JOIN (Jurnal Online Informatika)
Volume 7 No. 2 | December 2022: 226-235
DOI: 10.15575/join.v7i2.925
p-ISSN: 2528-1682
e-ISSN: 2527-9165

Delineation of the Early 2024 Election Map: Sentiment Analysis Approach to Twitter Data

Nur Ulum Rahmanulloh[1], Ibnu Santoso[2]
[1,2]Department of Statistical Computing, Politeknik Statistika STIS, Indonesia

Article Info

Article history:

Received September 19, 2022
Revised November 24, 2022
Accepted November 24, 2022
Published December 26, 2022

Keywords:

Politics
2024 Election
Sentiment Analysis
Twitter

ABSTRACT

As a democratic country, the people hold an important role in determining power in Indonesia. The closest political agenda in Indonesia is the 2024 Election. A survey has been conducted by several private survey agencies regarding the 2024 political map which has revealed the top five names, namely Prabowo Subianto, Ganjar Pranowo, Anies Baswedan, Sandiaga Uno, and Ridwan Kamil. This study aims to describe the initial map of the 2024 Election through a sentiment analysis approach to Twitter data. This study uses tweet data that mentions five political figures during 2021. In general, the demographic condition of Twitter users that pros or cons to five political figures, among them: located on the Java, in the age group 19–29 years old, and male. The sentiment analysis method used is supervised learning with different methods for each figure. The difference in methods adjusts the best evaluation value given in each figure. The results showed that the highest positive sentimental tweets and the highest number of pro accounts was about Ganjar Pranowo. On the other hand, the highest negative sentiment and the highest number of contra accounts was about Prabowo Subianto. Many words that often appear on a figure's positive sentiment are expressions of hope, prayer, and support. On negative tweets, the word that comes up a lot relating to the work field or work region of the figures.

Corresponding Author:

Nur Ulum Rahmanulloh,
Departmen of Statistical Computing, Politeknik Statistika STIS,
Jl. Otto Iskandardinata No.64C, Jakarta 13330
Email: nur.ulum.1999@gmail.com

1. INTRODUCTION

Political science is defined as a science that studies efforts to achieve a good life [1]. In a state context, politics is about power, public policy, and distribution. Power is the most crucial concept in politics. Power will determine the public policy and distribution [1]. Indonesia adheres to a democratic form of government [2]. Democracy derives from the Greek Language, namely *demos* which means the people and *kratein* which means power [1]. It shows that the people play an important role in determining power in Indonesia through the General Election (Pemilu). According to article 1 of the Act Number 7 of 2017 (UU NO. 17 2017), one of the purposes of the Election is to elect the President and Vice President [3]. Elections of the president and vice-president directly by the people were first conducted in 2004 and conducted regularly once every five years [4].

According to the data from National Medium-Term Development Plan (RPJMN) and the General Election Commission of Indonesia (KPU), the Election participation rate, since the presidential election in 2014, experienced an increasing trend until it peaked at 82% in the 2019 presidential election. It shows that the people in Indonesia are increasingly aware of the importance of politics in the life of the state. In the 2020 Regional Election (Pilkada), the Election participation rate decreased again. This is due to public's fear of the Covid-19 pandemic that is hitting the world, including Indonesia.

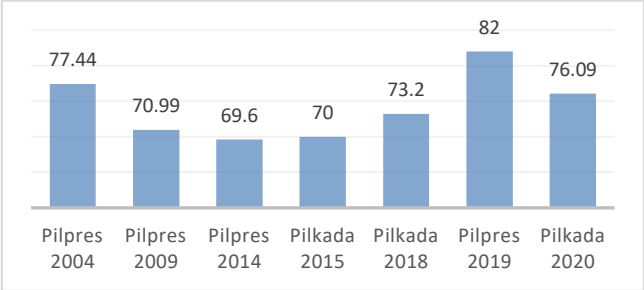

Figure 1. Voter participation in the presidential and regional elections 2004 – 2020

The government's agenda in political field continues, one of the closest agenda is the Presidential Election (Pilpres) in 2024. Although the presidential election is still two more years, some figures are starting to appear among the people as presidential candidates [5]. According to the Constitution, presidential candidates can only be nominated by political parties. However, as an aggregation of the people's interests, political parties should keep in mind the political preferences of the people if they do not want to be abandoned by the people [5]. Responding to a growing phenomenon among the public about 2024 Elections, several survey agencies have conducted surveys about people's choice preferences. The results of the survey revealed five main candidates, namely Prabowo Subianto, Ganjar Pranowo, Anies Baswedan, Sandiaga Uno, and Ridwan Kamil [5] [6] [7] [8] [9] [10].

In the era of fourth industrial revolution, human life is inseparable from the internet. Based on data published by Hootsuite (We are Social) as of January 2021, about 73.7% of Indonesian population are internet users. About 61.8% of the population in Indonesia are active users of social media with 85.5% Twitter users [11]. With the number of Twitter users in Indonesia reaching 108.12 million users, many analyzes have been carried out on Twitter's data. The research has been carried out in various fields, one of them is in the political field. Some of these researches include: research by Fitriyyah et al [12] that analyzes sentiment of presidential candidates in 2019 Elections using naïve bayes method. Research by Sunjana [13] which tried to see the polarization related to the 2019 Elections using sentiment analysis and social network analysis approach. Research by Gustomy [14] which tried to see political polarization, especially in the Covid-19 pandemic discourse.

This research aims to describe the early map of the 2024 Election according to sentiments and direction of Twitter users' support for five political figures. From the result of sentiment analysis, modelling of the topics that are widely discussed about each figure was carried out on positive and negative sentiments. From the data direction of Twitter users' support, an analysis of the demographic condition of Twitter users was carried out. The demographic conditions included the location distribution, age group, and gender of Twitter users.

2. METHOD

This research was conducted through several steps which are presented in Figure 2.

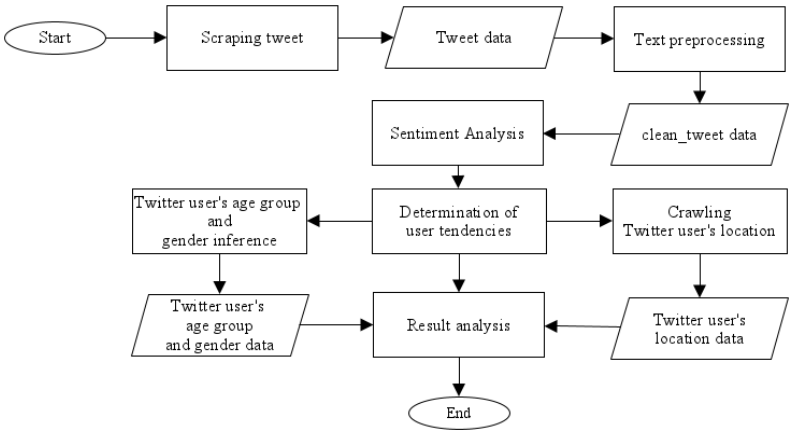

Figure 2. Research flowchart

2.1. Data Collection

There are two methods used in collecting data, namely scraping and crawling. Scraping is a method of collecting data without interacting with a program's API [15]. Whereas Crawling is the process of obtaining

large-scale data from a web page by utilizing the web API [16]. Scraping is done to retrieve tweet data by using the snscrape library. The keywords used are '@prabowo', '@ganjarpranowo', '@aniesbaswedan', '@sandiuno', and '@ridwankamil'. The data collected are tweets tagging these five Twitter accounts during 2021. Past study has shown the feasibility of using tweet that mentioned specific political figure to show the election map [17]. In addition, crawling is also used which focuses on the location of Twitter users available in the user's profile. The crawling process is carried out using the tweepy library.

2.2. Data Preprocessing

After data collection, then data preprocessing is the next step. Data preprocessing consists of two steps, namely: data filtering and tweet cleaning. Data filtering aims to filter tweets that specifically tag one figure's account or tag that figure's account and accounts that associated with the figure. The list of accounts that associated with the figures is presented in the table 1.

Table 1. List of figure's Twitter account and accounts that associated with that figure

No.	Political Figure	Figure's Account	Associated Accounts
1.	Prabowo Subianto	@prabowo	@Kemhan_RI, @Gerindra
2.	Ganjar Pranowo	@ganjarpranowo	@TajYasinMZ, @provjateng, @humasjateng
3.	Anies Baswedan	@aniesbaswedan	@BangAriza, @DKIJakarta
4.	Sandiaga Uno	@sandiuno	@Kemenparekraf
5.	Ridwan Kamil	@ridwankamil	@PaUuRuzhanul, @humasjabar

The cleaning process aims to clean the data that is not necessary in the analysis. Cleaning process is done in several steps, namely: case folding or the process of converting characters into lowercase format, removing username, retrieving hashtag by removing hash mark ('#'), removing url, removing symbols, normalizing or converting text into standard format, stemming or converting text into its root word, eliminating the use of extra letters unnecessary, and removing words related to certain political figures, such as: '*gubernur*', '*menhan*', 'jakarta', '*menteri*', '*kang*', and so on.

2.3. Data Analysis

After the data is cleaned at the preprocessing step, sentiment analysis is carried out on the data. Sentiment analysis is carried out in several steps which are presented in figure 3.

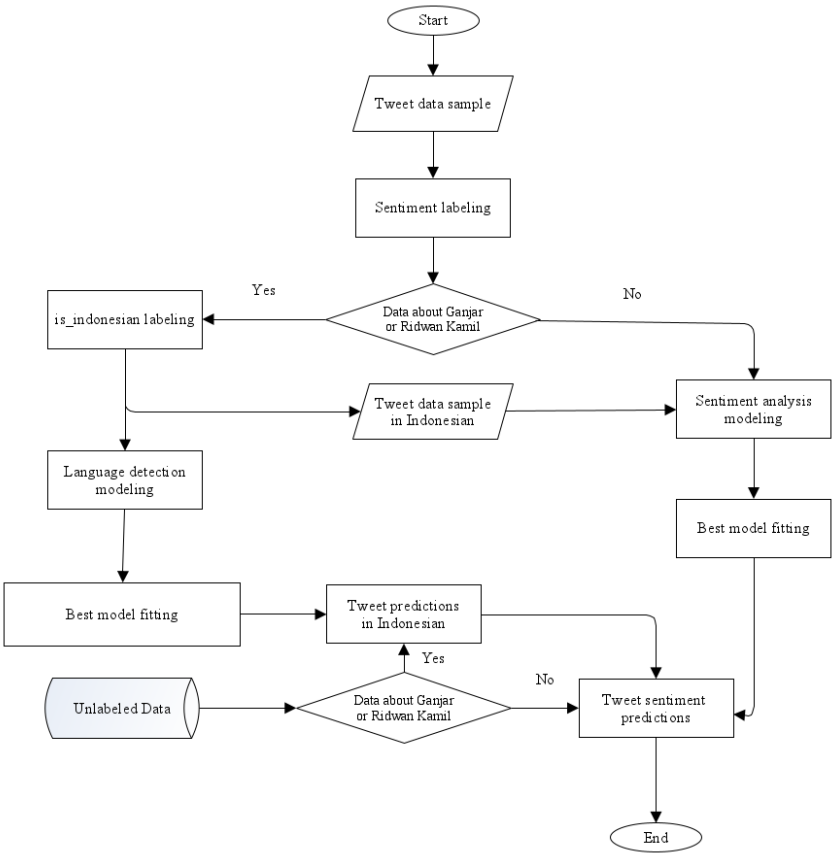

Figure 3. Sentiment analysis flowchart

The first step in sentiment analysis is labeling sample data that is used as training and testing data. The labeling process is carried out in exactly one tweet on each Twitter account and with balanced labels between categories. Before the model selection for sentiment analysis, second step on labelling was carried out on the tweet that mention Ganjar Pranowo and Ridwan Kamil to classify the language used in tweets. The labeling aims to retrieve the Indonesian-language tweet because many tweets use local languages, namely Javanese to Ganjar Pranowo and Sundanese to Ridwan Kamil, which can reduce the accuracy of sentiment prediction results.

After the labeling process, especially on Ganjar Pranowo and Ridwan Kamil's tweet data, supervised learning modeling was carried out to predict whether the tweet was in Indonesian or not. Some models that are tried to detect language on the tweet are random forest and logistic regression with the combination of feature extraction used are count vectorizer, tfidf vectorizer, and hashing vectorizer. The method used in this modeling is by choosing the method with the highest model evaluation value. Evaluation is done by looking at the accuracy, precision, recall, and f1-score of the model. In addition, average accuracy from cross validation model is also seen.

The process continues at the model selection for sentiment analysis. There are three supervised learning methods used, namely Long Short-Term Memory (LSTM), random forest, and logistic regression. On LSTM method, feature extraction is done by using embedding matrix. The LSTM method was tried until the minimum accuracy was still above the best accuracy of the next two methods. If the minimum accuracy is not achieved, the model selection process continues on the next two methods. In the next two methods, there are three feature extraction methods used, namely count vectorizer, tfidf vectorizer, and hashing vectorizer. The model with the best accuracy, precision, recall, and f1-scores are retested using cross validation methods. A method with average cross-validation accuracy at least 70% is used as the final model to predict tweet sentiments because model with accuracy above 70% can already be categorized as a model that has a fairly good performance [18]. However, if the average cross validation accuracy is still below 70%, the best model is replaced with the second model. It was done repeatedly until a model is obtained with an average cross validation accuracy at least 70% is obtained.

The next step after model selection is to make sentiment predictions on the tweet data relating to the figure. Before making sentimental predictions, on the tweet about Ganjar Pranowo and Ridwan Kamil, it was predicted that the tweet was in Indonesian or not by using a model that had been built previously.

Determination of Twitter user support based on comparison between positive and negative tweets in a certain time period. Determination of Twitter user support is done separately among political figures. There are three categories of support in this research, namely pros, cons, and neutral.

After obtaining sentiment data and Twitter users' support, then descriptive analysis was carried out on the data. In addition, demographic data related to the Twitter users is taken to see the demographic patterns of Twitter users who are pros or cons to certain figures. The demographic data consists of the location, age group, and gender of Twitter users. Twitter user location data is taken by crawling method using tweepy library, while the data of age group and gender is taken through prediction results using m3inference library. Some descriptive analysis on this research include: number of tweets per month for each political figure, distribution of sentiments to each figure in one year, number of sentiments for each figure per month, n-top words sentiment to each figure, Twitter user support for one year, n-top words location of users who are pros and cons, and the distribution of Twitter users who are pros or cons to figures based on age group and gender. The findings are presented in data visualization to make it easier to understand.

2.4. Long Short-Term Memory (LSTM)

LSTM is a type of Recurrent Neural Network (RNN). RNN is a type of neural network that specifically designed to solve sequential problems. RNN uses the state of previous stage as additional information to produce output at next stage [19]. Unlike other RNN methods, LSTM makes it possible to ignore problems that appear in the training model so the model can provide better results. The computing unit in LSTM is called as a memory cell. Memory cell has weighing parameters on input, output, and internal state. The key before going to the memory cell is gate. There are three types of gates, namely: forget gate, input gate, and output gate. Forget gate is responsible to determining the incoming information can be removed or not. Input gate is responsible to determining the value of the input used to update memory state. Output gate is responsible to determining the result based on the input and the memory [19].

2.5. Random Forest

Random forest is an ensemble method in machine learning. This method was built from the weakness of decision tree that has a low accuracy for new data. Random forest is built using n-tree. The n-tree can complement each other in the classification process. The classification process in the random forest is determined based on the highest number of classes that appear in n-tree [20].

2.6. Logistic Regression

According to Hosmer and Lemeshow (1989) [21], logistic regression comes from Bernoulli model. From a set x_n predictors, the probability value of the binary dependent variable y_n. The probability model can be defined as follows.

$$P(x_n) = \sigma(w \, . \, x_n)$$

With the logistics function as follows.

$$\sigma(\theta) = \frac{1}{1 + exp\,[-\theta]}$$

The logistic function can map the real number θ into the interval (0, 1). When Y has more than two classes, logistic regression can be used even it is more complex [21]. The probability model can be defined as follows.

$$P(x_n) = \pi(c, x_n, w) = \frac{exp\,(w_c . x_n)}{\Sigma_{c'} exp\,(w_{c'} . x_n)}$$

The multinomial model is a generalization of the binary model by defining $w_0 = 0$ and $w_1 = w$.

3. RESULTS AND DISCUSSION

3.1. Data Description

In figure 4, a trend visualization of tweets data that mention a certain figure or mention figure and accounts that associated with the figure.

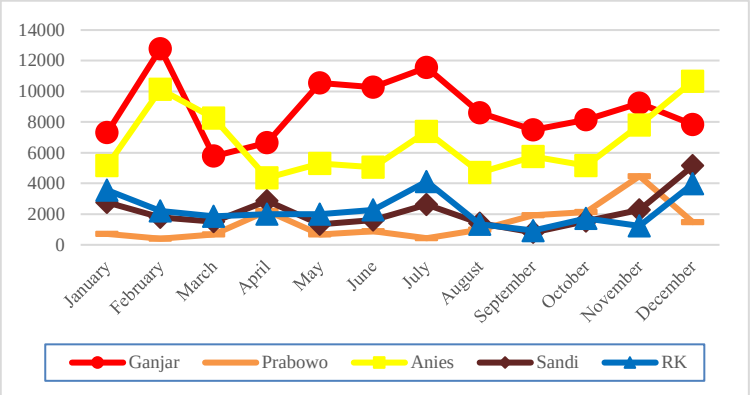

Figure 4. Trendlines of tweets that mention figures and account that associated with the figure

Based on figure 4, tweets about Ganjar Pranowo is the highest in 2021 followed by tweets about Anies Baswedan. Tweets about Anies Baswedan got higher than Ganjar Pranowo in March and December. Tweets about Ganjar Pranowo have been on a downward trend since July. On the contrary, tweets about Anies Baswedan on an upward trend since October and surpassed tweets about Ganjar Pranowo in December. The smallest number of tweets are tweets about Prabowo Subianto. The number of tweets that mention a figure or the figure with associated accounts during 2021 is 106,268 tweets for Ganjar Pranowo, 79,759 tweets for Anies Baswedan, 27,255 tweets for Ridwan Kamil, 25,700 tweets for Sandiaga Uno, and 16,968 tweets for Prabowo Subianto.

3.2. Sentiment Analysis

Model selection is needed to ensure that sentiment analysis for each political figure gives the best result. According to research by Jaidka [22], the effective method in different cases may be different. The selected model that used to predict the tweets of five political figures is presented in the table 2.

Table 2. Selected models for each figure

No.	Political Figure	Method	Accuracy
1.	Prabowo Subianto	Logistic regression + tfidf vectorizer	72%
2.	Ganjar Pranowo	Logistic regression + count vectorizer*	90%
		LSTM	79%
3.	Anies Baswedan	LSTM	76%
4.	Sandiaga Uno	Logistic regression + tfidf vectorizer	76%
5.	Ridwan Kamil	Logistic regression + tfidf vectorizer *	90%
		Logistic regression + tfidf vectorizer	72%

*Method to detect language on tweet

Based on the selected model, sentiment prediction was made for each figure. The percentage distribution of sentiment for each figure is presented in the figure 5.

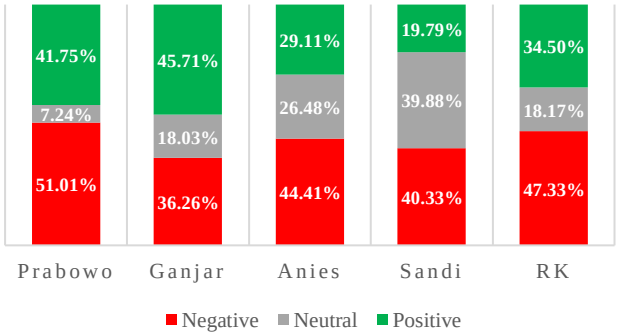

Figure 5. Distribution of sentiment for political figures during 2021

Based on figure 5, it can be noted that Ganjar Pranowo (45.71%) has the highest percentage of positive tweets as compared to other figures. Then, followed by Prabowo Subianto at 41.75%, Ridwan Kamil at 34.50%, Anies Baswedan at 29.11%, and Sandiaga Uno at 19.79%. The highest percentage of negative tweets is owned

by Prabowo Subianto (51.01%). Then followed by Ridwan Kamil at 47.33%, Anies Baswedan at 44.41%, Sandiaga Uno at 40.33%, and Ganjar Pranowo at 36.26%.

3.3. Word Cloud

Word cloud is used to see the frequently appearing word of each figure divided into positive and negative sentiments. Here's word cloud of positive sentiment for each figure.

Figure 6. Prabowo's positive sentiment

Figure 7. Ganjar's positive sentiment

Figure 8. Anies' positive sentiment

Figure 9. Sandi's positive sentiment

Figure 10. RK's positive sentiment

In general, positive tweets for each figure contain prayers and hopes, such as: '*moga'*, *'sehat'*, and *'selalu'*. In addition, it can also be a compliment, such as: *'mantap'* and *'keren'*. On tweets about Prabowo Subianto, the words that appeared were different from other figures. Words that appear related to his position as Minister of Defense of the Republic of Indonesia, such as: *'tahan'*, 'Indonesia', *'kerja'*, and *'kuat'*.

Here's word cloud of negative sentiment for each figure.

Figure 11. Prabowo's negative sentiment

Figure 12. Ganjar's negative sentiment

Figure 13. Anies' negative sentiment

Figure 14. Sandi's negative sentiment

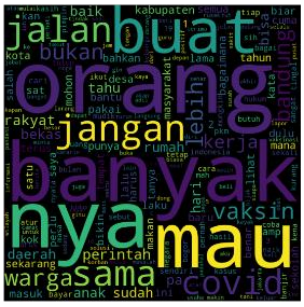

Figure 15. RK's negative sentiment

In general, the negative tweets mostly discuss the condition of the figure's region or field of work. For example, the word *'jalan'* is related to damaged road, and the word '*wisata'* is related to the opening of tourist attractions during the Covid-19 pandemic in Indonesia. On the tweets about Prabowo Subianto and Sandiaga Uno, there are words besides words related to the field of work. The words are related to disappointment with the decision of the two figures to join the government cabinet as Minister of Defense and Minister of Tourism and Creative Economy. The words include: *'rakyat'*, *'jangan'*, *'mau'*, and *'dukung'*.

3.4. Support Direction

Support direction is determined based on the highest sentiments on each Twitter user. Twitter users with the most positive tweets are classified as pro accounts. Twitter users with the most negative tweets are classified as contra accounts. Twitter users who have the most neutral tweets or same number of negative and positive tweets are classified as neutral accounts. Here's visualization of Twitter users' support for five political figures during 2021.

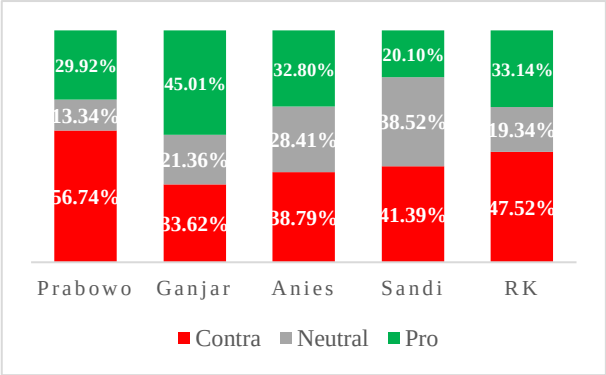

Figure 16. Twitter users' support direction for political figures during 2021

Based on figure 16, it can be noted that Ganjar Pranowo (45.01%) has the highest percentage of pro Twitter users compared to other figures. Then, followed by Ridwan Kamil at 33.14%, Anies Baswedan at 32.80%, Prabowo Subianto at 29.92%, and Sandiaga Uno at 20.10%. The highest percentage of contra accounts is owned by Prabowo Subianto (56.74%). Then followed by Ridwan Kamil at 47.52%, Sandiaga Uno at 41.39%, Anies Baswedan at 38.79%, and Ganjar Pranowo at 33.62%.

3.5. N-top Words Location

N-top words location is used to see the distribution of pros and cons Twitter users' locations. Here's visualizations of five locations with the highest number of pros from each political figure.

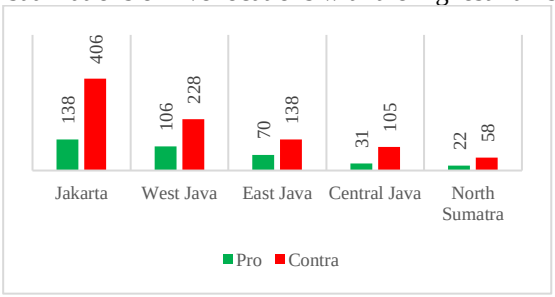

Figure 17. Prabowo's five highest location

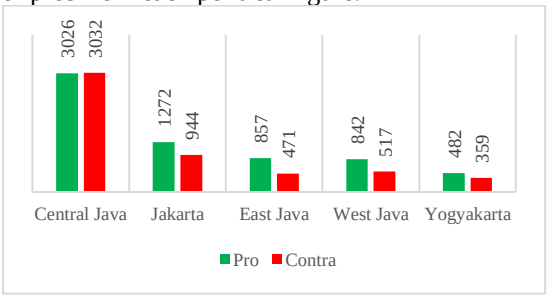

Figure 18. Ganjar's five highest location

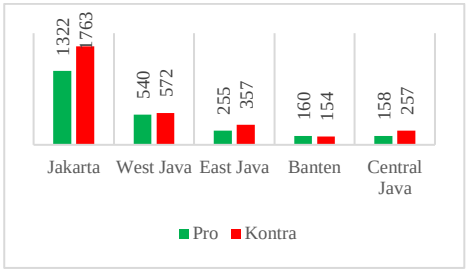

Figure 19. Anies' five highest location

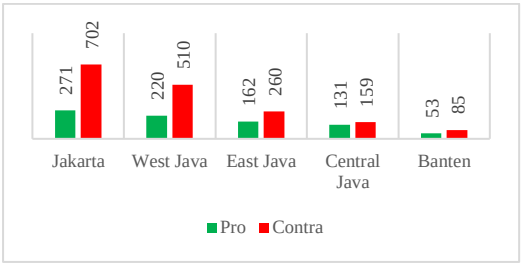

Figure 20. Sandi's five highest location

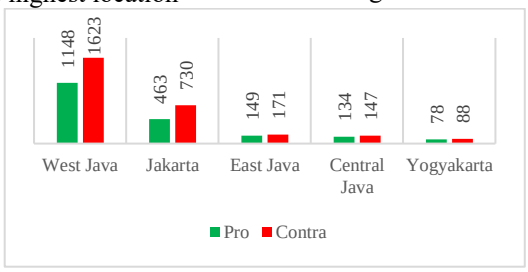

Figure 21. RK's five highest location

Based on the five visualizations above, it can be known that the majority of Twitter users who are pro and contra to the five figures are located in the Java. Only on the users who are pro to Prabowo Subianto, the fifth position is North Sumatra. Twitter users who are pro or contra to three political figures, namely Ganjar Pranowo, Anies Baswedan, and Ridwan Kamil, are mostly in his leadership area, namely: Central Java, Jakarta, and West Java. Whereas the most Twitter users who are pro and contra to Prabowo Subianto and Sandiaga Uno are located in Jakarta.

3.6. Age Group and Gender

The distribution of age groups and genders of Twitter users who are pro and contra to figures is based on prediction results using m3inference library [23]. The analyzed are limited to individual accounts that were also based on the prediction results. The prediction results had 53% accuracy in predicting age groups, 86% in predicting gender, and 94% in predicting account types. Here's a visualization of the distribution of the age groups of Twitter users who are pro and contra to each political figure.

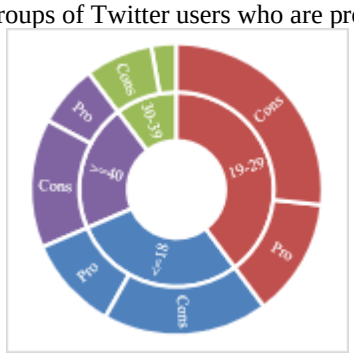

Figure 22. Prabowo's pro-cons account age group

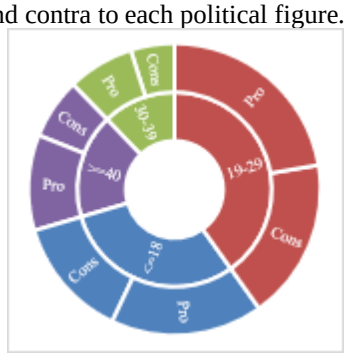

Figure 23. Ganjar's pro-cons account age group

Figure 24. Anies' pro-cons account age group

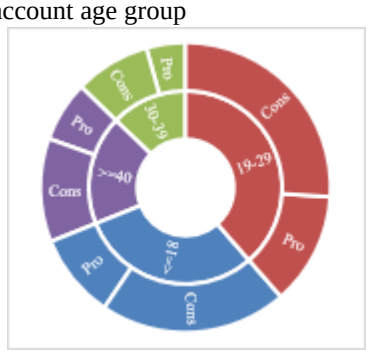

Figure 25. Sandi's pro-cons account age group

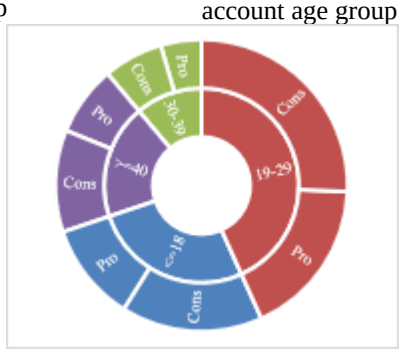

Figure 26. Rk's pro-cons account age group

In general, Twitter users who are pro or contra to each figure are mostly in the 19-29 years old, followed by the under-18 years old, over 40 years old, and the smallest is 30-39 years old. Here's visualizations of the gender distribution of Twitter users who are pro and contra to each political figure.

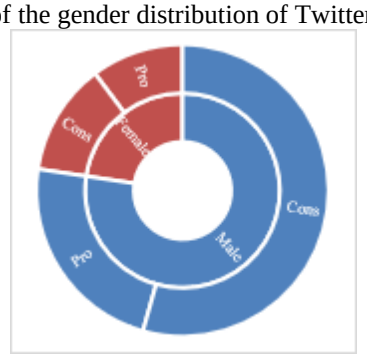

Figure 27. Prabowo's pro-cons account gender

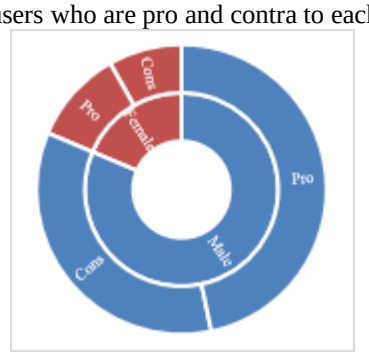

Figure 28. Ganjar's pro-cons account gender

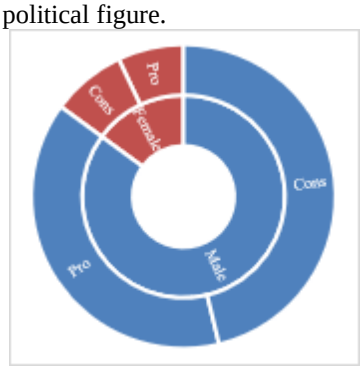

Figure 24. Anies' pro-cons account gender

Figure 25. Sandi's pro-cons account gender

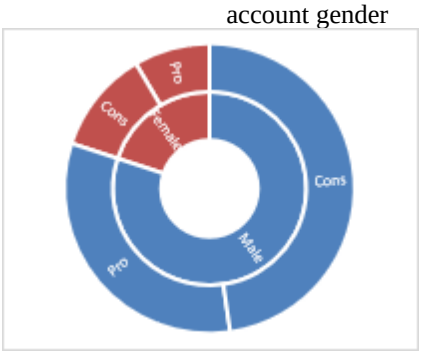

Figure 26. RK's pro-cons account gender

In general, Twitter users who were pro and contra to the five political figures are dominated by the male gender rather than the female. It shows that political topics are more attractive to men than women.

4. CONCLUSION

Based on the results, the conclusion can be drawn as follows: The highest positive tweets were found in tweets about Ganjar Pranowo (45.71%), while the highest negative tweets were found in tweets about Prabowo Subianto (51.01%). Words that come up a lot in positive tweets are prayers, hope, and support. Specifically, the tweets about Prabowo Subianto contained words that related to the national defence field. Words that come up a lot in negative tweets are about problems that arise in the area or field of work of the political figure. Negative tweets about Prabowo Subianto and Sandiaga Uno also contain a lot of disappointment with their decision to become ministers in the current government. The highest percentage of pro Twitter users are in tweets about Ganjar Pranowo (45.01%), while the highest contra Twitter users are in tweets about Prabowo Subianto (56.74%). The location of Twitter users who are pro and contra to five political figures is dominated by Java. The age groups of Twitter users who are pro and contra to the five political figures mostly in the 19 – 29 years old. The gender of Twitter users who are pro and contra to five political figures is dominated by male. It shows that political topics are more attractive to men than women.

Based on the results, several suggestions can be given for further study as follows: Follow-up study can be carried out on the tweet data for the next years, especially ahead of the election. Further study could be carried out related to some of the findings in this study, such as: the majority locations in Java, the domination of 19-29 years old group compared to other age groups, and the majority who discuss politics are male.

5. REFERENCES

[1] M. Budiardjo, Dasar-Dasar Ilmu Politik, Jakarta: Gramedia, 2008.
[2] Pemerintah Indonesia, UNDANG-UNDANG DASAR NEGARA REPUBLIK INDONESIA TAHUN 1945.
[3] S. Raharjo, Undang-Undang Pemilu 2019 Berdasarkan Undang-Undang Nomor 7 Tahun 2017 Tentang Pemilihan Umum, Jakarta: Bhuana Ilmu Populer, 2018.
[4] BPS, Statistik Politik 2019: Pemilu 1955-2019, Jakarta: Badan Pusat Statistik, 2019.

[5] SMRC, "Partai dan Calon Presiden: Kecenderungan Sikap Pemilih Menjelang 2024," Saiful Mujani Research & Consulting, Jakarta, 2021.

[6] Charta Politika, "Rilis Survei Nasional: Evaluasi Kebijakan & Peta Politik Masa Pandemi," Charta Politika Indonesia, Jakarta, 2021.

[7] Indikator, "Evaluasi Publik Terhadap Kinerja Penanganan Pandemi, Vaksinasi, dan Peta Elektoral Terkini," Indikator Politik Indonesia, Jakarta, 2021.

[8] KedaiKOPI, "Laporan Hasil Survei Calon Pemimpin Indonesia 2024: Banjir Tokoh Menuju 2024," Lembaga Survei KedaiKOPI, Jakarta, 2021.

[9] LSI, "Evaluasi Publik Terhadap Kondisi Nasional dan Peta Awal Pemilu 2024," Lembaga Survei Indonesia, Jakarta, 2021.

[10] Poltracking, "Survei Nasional: Evaluasi 2 Tahun Pemerintahan Joko Widodo - Ma'ruf Amin & Peta Politik Elektoral Pilpres 2024," Poltracking Indonesia, Jakarta, 2021.

[11] S. Kemp, "Digital 2021: Indonesia," Datareportal, 11 Februari 2021. [Online]. Available: https://datareportal.com/reports/digital-2021-indonesia. [Accessed 6 November 2021].

[12] S. N. J. Fitriyyah, N. Safriadi and E. E. Pratama, "Analisis Sentimen Calon Presiden Indonesia 2019 dari Media Sosial Twitter Menggunakan Metode Naive Bayes," *Jurnal Edukasi dan Penelitian Informatika*, pp. 279-285, 2019.

[13] M. N. Habibi and Sunjana, "Analysis of Indonesia Politics Polarization before 2019 President Election Using Sentiment Analysis and Social Network Analysis," *Modern Education and Computer Science*, vol. 11, pp. 22-30, 2019.

[14] R. Gustomy, "Pandemi ke Infodemi: Polarisasi Politik dalam Wacana Covid-19 Pengguna Twiiter," *Jurnal Ilmiah Ilmu Pemerintahan*, vol. 5, no. 2, pp. 190-205, 2020.

[15] R. Mitchell, Web Scraping with Python, California: O'Reilly Media Inc., 2018.

[16] C. Olston and M. Najork, Web Crawling, Massachusetts: now Publishers, 2010.

[17] R. Liu, X. Yao, C. Guo and X. Wei, "Can We Forecast Presidential Election Using Twitter Data? An Integrative Modelling Approach," *Annals of GIS*, pp. 43-56, 2021.

[18] D. Kurniawan, Pengenalan Machine Learning dengan Python, Jakarta: Elex Media Komputindo, 2022.

[19] J. Brownlee, "A Gentle Introduction to Imbalanced Classification," 23 Desember 2019. [Online]. Available: https://machinelearningmastery.com/what-is-imbalanced-classification/.

[20] T. Ho, "Random decision forests," in *International Conference on Document Analysis and Recognition (ICDAR)*, Montreal, 1995.

[21] A. I. Schein and L. H. Ungar, "Active Learning for Logistic Regression: An evaluation," *Machine Learning*, vol. 68, no. 3, pp. 235-265, 2007.

[22] K. Jaidka, M. M. Skoric, S. Ahmed and M. Hilbert, "Predicting Elections from Social Media: A Three-Country, Three-Method Comparative Study," *Asian Journal of Communication*, 2018.

[23] Z. Wang, S. Hale, D. I. Adelani, P. Grabowicz, T. Hartman, F. Flock and D. Jurgens, "Demographic Inference and Representative Population Estimates from Multilingual Social Media Data," in *WWW'19: The World Wide Web Conference*, San Francisco, 2019.