Journal of Sisfotek Global
ISSN (Online): 2721 - 3161, ISSN (Print): 2088 - 1762
DOI: http://dx.doi.org/10.38101/sisfotek.v12i2.552
vol. 12, No. 2, September 2022, pp. 114-119

Sentiment Analysis of Presidential Candidates Anies Baswedan and Ganjar Pranowo Using Naïve Bayes Method

Nurirwan Saputra[1], Karandi Nurbagja[2] & Turiyan[3]

[1,2,3]Informatics Study Program, Faculty of Science and Technology, University PGRI Yogyakarta
Email: [1]nurirwan@upy.ac.id, [2]karandinurbagja@gmail.com, [3]turiyan295@gmail.com,

HISTORY OF ARTICLES

Received: September 5th, 2022
Revised: September 27th, 2022
Accepted: September 27th,2022

KEYWORDS

Anies Baswedan
Pranowo reward
Quadgram
Naive Bayes
Presidential candidate

ABSTRACT

Presidential elections in Indonesia are carried out in a democratic manner in which the people choose the figures who will nominate themselves for president. With the presidential nomination, there will be many surveys of several figures who have electability to become presidential candidates. Based on a survey that has been issued by several figures who are running for president, namely Anies Baswedan and Ganjar Pranowo, who are the benchmarks for the community to be able to express their opinions from existing social media, one of which is Facebook. This study takes data through a scraping process which is then cleaned or cleaned, then five labels are given, namely: 1 (very negative), 2 (negative), 3 (neutral), 4 (positive), and 5 (very positive). aims to see which sentiment is the highest given by warganet to the upcoming presidential election. This study concludes that netizens have negative sentiments towards figures in the upcoming presidential election. seen from the data randomly generated 49% negative comments, 35% positive comments and 16% neutral. In addition, from 510 data taken by classification using the Naïve Bayes method, as well as testing using the 10-fold cross validation method with Quadgram tokenization resulted in an accuracy of 42.75%, precision 42.10%, and recall 42.70%.

1. Introduction

The presidential election in Indonesia is an important moment where people can go through the democratic process, namely the presidential and vice presidential elections which are held every five years [1]. To become president, there is a requirement that a person is not allowed to become president if that person has been president for two consecutive terms [2]. Based on this, there are many surveys of figures who want to run for president. One of the figures who run for president is Anies Baswedan who was elected governor of DKI Jakarta who has received awards and also has policies and work programs that are quite effective in problems in the DKI Jakarta area [3]. Besides Anis Baswedan, there are also figures who are running for president, namely Ganjar Pranowo, the governor of Central Java who has served for two terms. Ganjar Pranowo is a governor who has flexibility for his people because he has solved problems in the Central Java area [4].

Facebook is a potential media choice for communication about political information due to the high number of social media usage in Indonesia [5]. The number of Internet penetration in Indonesia according to the Association of Indonesian Internet Service Providers or APJII was 54.68% in 2017 with a total of 143.26 million Internet users. Of that number, 87.13% are social media users, 56.01% use the internet to access political information [6].

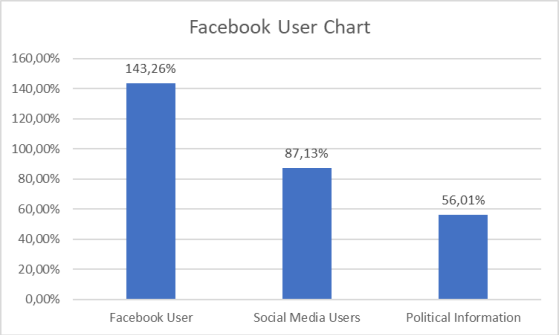

Figure 1. Facebook User Graph

It is not surprising that many people cast positive or negative opinions about presidential candidates who have high flexibility and also react to figures chosen by netizens. Netizens' opinions will be taken through a cleaning and scraping process. The cleaning process is used to clean words or punctuation that are not needed from datasets taken from Facebook [7]. And also doing scraping by extracting data or information from the intended site in retrieval of 510

datasets that have negative sentiments with as many as 49% of netizens' opinions about presidential candidates.[8]. The following is an image containing sentiment data provided by netizens:

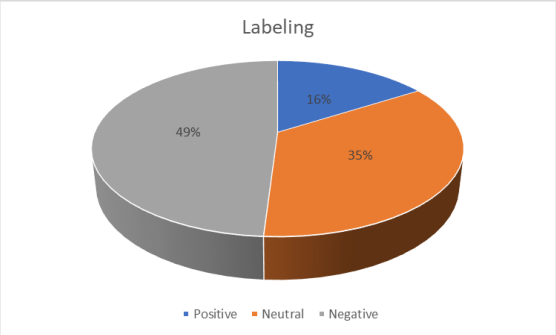

Figure 2. Netizen's Sentiment Graph

In the process of scraping the data, it is hoped that this opinion can be concluded as an accumulation of reactions using the sentiment analysis method of the Naïve Bayes algorithm. Naïve Bayes is a method that is suitable for calculations in this study because this algorithm is very easy to understand, faster in calculating accuracy and only requires a little training data used [9], and also the Naïve Bayes method is very influential for this study because it has a high enough accuracy than the C45 method [10].

The software that will be used is WEKA which is an application to identify information from datasets that are entered into the software with a choice of methods that have been provided in the software, one of which is the Naïve Bayes method [11]. The dataset used in this study was obtained from netizens comments on presidential candidates Anies Baswedan and Ganjar Pranowo whose names were included in the presidential election survey nominations. Through a dataset that has been taken from social media, namely Facebook, it can show a pattern of how citizens respond to two figures, namely Anies Baswedan and Ganjar Pranowo. The dataset obtained will be processed using the Python programming language and using Jupyter tools which are work tools that can be executed simply. By using the Jupyter application, it can be used for stemming programming from literary masters, namely the process of getting basic words in datasets that have been taken from citizen comments by removing affixes [12].

After the dataset is converted into basic words through the stemming process, after that the dataset is given guidelines or labels containing numbers that will make labeling easier. This number is a label that contains information Very Negative for number 1, Negative for number 2, Neutral for number 3, Positive for number 4, and Very positive for number 5. Which will determine which accuracy is the highest from the sentiment given by the warganet. After that, the dataset is entered into the WEKA application using the Naïve Bayes method which will produce the accuracy of the given label.

From the explanation above, it can be concluded that this study will discuss the analysis of the accuracy of the netizen's sentiment regarding the news of Anies Baswedan and Ganjar Pranowo who will run for president using the Naïve Bayes method using the WEKA application.

2. Research Methodology

The method that will be used in this research is to take a dataset about "citizen comments of presidential candidates Anies and Ganjar" which predicts the accuracy of netizen responses using the Naïve Bayes method:

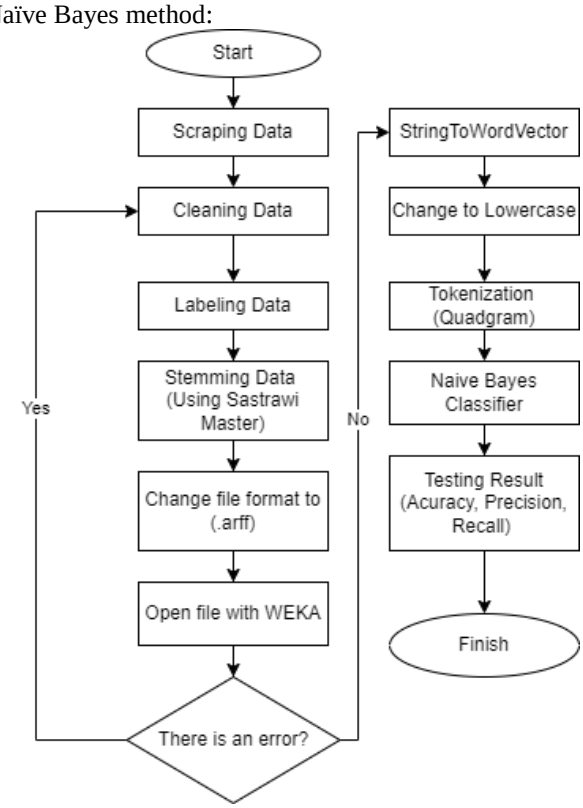

Figure 3. Methodology Flowchart

2.1 Data Collection

In this research, the first thing to do is collect data, sourced from Facebook by scraping comments without using an API from the retrieval of the "CNN Indonesia News about presidential candidate competition" related to the opinions of netizens on Anies and Ganjar who are running for president. This study took a random sample of datasets obtained as many as 510 citizen response data taken through Facebook. This news was uploaded on June 1, 2022 and contains the most comments until June 2022.

2.2 Preprocessing Data

Preprocessing data is the first step in preparing data before the classification process is carried out for

cleaning, or changing the source of data [13]. The following data preprocessing is carried out.

Cleaning: Cleaning is a process that cleans words that are not needed. This aims to parse the noise that will cause data that is not optimal [13].

Labels: A process that labels words such as Very Negative, Negative, Neutral, Positive, and Very Positive in order to calculate the accuracy of the sentiment results obtained automatically [14]. The labeling given to the datasets used include:

- The word labeling is very negative given the number (1): giving very obscene comments to presidential candidates that contain pornography and *SARA* which will have a negative impact on readers and will trigger debate between each other.
- The labeling of negative words is given a number (2): Negative comments provide comments that vilify and bring down presidential candidates by mentioning the problems of presidential candidates and comparing the achievements of each presidential candidate.
- The labeling of the word neutral is given a number (3): Neutral comments contain comments that do not take sides with each other, and provide good knowledge and information to the people who will choose the presidential candidate so that there will be no debate and conflict against both parties.
- Positive word labeling is given a number (4): Positive comments provide comments that support the hard work of each presidential candidate who will be elected.
- Very positive labeling is given a number (5): Comments containing appreciation and best wishes to the presidential candidate from the Indonesian people in leading the Indonesian state and showing support among presidential candidates.

Stemming: After labeling the dataset, the next step is to stem the python programming which will make the dataset the basis for netizens' opinions. After stemming the dataset will be entered into WEKA formatted (.arff). Literature master is a library for generating uniform tenses [15]. The main literature also contains 29932 root words [16]. The results of stemming following the table. table. 1:

Table1.Stemming Data

Label	Comment	Stemming
Positive	*Whatever the story, Anis will still be the president*	*What's the story of Anis, who becomes president?*
Negative	*the story is just dreaming kwkwkw*	*the story is just dreaming kwkwkw*
Neutral	*Who said the choice of the Indonesian people*	*Who said vote for Indonesian people*

Unsupervised Learning: Furthermore, the data through unsupervised learning will be grouped or data categories will be converted into StringToWordVector [17].

Term Frequency-Inverse Document Frequency (TF-IDF): After that, process TF-IDF, Lowercase, and Tokenization. The TF-IDF process is the process used to assign weights to the words in the dataset [18]. The term Frequency uses the following formula Eq. 1.

$$tf(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

The number of occurrences of the term (t) in the document (d). The more often a term occurs, the larger the tf value.

To find the idf can be seen in the following formula. Eq. 2.

$$idf(t,D) = log\frac{N}{|\{d \in D: t \in d\}|}$$

A measure of the information provided by the term t, ie how often or infrequently the term appears throughout the document. The rarer the terms are between documents, the larger the idf value. This value is the inverse logarithm of the number of documents having the t term divided by the total number of documents (N).

Lowercase: After generating numbers from the TF-IDF, a lowercase process is carried out which makes the word format lowercase or vice versa. After the dataset is converted to lowercase with lowercase the dataset is entered into Tokenization [19].

Tokenization: Tokenization is the process of cutting the input string based on each compiled word [7]. Tokenization carried out in this research uses Quadgram tokenization which is further classified for the accuracy value of the Naïve Bayes method. The results of Quadgram tokenization can be seen in the table. table.2:

Table 2. Quadgram Tokenization

Label	Comment	Quadgram
Positive	*Whatever the story, Anis will still be the president*	*"Whatever the story is, Anis will always be the president", "Anis will still be president"*
Negative	*the story is just dreaming kwkwkw*	*"the story is just imaginary kwkwkw"*

Label	Comment	Quadgram
Neutral	*Who said the choice of the Indonesian people*	*"Who said the people's choice","said the Indonesian people's choice"*

2.2 Classification using the Naïve Bayes method

This study performs a classification using the Naïve Bayes method to determine the values of accuracy, precision, and recall. And also determine the confusion matrix by dividing the testing data and training data whose results are normalized using the formula: "IF (normalization "Negative"> Normalization "Positive"> Normalization "Neutral")".

3. Results and Discussion

3.1 Results of the implementation of the Naïve Bayes method

The prediction model using the Naïve Bayes method is done by looking for comments from the dataset that has gone through the scraping and cleaning process. Then look for the average (mean) with the tokenization process that has been determined from the training data. The formula used to predict the Naïve Bayes method using the equation [20]. Eq 3.

$$p(c|E) = p(E|c)p(c)p(E)$$

In classification using the Naive Bayes method, the purpose of this algorithm is to build a classifier with a given label, especially in this study as many as 5 labels. In the formula, the character "E" is represented by a tuple of attribute values (x1, x2, ,⋯ , xn), where "xi" is the attribute value "Xi" and "C" represents the classification variable [21].

3.2 Test

The Naive Bayes classification was tested using a validation method, namely 10-fold cross validation, which is a process that divides testing and training data in stages by 10 in the Naïve Bayes method [22]. And after testing, it will produce accuracy, precision, recall values from the confusion matrix table. 3 is generated.

Table 3. Confusion Matrix

	VN	NEG	NEU	POST	VP
VN	0	11	2	13	1
NEG	7	124	26	39	28
NEU	1	29	37	8	6
POST	4	33	20	37	19
VP	1	20	7	17	20

Notes Table. 3:

VN = Very Negative
Neg = Negative
Neu = Neutral
Pos = Positive
VP = Very Positive
To get the accuracy value use the following formula. Eq 4.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}*100\%$$

To get the precision value use the following formula. Eq 5.

$$Precision = \frac{TP}{TP+FP}*100\%$$

Meanwhile, to get the recall value using the following formula. Eq 6.

$$Recall = \frac{TP}{TP+FN}*100\%$$

Information :

Table 4. Formula Description

Symbol	Information
TP :	is a sentence that is classified as VN/NEG/NEU/POS/VP and is a sentence VN/NEG/NEU/POS/VP
TN :	is a sentence that is not classified as VN/NEG/NEU/POS/VP and is not a VN/NEG/NEU/POS/VP sentence
FP :	is a sentence that is classified as VN/NEG/NEU/POS/VP but is not a VN/NEG/NEU/POS/VP sentence
FN :	is a sentence that is not classified as VN/NEG/NEU/POS/VP but is a VN/NEG/NEU/POS/VP sentence. The percentage results generated can be seen in table 4.

TP= is a sentence classified as VN/NEG/NEU/POS/VP and is a sentence VN/NEG/NEU/POS/VP
TN= is a sentence that is not classified as VN/NEG/NEU/POS/VP and is not a VN/NEG/NEU/POS/VP sentence
FP= is a sentence classified as VN/NEG/NEU/POS/VP but is not a VN/NEG/NEU/POS/VP sentence
FN= is a sentence that is not classified as VN/NEG/NEU/POS/VP but is a VN/NEG/NEU/POS/VP sentence.

The percentage results generated can be seen in table 4.

Table 5. Accuracy, Precision and Recall

Naive Bayes Classification		
Accuracy	**Precision**	**Recall**
42.75%	42.10%	42.70%

The prediction model of this study chose the tokenization Quadgram with 42.75% accuracy, 42.10% precision, and 42.70% recall.

4. Discussion

Sentiment analysis using quadgram tokenization has a significant effect on the resulting accuracy, precision and recall. The percentage tends to be small because using Quadgram tokenization groups sentences into 4 words, so dividing into 4 words can create different meanings compared to unigrams, bigrams and trigrams. In addition, by using Quadgram tokenization, it is very rare to find the same 4-word sequence pattern from a sentence.

5. Conclusion

The results of the Naïve Bayes classification in the CNN Indonesia News dataset on Scraping and Cleaning of the Presidential Candidate Competition. Providing labeling, namely: 1 (very negative), 2 (negative), 3 (neutral), 4 (positive), and 5 (very positive) The labeling that will determine the tokenization used is Quadgram which has accuracy, 42.10% precision, and 42.70% considering and producing the highest sentiment, namely negative sentiment.

References

[1] N. Hermawan, "Representasi Anies dan Ganjar pada Bursa Calon Presiden Indonesia 2024 dalam Berita Online Okezone.com," *Syntax Lit. ; J. Ilm. Indones.*, vol. 6, no. 1, p. 24, 2021, doi: 10.36418/syntax-literate.v6i1.4613.

[2] I. G. H. Kurniawan and H. Arianto, "Polemik Pembatasan Masa Jabatan untuk Jabatan Publik di Indonesia Terkait dengan Demokrasi dan Pancasila," *Lex Jurnalica*, vol. 17, pp. 264–270, 2020.

[3] N. Aziza, "Analisis Gaya Kepemimpinan Yang Diterapkan Anies Baswedan, Sebagai Gubernur Dki Jakarta," *J. Media Huk.*, vol. 21, no. June, 2021.

[4] A. Probowati, "Kepemimpinan ganjar pranowo," *Yogyakarta, Univ. Muhammadiyah*, no. June, pp. 0–15, 2021.

[5] S. Suratno, I. Irwansyah, N. F. Ernungtyas, G. F. Prisanto, and S. Hasna, "Pemanfataan Media Sosial Facebook Sebagai Strategi Komunikasi Politik," *SOURCE J. Ilmu Komun.*, vol. 6, no. 1, pp. 89–98, 2020, doi: 10.35308/source.v6i1.1552.

[6] W. Yasya, P. Muljono, K. B. Seminar, and H. Hardinsyah, "Pengaruh Penggunaan Media Sosial Facebook Dan Dukungan Sosial Online Terhadap Perilaku Pemberian Air Susu Ibu," *J. Stud. Komun. dan Media*, vol. 23, no. 1, p. 71, 2019, doi: 10.31445/jskm.2019.1942.

[7] N. Saputra, T. B. Adji, and A. E. Permanasari, "Analisis Sentimen Data Presiden Jokowi Dengan Preprocessing Normalisasi Dan Stemming Menggunakan Metode Naive Bayes Dan SVM," *J. Din. Inform.*, vol. 5, no. 1, pp. 1–12, 2015, [Online]. Available: http://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/4793%0Ahttps://jurnal.teknikunkris.ac.id/index.php/semnastek2019/article/view/343/342

[8] S. Satriajati, S. B. Panuntun, and S. Pramana, "Implementasi Web Scraping Dalam Pengumpulan Berita Kriminal Pada Masa Pandemi Covid-19," *Semin. Nas. Off. Stat.*, vol. 2020, no. 1, pp. 300–308, 2021, doi: 10.34123/semnasoffstat.v2020i1.578.

[9] A. Rahman, F. Rahmat, M. Y. Fariqi, and S. Adi, "Metode Naive Bayes untuk Menganalisis Akurasi Sentimen Komentar di Youtube," 2020. [Online]. Available: http://bit.ly/2u802Pe

[10] F. Fathonah and A. Herliana, "Penerapan Text Mining Analisis Sentimen Mengenai Vaksin Covid - 19 Menggunakan Metode Naïve Bayes," *J. Sains dan Inform.*, vol. 7, no. 2, pp. 155–164, 2021, doi: 10.34128/jsi.v7i2.331.

[11] U. Pujianto and P. Y. Ristanti, "Perbandingan kinerja metode C4.5 dan Naive Bayes dalam klasifikasi artikel jurnal PGSD berdasarkan mata pelajaran," *Tekno*, vol. 29, no. 1, pp. 50–67, 2019, doi: 10.17977/um034v29i1p50-67.

[12] N. J. M. Verdaningroem and A. Saifudin, "Penerapan Kamus Dasar Pada Algoritma Porter Untuk Mengurangi Kesalaham Stemming Bahasa Indonesia," *J. Teknol. Univ. Muhammadiyah*, vol. 10, no. 2, pp. 103–112, Jul. 2018, doi: 10.24853/jurtek.10.2.103-112.

[13] F. A. Muttaqin and A. M. Bachtiar, "Implementasi Teks Mining Pada Aplikasi Pengawasan Penggunaan Internet Anak 'Dodo Kids Browser,'" *J. Ilm. Komput. dan Inform.*, pp. 1–8, 2016.

[14] L. Ardiani, H. Sujaini, and T. Tursina, "Implementasi Sentiment Analysis Tanggapan Masyarakat Terhadap Pembangunan di Kota Pontianak," *J. Sist. dan Teknol. Inf.*, vol. 8, no. 2, p. 183, 2020, doi: 10.26418/justin.v8i2.36776.

[15] R. Riyadduloh and A. Romadhony, "Normalisasi Teks Bahasa Indonesia Berbasis Kamus Slang Studi Kasus: Tweet Produk

Gadget Pada Twitter," *eProceedings Eng.*, vol. 8, no. 4, pp. 4216–4228, 2021.

[16] A. Librian and R. Kukuh, "Sastrawi · GitHub."

[17] H. Abijono, P. Santoso, and N. L. Anggreini, "Algoritma Supervised Learning Dan Unsupervised Learning Dalam Pengolahan Data," *J. Teknol. Terap. G-Tech*, vol. 4, no. 2, pp. 315–318, 2021, doi: 10.33379/gtech.v4i2.635.

[18] R. Melita, V. Amrizal, H. B. Suseno, and T. Dirjam, "Penerapan Metode Term Frequency Inverse Document Frequency (Tf-Idf) Dan Cosine Similarity Pada Sistem Temu Kembali Informasi Untuk Mengetahui Syarah Hadits Berbasis Web (Studi Kasus: Hadits Shahih Bukhari-Muslim)," *J. Tek. Inform.*, vol. 11, no. 2, pp. 149–164, 2018, doi: 10.15408/jti.v11i2.8623.

[19] R. T. Susilo and S. Jaya, "Implementasi Web Mining dengan Metode Clustering pada Dokumen Akreditasi Program Studi," *Pros. Semnastek*, pp. 1–7, 2019.

[20] M. N. Tentua, T. Turiyan, and K. Nurbagja, "Komparasi Metode Naïve Bayes dan C45 pada Prediksi Pelanggan Deposito Berjangka," vol. 11, no. 1, pp. 92–99, 2022.

[21] H. Zhang, "The Optimality of Naive Bayes", Accessed: Sep. 17, 2022. [Online]. Available: www.aaai.org

[22] P. Pitria, "Analisis Sentimen Pengguna Twitter Pada Akun Resmi Samsung Indonesia Dengan Menggunakan Naïve Bayes," *Undergrad. Theses from JBPTUNIKOMPP*, 2019, Accessed: Sep. 17, 2022. [Online]. Available: https://123dok.com/document/8yde231q-analisis-sentimen-pengguna-twitter-resmi-samsung-indonesia-menggunakan.html