khazanah informatika
Jurnal Ilmu Komputer dan Informatika

Prediction of Presidential Election Results using Sentiment Analysis with Pre and Post Candidate Registration Data

Asno Azzawagama Firdaus[1*], Anton Yudhana[2], Imam Riadi[3]
[1]Master Program of Informatics
Universitas Ahmad Dahlan
Yogyakarta, Indonesia
[2]Departement of Electrical Enginering
Universitas Ahmad Dahlan
Yogyakarta, Indonesia
[3]Departement of Information System
Universitas Ahmad Dahlan
Yogyakarta, Indonesia
*2207048008@webmail.uad.ac.id

Abstract-Social-media is a solution for politicians as a campaign tool because it can save costs compared to conventional campaigns. The 2024 Indonesian presidential election has attracted public attention, especially among social media users. Twitter, as one of the most widely used social media platforms in Indonesia, has become an effective campaign platform. Sentiment analysis is one approach that can be used to measure public opinion on Indonesian presidential candidates based on Twitter data. The data was collected before the declaration of candidates in March 2023 and shortly after the registration of presidential and vice-presidential candidates in November 2023. The data obtained amounted to 15,000 in March 2023 collection and 11,569 in November 2023 collection and used manual labeling by linguists. After removing duplicated tweets, the data changed to 10,569 data with each candidate having 3,523 data for March 2023 and 4,893 data, with each candidate pair having 1,631 data for November 2023. The sentiment analysis classification model is determined using the Naïve Bayes and Support Vector Machine (SVM) methods with Term Frequency-Inverse Document Frequency (TF-IDF) feature extraction. Based on the data, the highest percentage of positive sentiment for the data obtained in March 2023 is for Ganjar Pranowo data by 77.94% and the highest percentage of negative sentiment is for Anies Baswedan data by 31.39%. Meanwhile, for the data obtained in November 2023, the highest positive sentiment was obtained for the candidate pair Ganjar Pranowo - Mahfud MD by 69.16%, and the highest negative sentiment was found in the data Prabowo Subianto - Gibran Rakabuming Raka by 52.12%. Words that frequently appeared in the positive sentiment for Ganjar Pranowo - Mahfud MD included "strong", "corruption", "support", "appreciation", and others. This research achieved the highest accuracy for SVM method which is 86% and Naive Bayes method which is 79%.

Keywords: Indonesia; Naïve Bayes; President; Sentiment Analysis; Support Vector Machine, Twitter

Article info: *submitted January 25, 2023, revised November 16, 2023, accepted December 03, 2023*

1. Introduction

In recent years, the integration of social media data and sentiment analysis has emerged as a powerful tool in understanding public opinion and predicting election outcomes [1], [2]. In the context of Indonesia, a country with a vibrant social media landscape and a significant presence on platforms like Twitter, utilizing sentiment analysis to gauge public sentiment before and after the registration of presidential candidates has become a pertinent research focus.

Long before the scheduled registration of Presidential candidates, social media has been enlivened with talk of the Indonesian Presidential Election. Even the scheduled year for the 2024 Presidential Election, two years earlier, has been widely discussed on social media, especially Twitter. This study aims to explore the potential of sentiment analysis on Twitter data to predict the outcome of the presidential election in Indonesia. By analyzing tweets both before and after the official registration of candidates, the study sought to uncover patterns, trends, and shifts in public opinion that could affect the electoral landscape. The study will utilize natural language processing techniques to analyze the textual content of tweets, categorizing them into positive and negative sentiments.

Understanding the dynamics of public sentiment on social media is crucial for political analysts, policymakers, and candidates themselves. This provides an additional dimension to traditional polling methods, offering a real-time and organic perspective on how the public perceives candidates and their platforms. Next, the study will explore whether there were any notable changes in sentiment after the formalization of candidates, shedding light on the impact of official candidacy on public perception.

As Indonesia continues to navigate its democratic journey, the study contributes to the growing field of election prediction and the use of social media data for political analysis. The findings may not only offer insight into potential election outcomes but also help inform candidates to connect better with voters in the digital age.

Recent research [2] on sentiment analysis used to analyze public opinion on his alignment with the Presidential candidate in the 2023 Nigerian Presidential Election. The alignment of the community was analyzed based on the sentiment obtained on Twitter (X) data with negative, neutral and positive class categories. Similar to the research topic, another study [1] analyzed public sentiment towards candidates for the 2023 Nigerian Presidential Election using LSTM, Linear SVC and BERT methods.

The results of the General Election prediction based on sentiment analysis used using Twitter (X) data for the period January to March 2019 in India correspond to the actual election results available [3]. Other research [4] and the success of presidential campaigns is frequently attributed to SM performance. Within this new scenario, many methodological proposals that use SM data have been put forward for predicting election results. However, the most common approach, based on the volume and sentiment analysis of mentions on Twitter, has been frequently criticized and challenged. Thus, recent surveys have indicated new directions, such as the use of data from more than one SM platform, the adoption of nonlinear machine learning (ML also suggests the chances of predicting the outcome of the Presidential Election based on a social media framework using Machine Learning. The data comes from social media Twitter (X), Facebook and Instagram is more than 65,000 posts. A total of 195 polls for presidential predictions in Latin America such as Argentina (2019), Brazil (2018), Colombia (2018) and Mexico (2018). The results showed that there was a high degree of accuracy in predicting the results of the vote for various candidates as well as providing daily predictions. This is better than traditional polls and can be applied using predictions of upcoming elections to similar scenarios. A similar scenario in another study [5] suggested that social media data can predict election outcomes. Analyze the correlation between social media performance (Facebook, Twitter (X) and Instagram) with the votes obtained in the election. More than 40,000 posts from January to October 2018 in Brazil have been able to provide a strong correlation between the proposed model and the actual votes obtained.

The application of sentiment analysis requires a precise classification method in its case. Support Vector Machine and Naive Bayes are the best methods in terms of sentiment analysis. The Support Vector Machine was chosen because it excels [6] in several studies [7]-[9] in the case of sentiment analysis. Likewise, Naive Bayes which has been widely used in its application [10]-[12] because of its high accuracy when compared to other methods [13].

Finally, this study focuses on the scope of sentiment analysis of Twitter data in the time span before and after candidate registration. The sentiment analysis approach uses the Support Vector Machine and Naive Bayes methods so as to get the best model to develop on the issue of the Presidential Election.

2. Methods

General Election organizers announced the voting schedule for the President and Vice President of Indonesia on February 14, 2024, and the candidates' campaign schedule for November 28, 2023 to February 10, 2024. However, each political party has conducted a political process to determine its own candidates. Even in the period from late 2022 to early 2023, candidates have been nominated to be presidential candidates by each party. This was also validated by survey agencies [14], candidates who have the potential to become presidential candidates are Anies Baswedan, Ganjar Pranowo, and Prabowo Subianto.

When the General Election organizer opens registration for Presidential Candidates and Vice-Presidential Candidates, the three candidates register with their respective Vice-Presidential Candidates. So, this study took data on the time span before candidates officially register and after officially registering with their spouses (Vice Presidential candidates). This was done to see a comparison of the potential alignment of the Twitter user community before and after candidate registration.

Sentiment analysis is a method to measure the results of the presidential election based on social media data. Sometimes some studies only take one time to assess the sentiment that occurs in an object under study. The study compared the two data with the timing of each collection before and after candidate registration. In addition, unbalanced data on each class can result in inappropriate results. Therefore, this study applies the use of stratified K-Fold Cross Validation to overcome this. The framework of this study is shown as in Figure 1.

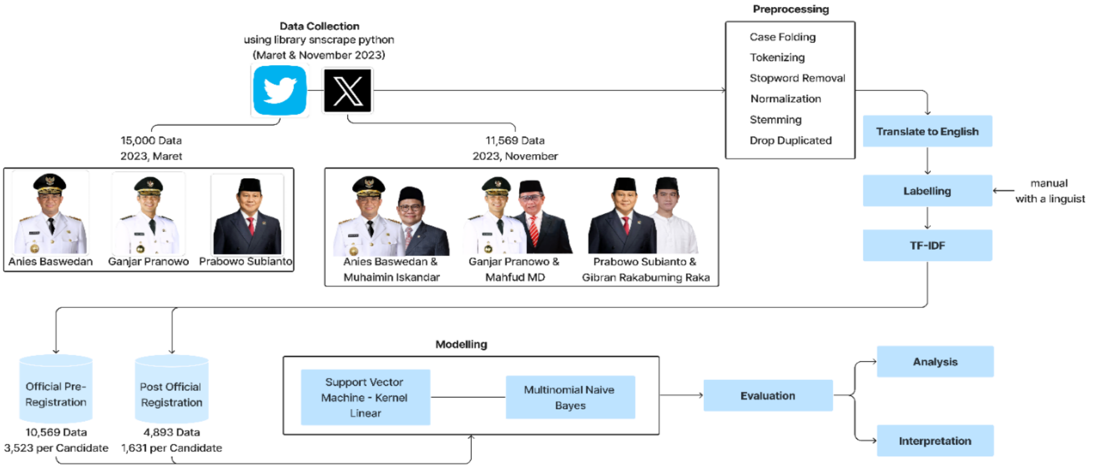


Figure 1. Framework for Twitter based sentiment analysis for Indonesia 2024 election

Figure 1 shows the research framework from data collection to research analysis and interpretation. Data collection will be carried out in March 2023 as many as 15,000 and November 2023 as many as 11,569. The data then goes through the preprocessing stage and is translated into English for manual labeling by experts. Weighting techniques were carried out using TF-IDF to then create classification models using Linear SVM and Multinomial Naive Bayes. Until finally an evaluation of the model is carried out to be further analyzed and given research interpretation.

Data collection is done using SNScrape in python programming. Users are required to have a developer account to retrieve data [15]. The data collected is 15,000 tweets in April 2023. The data collected consists of several Twitter (X) attributes as shown in Table 1.

Table 1. Feature item description Twitter (X)

Item	Description
Tweet Data	Date tweet was posted Twitter (X)
Created Account	The date the user joined Twitter (X)
Username and User ID	Username naming on Twitter (X) and ID
Following and Followers	Number of Accounts Followed and Who Followed
Tweet Count	Number of posts on Twitter (X)
Tweet Location	Name of the location where the tweet was posted
Tweet	The Twitter (X) post

An important part is to clean raw data [16] and reconstruct text into a more processable form for machine learning algorithms [17]. Preprocessing has stages such as tokenization, stop word removal, lower case conversion, stemming, removing number, etc. [18] everyone is expressive in one way or other. Many social websites and android applications whether being Facebook, WhatsApp or Twitter, in this highly advance and the modernized world is flooded with views and data. One of the most global and popular platforms is Twitter. This is seen as the main source of sentiments where almost every enthusiastic or social person tends to express his or her views in form of comments. These comments not only express the people but also give the understanding of their mood. Text present on these medias are unstructured in nature, so to process them firstly we need to pre-process, six pre-processing techniques are used and then features are extracted from the pre-processed data. There are so many feature extraction techniques such as Bag of Words, TF-IDF, word embedding, NLP(Natural Language Processing. This research uses stages such as case folding, tokenizing, stop word removal, normalization and stemming on preprocessing data.

Term Frequency - Inverse Document Frequency (TF-IDF) is a weighting method used to calculate the value of each word in the entire document. The formula for TF-IDF is in Equation 1 – 3 [19].

$$\mathrm{tf}_i = \frac{n_i}{\sum_k n_k} \tag{1}$$

$$\mathrm{idf}_i = log \frac{|D|}{|\{d_j : t_i \in d_j\}|} \tag{2}$$

$$\mathrm{tfidf}_i = \mathrm{tf}_i \mathrm{idf}_i \tag{3}$$

TF indicates the number of occurrences in the corpus (Function 1). The IDF is a measure of the importance of an entire corpus term. It consists of the calculation of the logarithm of the inverse relationship of corpus documents (Function 2). The weight of the TF-IDF is calculated by multiplying the two (Function 3). The greater the weight indicates the more important words are relevant on the corpus [20] a rule-based sentiment analysis lexicon and the Term Frequency-

Inverse Document Frequency weighting method. These three (input.

Naive Bayes is a machine learning algorithm with a probabilistic approach, relying on Bayes' Theorem, and commonly employed for classification purposes [21]. Naive Bayes has the lowest error rate when compared to other classifier algorithms [22]. This algorithm is based on Bayes' theorem formula in the Equation 4

$$(A|B) = \frac{P(B|A)P(A)}{\mathrm{P(B)}} \tag{4}$$

The equation (Function 4) outlines how to compute the probability of hypothesis A given condition B. This formula incorporates the initial class probability P(A) derived from training data, the conditional probability of attribute distribution P(A|B), and the initial probability within each class P(B|A) * P(B).

Support Vector Machine (SVM) is a very effective algorithm in advanced machine learning [23]notably from climate change and, for that purpose, remote sensing is routinely used. However, identifying specific crop types, cropland, and cropping patterns using space-based observations is challenging because different crop types and cropping patterns have similarity spectral signatures. This study applied a methodology to identify cropland and specific crop types, including tobacco, wheat, barley, and gram, as well as the following cropping patterns: wheat-tobacco, wheat-gram, wheat-barley, and wheat-maize, which are common in Gujranwala District, Pakistan, the study region. The methodology consists of combining optical remote sensing images from Sentinel-2 and Landsat-8 with Machine Learning (ML. The goal of the SVM for planning a hyperplane is ideal and is called a decision limit using the distance between the closest samples [24]. In addition, SVM is highly dependent and easier to implement using TF-IDF to calculate weights on documents [25]text summarization process diminishes the redundant information and retrieves the useful and relevant information from a text document to form a compressed and shorter version which is easy to understand and time-saving while reflecting the main idea of the discussed topic within the document. The approaches of automatic text summarization earn a keen interest within the Text Mining and NLP (Natural Language Processing. The kernel function in SVM is a linear segmentation in the feature space for large amounts of undifferentiated data that indirectly affects the performance of SVM classification [26]. SVM Linear has the best accuracy compared to other kernels [27]. The linear SVM test in Equation 5 [28].

$$\mathrm{f(x)} = \mathrm{w.x} + \mathrm{b} \tag{5}$$

w is the weight vector, x is the feature vector, and b is the bias.

Measuring models in machine learning is critical [29] namely Convolutional Neural Network (CNN. Cross validation (K-Fold) is a method for evaluating predictions to training and testing samples. The data partition will perform partial testing and partial training so that it will be repeated for a certain time to determine errors each time [30]. However, there is an unbalanced division of each data in the class so that the solution to the problem is stratified [31]. Stratified K-Fold Cross Validation is a method of collecting data into k-folds to provide a similar proportion of classes in each sample [32]the solution to many practical problems relies on machine learning tools. However, compiling the appropriate training data set for real-world classification problems is challenging because collecting the right amount of data for each class is often difficult or even impossible. In such cases, we can easily face the problem of imbalanced learning. There are many methods in the literature for solving the imbalanced learning problem, so it has become a serious question how to compare the performance of the imbalanced learning methods. Inadequate validation techniques can provide misleading results (e.g., due to data shift. Evaluation methods for classification are used. Accuracy, Precision, Recall, and F1 scores proved useful for confusion matrices [24]. Each formula is shown in equations 6, 7, 8, and 9, respectively.

$$Accuracy = TP + \left(\frac{TN}{TP}\right) FP + FN + TN \tag{1}$$

$$Precision = \frac{TP}{TP} + FP \tag{2}$$

$$Recall = \frac{TP}{\mathrm{TP}} + FN \tag{3}$$

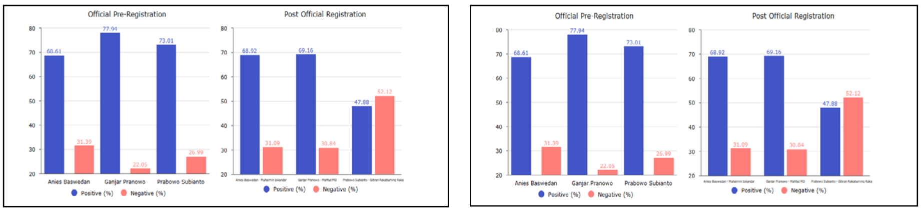
$$F1 = 2\ x\ precision\ x \frac{recall}{\mathrm{precision}} + recall \tag{4}$$

TP shows a positive result, FP shows a false positive result, TN shows a negative result and FN shows a false negative result.

3. Results

The political map in the 2024 Indonesian presidential election is projected using several approaches. This study determines the sentiment and interpretation of research for projected election results based on comparative data before and after official registration of candidates.

a. Sentiment Comparison

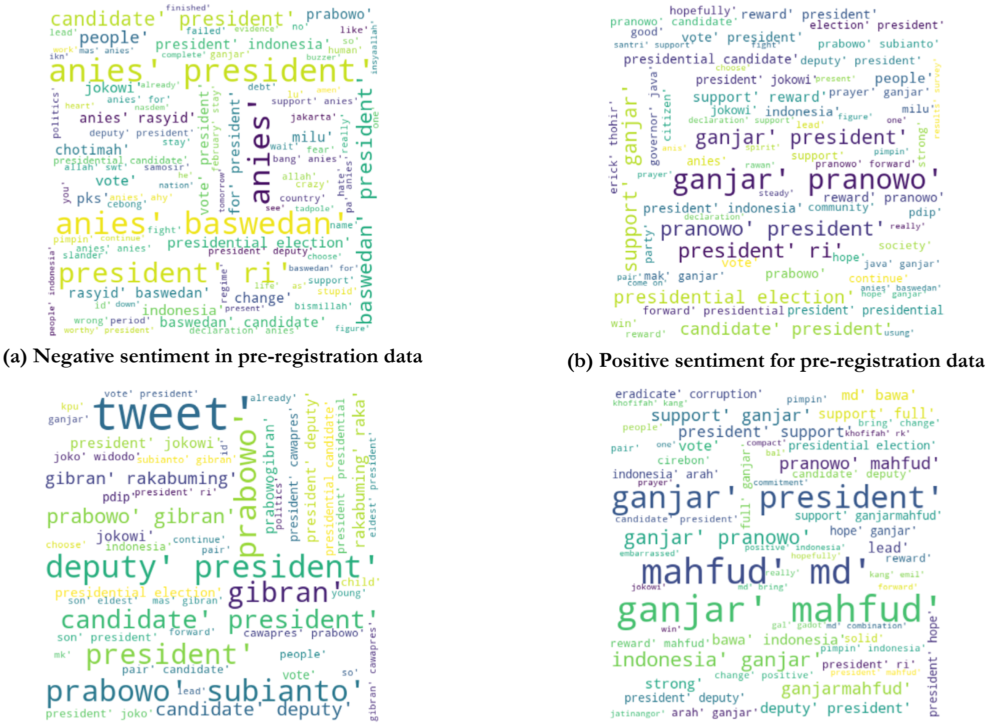


(a) Pre official registration candidates (b) Post official registration candidates

Figure 2. Comparison of sentiment pre and post official registration candidates

Figure 2a and 2b shows a comparison graph between the data with the collection before official registration and after the official registration of each candidate. After official registration, each candidate has determined the candidate for Deputy with their respective coalitions. The results of the data showed with the time of collection before registration with the largest percentage of positive sentiment, namely Ganjar Pranowo as much as 77.94% and the largest negative sentiment, namely Anies Baswedan as much as 31.39%. Meanwhile, after registration, the Presidential Candidate and Vice President pair with the largest positive sentiment were Ganjar Pranowo and Mahfud MD at 69.16%. While the biggest negative sentiment was on Prabowo Subianto and Gibran Rakabuming Raka by 52.12%.

Based on the sentiment graph obtained, the common words that emerge from each data for the most positive sentiment and the most negative are shown in Figure 3.



(a) Negative sentiment in pre-registration data

(b) Positive sentiment for pre-registration data

(c) Negative sentiment in post-registration data

(d) Positive sentiment for post-egistration data

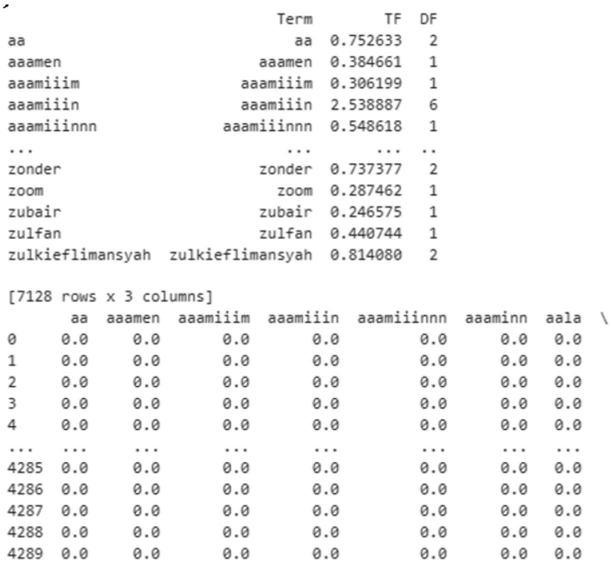
Figure 3. Word cloud of sentiment pre and post official registration candidates

Figure 3a – 3d shows a comparison of the set of words at the discussion before and after each candidate's official registration. Data displayed before official registration for the largest negative sentiment (Figure 3a) is found in Anies Baswedan data 31.39% and the largest positive (Figure 3b) in Ganjar Pranowo data 77.94%. Words that contain negative sentiments in Anies Baswedan's data are shown such as the words "failed", "change", "fear", "down", and other words. While the positive sentiment in Ganjar Pranowo's data shows the words "reward", "support", "hopefully", "prayer", and other words. Meanwhile, the data after registration showed the highest collection of negative words for Prabowo Subianto - Gibran Rakabuming Raka data at 52.12% (Figure 3c) and the most positive for Ganjar Pranowo - Mahfud MD data at 69.16% (Figure 3d). A

collection of words that appear in Prabowo Subianto – Gibran Rakabuming Raka data such as “child”, “eldest”, “mk”, “young”, and other words. While the data on Ganjar Pranowo – Mahfud MD contains words such as “strong”, “corruption”, “support”, “reward”, and other words.

b. TF-IDF

TF-IDF plays a role in giving values in the form of weights as input into the model. Figure 4 shows a snapshot of TF-IDF on data.



```
                              Term        TF  DF
aa                              aa  0.752633   2
aaamen                      aaamen  0.384661   1
aaamiiim                  aaamiiim  0.306199   1
aaamiiin                  aaamiiin  2.538887   6
aaamiiinnn              aaamiiinnn  0.548618   1
...                            ...       ...  ..
zonder                      zonder  0.737377   2
zoom                          zoom  0.287462   1
zubair                      zubair  0.246575   1
zulfan                      zulfan  0.440744   1
zulkieflimansyah  zulkieflimansyah  0.814080   2

[7128 rows x 3 columns]
       aa  aamen  aaamiiim  aaamiiin  aaamiiinnn  aaaminn  aala  \
0     0.0    0.0       0.0       0.0         0.0      0.0   0.0
1     0.0    0.0       0.0       0.0         0.0      0.0   0.0
2     0.0    0.0       0.0       0.0         0.0      0.0   0.0
3     0.0    0.0       0.0       0.0         0.0      0.0   0.0
4     0.0    0.0       0.0       0.0         0.0      0.0   0.0
...   ...    ...       ...       ...         ...      ...   ...
4285  0.0    0.0       0.0       0.0         0.0      0.0   0.0
4286  0.0    0.0       0.0       0.0         0.0      0.0   0.0
4287  0.0    0.0       0.0       0.0         0.0      0.0   0.0
4288  0.0    0.0       0.0       0.0         0.0      0.0   0.0
4289  0.0    0.0       0.0       0.0         0.0      0.0   0.0
```

Figure 4. Results TF-IDF

Figure 4 illustrates the outcome of computing TF-IDF for every word within the document. The word’s weight is established by its frequency, and TF reflects its occurrence rate. IDF evaluates the uniqueness of words in the document, and the resultant TF-IDF is utilized as input for the classification model in each respective method.

c. Stratified K-Fold Cross Validation

The distribution of training and testing samples on each data is divided as Figure 5.

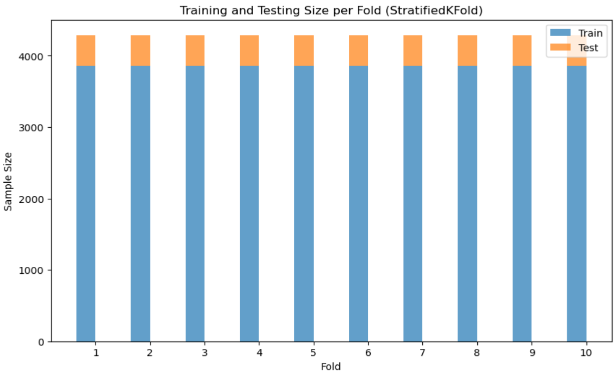
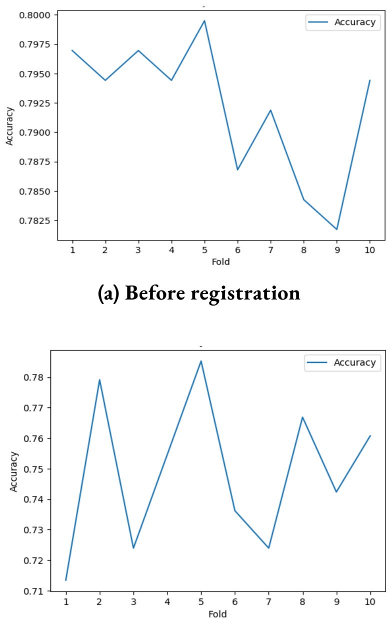


Figure 5. Training and testing size per fold (Stratified K-Fold)

Figure 5 shows the distribution of training and testing samples on each fold. The distribution of the sample is done using random state based on the number of folds available. The number of folds used is 10-fold.

d. Naïve Bayes

The Naïve Bayes Multinomial classification method is used to form a model of each data obtained. The data compared is data before the official registration of candidates and after the official registration of presidential candidates with the Vice President. The best model for the two data sets is shown in Figure 6 using a fold of 10.



(a) Before registration

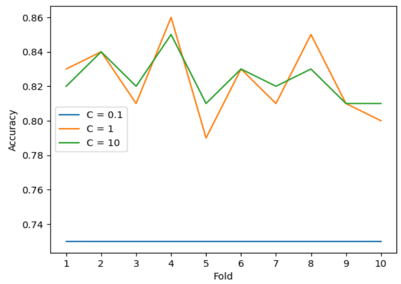
(b) After registration

Figure 6. Curve accuracy for each fold of Naïve Bayes

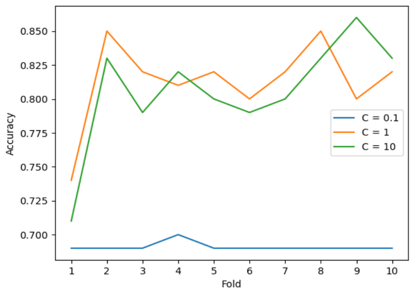
Figure 6a and 6b shows a graph of the best accuracy results for the Naive Bayes model from each data before and after registration for the three candidates. Figure 6a shows the data for the best Naive Bayes model before the official registration of the presidential and vice-presidential candidates. The best model obtained is on Ganjar Pranowo data on fold 5 with 80% accuracy. While the best Naive Bayes model after the official registration of the presidential and vice-presidential candidates is shown in Figure 6b. The best data shown is on the Ganjar Pranowo - Mahfud MD pair in fold 5 with 79% accuracy.

e. Support Vector Machine

The Linear Support Vector Machine method is used to form a classification model from two different data comparisons. The value of parameter C used for Support Vector Machine in this study is 0.1, 1, and 10 with a fold number of 10. Figure 7 shows the comparison of data between before (Figure 7a) and after registration (Figure 7b) for each candidate using the SVM linear method.



(a) Before registration



(b) After registration

Figure 7. Curve accuracy for each fold of Support Vector Machine

Figure 7a and 7b shows a graph of the best model for classification using linear SVM with each candidate's data before and after candidate registration. The best accuracy was obtained in the data before official registration, namely Prabowo Subianto (Figure 7a) at fold 4 with C=1, which is 86%. Meanwhile, the data after official registration obtained the best accuracy, namely the Ganjar Pranowo – Mahfud MD (Figure 7b) pair at fold 9 with C = 10, which is 86%

f. Evaluation

Comparison of Linear SVC and Multinomial Naive Bayes methods between data obtained before official registration and data obtained after the official registration of candidates. The method comparison on pre-registration data is illustrated with an example, using the data with the best accuracy, which is Prabowo Subianto's data with a Linear SVC accuracy of 86%, as shown in Figure 8.

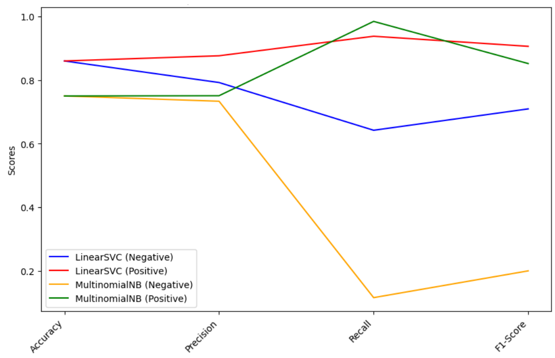


Figure 8. Comparison for Linear SVC and Multinomial Naïve Bayes for Prabowo Subianto (Pre-Registration)

Figure 8 shows the comparison between the two methods on Prabowo Subianto data. The accuracy of each class (positive and negative) for both methods was obtained by comparing model evaluations such as accuracy, precision, recall, and F1-Score. The accuracy for Linear SVC was 86% and Multinomial Naive Bayes was 75%.

The comparison of the two methods on the data obtained after the official registration of the respective candidates with pairs namely Presidential and Vice-Presidential Candidates is shown as Figure 9. The example shown is the Ganjar Pranowo - Mahfud MD data with the best accuracy.

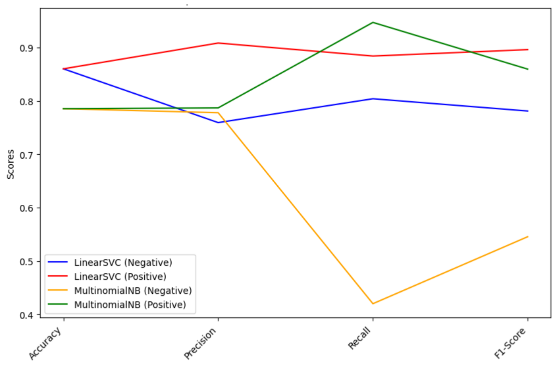


Figure 9. Comparison for Linear SVC and Multinomial Naïve Bayes for Ganjar Pranowo – Mahfud MD (Post-Registration)

Figure 9 shows the comparison between Linear SVC and Multinomial Naive Bayes methods on Ganjar Pranowo - Mahfud MD data obtained after the registration stage. The best accuracy obtained on Linear SVC is 86% and Multinomial Naïve Bayes is 79%.

g. Analysis

Among the two methods used, SVM has a higher average accuracy than Naïve Bayes. However, the SVM method has a longer computational speed, in contrast to Naïve Bayes which can do it quickly [33][34]. The accuracy of the Naïve Bayes method is lower than SVM because this method only requires small data in training [35]. Whereas the SVM method has good performance when used on large data [36]. The high accuracy results of SVM indicate that this method does not overfit [28], making it suitable for sentiment analysis approaches [37][38] with large and pertinent data for political content of presidential elections [39].

4. Discussion

Every election contestant, especially the presidential election, must carry out a campaign process. Campaigns are carried out to increase the popularity of candidates [40]. The popularity of candidates is shown in the mention of the number of words contained in the data obtained as in Table 2.

Table 2. Common Words

Common Words (Pre-Registration)	Count (Pre-Registration)	Common Words (Post-Registration)	Count (Post-Registration)
anies	3,695	anies	1,273
prabowo	3,533	ganjar	1,228
ganjar	2,272	mahfud	1,223
-	-	prabowo	939
-	-	gibran	852
-	-	muhaimin	483

Table 2 shows the common words that appear in the data by specializing in the names of Presidential candidates for data before registration and data of Presidential candidates and Vice-Presidential candidates for data after registration. The most common words that appear in the data before registration are anies as much as 3695, Prabowo as much as 3533, and ganjar as much as 2272. While the data after registration obtained the most words mentioned were anies as much as 1273, ganjar as much as 1228, mahfud as much as 1223, prabowo as much as 939, gibran as much as 852, and muhaimin as much as 483.

The performance of sentiment analysis is highly influenced by the method and data used [41]especially among social media users. Twitter, as one of the widely used social media platforms in Indonesia, functions as an effective campaign forum. However, the problem that arises is how to automatically collect social media data related to presidential discussions and provide conclusions on the analysis results. Of course, this is not easy if done manually. Sentiment analysis is one approach that can be used for this in order to draw conclusions and analysis related to the available data. Data was collected shortly after the registration of presidential and vice-presidential candidates in November 2023. This study aims to obtain sentiment results from the latest data obtained, get the best model from the Naive Bayes method, to conduct analysis in predicting presidential election results based on sentiment. However, at the time of data collection, candidate numbers had not been assigned by the Election organizers. The obtained data amounted to 11,569 records using the Valence Aware Dictionary for Sentiment Reasoning (VADER. However, the accuracy of the model is greatly influenced by factors such as the preprocessing used, parameters, and classification algorithms [42].

Projections of the results of the upcoming Indonesian Presidential Election based on the sentiment analysis approach can be misinterpreted because they have not been able to control bot/computer accounts and paid users/fake accounts. In addition, the limitations of the data used have not been able to accommodate based on the population or sample number of voters in Indonesia. However, this research can be a recommendation and a deeper review of the effect of sentiment analysis on the results of the 2024 Indonesian Presidential election when compared to actual results.

This research can also be implemented in other countries besides Indonesia with the concept of democratic presidential elections. Because some countries also apply the same General Election/Presidential Election system so that this research can be a reference to be applied to countries with the same election system. Especially countries that actively use social media as a campaign tool, this research can be developed with the same case.

5. Conclusion and Future Works

This study collected data on Twitter in March as many as 15,000 (before official registration) and November as many as 11,569 Tweets (after official registration). Data cleaning stages used such as case folding, tokenizing, stop word removal, normalization, stemming, and drop duplicated with the aim of facilitating the classification process. Labeling of the Tweets into positive and negative classes was done. This stage uses the VADER technique to proceed to the TF-IDF process which is useful for determining the value of word frequency in a document to help the method classification process. The classification methods used are SVM and Naïve Bayes with the division of test and training data samples using stratified K-Fold to anticipate data imbalance between classes. The validation models used are precision, recall, f1 score and accuracy. Based on data obtained before official registration, the largest percentage of positive sentiment was obtained, namely Ganjar Pranowo as much as 77.94% and the largest negative sentiment was Anies Baswedan as much as 31.39%. Whereas in post-registration, the presidential and vice-presidential pairs with the largest positive sentiment were Ganjar Pranowo and Mahfud MD at 69.16% and the largest negative sentiment was Prabowo Subianto and Gibran Rakabuming Raka at 52.12%. The best accuracy on data before registration using SVM is 86% and Naïve Bayes is 75%. While the data obtained after official registration obtained the best accuracy using SVM was 86% and Naïve Bayes was 79%. This proves that both methods are reliable for sentiment cases on the topic of presidential elections.

6. Acknowledgment

This work is supported by the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia based With Contract number: 181/E5/PG.02.00.PL/2023. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation. Author also gratefully for the cooperation as a partner of the Nusantara Research Institute, the Center for Human Development Studies Foundation.

References

[1] O. Olabanjo, A. Wusu, O. Afisi, M. Asokere, R. Padonu, O. Olabanjo *et al.*, "From Twitter to Aso-Rock: A Natural Language Processing Spotlight for Understanding Nigeria 2023 Presidential Election," *Heliyon*, vol. 9, no. 5, p. e16085, 2022, doi: 10.1016/j.heliyon.2023.e16085.

[2] D. O. Oyewola, L. A. Oladimeji, S. O. Julius, L. B. Kachalla, and E. G. Dada, "Optimizing sentiment analysis of Nigerian 2023 presidential election using two-stage residual long short term memory," *Heliyon*, vol. 9, no. 4, p. e14836, 2023, doi: 10.1016/j.heliyon.2023.e14836.

[3] A. Sharma and U. Ghose, "Sentimental Analysis of Twitter Data with respect to General Elections in India," *Procedia Computer Science*, vol. 173, no. 2019, pp. 325–334, 2020, doi: 10.1016/j.procs.2020.06.038.

[4] K. Brito and P. J. L. Adeodato, "Machine learning for predicting elections in Latin America based on social media engagement and polls," *Government Information Quarterly*, vol. 40, no. 1, p. 101782, Jan. 2023, doi: 10.1016/j.giq.2022.101782.

[5] K. Dos Santos Brito, S. R. De Lemos Meira, and P. J. L. Adeodato, "Correlations of social media performance and electoral results in Brazilian presidential elections," *Information Polity*, vol. 26, no. 4, pp. 417–439, 2021, doi: 10.3233/IP-210315.

[6] M. Yousef and A. ALali, "Analysis and Evaluation of Two Feature Selection Algorithms in Improving the Performance of the Sentiment Analysis Model of Arabic Tweets," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 6, pp. 705–711, 2022, doi: 10.14569/IJACSA.2022.0130683.

[7] A. M. Iddrisu, S. Mensah, F. Boafo, G. R. Yeluripati, and P. Kudjo, "A sentiment analysis framework to classify instances of sarcastic sentiments within the aviation sector," *International Journal of Information Management Data Insights*, vol. 3, p. 100180, 2023, doi: 10.1016/j.jjimei.2023.100180.

[8] P. Savci and B. Das, "Prediction of the customers' interests using sentiment analysis in e-commerce data for comparison of Arabic, English, and Turkish languages," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 3, pp. 227–237, 2023, doi: 10.1016/j.jksuci.2023.02.017.

[9] Z. A. Diekson, M. R. B. Prakoso, M. S. Q. Putra, M. S. A. F. Syaputra, S. Achmad, and R. Sutoyo, "Sentiment analysis for customer review: Case study of Traveloka," *Procedia Computer Sciences*, vol. 216, no. 2022, pp. 682–690, 2023, doi: 10.1016/j.procs.2022.12.184.

[10] W. M. Shaban, A. H. Rabie, A. I. Saleh, and M. A. Abo-Elsoud, "Accurate detection of COVID-19 patients based on distance biased Naïve Bayes (DBNB) classification strategy," *Pattern Recognition*, vol. 119, 2021, doi: 10.1016/j.patcog.2021.108110.

[11] H. Zhang, L. Jiang, and L. Yu, "Attribute and instance weighted naive Bayes," *Pattern Recognition*, vol. 111, p. 107674, Mar. 2021, doi: 10.1016/j.patcog.2020.107674.

[12] S. Wang, J. Ren, and R. Bai, "A semi-supervised adaptive discriminative discretization method improving discrimination power of regularized naive Bayes," *Expert System with Applications*, vol. 225, no. November 2022, 2023, doi: 10.1016/j.eswa.2023.120094.

[13] V. A. Fitri, R. Andreswari, and M. A. Hasibuan, "Sentiment analysis of social media Twitter with case of Anti-LGBT campaign in Indonesia using Naïve Bayes, decision tree, and random forest algorithm," *Procedia Computer Sciences*, vol. 161, pp. 765–772, 2019, doi: 10.1016/j.procs.2019.11.181.

[14] A. A. Firdaus, A. Yudhana, and I. Riadi, "Public Opinion Analysis of Presidential Candidate Using Naïve Bayes Method," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics and Control*, vol. 4, no. 2, pp. 563–570, May 2023, doi: 10.22219/kinetik.v8i2.1686.

[15] K. K. Agustiningsih, E. Utami, and O. M. A. Alsyaibani, "Sentiment Analysis and Topic Modelling of The COVID-19 Vaccine in Indonesia on Twitter Social Media Using Word Embedding," *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika*, vol. 8, no. 1, pp. 64–75, 2022, doi: 10.26555/jiteki.v8i1.23009.

[16] D. Arifah, T. H. Saragih, D. Kartini, and M. I. Mazdadi, "Application of SMOTE to Handle Imbalance Class in Deposit Classification Using the Extreme Gradient Boosting Algorithm," *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika*, vol. 9, no. 2, pp. 396–410, 2023, doi: 10.26555/

jiteki.v9i2.26155.

[17] M. Qorib, T. Oladunni, M. Denis, E. Ososanya, and P. Cotae, "Covid-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset," *Expert System with Applications*, vol. 212, p. 118715, Feb. 2023, doi: 10.1016/j.eswa.2022.118715.

[18] R. Ahuja, A. Chug, S. Kohli, S. Gupta, and P. Ahuja, "The impact of features extraction on the sentiment analysis," *Procedia Computer Sciences*, vol. 152, pp. 341–348, 2019, doi: 10.1016/j.procs.2019.05.008.

[19] M. Liang and T. Niu, "Research on Text Classification Techniques Based on Improved TF-IDF Algorithm and LSTM Inputs," *Procedia Computer Sciences*, vol. 208, pp. 460–470, 2022, doi: 10.1016/j.procs.2022.10.064.

[20] M. Chiny, M. Chihab, Y. Chihab, and O. Bencharef, "LSTM, VADER and TF-IDF based Hybrid Sentiment Analysis Model," *International Journal of Advanced Computer Sciences and Applications*, vol. 12, no. 7, pp. 265–275, 2021, doi: 10.14569/IJACSA.2021.0120730.

[21] D. Suleiman, A. Odeh, and R. Al-Sayyed, "Arabic Sentiment Analysis Using Naïve Bayes and CNN-LSTM," *Informatica*, vol. 46, no. 6, pp. 79–86, 2022, doi: 10.31449/inf.v46i6.4199.

[22] Z. Ye, P. Song, D. Zheng, X. Zhang, and J. Wu, "A Naive Bayes model on lung adenocarcinoma projection based on tumor microenvironment and weighted gene co-expression network analysis," *Infectious Disease Modelling*, vol. 7, no. 3, pp. 498–509, 2022, doi: 10.1016/j.idm.2022.07.009.

[23] A. Tariq, J. Yan, A. S. Gagnon, M. Riaz Khan, and F. Mumtaz, "Mapping of cropland, cropping patterns and crop types by combining optical remote sensing images with decision tree classifier and random forest," *Geo-Spatial Information Science*, vol. 00, no. 00, pp. 1–19, 2022, doi: 10.1080/10095020.2022.2100287.

[24] A. Tariq, Y. Jiango, Q. Li, J. Gao, L. Lu, W. Soufan *et al.*, "Modelling, mapping and monitoring of forest cover changes, using support vector machine, kernel logistic regression and naive bayes tree models with optical remote sensing data," *Heliyon*, vol. 9, no. 2, 2023, doi: 10.1016/j.heliyon.2023.e13212.

[25] R. Khan, Y. Qian, and S. Naeem, "Extractive based Text Summarization Using KMeans and TF-IDF," *International Journal of Information Engineering and Electronic Business*, vol. 11, no. 3, pp. 33–44, 2019, doi: 10.5815/ijieeb.2019.03.05.

[26] W. Liu, S. Liang, and X. Qin, "Weighted p-norm distance t kernel SVM classification algorithm based on improved polarization," *Scientific Reports*, vol. 12, no. 1, pp. 1–17, 2022, doi: 10.1038/s41598-022-09766-w.

[27] M. D. Prasetio, R. Y. Xavier, H. Rachmat, W. Wiyono, and D. S. E. Atmaja, "Sentiment analysis on myindihome user reviews using support vector machine and naïve bayes classifier method," *International Journal of Industrial Optimization*, vol. 2, no. 2, pp. 151–164, 2021, doi: 10.12928/ijio.v2i2.4449.

[28] A. D. Cahyani, "Aspect-Based Sentiment Analysis from User-Generated Content in Shopee Marketplace Platform," *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika*, vol. 9, no. 2, pp. 444–454, 2023, doi: 10.26555/jiteki.v9i2.26367.

[29] A. Peryanto, A. Yudhana, and R. Umar, "Convolutional Neural Network and Support Vector Machine in Classification of Flower Images," *Khazanah Informatika: Jurnal Ilmu Komputer dan Informatika*, vol. 8, no. 1, pp. 1–7, 2022, doi: 10.23917/khif.v8i1.15531.

[30] B. Alsari, R. Alkhaldi, D. Alsaffar, T. Alkhaldi, H. Almaymuni, N. Alnaim *et al.*, "Sentiment analysis for cruises in Saudi Arabia on social media platforms using machine learning algorithms," *Journal of Big Data*, vol. 9, p. 21, Dec. 2022, doi: 10.1186/s40537-022-00568-5.

[31] C. B. G. Allo, L. S. A. Putra, N. R. Paranoan, and V. A. Gunawan, "Comparing Logistic Regression and Support Vector Machine in Breast Cancer Problem," *Jambura Journal of Probability and Statistics*, vol. 4, no. 1, pp. 1–8, Jun. 2023, doi: 10.34312/jjps.v4i1.19246.

[32] S. Szeghalmy and A. Fazekas, "A Comparative Study of the Use of Stratified Cross-Validation and Distribution-Balanced Stratified Cross-Validation in Imbalanced Learning," *Sensors*, vol. 23, no. 4, p. 2333, Feb. 2023, doi: 10.3390/s23042333.

[33] A. Basuki, "Sentiment Analysis of Customers ' Review on Delivery Service Provider on Twitter Using Naive Bayes Classification," *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika*, vol. 9, no. 2, pp. 420–428, 2023, doi: 10.26555/jiteki.v9i2.26327.

[34] R. Wang, "Automatic Classification of Document Resources Based on Naive Bayesian Classification Algorithm," *Informatica*, vol. 46, no. 3, pp. 373–382, 2022, doi: 10.31449/inf.v46i3.3970.

[35] A. Yudhana and A. Dwi, "Spatial distribution of soil nutrient content for sustainable rice agriculture using geographic information system and Naïve Bayes classifier," *International Journal on Smart Sensing and Intelligent Systems*, vol. 16, no. 1, 2023, doi: 10.2478/ijssis-2023-0001.

[36] H. Syahputra and A. Wibowo, "Comparison of Support Vector Machine (SVM) and Random Forest Algorithm for Detection of Negative Content on Websites," *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika*, vol. 9, no. 1, pp. 165–173, 2023, doi: 10.26555/jiteki.v9i1.25861.

[37] N. Pavitha, V. Pungliya, A. Raut, R. Bhonsle, A. Purohit, A. Patel *et al.*, "Movie recommendation and sentiment analysis using machine learning," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 279–284, 2022, doi: 10.1016/j.gltp.2022.03.012.

[38] R. P. Pratama and A. Tjahyanto, "The influence of fake accounts on sentiment analysis related to COVID-19 in Indonesia," *Procedia Computer Sciences*, vol. 197, pp. 143–150, 2021, doi: 10.1016/j.procs.2021.12.128.

[39] A. A. Firdaus, A. Yudhana, I. Riadi, and Mahsun. "Indonesian presidential election sentiment : Dataset of response public before 2024," *Data in Brief*, vol. 52, p. 109993, 2024, doi: 10.1016/j.dib.2023.109993.

[40] U. Daxecker and M. Rauschenbach, "Election type and the logic of pre-election violence: Evidence from Zimbabwe," *Electoral Studies*, vol. 82, no. January, 2023, doi: 10.1016/j.electstud.2023.102583.

[41] A. A. Firdaus, A. Yudhana, and I. Riadi, "Prediction of Indonesian Presidential Election Results using Sentiment Analysis with Naïve Bayes Method," *Jurnal Media Informatika Budidarma*, vol. 8, no. 1, pp. 41–50, 2024, doi: 10.30865/mib.v8i1.7007.

[42] H. O. Ahmad and S. U. Umar, "Sentiment Analysis of Financial Textual data Using Machine Learning and Deep Learning Models," *Informatica*, vol. 47, no. 5, pp. 153–158, 2023, doi: 10.31449/inf.v47i5.4673.