Jurnal Informasi dan Teknologi
https://jidt.org/jidt
2024 Vol. 6 No. 1 Page: 176-181 e-ISSN: 2714-9730

Application of Naive Bayes Classifier Method to Analyze Social Media User Sentiment Towards the Presidential Election Phase

Firdaus Yuni Dharta[1]✉, Ardhana Januar Mahardhani[2], Sitti Rachmawati Yahya[3], Andika Dirsa[4], Elvira M. Usulu[5]

[1]Universitas Singaperbangsa Karawang
[2]Universitas Muhammadiyah Ponorogo
[3]Universitas Siber Asia
[4]UIN Imam Bonjol Padang
[5]Universitas Yapis Papua

firdaus.yunidharta@fisip.unsika.ac.id

Abstract

This research aims to analyze the sentiment of social media users towards the election. The author collected data in this research through a literature study and observation. The author uses a classification method with the Naïve Bayes Classifier Algorithm and Support Vector Machine to analyze sentiment results. Next, this research extracts word assessment features using TextBlob, which changes text into positive or negative classes. Based on the research results, after going through the text preprocessing stage of more than 15,000 tweets, 11,000 clean tweets were obtained, which were then labelled using the text blob library in Python. The labelling results show that 4,000 tweets are positive, and the rest are harmful, indicating that most social media users' sentiment towards the election is positive. Words that often appear in the positive class express support and confidence in implementing elections that are considered honest and fair. On the other hand, words in the negative class reflect negative sentiment towards implementing elections, which are considered unsuccessful and time-consuming. The Naïve Bayes method provides accuracy, precision, and recall values of 85%, 80%, and 75%. In the Support Vector Machine method, testing is carried out with three kernels (linear, RBF, and poly), where the poly kernel with the best parameter values C is ten and degree is 1 produces the highest accuracy, precision, and recall of 90%, 90%, and 85%, respectively.

Keywords: Social Media, Election, Naïve Bayes Classifier, Accuracy.

1. Introduction

Rapid technological developments have penetrated various aspects of life, including the economic, health, social, and political fields. Social media applications are a form of significant change in the way humans interact and communicate [1]. In Indonesia, there has been a substantial increase in the use of social media and smart devices, such as smartphones. Most Indonesians use social media to communicate with relatives, family, and others. Social media is a platform for socializing and for people to express their views and opinions on current issues [2]. Social media has evolved from its original purpose of social interaction to become a multifunctional tool, thanks to its ability to spread information quickly and without limits [3]. The function of social media extends from branding, product buying, and selling transactions to being a place for discussion. Social media can quickly disseminate information to various corners of the world due to technological advancements [4]. This significantly impacts information sharing and allows social media to become a means for multiple purposes, including contributing to the country's development. People can freely express their aspirations and opinions, making social media a solution to involve the community in developing the country [5].

Social media users generally actively discuss various things, issues, news, and current topics in the spotlight. One issue currently receiving widespread attention is the implementation of general elections, which have received various positive and negative responses from netizens. Elections in Indonesia are clear evidence of the implementing of a democratic system where citizens can elect their representatives every five years [6]. The importance of community participation in general elections is an effort to strengthen the democratic process at the level of society as a whole. General elections in Indonesia are a crucial part of the democratization process. The extent to which people actively participate in the general election process measures the success of democracy [7]. However, data from the General Election Commission shows that voter participation in Indonesia has decreased in each post-reform election period. One of the factors causing the decline in community participation, especially in urban areas, is the lack of information they receive about politics. This has resulted in apathy towards politics among the public, which then affects their desire to participate in elections [8].

Receipt: 23-06-2023 | Revision: 15-12-2024 | Publish: 04-02-2024 | doi: 10.60083/jidt.v6i1.494

Furthermore, negative perceptions of politics and politicians contribute to low voter participation, particularly among teenagers or young voters. A lack of positive knowledge and perception of the world of politics causes some people to be reluctant to be actively involved in the general election process [9].

Responding to the low level of public participation in general elections, this research aims to understand patterns of public perception regarding elections through social media. The main focus of this research is sentiment analysis utilizing the Naive Bayes Classifier algorithm and Support Vector Machine. Hopefully, the findings will aid in attempts to raise the voter turnout rate in the upcoming two years. To address current issues, we used various techniques in our study. It was decided to adopt the Naïve Bayes Classifier approach due to its advantages in accuracy, speed, and simplicity [10]. Naive Bayes has also been utilized extensively for sentiment classification in research. Scholars have previously investigated using the Naive Bayes and K-Nearest Neighbor algorithms to classify recipients of non-cash welfare cards. They discovered that Naive Bayes outperforms K-Nearest Neighbor in terms of accuracy [11]. The results of this research can provide valuable insight and information for interested parties to develop effective strategies and efforts to increase public participation in future general elections [12]. Naïve Bayes, neural networks, and C are some algorithms that can be used to predict creditworthiness. Naïve Bayes has proven more accurate than other algorithms in this context [13][14] [15] [16]. This statistical classification method is not only effective in predicting the possibility of creditworthiness but is also able to forecast data well, identify trend patterns, and provide forecasts for the next month. In addition, the Support Vector Machine is also a powerful classification method in the context of this research [17][18][19]. SVM uses hyperplanes to perform classification, making it easier to group positive and negative opinions [20][21][22].

2. Research Methods

In this research, the author collected data and information related to public sentiment towards the election. Literature studies and observations support the data collection process. The author utilizes a classification method with the Naïve Bayes Classifier Algorithm and Support Vector Machine to examine sentiment data. Data preprocessing was done before the analysis, which included several procedures to clean and prepare the data, including case folding, cleaning, tokenization, stopword removal, normalization, and stemming. Next, this study uses TextBlob to extract word assessment features by classifying text into positive and negative groups. This technique calculates a value for each word in a tweet. Next, a word cloud is used to visualize the labelled data set and identify the most frequently occurring words in each class. The data set is split into two categories: training and test data. After that, the training data is processed to perform modelling and decide sentiment classification on the test data using the Naïve Bayes Classifier and Support Vector Machine techniques. The outcomes of the two approaches are then compared using an f-score, accuracy, precision, recall, and confusion matrix. An insight into the model's efficacy in sentiment prediction on test data is provided by this evaluation.

3. Results and Discussion

The data crawling process begins with the first step, registering an account as a social media developer via the official website. Users that successfully register are granted access to the API and four unique codes: the access token, access token secret, access token key, and API secret for a single application. This authentication stage allows users to retrieve data sources from social media using various programming languages and applications following applicable regulations. After successful authentication, the next step is to retrieve data via the API, but with a time limit where standard users can only access data a maximum of one week back with a limit of 10,000 tweets per week. Data collection focuses on tweets containing election keywords, except retweets, to focus attention on the original data. The final step in the crawling process is saving the captured data into a CSV file. Once the data has been successfully obtained and saved, the next stage involves carrying out text processing to reduce noise and prepare the data for easier processing in the subsequent stage. This process ensures that the data captured can provide a clean and structured basis for further analysis regarding public sentiment towards the election.

After obtaining the data set from the crawling results, the initial process in the text preprocessing stage is called case folding. At this stage, we change all letter characters in the data set to lowercase to facilitate analysis and maintain consistency in data processing. The next step is to clean the data set from components that are irrelevant and have no meaning, such as mentions, links, hashtags, URLs, punctuation, and whitespace. This aims to make the data cleaner and more focused, allowing for more accurate analysis. The next stage in text preprocessing is tokenization, where sentences are broken down into chunks of words or tokens. We carry out the tokenization process to understand the occurrence of words in each tweet separately and facilitate further analysis of word patterns. After tokenizing the tweets, we perform a normalization step to standardize words with similar meanings. Normalization attempts to standardize their writing by providing abbreviated or non-standard words with a consistent meaning in sentiment analysis. Normalization ensures that the data used in the research has a uniform format and can be processed more effectively in subsequent steps.

Stopword elimination is a text preparation step that eliminates words often used but doesn't significantly affect a sentence's sentiment. The stopword method was conducted for this study using a stopword corpus from the Sastrawi library. The use of stop words helps focus analysis on words that are more meaningful in the context of sentiment towards the election. Next, we carry out the stemming stage to obtain the essential words for each word in the dataset. This process is carried out by removing affixes at the beginning, insertions, suffixes, and combinations of affixes from a word. The Python programming language's Sastrawi package facilitates the stemming process, making the dataset's words more uniform and manageable. Then, we carry out the translation process as a feature extraction stage using the TextBlob library. The TextBlob library processes data textually and expresses positive or negative emotions. It is important to note that TextBlob, by default, can only recognize English. This process helps assess public sentiment towards the election through feature extraction from data sets that have gone through previous stages in text preprocessing.

Labelling and word weighting are essential steps in the feature extraction stage to assign a value or weight to every word in every tweet that has passed the preprocessing stage. The Python TextBlob library processes data textually and expresses positive or negative emotions. Despite being designed to recognize English by default, TextBlob is applied to Indonesian language data sets in this research. Each tweet's resulting TextBlob object processes natural language learning and provides sentiment labels automatically. The advantage of this method lies in its ability to offer sentiment labels quickly, allowing the processing of extensive data sets to be carried out efficiently. This automation process also helps avoid distortions from personal opinions, providing objectivity in sentiment labelling. By applying the TextBlob process script, this research can extract sentiment features from each tweet more efficiently, providing a solid basis for further analysis regarding the perceptions and views of social media users regarding the election.

TextBlob labelling results indicate that 3,500 tweets exhibited positive sentiment, while 1,500 tweets displayed negative sentiment. The analysis of user data revealed that most social media users in that period expressed positive sentiments towards the election. After classifying the data set into two classes, positive and negative, we analyzed it to determine the frequency of the most frequently appearing words. Visualizing the analysis results in a word cloud effectively summarizes essential information from large data sets. Wordcloud provides a clear picture of the dominant keywords in positive and negative sentiment related to the election. This approach provides information about the distribution of sentiment and identifies the most significant keywords. Thus, this analysis offers more profound insight into people's views and sentiments towards the elections. Using TextBlob as a tool for sentiment labelling, word frequency analysis, and word cloud visualization helps understand the dynamics of public opinion more comprehensively based on processed data.

The analysis of word frequency and the word cloud visualization of the positive class yield several relevant conclusions about the positive sentiments expressed by the public. In this context, the word "support" indicates assistance or support provided to a particular candidate or political party, indicating positive solidarity from the community towards that political entity. Furthermore, the word "people" refers to citizens as voters in elections, reflecting society's awareness and active role in the democratic process. The abbreviation "RI" for the Republic of Indonesia recognizes and identifies Indonesia as where elections are held. The word "win" reflects optimistic hope and belief in the potential success of the candidate or political party that received support in the election. The "serial number" has become necessary as an identifier for a political party or candidate, highlighting public awareness of the significance of this identification in voting. Furthermore, "honest" and "fair" reflect the people's aspirations and hopes for implementing quality elections, where integrity and fair treatment are prioritized. Overall, this analysis provides a deeper understanding of the aspects that the public considers essential in responding to and supporting the election process, creating a positive picture regarding the implementation of elections and active participation in democracy.

Overall, the analysis results in the positive class illustrate strong support and confidence in implementing elections which are considered honest and fair. The public firmly voiced their hopes for the victory of the candidate or political party that received support. However, from the results of this analysis, several aspects of negative sentiment were identified, which provide further insight into opposing views. First, the word "postpone" reflects dissatisfaction with the scheduling or implementation of elections, which may be considered not following people's expectations or needs. Then, the word "failed" describes a negative view of the election process or results deemed unsatisfactory by some people. Finally, the word "don't" expresses rejection or a suggestion that a specific action or policy not be carried out, indicating disagreement with some aspect of the election. Overall, words in the negative class show people's dissatisfaction or disapproval of various aspects of elections, including scheduling, results, or specific policies. This analysis provides deep insight into variations in sentiment among the public, which can be the basis for improving and perfecting the election process in the future.

This study uses the Naïve Bayes approach and shares test data among multiple scenarios to evaluate the model's effectiveness. Next, we used the Naïve Bayes approach to handle the test and train data, and we ran the Naïve Bayes model's Python program seven times. The outcomes of the experiment demonstrate that accuracy varies in different data-sharing situations. For instance, the accuracy achieves 85% when the data set split is 90%:10%; it reaches 85% when the data set split is 80%:20%, and it comes to 85% when the data set split is 70%:30%. The fact that this gap exists indicates how sensitive the model is to data sharing. At a 90:10 data set ratio, the model's accuracy was at its best. Notably, the confusion matrix shows 1000 accurate predictions for positive sentiment and 300 correct predictions for negative sentiment, with a 70% to 30% ratio. We performed cross-validation to guarantee maximum accuracy after determining the ratio with the best accuracy value. The first fold's top values of accuracy, precision, and recall were 90%, 90%, and 85% when the 10-fold cross-validation method was applied. This process confirms that the Naïve Bayes model can provide good results in classifying sentiment towards the election. Optimal accuracy, especially after cross-validation, shows the model's reliability in dealing with data variations and increases confidence in interpreting election sentiment.

This study used grid search cross-validation and three kernels, linear, RBF, and poly, in the Support Vector Machine technique to determine the optimal parameter values. The researchers ran the Support Vector Machine Model's Python software seven times. According to the researchers, the optimal C parameter value for the linear kernel was 1, and its recall, accuracy, and precision scores were 90%, 90%, and 85%. The optimal values for the C and gamma parameters in the RBF kernel are 10 and 0.01 90%, respectively, yielding the highest accuracy, precision, and recall. In the meantime, the poly kernel produced 90%, 90%, and 90% outcomes for accuracy, precision, and recall, respectively, with a best C parameter value of 10 and a best degree parameter value of 1. 90% accuracy was the best result obtained by the poly kernel. After that, we used the poly kernel to carry out several ratio divisions. The findings indicate that an accuracy of 90% is achieved when the data set is split 90%:10%, 90% when the data set is split 80%:20%, and 90% when the data set is split 70%:30%. With an accuracy of 90%, the poly kernel performed best at a data set ratio of 70:30. There are 1000 accurate predictions for positive sentiment and 350 correct predictions for negative sentiment in the confusion matrix derived from the SVM algorithm at a 70%:30% ratio. After determining the ratio with the highest average accuracy, precision, and recall, we carried out cross-validation to attain maximum accuracy, precision, and recall. The 10-fold cross-validation approach yielded the best results in this study regarding accuracy, precision, and recall at the tenth fold. This demonstrates that the Support Vector Machine model performs optimally in identifying sentiment toward the election during cross-validation, mainly when using the best poly kernel and parameters.

The initial data crawling process revealed 15,000 raw data points, which went through the text preprocessing stage to be processed into 11,000 clean data points. The following process involved labelling the 5,000 data sets using TextBlob. The daily trend graphic visualization depicts the changes in the number of tweets per day during the data collection period. Based on the labelling results, there were 1,000 positive and 300 negative tweets, providing a beneficial picture of the trend in public sentiment towards the election during the observation period. This analysis is essential for understanding the dynamics of public opinion and can be the basis for better decision-making in the election context. The Naive Bayes approach had an accuracy of 85% according to the classification test results, whereas the Support Vector Machine method had a greater accuracy of 90%. We preprocessed 15,000 tweets in preparation for testing to remove sentiment-less characters from the data set. Next, we used the Python TextBlob package to label the data collection into positive and negative classes. When the classification approach, which used poly as the ideal kernel, was applied to the data set, it produced instead a decent accuracy, 85% for the Naive Bayes method and 90% for the Support Vector Machine method. These results show that the Support Vector Machine method, in the context of sentiment classification on election tweet data, performs better than the Naive Bayes method. This can provide important insights regarding model evaluation and selection in sentiment analysis on big data such as election tweets.

4. Conclusion

After going through the text preparation stage of 15,000 tweets, 11,000 clean tweets were received, which were then tagged using the text blob package in Python. These findings are based on the discussion outcomes. The labelling results indicate that the positive class includes 4,000 tweets, and the negative class consists of 1,500 tweets, suggesting that most social media users' sentiment towards the election is positive. Visualization as a word cloud of words in each positive and negative class illustrates essential aspects. Words that frequently appear in the positive class, such as "support," "win," "RI," "people," and "honest," express support and confidence in the implementation of elections that are considered honest and fair.

On the other hand, words in the negative class, such as "no," "postponed," and "failed," reflect negative sentiment towards the implementation of elections, which are considered to be unsuccessful and time-consuming. The Naïve Bayes method provides accuracy, precision, and recall values of 85%, 80%, and 80%. Three kernels are tested in the Support Vector Machine method: poly, RBF, and linear. The poly kernel with the best parameter values (C = 10 and degree = 1) yields the highest accuracy, precision, and recall results, which are 90%, 90%, and 85%, respectively. It is therefore intended that this research will be helpful and contribute to the electoral context, as well as serve as a reference in the natural language processing field in general and sentiment analysis in particular, employing the Naïve Bayes and Support Vector Machine techniques in particular.

Based on the findings of this research, several suggestions can contribute to future research. First, considering that the data set in this study is limited to Indonesian-language tweets, further research could expand the scope to compare sentiment based on language or analyze the location distribution of social media users. This can provide deeper insight into differences in sentiment across linguistic and geographic contexts. Furthermore, at the preprocessing stage, it is recommended to add Part of Speech information so that adjectives, negation words, and interjections can obtain a polarity that is more appropriate to the context they should be in. The choice of sentiment classification method can also be enhanced by using POS to increase accuracy. In providing label classes to the data set for future research, it is recommended to consider adding positive, neutral, and negative sentiment labels. The addition of this class can provide a more nuanced understanding of social media users' views on the topic under study. Finally, to increase the validity of classification results, future research can consider a comparison with other classification methods or carry out different feature selections. Thus, future research can further explore which method or feature best classifies sentiment in tweet data. These suggestions will serve as a helpful guide for developing further research in sentiment analysis on social media platforms.

References

[1] Aulia, G. N., & Patriya, E, "Implementation of Lexicon Based and Naive Bayes in Twitter User Sentiment Analysis on the 2019 Presidential Election Topic," *Jurnal Ilmiah Informatika Komputer*, vol. 24, no. 2, pp. 140–153, 2019.

[2] Mahardhani, A. J. (2023). The Role of Public Policy in Fostering Technological Innovation and Sustainability. *Journal of Contemporary Administration and Management (ADMAN)*, *1*(2), 47-53.

[3] Tannady, H., Dewi, C. S., & Gilbert. (2024). Exploring Role of Technology Performance Expectancy, Application Effort Expectancy, Perceived Risk and Perceived Cost On Digital Behavioral Intention of GoFood Users. *Jurnal Informasi Dan Teknologi*, *6*(1), 80-85. https://doi.org/10.60083/jidt.v6i1.477

[4] Madyatmadja, E. D., Marvell, M., Andry, J. F., Tannady, H., & Chakir, A. (2021, August). Implementation of big data in hospital using cluster analytics. In *2021 International Conference on Information Management and Technology (ICIMTech)* (Vol. 1, pp. 496-500). IEEE.

[5] Destari, D., Tannady, H., Zainal, A. G., Nurjanah, S., & Renwarin, J. M. (2021). The Improvement of Employee's Performance in Plastic Ore Industry: Mediating Role of Work Motivation. *Turkish Online Journal of Qualitative Inquiry*, *12*(7).

[6] Chaudhry, H. N., Javed, Y., Kulsoom, F., Mehmood, Z., Khan, Z. I., Shoaib, U., & Janjua, S. H, "Sentiment analysis of before and after elections: Twitter data of Election 2020," *Electronics (Switzerland)*, vol. 10, no. 17, pp. 1–26, 2021.

[7] Diawati, P., Gadzali, S. S., Mahardhani, A. J., Irawan, B., & Ausat, A. M. A. (2023). Analyzing the Dynamics of Human Innovation in Administration. *Jurnal Ekonomi*, *12*(02), 537-540.

[8] Parlika, R., Pradika, S. I., Hakim, A. M., & N M, K. R, "Twitter Sentiment Analysis of Bitcoin and Cryptocurrencies Based on Python Textblob," *Jurnal Ilmiah Teknologi Informasi Dan Robotika*, vol. 2, no. 2, pp. 33–37, 2020, https://doi.org/10.33005/jifti.v2i2.22

[9] Octiva, C. S., Israkwaty, Nuryanto, U. W., Eldo, H., & Tahir, A. (2024). Application of Holt-Winter Exponential Smoothing Method to Design a Drug Inventory Prediction Application in Private Health Units. *Jurnal Informasi Dan Teknologi*, *6*(1), 1-6. https://doi.org/10.60083/jidt.v6i1.464

[10] Hendy, T., Resdiansyah, R., Johanes, F. A., & Rustono, F. M. (2020). Exploring the role of ICT readiness and information sharing on supply chain performance in coronavirus disruptions. *Technol. Rep. Kansai Univ*, *62*, 2581-2588.

[11] Pintoko, B. M., & L., K. M, "Sentiment Analysis of Online Transportation Services on Twitter Using the Naive Bayes Classifier Method," *E-Proceeding of Engineering*, vol. 5, no. 3, pp. 8121–8130, 2018.

[12] Gunawan, F. E., Suyoto, Y. T., & Tannady, H. (2020). Factors affecting job performance of hospital nurses in capital city of Indonesia: Mediating role of organizational citizenship behavior. *Test Engineering and Management*, *83*, 22513-22524.

[13] Rokhman, K. A., Berlilana, B., & Arsi, P, "Comparison of Support Vector Machine and Decision Tree Methods for Sentiment Analysis Review Comments on Online Transportation Applications," *Journal of Information System Management (JOISM)*, vol. 3, no. 1, pp. 1–7, 2021.

[14] Sutrisno, S., Tannady, H., Ekowati, D., MBP, R. L., & Mardani, P. B. (2022). Analisis Peran Kualitas Produk Dan Visual Identity Terhadap Purchase Intention Produk Teh Dalam Kemasan. *Management Studies and Entrepreneurship Journal (MSEJ)*, *3*(6), 4129-4138.

[15] Andry, J. F., Tannady, H., & Nurprihatin, F. (2020, March). Eliciting requirements of order fulfilment in a company. In *IOP Conference Series: Materials Science and Engineering* (Vol. 771, No. 1, p. 012023). IOP Publishing.

[16] Solehati, A., Mustafa, F., Hendrayani, E., Setyawati, K., Kusnadi, I. H., Suyoto, Y. T., & Tannady, H. (2022). Analisis Pengaruh Store Atmosphere dan Service Quality Terhadap Brand Preference (Studi Kasus Pelanggan Gerai Ritel Kopi di DKI Jakarta). *Jurnal Kewarganegaraan*, *6*(2), 5146-5147.

[17] Basrah S & Samsul I, "The Influence of Product Quality and Service Quality on Consumer Satisfaction," *Jurnal Riset Manajemen Sains Indonesia (JRMSI)*, vol.3, no.1, 2022.

[18] D. Abdullah, "Perancangan Sistem Informasi Pelayanan Kapal," *J. Ilm. Teknol. Inf. Terap.*, 2015.

[19] Hasanun, D. Abdullah, and M. Daud, "Pengembangan Sistem E-Learning Politeknik Negeri Lhokseumawe dengan Model Vark ", *jidt*, vol. 5, no. 4, pp. 222-228, Dec. 2023.

[20] A. Faridhatul Ulva, D. Abdullah, Masriadi, Nurhasanah, N. Alimul Haq, and B. Ulumul Haq, "AROS(AgRO-Smart) : Smart City Pertanian dengan Track and Trace GPS berbasis Mobile", *jidt*, vol. 5, no. 4, pp. 78-91, Nov. 2023.

[21] D. . K. Pramudito, A. . Titin Sumarni, E. . Diah Astuti, B. . Aditi, and Magdalena, "The Influence of User Trust and Experience On User Satisfaction Of E-Commerce Applications During Transactions in Mini Markets Using Delon and McLean Method", *jsisfotek*, vol. 5, no. 4, pp. 1–7, Oct. 2023.

[22] S. Budi Utomo, J. P. Nugraha, E. Sri wahyuningsih, R. . Indrapraja, and F. A. . Binsar Kristian Panjaitan, "Analysis of The Effectiveness of Integrated Digital Marketing Communication Strategies in Building MSMEs Brand Awareness Through Social Media", *jsisfotek*, vol. 5, no. 4, pp. 8–13, Oct. 2023.