JURNAL INFOTEL
Informatics - Telecommunication - Electronics
Website: http://ejournal.st3telkom.ac.id/index.php/infotel
ISSN: 2085-3688; e-ISSN: 2460-0997

Opinion mining of Indonesian presidential election on twitter data using decision tree method

Nur Ghaniaviyanto Ramadhan[1,*], Merlinda Wibowo[2], Nur Fatin Liyana Mohd Rosely[3], Christoph Quix[4]
[1,2]Institut Teknologi Telkom Purwokerto
[3]INTI International University
[4]Hochschule Niederrhein
[1,2]Jl. D.I. Panjaitan, No. 128, Purwokerto 53147, Indonesia
[3]Persiaran Perdana BBN, Nilai 71800, Malaysia
[4]Reinarzstraße 49, Krefeld 47805, Germany
[*]Corresponding email: ghani@ittelkom.ac.id

Received 23 August 2022, Revised 19 September 2022, Accepted 27 September 2022

Abstract — Indonesia is a country led by a president. Every five years, the presidential election will be conducted democratically. The latest election was in 2019 and the elected president will be served for five years from that year until 2024. This process is repeated for the next five years term of service by holding a general election. As an example, the next president's term will be from 2024 to 2029. The election process itself is a thick process where each presidential candidate held campaigns to convince and gain support from the people. The campaign is carried out directly to village locations and on social media; Twitter/Facebook/YouTube. Since social media become a powerful platform nowadays, most of the updated campaign news is shared through social media. One of them is Twitter. The written news led to an exciting analyzing opportunity because it consists various of opinions. Additionally, the contents are continuously accessible. On Twitter, a tweet is used as a term to reflect a response to a post. The high number of tweets indicates a high number of post responses. The related post about the latest election displayed a big number of tweets. Therefore, this study aimed to examine the sentiment of a tweet by exploring the public statement of the 2024 presidential election. The result of sentiment categories which are positive, negative, and neutral, and the word tweet related to the sentiment category will be visualized. Then, the classification process by using the decision tree approach is used and the obtained results are compared with the regression-based method. The accuracy result show decision tree contributes up to 99.3% and is superior to the regression-based by 2.5%.

Keywords – Decision tree, indonesian presidential, opinion mining, twitter data

Copyright ©2022 JURNAL INFOTEL
All rights reserved.

I. Introduction

Indonesia is a country led by a president. The president has the duty as head of state and head of government. To date, there have been seven presidents who have led the Indonesian state. The president in Indonesia is elected through the community through a democratic process, namely the presidential election (pilpres), which is held every five years. Becoming a president has several requirements. One of them, the person who served as president for two consecutive terms is not allowed to participate in the next election [1].

Indonesia also has many survey institutions to determine the electability of a presidential candidate. In addition, the related presidential comments can be used to determine the favoritism of the running presidential candidate and sentiment analysis will be a good approach for this purpose [2]. Plus, data collection could be from different social media such as Twitter, Facebook, and YouTube. Several studies have been conducted on sentiment analysis on the presidential election in 2019. Buntoro *et al.* [3] researched the estimation of presidential election sentiment using data from Twitter, namely #Jokowi and #Prabowo, for machine learning methods, namely SVM and Naive

Bayes.

Ismail and Lhaksamana [4] conducted a study on sentiment analysis of the 2019 presidential election using the nave Bayes method with the data from Twitter, namely Jokowi-Ma'aruf and Prabowo-Sandi. Pratama *et al.* [5] conducted a study to find out the conversations on Twitter in the first debate of the presidential candidates of the Republic of Indonesia through hashtags from the two pairs of candidates using the fined-gained method. Kristiyanti *et al.* [6] aim to predict the election of the President and Vice President of the Republic of Indonesia for the period 2019-2024 through a public opinion mining process on Twitter and test it accurately with a classification algorithm, namely Support Vector Machine (SVM) with Particle Swarm Optimization (PSO) and Genetic Algorithms (GA).

Zuhdi *et al.* [7] hope their research can help public research opinion on Twitter social media that contains positive, negative, and neutral sentiments using the KNN method. Prianto *et al.* [8] focused on extracting data from text generated from twitter social media that responded to the accounts of the Indonesian presidential and vice presidential candidates in the 2019 election using the lexicon method. Budi and Nugroho [9] conducted a sentiment analysis of the 2019 Indonesian presidential candidate based on public comments on the Facebook social network using the Naive Bayes model.

Didik *et al.* [10] discussed public opinion in the 2019 presidential election using the nave Bayes classification and TF-IDF weighting. Sitti *et al.* [11] discuss various opinions of Twitter users with positive and neutral sentiments. However, determining the idea of Twitter users requires considerable effort and time due to the many tweets used. In addition to the above research, other studies also discuss the same thing, namely the opinion of the general public using the most chosen social media, Twitter [12]–[16].

Based on the concerns in previous research related to the presidential election sector in Indonesia, this study conducts opinion mining on the public. Twitter is used as a social media platform to find out opinions that have been spread for the 2024 presidential election. Table 1 is a comparison of the contribution of this study to previous studies.

Table 1. Comparison Contribution

Authors	**Methods**
Buntoro *et al.* [3]	Naïve Bayes and SVM
Ismail *et al.* [4]	Naïve Bayes
Pratama *et al.* [5]	Finned-gined
Kristiyanti *et al.* [6]	SVM, PSO, Genetic Algorithm
Zuhdi *et al.* [7]	KNN
This Research	Decision tree

II. Research Method

This study will use the research flow as shown in Fig. 1.

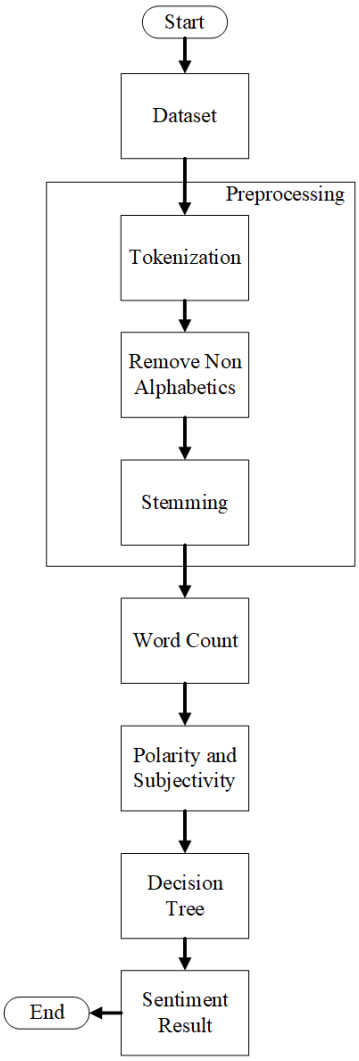

Fig. 1. System diagram.

A. Dataset

The first step of this research is to extract data from social media Twitter by using the Twitter API. Tweet data is carried in the form of tweets: (Presidential Election 2024, Pilpres 2024, #Pilpres2024) and retweets from Twitter users discussing the 2024 presidential election. The amount of tweet data used is 2475, with 30 days of data collection starting from July 16, 2022, to August 15, 2022. Table 2 is an example of the form of tweet data used.

While in Table 3 shows the distribution of the number of tweets in words used.

B. Clean Tweet

In this process, the tweet data that has been crawled is then cleaned up. The cleaned data includes existing special characters such as (?!(@) and others. We also check whether there are duplicate tweet data, and if there is one, the exact tweet will be deleted. As a result,

Table 2. Tweet Data

Tweet in Bahasa	**Tweet in English**
Koalisi Parpol Jelang Pilpres 2024 Mulai Terbentuk, Makin Dinamis Saat Pencalonan.	A coalition of Political Parties Ahead of the 2024 Presidential Election Begins to Form, More Dynamic During Nominations.
Dukungan untuk Menteri BUMN @erickthohir maju dalam pemilihan presiden 2024 terus bermunculan.	Support for the Minister of BUMN @erickthohir show in the 2024 presidential election continues to emerge.
Sahabat Ganjar Gandeng Kaum Ibu di Bogor Deklarasi Dukungan untuk Pilpres 2024.	Friends of Ganjar Collaborate with Mothers in Bogor Declaration of Support for the 2024 Presidential Election.
Prabowo Maju di Pilpres 2024, Jhon Sitorus Singgung Janji Anies.	Prabowo Runs in the 2024 Presidential Election, Jhon Sitorus Alludes to Anies Promise.
Don Adam Bilang Capres 2024 Harus Tegas dan Pro Rakyat.	Don Adam Says 2024 Presidential Candidates Must Be Firm and Pro People.
Prabowo Siap Bersaing dengan Airlangga di Pilpres 2024.	Prabowo Ready to Compete with Airlangga in the 2024 Presidential Election.

Table 3. Data Distribution

Word	**Tweet Amount**
Pemilihan Presiden 2024 Presidential Election 2024	2475
Pilpres 2024	2475
#Pilpres2024	2475

the tweet data is ready to be weighted polarity and subjectivity values.

C. Polarity and Subjectivity

This step for tweet data will be given weighting using polarity and subjectivity values. Where polarity is the value that is returned in the form of a value from the range -1 to 1 by counting the number of words in that one sentence. Meanwhile, subjectivity measures the amount of opinion and information in the text. Higher subjectivity means the text contains personal statements rather than accurate information. Polarity and subjectivity values are values obtained from the system where the calculation results from the data are the number of words in the data.

D. Labelling

The sentence will be labeled based on the previous polarity and subjectivity values in this process. The labeling provisions are based on Algorithm 1.

Algorithm 1 Labelling Algorithm

Require: $Polarity$, $Subjectivity$
Ensure: $Sentiment$
0: **if** $Polarity > 0.0$ and $Subjectivity > 0.0$ **then**
0: $Sentiment = Positive$
0: **else if** $Polarity == 0.0$ and $Subjectivity == 0.0$ **then**
0: $Sentiment = Neutral$
0: **else**
0: $Sentiment = Negative$
0: **end if**=0

So, all tweets result already have a positive/negative/neutral sentiment label.

E. Decision Tree

After the labeling is complete, testing will be carried out using a tree-based machine-learning model. The decision tree was chosen here because this algorithm has been widely used in various problems related to text mining, especially on social media [17]–[20]. The decision tree algorithm has a scoring technique called information gain and the Gini index. This study uses the calculation of the score of information gain in (1) and (2) [21].

$$Info = \sum_{j} \left(\frac{Nj(t)}{N(t)} \right) \log 2 \left(\frac{Nj(t)}{N(t)} \right) \quad (1)$$

where Nj is the number of samples from class j, $N(t)$ the number of samples from node t, and $Nj(t)$ the number of samples of class j at node t.

$$InfoGain = Info(Parent) - \sum_{k} (pk) Info(child) \quad (2)$$

where Nj is the number of samples from class j, $N(t)$ the number of samples from node t, and $Nj(t)$ the number of samples of class j at node t.

Meanwhile, to measure the accuracy results using (3) [22].

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

where TP is an actual positive value that is entirely true, TN is a true negative where all values are false. FP is a false positive where the real value is incorrect, but the detected is true. FN is a false negative, the opposite of a false positive.

III. RESULT

In this research, we create a tweet data visualization for each sentiment category. As in Fig. 2 positive category word cloud, Fig. 3 negative category word cloud, and Fig. 4 neutral category word cloud.

Table 4 is an example of tweets results after labeling positive, negative, and neutral. Meanwhile, Table 5 is the distribution of the number of tweets resulting from each word for each sentiment category.

Table 4. Result of Labelling

Word in Bahasa	**Word in English**	**Sentiment Result**
bukan cak Imin, tiga kepala daerah ini dinilai paling ideal jadi cawapres prabowo.	not cak Imin, these three regional heads are considered the most ideal to be prabowo cawapres.	Positive
siang ini dibikin heboh dengan berita bersatunya pemain politik identitas but its ok kalian mempermudah pilihan.	this afternoon was made excited by the news of the unification of identity politics players but it's ok you guys made the choice easier.	Positive.
berkas lengkap 24 parpol masuk tahap verifikasi administrasi.	complete files of 24 political parties enter the administrative verification stage.	Neutral.
pilpres berat PDIP disarankan berkoalisi walau bisa usung capres sendiri.	tough presidential election PDIP is advised to compete even though it can carry its own capres.	Neutral.
pengamat politik Dedi Kurnia Syah menilai belum ada koalisi partai politik yang solid dan bisa langgeng.	political observer Dedi Kurnia Shah assessed that there is no coalition of political parties that is solid and can last.	Negative.
peluang wan abud untuk maju pada kontestasi pilpres 2024 telah pupus. kurangnya partai pengusung.	wan abud chances of running in the contestation of the 2024 presidential election have been dashed. lack of a carrying party.	Negative.

Table 5. Result Distribution Tweet of Sentiment Category

Word	**Sentiment Category**	**Tweet Amount**
Pemilihan Presiden 2024	Positive	304
	Neutral	1473
Presidential Election 2024	Negative	698
Pilpres 2024	Positive	615
	Neutral	1245
	Negative	615
#Pilpres2024	Positive	489
	Neutral	1198
	Negative	788

Table 6 is the result of applying the decision tree method for each word used in the search on Twitter.

Fig. 2. Sentiment positive word count.

Fig. 3. Sentiment negative word count.

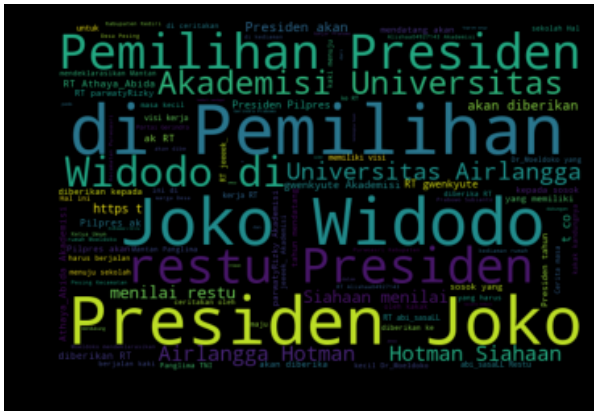

Fig. 4. Sentiment neutral word count.

IV. Discussion

Based on the results obtained in the previous section, Fig. 2, Fig. 3, and Fig. 4 are the category results of the words produced by each category. In the positive class, one example of a comment can be observed, namely resurrection, where the word has a positive meaning. One example of a word in the negative category can be followed, namely fire. In the neutral type, one example of a comment that can be observed is a blessing.

Table 4 is the result of labeling where the results of the tweet sentence have been clean. Cleaning the character in this tweet is very important and affects the accuracy results produced. The experiment proved that without cleaning, the accuracy result showed 6% difference. The difference in accuracy is considered vast and can lead to misinterpretation by the public.

Meanwhile, the results from Table 5 show that the number of categories of tweet sentiment is different.

Table 6. Result of Sentiment

Word Twitter	**Method**	**Accuracy %**
Pemilihan Presiden 2024	Decision Tree	99.2
Pilpres 2024		98.8
#Pilpres2024		100
Pemilihan Presiden 2024	Logistic Regression	98.5
Pilpres 2024		95.9
#Pilpres2024		96.2

Among the three words used, the neutral class was the most produced. Meanwhile, the minor category results are positive. This shows that not many people have expressed their opinions this month regarding the presidential election, especially in the positive and negative categories.

Table 6 is the result of applying the decision tree method where the highest accuracy is produced by the word #Pilpres2024 of 100%. At the same time, the lowest accuracy is the 2024 presidential election at 98.8%. The differences in the words used also affect the obtained accuracy. Compared to the logistic regression method, the word that produces the highest accuracy is the '2024 Presidential Election' at 98.5%, while the lowest is the '2024 Presidential Election' at 95.9%.

Differences in the results of the word also produce different accuracy with different methods. This is because the decision tree method has a way of working in the form of decision trees. This is certainly more accurate in predicting than the regression-based method, which works based on the data type.

V. Conclusion

This study discusses opinion analysis on Twitter social media. The obtained results in this study by applying the tree-based method are superior to the regression-based way. The word that produces the highest accuracy is #Pilpres2024 at 100%, while the lowest word accuracy is 'Pilpres 2024'. The number of distributions of each tweet word also affects the accuracy with a difference of 6%.

Acknowledgement

The author would like to thank the Institut Teknologi Telkom Purwokerto, which has fully supported the writing of this scientific paper. As the first author to contribute in terms of research ideas and background. The second author contributed in terms of proofreading and research methods. The third author contributes to writing-related works and discussions, and the fourth author contributes to writing conclusions and reviewing the overall contents.

References

[1] S. Negara, "Undang - Undang Republik Indonesia Nomor 42 Tahun 2008 Tentang Pemilihan Umum Presiden dan Wakil Presiden," 2008.

[2] M. R. F. Sya'bani, U. Enri, and T. N. Padilah, "Analisis sentimen terhadap bakal calon presiden 2024 dengan algoritme naive bayes," *JURIKOM (Jurnal Ris. Komputer)*, vol. 9, no. 2, pp. 265–273, 2022.

[3] G. A. Buntoro, R. Arifin, G. N. Syaifuddiin, A. Selamat, O. Krejcar, and F. Hamido, "The Implementation of the machine learning algorithm for the sentiment analysis of Indonesia's 2019 Presidential election," *IIUM Eng. J.*, vol. 22, no. 1, pp. 78–92, 2021.

[4] M. M. Ismail and K. M. Lhaksmana, "Sentimen analisis pada media online mengenai pemilihan presiden 2019 dengan menggunakan metode naive bayes," *eProceedings Eng.*, vol. 6, no. 2, 2019.

[5] S. F. Pratama, R. Andrean, and A. Nugroho, "Analisis sentimen twitter debat calon presiden indonesia menggunakan metode fined-grained sentiment analysis," *JOINTECS (Journal Inf. Technol. Comput. Sci.*, vol. 4, no. 2, pp. 39–44, 2019.

[6] D. A. Kristiyanti, Normah, and A. H. Umam, "Prediction of indonesia presidential election results for the 2019-2024 period using twitter sentiment analysis," in *2019 5th International Conference on New Media Studies (CONMEDIA)*, 2019, pp. 36-42, doi: 10.1109/CONMEDIA46929.2019.8981823.

[7] A. M. Zuhdi, E. Utami, and S. Raharjo, "Analisis sentiment twitter terhadap capres Indonesia 2019 dengan metode K-NN," *J. Inf. J. Penelit. dan Pengabdi. Masy.*, vol. 5, no. 2, pp. 1–7, 2019.

[8] C. Prianto, N. H. Harani, and I. Firmansyah, "Analisis sentimen terhadap kandidat presiden republik indonesia pada pemilu 2019 di media sosial twitter," *J. MEDIA Inform. BUDIDARMA*, vol. 3, no. 4, pp. 405–413, 2019.

[9] S. E. Budi and A. Nugroho, "Analisis sentimen calon presiden indonesia 2019 berdasarkan komentar publik di facebook," *J. Eksplora Inform.*, vol. 9, no. 1, pp. 60–69, 2019.

[10] M. D. R. W. Wahyudi, "Analisis sentimen ujaran kebencian pemilihan presiden 2019 menggunakan algoritme Naïve Bayes," *JNANALOKA*, vol. 1, no. 1, pp. 25–33, 2020.

[11] S. N. J. Fitriyyah and E. E. Pratama, "Analisis sentimen calon presiden indonesia 2019 dari media sosial twitter menggunakan metode naive bayes," *JEPIN (Jurnal Edukasi dan Penelit. Inform.*, vol. 5, no. 3, pp. 279–285, 2019.

[12] A. F. Sabily, P. P. Adikara, and M. A. Fauzi, "Analisis sentimen pemilihan presiden 2019 pada twitter menggunakan metode maximum entropy," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2548, p. 964X, 2019.

[13] L. Septiani and Y. Sibaroni, "Sentiment analysis terhadap tweet bernada sarkasme berbahasa indonesia," *J. Linguist. Komputasional*, vol. 2, no. 2, pp. 62–67, 2019.

[14] I. Santoso, W. Gata, and A. B. Paryanti, "Penggunaan feature selection di algoritma support vector machine untuk sentimen analisis komisi pemilihan umum," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 3, no. 3, pp. 364–370, 2019.

[15] R. Ardiansyah, "Analisis sentimen calon presiden dan wakil presiden periode 2019-2024 pasca debat pilpres di Twitter," *Sci. Comput. Sci. Informatics J.*, vol. 2, no. 1, pp. 21–28, 2019.

[16] A. C. Najib, A. Irsyad, G. A. Qandi, and N. A. Rakhmawati, "Perbandingan metode lexicon-based dan SVM untuk analisis sentimen berbasis ontologi pada kampanye pilpres indonesia tahun 2019 di twitter," *Fountain of Informatics*, vol. 4, no. 2, pp. 41–48, 2019.

[17] S. S. I. Ismail, R. F. Mansour, A. El-Aziz, M. Rasha, and A. I. Taloba, "Efficient e-Mail spam detection strategy using genetic decision tree processing with NLP features," *Comput. Intell. Neurosci.*, vol. 2022, 2022.

[18] H. A. Bouarara, "Sentiment analysis using machine learning algorithms and text mining to detect symptoms of mental difficulties over social media," in *Research Anthology on Implementing Sentiment Analysis Across Multiple Disciplines*, IGI Global, 2022, pp. 581–595.

[19] K. Zerrouki, R. M. Hamou, and A. Rahmoun, "Sentiment analysis of tweets using naïve bayes, KNN, and decision tree," in *Research Anthology on Implementing Sentiment Analysis Across Multiple Disciplines*, IGI Global, 2022, pp. 538–554.

[20] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, "An introduction to decision tree modeling," *J. Chemom. A J. Chemom. Soc.*, vol. 18, no. 6, pp. 275–285, 2004.

[21] N. G. Ramadhan and T. I. Ramadhan, "Analysis sentiment based on IMDB aspects from movie reviews using SVM," *Sink. J. dan Penelit. Tek. Inform.*, vol. 7, no. 1, pp. 39–45, 2022.