Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th February 2024. Vol. 3. No. 2; e-ISSN : 2808-4519

Classification of Public Sentiment Toward 2024 Presidential Candidates on Social Media Platform X Using Naïve Bayes Algorithm

Ramdhan Hakiki[1]*, Agung Pambudi[2], Asriyanik[3]

[1,2,3]Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Muhammadiyah Sukabumi
ramdhan047@ummi.ac.id [1], agungpambd@ummi.ac.id [2], asriyanik263@ummi.ac.id [3]

Abstract

This research examines the use of Naïve Bayes algorithm to classify public sentiment on social media X towards Indonesia's 2024 presidential candidates. Against the backdrop of the importance of presidential elections in a democracy, this research focuses on analyzing public sentiment from June to August 2023. The Naïve Bayes method was chosen to process review data about the three main candidates. The classification results provide insight into the positive and negative sentiments of the public, providing benefits for political parties and researchers in understanding public opinion. This research also enhances the understanding of sentiment classification in a political context and provides readers with a useful reference on the Naïve Bayes approach to sentiment classification. In terms of accuracy, the developed naïve bayes model shows a success rate with an accuracy of 74% for Anies Baswedan, 74% for Ganjar, and 88% for Prabowo.

Keywords: *Sentiment Classification, Naïve Bayes, Presidential Candidates*

1. Introduction

General elections have been held many times in Indonesia's history, but direct elections by the Indonesian people began for the first time during the reform period after the collapse of the New Order era, namely in 2004. The 2024 Indonesian presidential and vicepresidential elections, also known as the 2024 presidential election, is a democratic stage to elect the president and vice president of the republic of Indonesia for the 2024-2029 service period, which will be held on Wednesday, February 14, 2024. The presidential election is a very important political event in a country. The election will be the fifth direct presidential and vicepresidential election in Indonesia. According to the 1945 Constitution, the president and vice president are elected in one pair and filled by direct election by the people in article 6A paragraph (1) of the third amendment of the 1945 Constitution "The president and vice president are elected in one pair and filled by direct election by the people" [1].

Looking at public opinion about the 2024 presidential candidate, it is not lost on the attention expressed by the public on social media, one of which is social media platform X related to the presidential candidate in the 2024 presidential election. Twitter, which is now social media platform X, has seen its users in Indonesia grow rapidly in recent years. Social media X is very popular in Indonesia with millions of active users who use social media X to interact, express opinions and also share or search for information with current news and trends [2].

According to [3]sentiment classification is a method to find out the opinion of a person or group of people on certain issues, products, services or groups. Sentiment classification is the process of identifying and categorizing views or feelings represented in text or written language into positive, negative, or neutral sentiment categories.

There are several methods that can be used in performing sentiment classification, including K-Nearest Neighbor (KNN), Support Vector Machine (SVM) and also Naïve Bayes. Previous research on sentiment analysis of online loans on twitter used the Support Vector Machine (SVM) method which resulted in an accuracy of 62% [4]. Another study, namely sentiment analysis of perceptions of the elected government in the 2019 presidential election on twitter using Naïve Bayes, resulted in an accuracy value of 81% [5]. And other research using the K-Nearest Neighbor method with the title "sentiment analysis of twitter users on the Indonesian football polemic using TF-IDF weighting and K-Nearest Neighbor" with an accuracy value of 79.99% [6].

With the comparison of methods for sentiment classification above, researchers will use the Naïve Bayes algorithm because it obtained the largest accuracy value with 81%. According to [7], Naïve Bayes is a classification method that is widely used in processing data mining or text mining. By applying the Naïve Bayes technique to sentiment classification with case studies of opinions or reviews of the 2024 presidential candidates on social media.

2. Research methods

This research method aims to describe the stages of the method and research design based on data collection or information obtained, as well as the research tools to be used and the schedule of research activities to be carried out. The implementation of this research also aims to analyze public sentiment towards presidential candidates on the X- social media platform using the Naïve Bayes algorithm. The stages of this research method are as follows:

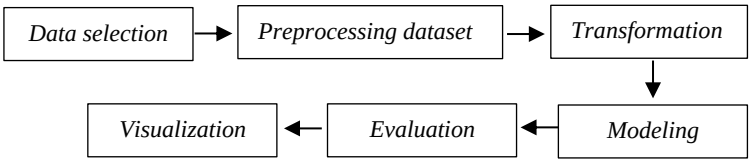

Fig. 1: Stages of Sentiment Classification in research

In this research method, the Knowledge Discovery in Database (KDD) method is used by researchers to apply the Naïve Bayes algorithm to process and analyze tweet data. This method is used with the aim that researchers can extract valuable information hidden in the review data set so that previously unknown, useful and potential knowledge can be found for future decision making. Based on the previous Knowledge Discover In Database (KDD) explanation, here are the steps for the KDD method:

2.1. Data selection

Data Selection is the initial stage of this research which will select and take reviews from social media user X-'s tweets regarding the 2024 presidential candidates and the process is collected using data crawling techniques via the Twitter API using Google Colaboratory. Data collected on Presidential Candidate Prabowo Subianto amounted to 1000 tweets, Ganjar Pranowo 1000 tweets and Anies Baswedan 1000 tweets from June 2023 to August 2023 which will then be processed again according to the sentiment classification stage.

2.2. Pre-processing

The pre-processing dataset stage is the stage of processing the dataset which was originally in the form of raw data or dirty data which will be processed into data according to needs for further classification because in this preprocessing there are many stages that will be carried out to process the data, namely cleaning, casefolding, tokenizing, filtering, slang words replacement and stemming. The aim of this stage is also to produce quality and relevant data to carry out the desired sentiment classification process.

2.3. Transformation

The transformation stage is part of the KDD process which aims to change or modify review data so that it is easier to process and analyze. This stage is usually carried out after the dataset preprocessing stage and can include various processes, such as word weighting using Term Frequency inverse document frequency (TF-IDF). By carrying out transformations, tweet data will be easier to process and analyze so that the analysis results are more accurate and useful.

2.4. Modeling

The modeling stage is a stage in the Knowledge Discovery in Database (KDD) process which aims to build a model that is able to apply patterns or relationships to review data. In this step, the training data that has been previously processed and transformed will be used to build the model. Many algorithms can be used in the modeling stage, one of which is Naïve Bayes.

2.5. Evaluazation

The evaluation stage in Knowledge Discovery in Database (KDD) is the stage of assessing the results of the model that has been built at the modeling stage. In this step the model has been built and will be applied to the test data that has been prepared previously. The results of applying the model will be analyzed and compared with the expected results, so that the level of accuracy of the model can be known. There are several metrics that can be used in evaluation, such as accuracy, precision, recall, and f1-score. Apart from that, the confusion matrix can also be used to find out how well the model can predict actual results. It is important to carry out this evaluation stage in order to find out to what extent the model being built can capture the patterns or relationships contained in the tweet data.

2.6. Visualization

The visualization stage in Knowledge Discovery in Database (KDD) is the final stage of the KDD process which aims to display the results of the model that has been built in the previous stage in a visual form that is easy for other people to understand. After the tweet data has been successfully processed in the previous step, the next step is to display it in the form of a diagram or graph, making it easier to interpret and analyze the tweet data. Some examples of visualizations that can be used in KDD are scatter plots, bar charts or wordclouds. It is important to carry out this visualization stage so that it can make it easier to make decisions based on the results of the model that has been created.

3. Result and discussion

3.1. Data selection

Review data was obtained from social media platform X using search keywords such as #capres2024, #prabowo, #anies, and #ganjar. The data is in the form of tweets in Indonesian that were crawled from social media X, totaling 1000 tweets from each presidential candidate. After that, the data will be saved in CSV format.

3.2. Preprocessing dataset

Next is the pre-processing stage, which is the process of preparing data before it is applied to machine learning models or algorithms. The goal is to clean, format, and prepare the raw data so that it can be analyzed in modeling. The following are the stages of preprocessing in this research.

3.2.1. Cleaning

The first step is to clean the dataset to remove symbols, emoticons, or noise words in the tweet data.

Table 1: Cleaning

Appearance	
Input	**Output**
Semangat pak…semoga bisa memimpin negeri ini	Semangat pak semoga bisa memimpin negeri ini

3.2.2. Case folding

Case folding is a step to replace or change the entire case of the text to lowercase because in the dataset of tweets of 2024 presidential candidates there are uppercase and lowercase letters. Therefore, a case folding step is needed so that all words used are lowercase.

Table 2: Case folding

Appearance	
Input	**Output**
Semangat pak semoga bisa memimpin negeri ini	semangat pak semoga bisa memimpin negeri ini

3.2.3. Tokenizing

At this stage, tokenization is used to break sentences into words, on a space-separator basis.

Table 3: Tokenizing

Appearance	
Input	**Output**
‘semangat’ ‘semoga’ ‘bisa’ ‘memimpin’ ‘negeri’	‘semangat’ ‘pak’ ‘semoga’ ‘bisa’ ‘memimpin’ ‘negeri’ ‘ini’

3.2.4. Slang words replacement

The process of consistent terms that have the same meaning but are written differently may be necessary, due to misspellings, use of abbreviations, or informal language.

Table 4: Slang words replacement

Appearance	
Input	**Output**
krg	kurang

3.2.5. Stopword

The next step is to implement stopwords, which aims to remove words that have no meaning (stoplist) or have no effect on accuracy in the classification process, such as linking words. The results of the stopword process can be checked in the following table.

Table 5: Stopword

Appearance	
Input	**Output**
‘semangat’ ‘pak’ ‘semoga’ ‘bisa’ ‘memimpin’ ‘negeri’ ‘ini’	‘semangat’ ‘semoga’ ‘bisa’ ‘memimpin’ ‘negeri’

3.2.6. Stemming

Stemming in this pre-processing step is done to convert words into their base word form by removing the affixes attached to the word.

Table 6: Stemming

Appearance	
Input	**Output**

'semangat' 'semoga' 'memimpin' 'negeri'	'semangat' 'semoga' 'pimpin' 'negeri'

3.3. Tf-Idf

In this research, the TF-IDF technique, which is a term weighting method, is applied to assign a value or weight to each word in visitor reviews. This facilitates the classification process. Initially, the word frequency in a document is calculated, followed by the calculation of the inverse document frequency. The results of the Term Frequency and Inverse Document Frequency calculations can be seen in Table 7.

Table 7: The result of calculating TF and calculating IDF

Appearance		
Term	Tf	Idf
semangat	1	0.90308
semoga	1	0.90308
pimpin	1	0.90308
negeri	1	0.90308

3.4. Implementation of naive bayes

The implementation of classification with the Naïve Bayes method is carried out using data that has been processed through the text preprocessing stage and has been weighted using the TF-IDF method. The implementation of performing the classification stage using naïve bayes consists of 2 stages as follows:

3.4.1. Naïve Bayes modeling

The first step in performing classification is modeling using the Naïve Bayes method. This process begins with the creation of a Naïve Bayes model using the sklearn library in the Python programming language, specifically using the multinomial Naïve Bayes method. Afterwards, the classification model is formulated using the fit function to train the model using the training data. The following is the modeling program code:

```
from sklearn.naive_bayes import MultinomialNB
#pembentukan model Naïve Bayes
clf=MultinomialNB()
model = clf.fit(vec_latih,y_train)
```

Fig. 2: Program code for naïve bayes modeling

3.4.2. Test data prediction

The data set previously separated by 20%, referred to as x_test in the implementation code, will be used to test the data. This testing process is carried out using the Naïve Bayes model that has been created previously. The following is the program code to test the data:

```
#term frequency(pembobotan kata) data uji
vec_uji=vec.transform(x_test)
vec_uji
#klasifikasi data uji
predicted=clf.predict(vec_uji)

data_uji.insert(2, column='label_bayes',
value=predicted)
data_uji
```

Fig. 3: Program code for data testing

After passing the data testing stage using the test data as described earlier, the results can be seen in the "hasil_stemming" column including the test data, while "label" refers to the initial label of the test data. Meanwhile, the "label_bayes" column displays the classification results of the data that has been tested using the prepared model. The result display is as follows:

	hasil_stemming	label	label_bayes
924	['kritik', 'keras']	Negatif	Negatif
526	['usung', 'gelora', 'kini', 'abis', 'pilpres',...	Positif	Positif
568	['tolak', 'lupa', 'jahat']	Negatif	Negatif
659	['rawan', 'sambut', 'kunjung']	Negatif	Positif
635	['sambut', 'hangat', 'peci', 'pemuda', 'acara'...	Positif	Positif
...	...	...	...

Fig. 4: Data testing results using naïve bayes classifier

3.5. Evaluation

The performance evaluation of the program model in this study involves the use of confusion matrix. In this evaluation stage, the goal is to assess the performance of the algorithm used by calculating the accuracy, recall, precision, and f1-score values. The implementation of evaluation on test data to measure the performance of the Naïve Bayes algorithm model that has been developed is implemented using the sklearn metrics and seaborn libraries. The data results for the sentiment classification process using confusion matrix can be found in Table 8.

Table 8: Confusion matrix prabowo subianto

	Positif	**Negatif**
Positif	48	4
Negatif	21	126

Table 9: Confusion matrix anies baswedan

	Positif	**Negatif**
Positif	33	6
Negatif	54	106

Table 10: Confusion matrix ganjar pranowo

	Positif	**Negatif**
Positif	10	0
Negatif	44	145

Next to display the accuracy, precision, recall and f1score values by running the program code as follows:

```
from sklearn.metrics import
accuracy_score,precision_score,recall_score,f1_score
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix

target_names = ['Positif', 'Negatif']
 print('=========================================\n')
print(classification_report(y_test,predicted,
target_names=target_names))
```

Fig. 5: Program code for finding model performance values

The results of the above program code for each dataset can be seen as follows:

Table 11: Performance value for each dataset

Dataset	**Precision**		**Recall**		**F1-score**	
	Positif	**Negatif**	**Postif**	**Negatif**	**Positif**	**Negatif**
Anies baswedan	0.85	0.66	0.38	0.95	0.52	0.78
Ganjar pranowo	1.00	0.77	0.19	1.00	0.31	0.87
Prabowo subianto	0.92	0.86	0.70	0.97	0.79	0.91

4. Conclusion

This research successfully uses the Naïve Bayes algorithm to classify public sentiment towards 2024 presidential candidates on X social media. The results demonstrate the effectiveness of this algorithm in the context of sentiment classification, providing insights to political parties and presidential candidates. Through Knowledge Discovery in Databases (KDD) methodology, the research identified public perceptions of Prabowo (64% positive, 36% negative), Anies Baswedan (50.9% positive, 40.1% negative), and Ganjar (74.2% positive, 25.8% negative). The Naïve Bayes model achieved significant accuracy, namely 74% for Anies Baswedan, 74% for Ganjar, and 88% for Prabowo. In conclusion, this algorithm is useful in understanding public views on political figures, providing a basis for political parties and presidential candidates to design strategies based on social media responses.

References

[1] J. T. Nugraha and UUD, "No Analisis struktur kovarians indikator terkait kesehatan pada lansia yang tinggal di rumah, dengan fokus pada rasa subjektif terhadap kesehatan Title," vol. 105, no. 3, pp. 129–133, 1945, [Online]. Available: https://webcache.googleusercontent.com/search?q=cache:BDsuQOHoCi4J:https://media.neliti.com/media/publications/9138-ID-perlindungan-hukum-terhadap-anak-dari-konten-berbahaya-dalam-media-cetak-dan-ele.pdf+&cd=3&hl=id&ct=clnk&gl=id

[2] Rangga, "No Title," *kompas.id*, 2023. https://www.kompas.id/baca/riset/2023/12/14/media-sosial-pengaruhi-pemilih-pada-pemilu-2024

[3] S. Suryono, E. Utami, and E. T. Luthfi, "Klasifikasi Sentimen Pada Twitter Dengan Naive Bayes Classifier," *Angkasa J. Ilm. Bid. Teknol.*, vol. 10, no. 1, p. 89, 2018, doi: 10.28989/angkasa.v10i1.218.

[4] D. S. Utami and A. Erfina, "Analisis Sentimen Pinjaman Online di Twitter Menggunakan Algoritma Support Vector Machine (SVM)," *SISMATIK (Seminar Nas. Sist. Inf. dan Manaj. Inform.*, vol. 1, no. 1, pp. 299–305, 2021.

[5] F. A. Wenando, R. Hayami, and A. J. Anggrawan, "Analisis Sentimen Pada Pemerintahan Terpilih Pada Pilpres 2019 Ditwitter Menggunakan Algoritme Naïvebayes," *JURTEKSI (Jurnal Teknol. dan Sist. Informasi)*, vol. 7, no. 1, pp. 101–106, 2020, doi: 10.33330/jurteksi.v7i1.851.

[6] J. A. Septian, T. M. Fachrudin, and A. Nugroho, "Analisis Sentimen Pengguna Twitter Terhadap Polemik Persepakbolaan Indonesia Menggunakan Pembobotan TF-IDF dan K-Nearest Neighbor," *J. Intell. Syst. Comput.*, vol. 1, no. 1, pp. 43–49, 2019, doi: 10.52985/insyst.v1i1.36.

[7] R. L. Atimi and Enda Esyudha Pratama, "Implementasi Model Klasifikasi Sentimen Pada Review Produk Lazada Indonesia," *J. Sains dan Inform.*, vol. 8, no. 1, pp. 88–96, 2022, doi: 10.34128/jsi.v8i1.419.