. Cisaat, Kab. Sukabumi, Jawa Barat 43155, Indonesia
destri.komalasari_si20@nusaputra.ac.id

Nopita Amelia
Information System Study Program, Nusa Putra University
Jl. Raya Cibatu Cisaat No.21, Kec. Cisaat, Kab. Sukabumi, Jawa Barat 43155, Indonesia
nopita.amelia_si20@nusaputra.ac.id

Novita Damayanti
Information System Study Program, Nusa Putra University
Jl. Raya Cibatu Cisaat No.21, Kec. Cisaat, Kab. Sukabumi, Jawa Barat 43155, Indonesia
novita.damayanti_si20@nusaputra.ac.id

Habi Baturohmah
Information System Study Program, Nusa Putra University
Jl. Raya Cibatu Cisaat No.21, Kec. Cisaat, Kab. Sukabumi, Jawa Barat 43155, Indonesia
habi.baturohmah@nusaputra.ac.id

ABSCTRAK - In 2024 Indonesia will hold a democratic party to elect the Indonesian head of state. Every political figure who is nominated as head of state will consider their popularity based on public opinion. Currently, it has had a big impact in building public political opinions, views, sentiments and preferences (ahead of the General Election). Especially on social media, one of which is Facebook. There are several uses for social media, such as meeting new friends, finding out information about sports, economics, tourism and also political matters. One of them is the leader figure of the 2024 presidential candidate, so the author wants to know what information can be taken from public opinion on Facebook social media regarding the leader figure of the 2024 presidential candidate. This problem can be overcome by conducting research in the field of Sentiment Analysis, which is a field of research that focuses on the computational study of opinions, behavior and emotions towards an entity expressed in text form. This research was conducted to find out the results of sentiment analysis regarding the public's response to news about the 2024 presidential candidates and to classify them into three classes, namely positive, negative and neutral using the Naive Bayes method. From 700 comment data on Facebook for each 2024 presidential candidate's fanpage, several duplicate and unimportant data were deleted, resulting in 554 comment data. The accuracy results using the naïve Bayes classifier algorithm were 90.09% which was obtained by using a comparison of 20% test data and 80% training data.

Keywords— Presidential Candidates, Social Media, Facebook, Sentiment Analysis, Naive Bayes

I. INTRODUCTION

In this case, Indonesia is a democratic country that holds general elections to elect a head of state. Every five years, elections are usually held simultaneously in Indonesia. In 2024 there will be an official Indonesian political race. As a member of parliament who wants to run for office, of course he will consider the level of distrust seen from the public's reaction. People use social media platforms, one of which is Facebook, to voice their opinions.[1]

As a general rule, postings on Facebook are used to transfer data about clients and offer data, as well as convey news. User feelings can also be expressed in post content. In order to process text data and extract useful information about one's opinion about a political or public figure, proper analysis is required. Sentiment analysis which examines opinions or opinions that state or reveal perceptions that contain positive or negative sentiments, can be used to process a collection of comments in the form of text. The sentiment that is expected to complete the examination comes from comments on the fanpage page of the 2024 Indonesian presidential candidate on Facebook.[1]

Several previous tests have been carried out using the Naive Bayes method on social media data to analyze sentiment towards Indonesian presidential candidates. In this study, the data will be classified using the Naive Bayes classification method, which has been collected. This classification technique relies on direct statistical probability values and assumes a high degree of

independence from Bayesian rules. Initial information and sample data are used in this Bayesian analysis technique. The Naive Bayes Classifier method itself is used by testing data testing by utilizing training data. This Naive Bayes Classifier method provides high precision when applied to very large data sets and various information. This is demonstrated by research showing that this approach has a number of additional capabilities, including speed, simplicity, and high accuracy.

II. RESEARCH METHODS

2.1 Research Stages

This research was conducted through several stages of the process according to the data processing rules used in previous studies. That includes data scrapping, preprocessing, labeling, as well as analysis and evaluation.[2] As in Figure 1 below.

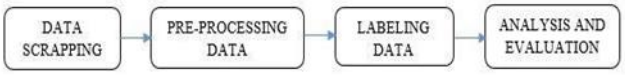

Figure 1. Research workflow

2.2 Preprocessing Data

Data pre-processing is a data mining technique involves transforming raw data into a format easy to understand. The data preprocessing steps are:

1. Cleaning, namely to reduce process noise during the classification process, cleaning refers to the process of reviewing words that are not needed. Character is the omitted word.
2. Case Folding, which is the process of converting all letters in a document into lowercase letters, only the letters a to z are changed.
3. Tokenizing, which is the process of cutting sentences into words by separating the words in a data set and figuring out how they are structured.
4. Filtering, namely the stage of taking important words from tokenizing consequences. A stoplist algorithm, which removes unimportant words, or a wordlist, which retains important words, can be used in the filtering process. [1] [3]

2.3 Naïve Bayes Classifier Method

The Naive Bayes method is a method used to calculate the probability of the appearance of words in a document and uses that value to determine whether the document is included in the positive, negative or neutral category. The basic formula for the Naive Bayes method is as follows:

$$P(C|W) = P(W|C)*P(C)/P(W)$$

Where:

P(C|W) is the probability of class C

(positive/negative/neutral) when it is known that there is a

W feature in the document

P(W|C) is the probability of occurrence of feature W in the training data set labeled C

P(C) is the prior probability of class C which is calculated from the total training data set labeled C divided by the total training data set

P(W) is the probability of occurrence of the W feature as a whole

In this study, the method that will be used is the Naive Bayes classification method to carry out sentiment analysis of texts related to the 2024 presidential election on Facebook social media. Sentiment analysis is a classification task that determines positive and negative sentiments.[10]

III. RESULT AND DISCUSSION

3.1 Data Collection

This method makes it possible to retrieve data in the form of desired Facebook status or comments on the Facebook fanpage of the Indonesian presidential candidate in 2024 as a data source. At the data collection stage, a level of accuracy and patience is required. Because the data collected must meet every required criteria, such as avoiding comments containing emoticons and other characters. Regarding the amount of data required, it must meet predetermined targets. The data collection itself was carried out from May 29 2023 to June 5 2023. Data collection in this study was carried out by the crawling method using the Instant Data Scraper tool.

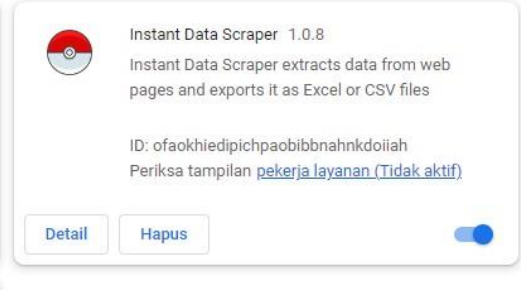

Figure 2. After the tool is installed

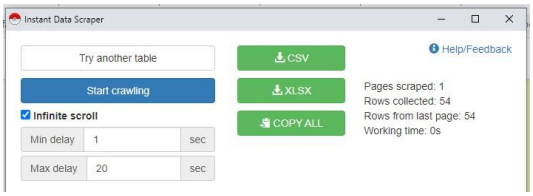

Figure 3. Data collection process

Facebook collects a wide range of data from its users, including demographic information, interests, social connections, and user-generated content such as posts, comments, and likes. This wealth of data allows researchers to explore various research questions across different domains, such as social sciences, psychology, marketing, and public opinion analysis. The collection of status data in this study was taken from the Facebook accounts of Anies Baswedan, Prabowo Subianto, and Ganjar Pranowo. The following is the amount of status data from each Facebook page for the 2024 Indonesian presidential candidates. Comments are selected from users of the Facebook social network for each presidential candidate at random. The following results of data collection can be seen in Table 1 below.

Table 1. Results of data collection

Name	Number Status	Number of comment
Anies	6	200
Prabowo	6	300
Ganjar	6	200
Total	**18**	**700**

3.2 Data Correction

The data format obtained from the Instant Data Scraper tool is in the form of xlxs so that the correction process can be done easily. then the data will go through the next stage, namely removing attributes that are not needed and deleting data that is detected as duplication. In this study the authors only used commentary data which contained text data in Indonesian. In Figure 4 below is the result of the data correction.

	A	B	C
1	Link Profil User	Username	Comment
2	https://www.facebc	Sis ArpacSt	Sehat selalu calon Presidenku
3	https://www.facebc	Mediyawati Irdiyawan	Semoga sehat selalu pak Anies insya Allah presiden 2024
4	https://www.facebc	Ilma Safitri Zunnuraini	Wah, jurusan busway di Jaksel ya Paaak? Saya yang domisili Jaktim iri deh kok jalur busway c
5	https://www.facebc	Sumarno Jogja	Mantabbb Pak Lanjutkan ke tingkat national...
6	https://www.facebc	Nadia Kirana Larasati	Sukses "di" Jakarta lanjut Indonesia
7	https://www.facebc	Suci Wulansari	Love sekebon buat pak Anies Baswedan
8	https://www.facebc	Ahmad Busairi	Super
9	https://www.facebc	Dedi Maolana	Tetap semangat p
10	https://www.facebc	Zahrawi Mochtar	Barokallah
11	https://www.facebc	Kiki Saragih	Bismillahirrahmanirrahim... Pak Anies Baswedan insya Allah 2024 jdi presiden RI. Aamiin
12	https://www.facebc	Abu Yayang Asri	Barakallahu fikum pak Anies Baswedan
13	https://www.facebc	Januar Wakhiddin	Allahu akbar..Allohu Akbar...Allohu Akbar
14	https://www.facebc	Desi Kamariyah	Barakallahu fiikum, Pak Anies dan keluarga..
15	https://www.facebc	Marlein Tampi	Mga sehat sllu lancar aktifitasnya Allah lindungi bpk dn kel dmn pun br ada ,Salam Restorasi .
16	https://www.facebc	Rinaririn Puspita	Masya ALLAH SWT, ga kalah ama negara Korea.
17	https://www.facebc	Abine Luqman Rayhana	Semoga Allaah memudahkan urusan anda
18	https://www.facebc	Ridho Hidayat-muzdalifah	Sehat selalu Pak . Semoga dimudahkan segala urusannya.
19	https://www.facebc	Budi Marwanto	Pk hati2 pk bpk prlu pngwal ancmn tror dicelkakan,jegal makin hrs lawn jngn diam dicurangin
20	https://www.facebc	Olle Hamid	Subsidi itu untuk rakyat, bukan konglomerat.

Figure 4. Data correction results.

3.3 Preprocessing Data

The data that is processed in the preprocessing stage is only commentary data that has been successfully collected and has gone through the data correction stage. The preprocessing steps used are the cleansing, case folding, tokenizing, and filtering stages.

Then at another stage there is the elimination of duplicate data with one form of duplicate examples such as "RI1" "Always healthy" and "Steady", then the results of changing the quantity of the amount of data are presented in table 2.

Table 2. Duplicate Data Removal Results

Name	Before	After	Difference
Anies	200	166	34
Prabowo	300	230	70
Ganjar	200	158	42
Total	**700**	**554**	**146**

3.4 Sentiment Class Labeling

The labeling process is basically divided into three categories of sentiment classes, namely positive sentiment, negative sentiment and neutral sentiment by calculating the value. A document can be classified as positive or negative by using Indonesian word collections or a dictionary consisting of positive and negative word groups. Then the labeling process is carried out on the excel data manually by calculating the score of the number of positive, negative and neutral words in the sentence.

Table 3. Comparison of the amount of data in the sentiment class

Name	Sentiment	Number of comments
Anies	Positive	151
	Negative	9
	Neutral	6
Prabowo	Positive	207
	Negative	12
	Neutral	11
Ganjar	Positive	141
	Negative	9
	Neutral	8

Furthermore, a reduction process was carried out in the class by entering the neutral sentiment class category into positive sentiment which was done manually. The following is a display of sentiment class presentations from each Indonesian presidential candidate in 2024.

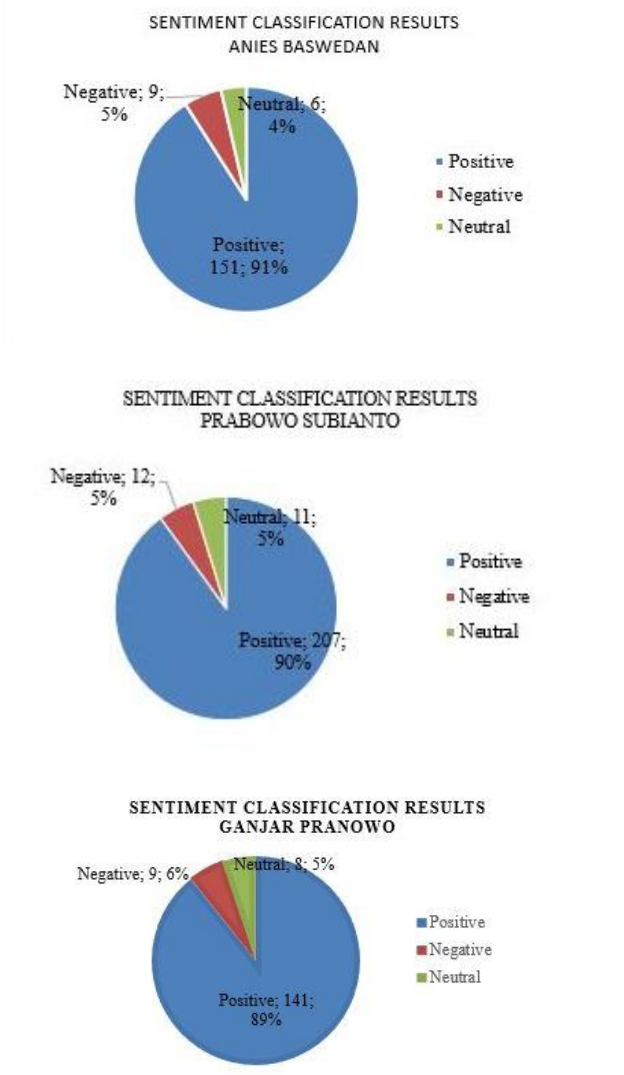

Figure 5. Sentiment class percentage

3.5 Naïve Bayes Classifier

At this stage of the classification process in the research carried out, this is done by using training data and test data from all data. The following shows the results of the Naïve Bayes Classifier classification stage using RapidMiner. Naïve Bayes is a simple learning algorithm that utilizes Bayes' rule together with a strong assumption that the attributes are conditionally independent given the class. While this independence assumption is often violated in practice, naïve Bayes nonetheless often delivers competitive classification accuracy. Coupled with its computational efficiency and many other desirable features, this leads to naïve Bayes being widely applied in practice. [4] Naïve Bayes is a relatively simple and computationally efficient algorithm. It is easy to implement and can handle large datasets with many features. The algorithm's simplicity makes it a good choice for quick prototyping, baseline models, or situations where interpretability is important.

It's important to note that the specific split between training and test data is not fixed and can vary depending on factors such as the dataset size, the complexity of the problem, and the availability of data. Allocating 80% of the data for training provides a relatively large amount of data for the model to learn from. Having an adequate amount of training data helps the model capture patterns and relationships present in the data, leading to better generalization and improved performance. While the training set needs to be large enough for effective learning, it's also important to have a substantial test set to assess the model's performance accurately. Allocating 20% of the data for testing provides a sizable test set to evaluate the model's generalization ability without overly sacrificing the training data size.

The choice of the training and test set split involves a trade-off between bias and variance. A smaller training set may lead to higher bias as the model has less data to learn from, potentially resulting in underfitting. Conversely, a smaller test set may lead to higher variance as the evaluation may be more sensitive to the particular instances in the test set. The 80/20 split strikes a balance between these two factors.

SimpleDistribution

Distribution model for label attribute Sentimen

Class Positif (0.901)
1 distributions

Class Netral (0.045)
1 distributions

Class Negatif (0.054)
1 distributions

accuracy: 90.09%

	true Positif	true Netral	true Negatif	class precision
pred. Positif	100	5	6	90.09%
pred. Netral	0	0	0	0.00%
pred. Negatif	0	0	0	0.00%
class recall	100.00%	0.00%	0.00%	

Figure 6. Classification results using The Naïve Bayes Classifier

IV. CONCLUSION

Shows the classification results obtained using the Naïve Bayes Classifier with a comparison of 80% training data and 20% of test data from 554 commentary data. The author also uses the confusion matrix method in the evaluation process to find accuracy values, precision values and recall values. In measuring performance using the confusion matrix, there are 4 (four) terms to represent the results of the classification process. The four terms are True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). The True Negative (TN) value is the number of negative data that is correctly detected, while False Positive (FP) is negative data but is detected as positive data. Meanwhile, True Positive (TP) is positive data that is correctly detected. False Negative (FN) is the opposite of True Positive, so data is positive, but is detected as negative data. [5] Precision is data taken based on insufficient information. In binary classification, the precision can be made equal to the positive predictive value. With an accuracy rate of 90.09%, it can be said that the Naïve Bayes Classifier method is good for use in the process of classifying comments on the 2024 Indonesian presidential candidate.

REFERENCES

[1] E. B. Santoso and A. Nugroho, "Analisis Sentimen Calon Presiden Indonesia 2019 Berdasarkan Komentar Publik Di Facebook," *Eksplora Inform.*, vol. 9, no. 1, pp. 60–69, 2019, doi: 10.30864/eksplora.v9i1.254.

[2] P. Growth, A. Cahyati, S. Arifin, M. Abror, and U. M. Sidoarjo, "Procedia of Engineering and Life Science Vol . 2 . No . 2 June 2022 Seminar Nasional & Call Paper Fakultas Sains dan Teknologi (SENASAINS 4 th) Universitas Muhammadiyah Sidoarjo Procedia of Engineering and Life Science Vol . 2 . No . 2 June 2022 Seminar," vol. 2, no. 2, 2022.

[3] B. Gunawan, H. S. Pratiwi, and E. E. Pratama, "Sistem Analisis Sentimen pada Ulasan Produk Menggunakan Metode Naive Bayes," *J. Edukasi dan Penelit. Inform.*, vol. 4, no. 2, p. 113, 2018, doi: 10.26418/jp.v4i2.27526.

[4] G. I. Webb, "Encyclopedia of Machine Learning and Data Science," *Encycl. Mach. Learn. Data Sci.*, no. April, 2020, doi: 10.1007/978-1-4899-7502-7.

[5] M. Alfi, "Analisis Sentimen Berdasarkan Knowledge Pattern dan Learning Vector Quantization," pp. 9–18, 2019.

[6] R. Masud, M. Syamsurrijal, T. Baharuddin, and M. Azizurrohman, "Forecasting political parties and candidates for Indonesia's presidential election in 2024 using twitter," *Int. J. Health Sci. (Qassim).*, no. August 2022, pp. 1323–1333, 2022, doi: 10.53730/ijhs.v6ns6.10691.

[7] B. Delvika, N. Abror, and U. R. Gurning, "Perbandingan Algoritma NBC dan C4. 5 Dalam Analisa Sentimen Pemilihan Presiden 2024 Pada Twitter: Comparison of the NBC and C4. 5 Algorithms in Sentiment …," *SENTIMAS Semin. Nas. …*, pp. 41–48, 2023, [Online]. Available: https://journal.irpi.or.id/index.php/sentimas/article/view/548%0Ahttps://journal.irpi.or.id/index.php/sentimas/article/download/548/336

[8] B. M. Iqbal, K. M. Lhaksmana, and E. B. Setiawan, "2024 Presidential Election Sentiment Analysis in News Media Using Support Vector Machine," *J. Comput. Syst. Informatics*, vol. 4, no. 2, pp. 397–404, 2023, doi: 10.47065/josyc.v4i2.3051.

[9] J. Cover, “Table of Content,” *Chinese J. Aeronaut.*, vol. 36, no. 8, pp. i–ii, 2023, doi: 10.1016/s1000-9361(23)00260-1.

[10] D. Nurcahyono, W. P. Putra, A. Najib, and T. R. Tulili, “Analysis sentiment in social media against election using the method naive Bayes,” *J. Phys. Conf. Ser.*, vol. 1511, no. 1, 2020, doi: 10.1088/1742-6596/1511/1/012003.

[11] M. Rizki, M. Fikri Hidayattullah, and Dwi Intan Af’idah, “Klasifikasi Opini Publik di Twitter Terhadap Bakal Calon Presiden Indonesia Tahun 2024 Menggunakan LSTM Secara Realtime Berbasis Website,” *Infotekmesin*, vol. 14, no. 2, pp. 285–295, 2023, doi: 10.35970/infotekmesin.v14i2.1908.

[12] I. P. A. E. D. Udayana, I. G. A. Indrawan, and I. P. D. G. A. Putra, “Decision Support System for Sentiment Analysis of Youtube Comments on Government Policies,” *J. Comput. Networks, Archit. High Perform. Comput.*, vol. 5, no. 1, pp. 27–37, 2023, doi: 10.47709/cnahpc.v5i1.1999.

[13] L. Dey, S. Chakraborty, A. Biswas, B. Bose, and S. Tiwari, “Sentiment Analysis of Review Datasets Using Naïve Bayes‘ and K-NN Classifier,” *Int. J. Inf. Eng. Electron. Bus.*, vol. 8, no. 4, pp. 54–62, 2016, doi: 10.5815/ijieeb.2016.04.07.

[14] A. D. Akmal, I. Permana, H. Fajri, and Y. Yuliarti, “Opini Masyarakat Twitter terhadap Kandidat Bakal Calon Presiden Republik Indonesia Tahun 2024,” *J. Manaj. dan Ilmu Adm. Publik*, vol. 4, no. 4, pp. 292–300, 2022, doi: 10.24036/jmiap.v4i4.160.

[15] R. Gustiawan, R. Rafzan, D. Ropawandi, S. Suswanta, and U. Pribadi, “Analyzing public sentiment on implementing the presidential threshold in Indonesia’s presidential election system,” *J. Civ. Media Kaji. Kewarganegaraan*, vol. 20, no. 1, pp. 156–166, 2023, doi: 10.21831/jc.v20i1.57295.