Research Article

Sentiment Analysis of Twitter Users Ahead of the 2024 Election Using the Naive Bayes Method

[1*]**Andi Subasar**

[1] Faculty of Engineering, Universitas Muhammadiyah Bengkulu, Bengkulu City, Indonesia

* Corresponding Author: andissr14@gmail.com

Abstract: Elections are pivotal moments in a democratic nation, where citizens can express their opinions and political preferences. In today's digital era, social media, particularly Twitter, has become a crucial platform for expressing sentiments related to elections. This research aims to analyze Twitter users' sentiments towards the 2024 election using the Naive Bayes method, Knowing the public's views, especially on the Twitter platform, regarding the 2024 election and implementing the Naive Bayes method to classify sentiment. The research method itself consists of data collection from Twitter using the Twitter API, data preprocessing including data cleaning, removal of URLs, hashtags, duplicate words, normalization of words, tokenization, and removing meaningless words using the Rapid Miner application, then testing using training data and testing data in the Naive Bayes method., the data obtained from the keyword "2024 election" on Twitter for the initial data amounted to 2991 data. After going through the cleaning process, clean data amounting to 1069 data was obtained. From the tested data, the results obtained are as follows: The precision class produces an average percentage of true positives of 100.00% while negatives of 81.48%. Class recall produces a percentage of 98.95%, and the accuracy of testing the model is 99.00%. The research results show that the Naive Bayes method has been successfully applied to analyze Twitter user sentiment.

Keywords: Sentiment Analysis; Twitter; 2024 Election; Naive Bayes Method; RapidMiner

check for updates

Citation: A.Subasar, "Sentiment Analysis of Twitter Users Ahead of the 2024 Election Using the Naive Bayes Method", *Iota*, **2024**, ISSN 2774-4353, Vol.04, 03. https://doi.org/10.31763/iota.v4i3.784

Academic Editor : Adi, P.D.P

Received : July, 14 2024

Accepted : July, 22 2024

Published : August, 22 2024

Publisher's Note: ASCEE stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2024 by authors. Licensee ASCEE, Indonesia. This article is an open access article distributed under the terms and conditions of the Creative Commons Atribution Share Alike (CC BY SA) license(https://creativecommons.org/licenses/by-sa/4.0/)

1. Introduction

Social media has a vital role in the digital era of human life today because it provides many platforms for humans to share news and information and express their opinions in public. Platforms such as Facebook, Twitter, TikTok, and Instagram provide features for humans to exchange opinions and information supported by a wide range of areas, making it easy for us to connect. Social media platforms are very important in determining policies and public reactions to these elections in politics, especially general elections.

Elections are an important political process in a democratic country, and Indonesia is no exception. In today's digital era, social media platforms such as Twitter provide a platform for users to share opinions, voice political preferences, and express sentiments related to elections. Therefore, analyzing Twitter user sentiment towards elections can provide valuable insights for political stakeholders in understanding voter preferences and measuring the level of support for certain candidates or political parties. There are many methods in Rapid Miner Software, and one of the most effective and popular sentiment analysis methods [1,2, 3, 4, 5, 6] is the Naive Bayes method.

This method is based on Bayes' theorem and uses probability models to classify Twitter users' sentiments into positive, negative, and neutral categories. The Naive Bayes method is suitable for sentiment analysis because it quickly models the relationship between text features and sentiment labels and provides fairly accurate results in various cases [7, 8, 9, 10, 11, 12].

The formulation of the problem from the existing background is how to identify and analyze the sentiment of Twitter users towards the 2024 election. How many are positive, negative, and neutral, and how is the data tested for accuracy and effectiveness of sentiment analysis [13, 14, 15, 16] using the Naive Bayes classification method? What is the effectiveness of the Naive Bayes method in classifying Twitter user sentiment towards the 2024 Election, and what are the factors that influence the polarization of Twitter user sentiment towards the 2024 Election? This research aims to find out the views or opinions of Twitter users on the 2024 election, how many have positive opinions, and how many are negative using the Naive Bayes method in naive Bayes software. This research will help understand the sentiment patterns of Twitter users related to elections. It can provide important insights for related parties, such as political parties and candidates, in understanding public opinion.

2. Method

This research aims to classify the sentiment of Twitter users towards the 2024 election so that it can provide an analysis of the community, especially the Twitter platform about this election, how many percent gave a positive response, and how many percent gave a negative response. The scope of this research itself is only limited to the Twitter platform, other social media platforms will not be discussed.

2.1 System Design

The system design that will be carried out in this research is divided into data crawling, preprocessing, classification method, and results. Data retrieval is done using the Twitter API token process in RapidMiner software. The tweet data collected includes ID from the user, user ID, and text. In the Twitter data retrieval stage, using the API in Rapidminer, the #pemilu2024 hashtag data is automatically retrieved. The raw data is then converted to Excel format and will go through data processing by labeling. The labeling is done manually according to the predetermined sentiment, namely positive sentiment for hashtag #pemilu2024 positive comments and negative sentiment for hashtag #pemilu2024 negative comments.

After the data has its label, it is separated into two parts, namely training data and testing data. Training data is used to classify tweets in their sentiment class, and the amount of data is greater while testing data is used to test how accurate the data that has been trained is. Moreover, System Design can be seen in Figure 1.

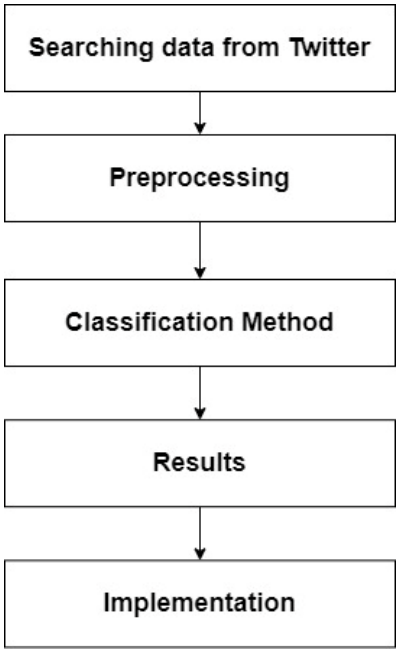


Figure 1. System Design

2.2 Preprocessing

Preprocessing is done to avoid invalid data, interference with data, and inconsistent data. This study's text processing stages include: 1. Removing URLs (http://www.*situs*.com), and email will be deleted at this stage 2. Removing Twitter Special Characters: This stage removes Twitter special characters such as hashtags and usernames and special characters such as retweets. 3. Tokenization Tokenize or tokenization is the process of collecting all the words that appear and removing any read or symbol that is not a letter. (Muthia, 2017) 4. Case Folding Case Folding is the process of uniformizing the shape of letters into uppercase or lowercase letters 5. Stopword Is removing words that do not affect the classification process.

2.3 Classification Method

The method used in this research is the naive Bayes method using RapidMiner software. Naive Bayes is a popular machine learning method, the easiest to understand, and also has good performance.

3. Result and Discussion

The results of this research will be explained in detail in this chapter using the naive Bayes method and the implementation of the classification. Retrieving data from Rapidminer software to Twitter can only be done after getting API key connections from Twitter. After the Twitter API connection has been connected, the next step is to search for data or what is commonly called crawling data by clicking on the #pemilu2024 hashtag query in Rapidminer; keep in mind that data retrieval using the keyword method can only take a maximum of 100 data in one search, meaning that if we want to retrieve any data, we have to crawl the data on twitter several times. The data obtained from the keyword election 2024 on Twitter for the initial data is 2991 data that has not been cleaned and must be converted into Excel data format to be processed. Then, the data is cleaned by removing unnecessary words such as retweet, URL, hashtag, mention, and others using the "Replace" operator in Rapidminer can be seen in Figure 2.

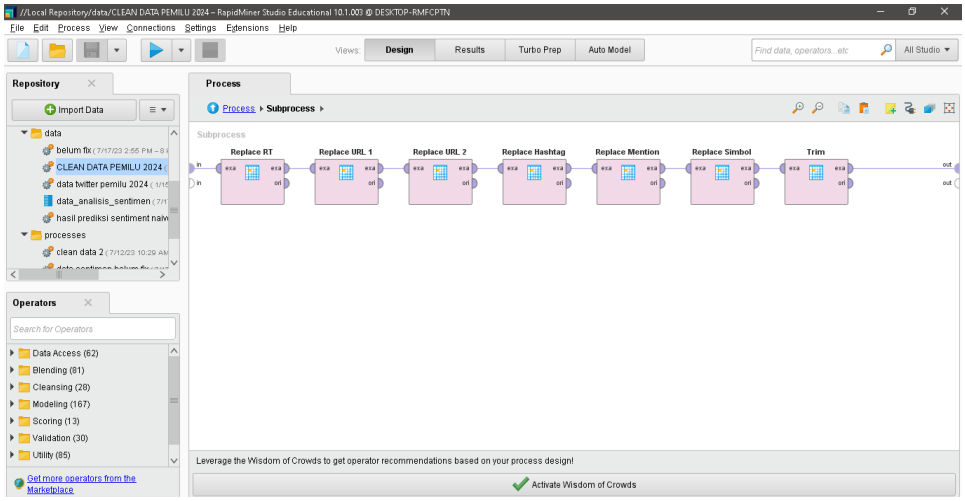


Figure 2. 'Replace' Operator

3.1 Classification Method

After going through the fireplace process, 1069 clean data were obtained, which were then processed in Rapidminer to determine the sentiment of Twitter user comments on the 2024 election. Then, to improve the quality of sentiment and ensure data security, data processing text is carried out using the "Tokenize" operator; tokenization replaces actual values with opaque values to obtain data security. Tokenize is also often called collecting all data that appears and eliminating any read or symbol that is not a letter. This process is continued by changing the letters that are still capitalized into lowercase letters by clicking the "Transform cases" operator, then removing English stopwords from the data using the "Filter Stopword" operator

and replacing them with language stopwords, then continuing filtering based on their respective content by clicking the "Filter token by length" operator. After completing the data tokenization process, the next step is manual labeling to become a reference material in predicting sentiment. Then, the applied model process was performed to see sentiment predictions in Rapidminer using 570 data samples manually labeled. For the prediction of the applied model in Rapidminer using 570 data samples, the following results were obtained, Example of Set Prediction Apply Model can be seen in Figure 3.

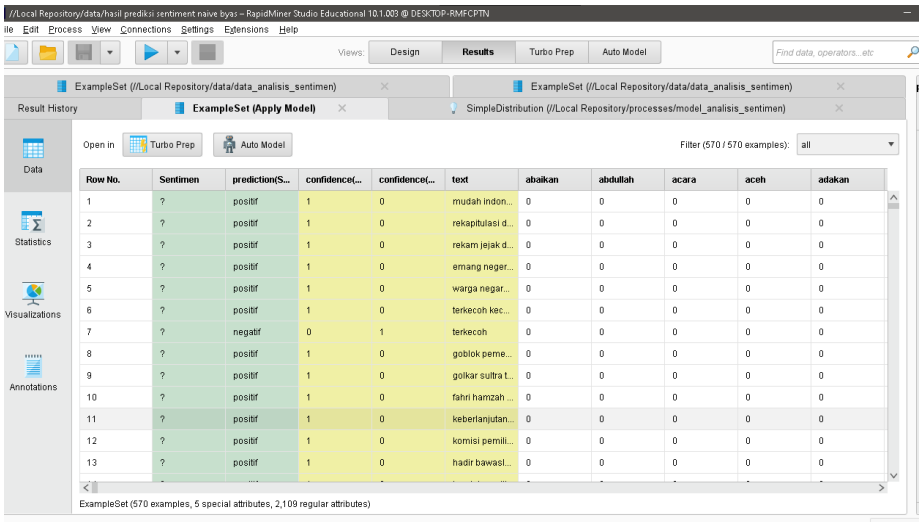


Figure 3. Example of Set Prediction Apply Model

Description: After testing with data samples, it was found that the more samples tested, the more accurate the sentiment analysis prediction [17, 18, 19, 20] in Rapidminer. This process means that the more data samples tested, the better the accuracy of testing in RapidMiner.

3.2 Naive Bayes Application Results

Naive Bayes is the most popular classification method and is highly accurate. The basic concept of the Naive Bayes method is to classify data based on simple probabilities. It is designed to be used with the assumption that the explanatory variables are independent of each other. Referring to the data processed by the author on the previous page in the Naive Bayes method, there are several sentiment categories, including positive, negative, and neutral. However, in this study, the authors only took positive and negative categories because they were more inclined to find out hate speech and good speech about the 2024 election. The naive Bayes Classification Model can be seen in Figure 4.

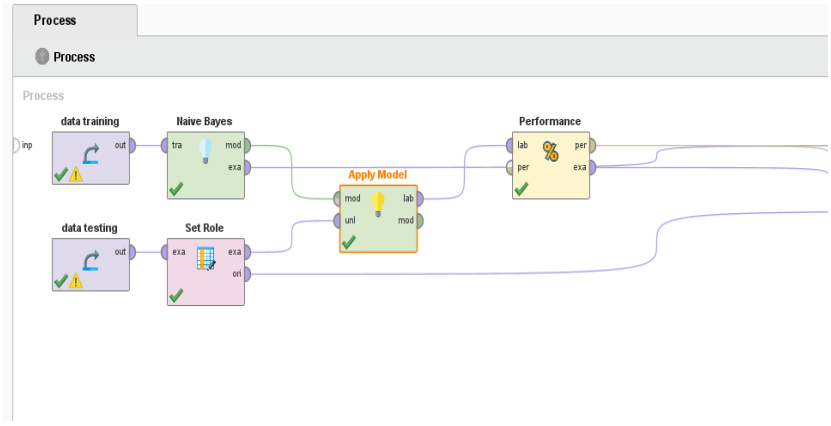
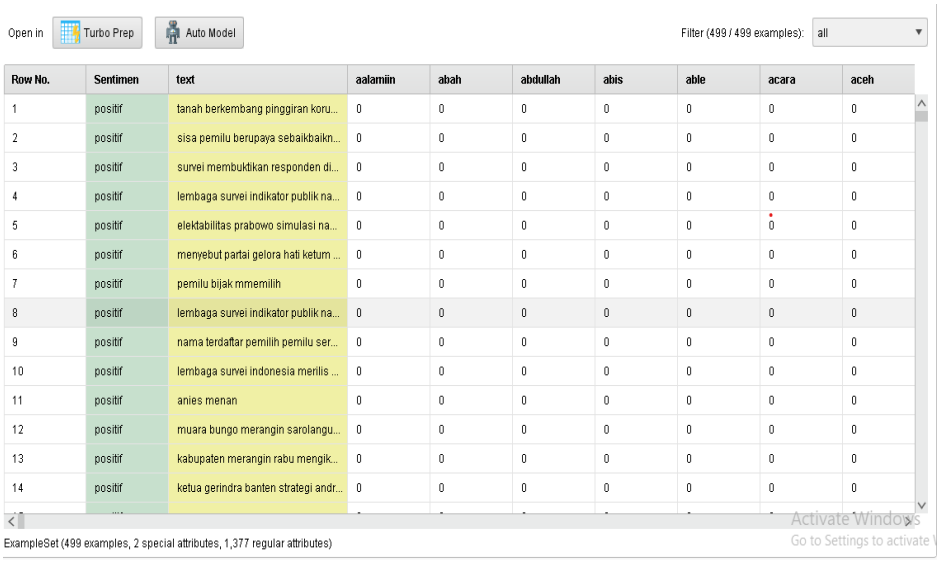


Figure 4. Naive Bayes Classification Model

Classification modeling using the Naive Bayes method, as in Figure 4 above, aims to achieve positive and negative sentiment by classifying 2 types of data, namely training data and testing data. In this study, the algorithm produces 99.9% accuracy using the Naive Bayes method. From testing the results of the data with Rapid Miner, we will get class precision and class recall; what is meant by class precision is the ratio of relevant documents to the total query result documents, while class recall is the ratio of the number of relevant documents retrieved to the total relevant documents, for the display of training data and testing data can be seen in Figure 5.


Open in Turbo Prep Auto Model Filter (499 / 499 examples): all

Row No.	Sentimen	text	aalamiin	abah	abdullah	abis	able	acara	aceh
1	positif	tanah berkembang pinggiran koru...	0	0	0	0	0	0	0
2	positif	sisa pemilu berupaya sebaikbaikn...	0	0	0	0	0	0	0
3	positif	survei membuktikan responden di...	0	0	0	0	0	0	0
4	positif	lembaga survei indikator publik na...	0	0	0	0	0	0	0
5	positif	elektabilitas prabowo simulasi na...	0	0	0	0	0	0	0
6	positif	menyebut partai gelora hati ketum ...	0	0	0	0	0	0	0
7	positif	pemilu bijak mmemilih	0	0	0	0	0	0	0
8	positif	lembaga survei indikator publik na...	0	0	0	0	0	0	0
9	positif	nama terdaftar pemilih pemilu ser...	0	0	0	0	0	0	0
10	positif	lembaga survei indonesia merilis ...	0	0	0	0	0	0	0
11	positif	anies menan	0	0	0	0	0	0	0
12	positif	muara bungo merangin sarolangu...	0	0	0	0	0	0	0
13	positif	kabupaten merangin rabu mengik...	0	0	0	0	0	0	0
14	positif	ketua gerindra banten strategi andr...	0	0	0	0	0	0	0

ExampleSet (499 examples, 2 special attributes, 1,377 regular attributes)

Figure 5. Data Training

Training data has been manually labeled as a reference when it is tested for accuracy in Naive Bayes. This data teaches the Naive Bayes model how to classify new data. Usually, training data consists of features (attributes) and labels (desired class or category). Data Testing can be seen in Figure 6.

Row No.	Sentimen	Text
1	?	2024 abis pemilu sama juga biar mereka banyak istirahat juga
2	?	Partai Gerakan Indonesia Raya Gerindra Kabupaten Pohuwato mengajukan Perbaikan Dokumen P...
3	?	Komisi Pemilihan Umum Kabupaten Karangasem menerima Pengajuan Perbaikan Bakal Calon An...
4	?	Komisi Pemilihan Umum Kabupaten Tulang Bawang menerima Pengajuan Perbaikan Dokumen B...
5	?	berikut kami sampaikan jumlah pemilih berdasarkan generasi pada Pemilu 2024 di Kabupaten Ma...
6	?	Hayoo kalian masuk generasi mana
7	?	Komisi Pemilihan Umum Kabupaten Tulang Bawang menerima Pengajuan Perbaikan Dokumen B...
8	?	Demokrat S14P Memperjuangkan Perubahan dan Perbaikan
9	?	PANGerindra Gelar Pertemuan Tanpa Ketum Zulkifli Hasan dan Prabowo Subianto Bahas Pemilu 2...
10	?	KPU telah menerima dokumen perbaikan syarat bakal calon anggota legislatif dari 18 parpol pesert...
11	?	Komisi Pemilihan Umum Kabupaten Tulang Bawang menerima Pengajuan Perbaikan Dokumen B...
12	?	Yaaa betul kita semua setara
13	?	Mari sahabat semua tanpa terkecuali kita salurkan hak pilih dan ayo ikut awasi bersama Pemilu mili...
14	?	2024 sih biar selesai masa pemilu nya jg trus lagian belom punya duittttt wkwkkw
15	?	Komisi Pemilihan Umum Kabupaten Karangasem menerima Pengajuan Perbaikan Bakal Calon An...

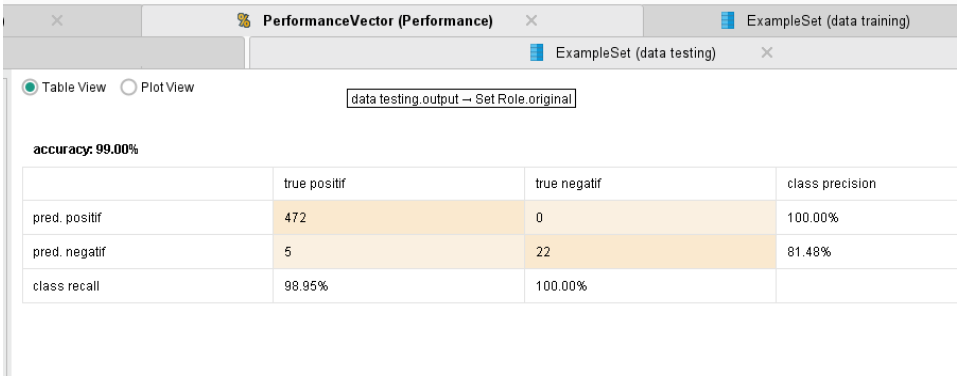
ExampleSet (593 examples, 1 special attribute, 1 regular attribute)

Figure 6. Data Testing

After the training data has been created and the data analysis model is obtained, the next step is determining the testing data or test data. The testing data itself is a collection of data used to test or evaluate the model's performance from the training data and, of course, has not been labeled positive or negative. The testing data itself tests machine learning after the model has been trained using training data.

3.3 Analysis Classification Results

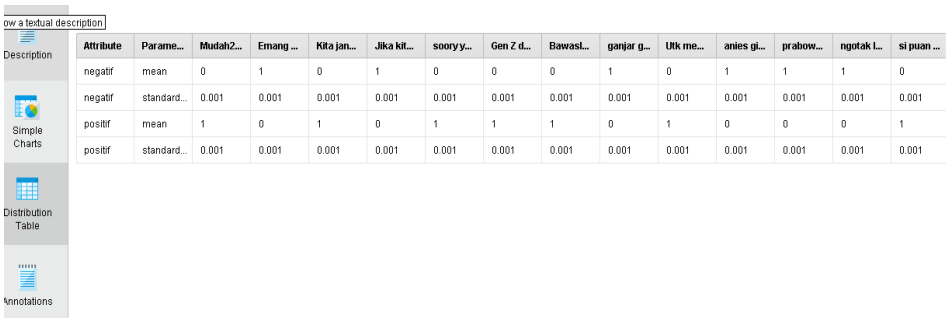
The classification results were obtained after testing sentiment analysis using the naive Bayes method. As many as 472 people gave positive comments on the 2024 election, while 22 others gave negative comments with an accuracy reaching 99.00%. With the following prediction levels: For the precision class, the average percentage of true positive is 100.00% while negative is 81.48%. Meanwhile, the recall class produces 98.95% for true positive and 100 for true negative. The following classification results are seen from the performance operator. Performance Tabel can be seen in Table 7.



	true positif	true negatif	class precision
pred. positif	472	0	100.00%
pred. negatif	5	22	81.48%
class recall	98.95%	100.00%	

Figure 7. Performance Tabel

From the performance testing table in Figure 7, it can be seen that class precision produces a percentage of 81.48% while class recall is 98.95%, with accuracy reaching 99.00%. The distribution of words can be seen in Figure 8.



Attribute	Parame...	Mudah2...	Emang ...	Kita jan...	Jika kit...	soory y...	Gen Z d...	Bawasl...	ganjar g...	Utk me...	anies gi...	prabow...	ngotak l...	si puan ...
negatif	mean	0	1	0	1	0	0	0	1	0	1	1	1	0
negatif	standard...	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
positif	mean	1	0	1	0	1	1	1	0	1	0	0	0	1
positif	standard...	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001

Figure 7. Table Distributor

In the context of Naive Bayes, the distributor table usually refers to the word probability distribution table calculated based on the training data. This table presents the probabilities of words appearing in each class or category. It is an important part of the Naive Bayes algorithm for text classification, as it provides the information needed to calculate the conditional probabilities of words in the text to be classified. The implications of this research lead to several aspects where we can see responses to find out the public's opinion on the 2024 election, how many are positive and also how many have negative sentiments so that we can know the tendency of the results of the classification with the Naive Bayes method with data on Twitter with the application of the classification method that makes us know how Twitter users view the general election that will take place this year, and because of

the limited time and data in this study, further research is expected to take and use more data coverage from the results of crawling data on Twitter about elections in Indonesia and the application of other classification methods by making comparisons so that they can see which one is better. It can also provide a more accurate picture of the results of this study.

5. Conclusion

This research aims to determine the public sentiment towards the 2024 election by applying the Naive Bayes Classifier method. How many people have positive or negative sentiments? Through data collection from Twitter processed in Rapidminer software, this research more or less answers how people's reactions and opinions, especially Twitter users, towards this election. The results showed that the Naive Bayes method was successfully applied to analyze the sentiment of Twitter users. The results of this study show how the public's opinion on this election is that most people are positive; it can be seen from the tests carried out with training data totaling 500 data and testing data totaling 592 data taken from Twitter using rapid miner software that it is easier to provide an overview of the sentiment analysis classification results of this study with the results as many as 472 people gave positive responses while 22 people gave negative responses. The accuracy obtained from this test is 99.00%; class precision produces a percentage of 100.00%, while class recall produces a percentage of 98.95%. The model can be applied to see the sentiment of Twitter users towards elections in positive and negative forms.

Acknowledgments: Thanks to the Faculty of Engineering, Universitas Muhammadiyah Bengkulu, Bengkulu City, Indonesia, for providing all the resources I used to complete this research. Hopefully, this research will continue to be developed and can be a good reference in the future.

Author contributions: The author is responsible for Conceptualization, Methodology, analysis, investigation, data curation, writing—original draft preparation, writing—review and editing, visualization, supervision of project administration, and funding acquisition. The author has read and agreed to the published version of the manuscript.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Rasyadi, Muhammad Hadiyan., and Khotimah, Husnul. (2017). *Analisis Sentimen pada Twitter Menggunakan Metode Naïve Bayes (Studi Kasus Pemilihan Gubernur DKI Jakarta 2017).* Scientific Repository IPB University, http://repository.ipb.ac.id:8080/handle/123456789/90127
2. Muthia, Dinda Ayu. (2017). *Analisis Sentimen Pada Review Restoran Dengan Teks Bahasa Indonesia Mengunakan Algoritma Naive Bayes*. J.T.I.K, Jurnal Ilmu Pengetahuan dan Teknologi Komputer, Vol 2 No 2, pp. 39-45, E-ISSN: 2527-4864
3. Ruhyana, Nanang. (2019). *Analisis Sentimen Terhadap Penerapan Sistem Plat Nomor Ganjil/Genap Pada Twitter Dengan Metode Klasifikasi Naive Bayes*. Ikraith-Informatika, Universitas Persada Indonesia Y.A.I, Vol. 3 No. 1, ISSN 2580-4316
4. *Schrape JF. (2016). Social Media, Mass Media and the Public Sphere*. Differentiation, Complementarity and Co-Existence. [diunduh 2017 Agu 29]. Tersedia pada: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2858891
5. Sidik, F., Suhada, I., Anwar, A. H., & Hasan, F. N. (2022). *Analisis Sentimen Terhadap Pembelajaran Daring Dengan Algoritma Naive Bayes Classifier*. Jurnal Linguistik Komputasional, 5(1), 34-43.
6. Yulita, Winda. (2021). *Analisis sentimen terhadap opini masyarakat tentang vaksin covid-19 menggunakan algoritma naïve bayes classifier*. JDMSI, Jurnal Data Mining dan Sistem Informasi, Vol 2, No 2. E-ISSN : 2775-9660
7. F. I. Maulana, P. D. P. Adi, D. Lestari, A. Purnomo and S. Y. Prihatin. (2022). *Twitter Data Sentiment Analysis of COVID-19 Vaccination using Machine Learning*. 2022 5th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), Yogyakarta, Indonesia , pp. 582-587, doi: 10.1109/ISRITI56927.2022.10053035.

8. S. Tiwari, A. Verma, P. Garg and D. Bansal. (2020). *Social Media Sentiment Analysis On Twitter Datasets*. 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, pp. 925-927, doi: 10.1109/ICACCS48705.2020.9074208.
9. S. A. El Rahman, F. A. AlOtaibi and W. A. AlShehri. (2019). *Sentiment Analysis of Twitter Data*. 2019 International Conference on Computer and Information Sciences (ICCIS), Sakaka, Saudi Arabia, pp. 1-4, doi: 10.1109/ICCISci.2019.8716464.
10. L. Wang, J. Niu and S. Yu. (2020). *SentiDiff: Combining Textual Information and Sentiment Diffusion Patterns for Twitter Sentiment Analysis*.IEEE Transactions on Knowledge and Data Engineering, vol. 32, no. 10, pp. 2026-2039, doi: 10.1109/TKDE.2019.2913641
11. R. Patil, N. Gada and K. Gala. (2019). *Twitter Data Visualization and Sentiment Analysis of Article 370*. 2019 International Conference on Advances in Computing, Communication and Control (ICAC3), Mumbai, India, pp. 1-4, doi: 10.1109/ICAC347590.2019.9036800.
12. A. Sholihat, F. Bei, R. Ainaya, F. Sembiring and A. Lattu, (2022). *Twitter Tweet: Sentiment Analysis on Illegal Investment using Naive Bayes Algorithm*. 2022 IEEE 8th International Conference on Computing, Engineering and Design (ICCED), Sukabumi, Indonesia, pp. 1-5, doi: 10.1109/ICCED56140.2022.10010383.
13. V. Ramanathan, H. Al Hajri and A. Ruth. (2024). *Conceptual Level Semantic Sentiment Analysis Using Twitter Data*. 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), Chennai, India, pp. 1-8, doi: 10.1109/ADICS58448.2024.10533498.
14. L. Sandra and F. Lumbangaol. (2021). *When Homecoming is not Coming: 2021 Homecoming Ban Sentiment Analysis on Twitter Data Using Support Vector Machine Algorithm*. 2021 International Conference on ICT for Smart Society (ICISS), Bandung, Indonesia, pp. 1-5, doi: 10.1109/ICISS53185.2021.9533255.
15. R. A. Rahman, F. Haziqah Mohamad Zaini, M. S. Nizam Mohd Danuri and A. Amin. (2022). *The Sentiment Analysis on Mental Health Awareness by Non-Governmental Organisation's Twitter*. 2022 International Visualization, Informatics and Technology Conference (IVIT), Kuala Lumpur, Malaysia, pp. 185-190, doi: 10.1109/IVIT55443.2022.10033345.
16. V. Singh, H. V. Kaushik and Reshma. (2024). *Social Media Sentiment Analysis Using Twitter Dataset*. 2024 Second International Conference on Data Science and Information System (ICDSIS), Hassan, India, pp. 1-5, doi: 10.1109/ICDSIS61070.2024.10594648.
17. B. Siswanto, F. L. Gaol, B. Soewito and H. L. H. S. Warnars. (2021). *Sentiment Analysis of Big Cities on The Island of Java in Indonesia from Twitter Data as A Recommender System*. 2021 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS, Jakarta, Indonesia, pp. 124-128, doi: 10.1109/ICIMCIS53775.2021.9699147.
18. A. Madan and U. Ghose. (2021). *Sentiment Analysis for Twitter Data in the Hindi Language*. 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, pp. 784-789, doi: 10.1109/Confluence51648.2021.9377142.
19. Pablo A. H. and Alessandri.F. (2024). *Analyzing digital societal interactions and sentiment classification in Twitter (X) during critical events in Chile*. Heliyon, Volume 10, Issue 12, doi. 10.1016/j.heliyon.2024.e32572
20. Ba Alawi, A., and Bozkurt, F.(2024). *A hybrid machine learning model for sentiment analysis and satisfaction assessment with Turkish universities using Twitter data*. Decision Analytics Journal, Volume 11, doi.10.1016/j.dajour.2024.100473