The Indonesian Journal of Computer Science
www.ijcs.net
Volume 13, Issue 4, August 2024
https://doi.org/10.33022/ijcs.v13i4.4126

Uncovering the Reasons Behind Abstain Voters' Stances in the 2024 Indonesian Presidential Election: Social Media X Study Cases

Irzanes Putri[1], Faiz Nur Fitrah Insani[2], Indra Budi[3], Aris Budi Santoso[4], Prabu Kresna Putra[5]
irzanes.putri@ui.ac.id[1], faiz.insani@ui.ac.id[2], indra@cs.ui.ac.id[3], aris.budi@ui.ac.id[4], prabu.kresna@ui.ac.id[5]
[1,2,3,4,5] Faculty of Computer Science, University of Indonesia

Article Information

Submitted : (date)
Reviewed : (date)
Accepted : (date)

Keywords

presidential election, stance analysis, topic modelling, coherence vote, abstain voters

Abstract

The Indonesian Government expects the participation of all Indonesian people in holding General Elections. However, according to the 2019 Political Statistics by BPS, there were 34.75 million people who did not exercise their right to vote or were abstain voters (golput) in the 2019 Election. This research aims to analyze individual attitudes towards abstaining voters using stance analysis and topic modelling. From 9,045 collected tweets, subsequent manual annotation revealed 2,566 pro stances, 5,264 neutral stances, and 1,215 contra stances. The classification models utilized are Random Forest, Decision Tree, Logistic Regression, Support Vector Machine, K-Nearest Neighbor, and Gradient Boosting. The classification outcomes will be analyzed by comparing the accuracy, precision, recall, and F1-score results based on their algorithms and n-grams. The results obtained from the stance analysis show that Random Forest achieved the highest accuracy and precision scores, with values of 84% and 83%, respectively. The discussion topic among those supporting golput due to low trust in the presidential and vice-presidential candidates. Other topics mentioned public feels dissatisfied with the pairs of candidates.

Attribution-ShareAlike 4.0 International License

A. Introduction

The 2024 general election in Indonesia will soon be held, and presidential candidates are starting to take action to find their supporters. The president is a significant figure in a country's government and often receives public attention. Various policies, actions, and speeches have an important impact in terms of the economy, social issues, and politics. The public's assessment of presidential candidates is necessary to know their views on a candidate for the country's next leader.

Based on Law No. 7 of 2017, the Indonesian Government expects the participation of all Indonesian people in holding General Elections [1]. It means that the 2024 Election must receive the entire attention of the Indonesian people. However, according to the 2019 Political Statistics by BPS, there were 34.75 million people who did not exercise their right to vote or were abstain voters (golput) in the 2019 Election. This number is equivalent to 18.02% of the 2019 Election permanent voter list (DPT) of 192.77 million people [2]. It indicates that the number of abstain voters in Indonesia is still high. Two major groups cause the public not to exercise their right to vote, namely internal factors consisting of technical and occupational factors, and external factors consisting of administrative, socialization, and political factors [3]. This research focuses on external factors, specifically socialization, where the main cause is that many people do not receive information about the candidates for national leadership. Therefore, voters choose not to exercise their right to vote because they do not understand anything about the presidential candidates they choose.

Over the last two decades, the use of social media has become a daily routine for everyone. One of the frequently used social media platforms is X (formerly Twitter). Besides being used to express opinions and engage in discussions on various topics, X has also become a popular source, garnering increased attention for research in recent years [4]. Public opinion data from social media X can be utilized to analyze the positions or stances that the public will take on the 2024 Indonesian Presidential Election using a stance analysis.

Previous research, including the study conducted by Gunhal et al. [5], classified Twitter users' stances toward the 2020 California election propositions. The results of the stance detection model can offer a different perspective to assist politicians in making decisions that reflect the interests of the communities they serve. The study [6] analyzed the factors influencing individuals' decisions to get vaccinated in Japan by collecting Japanese-language tweets related to vaccines and classifying the vaccination stances of the users who posted those tweets. Another study [7] analyzes Saudi society's stance towards distance education during the COVID-19 pandemic, especially in the 2020 academic year. This study assessed public opinions on distance education that could be used by the Ministry of Education for decision-making. The study [8] conducted stance analysis to detect tweets expressing vaccine refusal using the Bidirectional Encoder Representations from Transformers (BERTs).

Previous studies have implemented stance analysis to assess public opinions and stances towards COVID-19. However, there is a lack of studies that apply stance analysis to examine presidential election attitudes on social media, especially X. This research aims to analyze individual attitudes towards abstaining voters, whether they pro, contra, or are neutral. Furthermore, this study will identify the factors that lead someone to become an abstinent voter. Therefore, this research is expected to answer the following research questions (RQ):

1. What is the public's tendency towards the stance on abstain voters in the 2024 Indonesian presidential election?
2. What are the factors that lead individuals to choose to become abstain voters?

B. Research Method

1. Data Collection

In this research, data collection is carried out by scraping social media X using a web scraping method in Jupyter Notebook, aided by the Chromium browser in the Node.js library, and further processed using Python to convert the data into a CSV file. We reviewed 16 research papers on sentiment analysis during presidential elections to decide which social media platform to focus on for our research. Surprisingly, we found that 14 out of these 16 papers used X (formerly Twitter) as their main data source. As a result, we've chosen to concentrate our research on X as the primary social media platform for data collection and analysis.

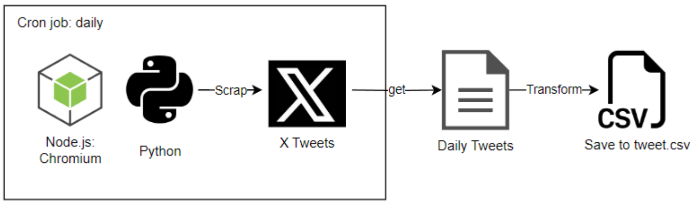
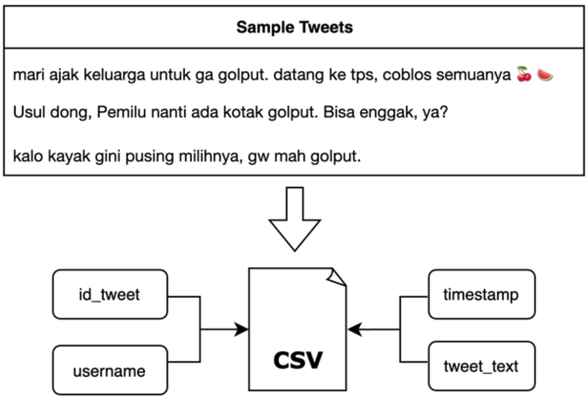


Figure 1. Data Collection on X

Figure 1 illustrates that the crawling process is automatically conducted every day. Subsequently, this data is compiled into a single document in CSV format and saved on the researcher's local file system. The collected data pertains to tweets mentioning "golput" words. An example of this data can be observed in Figure 5. In detail, the collected data includes tweet ID, timestamp, username, and the text of the tweet.


Sample Tweets

mari ajak keluarga untuk ga golput. datang ke tps, coblos semuanya 🍒🍉

Usul dong, Pemilu nanti ada kotak golput. Bisa enggak, ya?

kalo kayak gini pusing milihnya, gw mah golput.

Figure 2. Collected Data stored in CSV

2. Data Labeling

After the data collection process, the data is consolidated into a set of tweets ready for labeling. The labeling process involves using "Pro" and "Contra," where "Pro" refers to individuals who are in favor of abstaining from voting (“golput”), and "Contra" refers to those who are against abstaining from voting. "Neutral" serves as the middle point, where tweets do not have any specific inclination, such as unclear tweets, short tweets, or tweets that are questions.

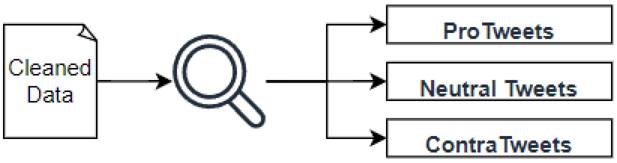


Figure 3. Data Labeling Process

3. Preprocess Data

After labelling the data, the next step is to pre-process the data. In this process, the data is prepared for machine learning. Data cleaning is part of data preprocessing, transforming raw data into usable data. According to Clark, data preprocessing is the stage where we can identify anomalies in our data [18]. The data preprocessing stage includes cleaning the data of tags or specific characters. The steps in data pre-processing are as follows:

a. lowercase or uppercase. In this study, lowercase is used. The target feature will be converted to binary, 1 for pro-stance and 0 for contra-stance.
b. Deduplication: This step involves removing duplicate data entries.
c. Null Removal: Null removal entails deleting data entries that are entirely empty due to previous cleaning processes.
d. Stemming: Stemming is the process of reducing words to their base form. For this research, stemming is performed using the Python library Sastrawi, as the data is in the Indonesian language.
e. Stopword Removal: It entails removing words that are deemed not significant for analysis. A list of stopword is used from the Sastrawi library, along with some customized words that are relevant to the current state of tweets in Indonesia.
f. Tokenization: Tokenization is the process of splitting text into individual tokens or words.

For a more detailed explanation, please refer to Figure 4:

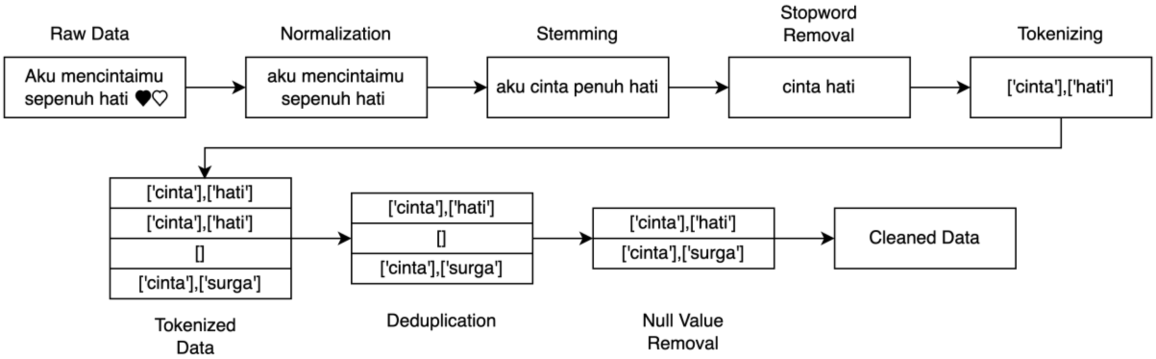


Figure 4. Data Preprocessing

4. Feature Extraction

In this stage, features from the text are extracted. This may involve techniques such as keyword extraction, word frequency counting, or text vectorization using the TF-IDF (Term Frequency-Inverse Document Frequency) method. The TF-IDF process is as follows:

a. TF (Term Frequency): It measures how often a keyword appears in a specific document. The more frequently the keyword appears, the higher its value.
b. IDF (Inverse Document Frequency): It measures how unique or rare the keyword is within the collection of documents. Keywords that appear in many documents have a low IDF value, while keywords that appear in few documents have a high IDF value. The IDF value helps in identifying the importance of the keyword across the entire collection of documents.
c. TF-IDF: To combine information from TF and IDF, we multiply TF by IDF:

$$\text{TF IDF}_{(\text{term,document,corpus})} = \text{TF}_{(\text{term,document})} \text{ x } \text{IDF}_{(\text{term,corpus})}$$

Equation 1. The Multiplication Formula of TF-IDF

This value provides a relative score for a keyword within a specific document compared to the entire collection of documents.

d. Document Ranking: Documents with high TF-IDF scores for specific keywords are considered more relevant to those keywords. Therefore, this method is used in search engines, text classification, and text analysis to identify the most relevant documents for a given keyword.

5. Stance Analysis

In this stage, after data has been cleaned and undergone feature extraction, it is ready for the classification process. The goal is to predict tweets as either pro, neutral, or contra stance. The algorithms used are Random Forest, Decision Tree, Logistic Regression, Support Vector Machine, K-Nearest Neighbor, and Gradient Boosting. The library used is sklearn from Python, which will import classification using the mentioned algorithms. For a more detailed view, we can refer to Figure 5.

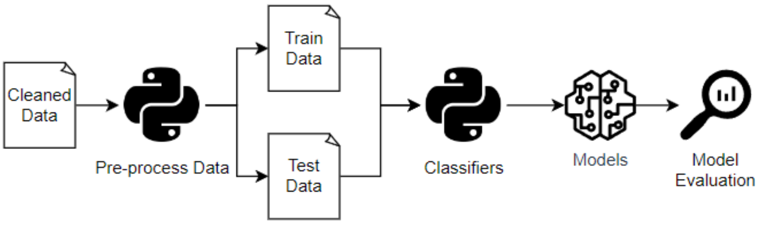


Figure 5. Stance Analysis Process

6. Topic Modelling

To identify the topics related to both the pro and contra stances on the issue of abstaining from voting (golput), we can use topic modelling techniques. The topic modelling process uses the LDA (Latent Dirichlet Allocation) algorithm available in the Gensim library in the Python programming language. We can observe the topic modelling process in Figure 6.

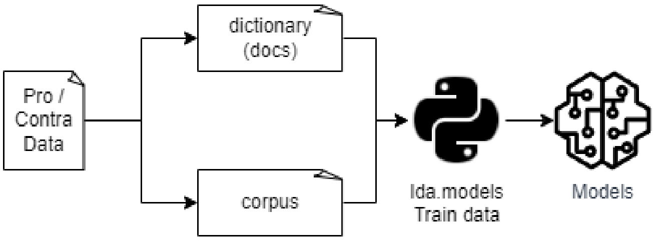


Figure 6. Topic Modelling Process

The cleaned data will be processed using LDA with the Python library. The cleaned tweets will be divided into two components:

a. Dictionary: This contains the indexing of the cleaned tweets, which aids in the analysis process.
b. Corpus: This is created using the corpora library and contains the words found within the tweets. It simplifies the analysis process.

Subsequently, all this data will be repeatedly trained until the appropriate model is found. We can observe the model formation process in Figure 7.

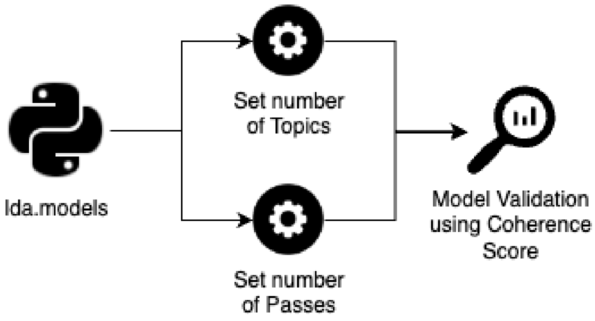


Figure 7. Model Formation Process

In the process of forming a model, it often involves iterative steps to achieve the best model. In the case of the LDA, coherence is used as a reference. Higher coherence values indicate a better-performing model [15]. To determine the appropriate number of topics, the researcher typically starts with a range from 1 to 10 and iterates through 100 iterations to achieve optimal results.

During each iteration, a coherence score is calculated. Each model is run, and the results are documented until the best result is obtained. This approach helps in selecting the number of topics that best fit the data and the model with the higher coherence, indicating a good model fit.

C. Result and Discussion

1. Data Collection

The data were obtained from social media X using the scraping method and resulted in 9,045 tweets related to undecided voters using the term "golput" (abstention). From this data, reviews were selected for manual stance annotation, which was conducted by 2 annotators. The manual annotation process yielded 2,566 pro stances, 5,264 neutral stances, and 1,215 contra stances.

2. Data Preprocessing

The collected data will undergo preprocessing stages such as normalization, deduplication, null value removal, stemming, stop word removal, and tokenizing. The neutral target feature will be removed initially for the stance analysis process, resulting in the utilization of 3.781 data points containing pro and contra stances.

3. Stance Analysis

This process employs 3.781 manually annotated data points. The tweets segmented into multiple sentences for modelling purposes, such as unigrams, bigrams, and trigrams. Subsequently, k-fold cross-validation used, where *k=5* divide the data into training and testing sets. The classification models used are Random Forest, Decision Tree, Logistic Regression, Support Vector Machine, K-Nearest Neighbor, and Gradient Boosting. The classification outcomes will be analyzed by comparing the accuracy, precision, recall, and F1-score results based on their algorithms and n-grams. The results of the stance analysis can be observed in Table 1.

Table 1. Classifier Performance Result

Classifiers	N-grams	Acc	Prec	Rec	F1
Random Forest	Unigram	**0.84**	**0.83**	0.95	0.88
	Bigram	0.82	0.81	0.96	0.88
	Trigram	0.76	0.73	1.00	0.85
Decision Tree	Unigram	0.76	0.81	0.83	0.82
	Bigram	0.83	0.83	0.93	0.88
	Trigram	0.77	0.74	1.00	0.85
Logistic Regression	Unigram	0.82	0.82	0.93	0.87
	Bigram	0.83	0.80	0.98	0.88
	Trigram	0.75	0.73	1.00	0.84
Support Vector Machine	Unigram	0.83	0.81	**0.98**	**0.89**
	Bigram	0.81	0.78	0.99	0.87

	Trigram	0.74	0.72	1.00	0.83
K-Nearest Neighbor	Unigram	0.78	0.76	0.97	0.86
	Bigram	0.75	0.73	0.99	0.84
	Trigram	0.70	0.69	1.00	0.82
Gradient Boosting	Unigram	0.81	0.79	0.98	0.87
	Bigram	0.80	0.78	0.97	0.86
	Trigram	0.71	0.70	1.00	0.82

According to Table 1, Random Forest achieved the highest accuracy and precision scores, with values of 84% and 83%, respectively. As for recall, all classifiers using trigrams obtained a recall value of 100%. However, the best results for recall and F1-score were attained by SVM, scoring 98% and 89%, respectively.

Table 2. Stance Classification Result

Label	Total	Percentage
Pro	3.112	30.17%
Neutral	5.622	54.50%
Contra	1.581	15.33%

4. Topic Analysis

After conducting stance analysis, topic modelling was performed on pro and contra tweets. The aim of this topic modelling was to identify the primary topics influencing an individual's inclination towards either pro or contra stances regarding abstaining from voting (golput). Latent Dirichlet Allocation (LDA) was employed in this study and validated using coherence scores. A higher coherence score indicates a better model. The initial testing involved determining the optimal number of topics based on the coherence value. The results of the experiment on the number of topics can be observed in Table 3.

Table 3. Topics and Coherence Score

Topics	**Coherence score**
1	0.330474
2	0.371572
3	0.367612
4	0.318945
5	0.340520
6	0.363217
7	0.374229
8	0.349403
9	0.380311
10	0.356187

After obtaining the most optimal topic results, topic modelling was conducted to identify the main topics within the pro and contra stances regarding abstaining from voting. This can be observed in Table 4 and Table 5.

Table 4. Main Topic of Pro Stance

Topic 1		Topic 2		Topic 3		Topic 4	
Word	**Prob**	**Word**	**Prob**	**Word**	**Prob**	**Word**	**Prob**
golput	0.119	golput	0.127	golput	0.242	golput	0.063
pilih	0.017	pilihan	0.031	kali	0.007	paslon	0.007
mending	0.017	mending	0.008	rakyat	0.006	coblos	0.007
prabowo	0.010	orang	0.007	mending	0.007	politik	0.007
suara	0.009	pemilu	0.006	pemilu	0.006	orang	0.007
ganjar	0.009	coblos	0.006	menang	0.005	hukum	0.006
maaf	0.006	calon	0.005	parpol	0.004	koruptor	0.005
tps	0.006	hak	0.005	dukung	0.004	mending	0.005
gibran	0.006	dukung	0.004	pusing	0.004	pemilu	0.005
capres	0.006	pilpres	0.004	putaran	0.004	rakyat	0.005

Table 4 shows 4 topics from the overall topics for the pro-abstention category. Each topic shows 10 words with the highest probabilities in each topic.

Table 5. Main Topic of Contra Stance

Topic 1		Topic 2		Topic 3		Topic 4	
Word	**Prob**	**Word**	**Prob**	**Word**	**Prob**	**Word**	**Prob**
golput	0.062	golput	0.024	golput	0.111	golput	0.061
pemilu	0.027	pemilu	0.011	pemilu	0.028	tolak	0.016
pemiludamai	0.012	warga	0.006	pilih	0.027	prabowo	0.012
tolak	0.011	politik	0.005	hak	0.017	keren	0.009
pemilumenuju indonesiamaju	0.009	nyoblos	0.005	suara	0.011	anies	0.008
pemiludamai negerimaju	0.009	demokrat	0.004	polri	0.009	ganjar	0.007
pilih	0.007	suara	0.004	hoax	0.008	suara	0.007
hoaks	0.007	sosialisasi	0.004	indonesia	0.008	dukung	0.007
pilpres	0.007	kpumelayani	0.004	pilihan	0.008	jokowi	0.007
pemilu bebashoaks	0.006	tps	0.004	muda	0.006	pilihan	0.006

Similar to the results of the pro-abstention category, Table 5 shows the top 10 words with the highest probabilities in each topic. In both the pro-abstention and contra-abstention categories, the words "*golput*" is the word with the highest probabilities, indicating that these word frequently appear in the set of tweets.

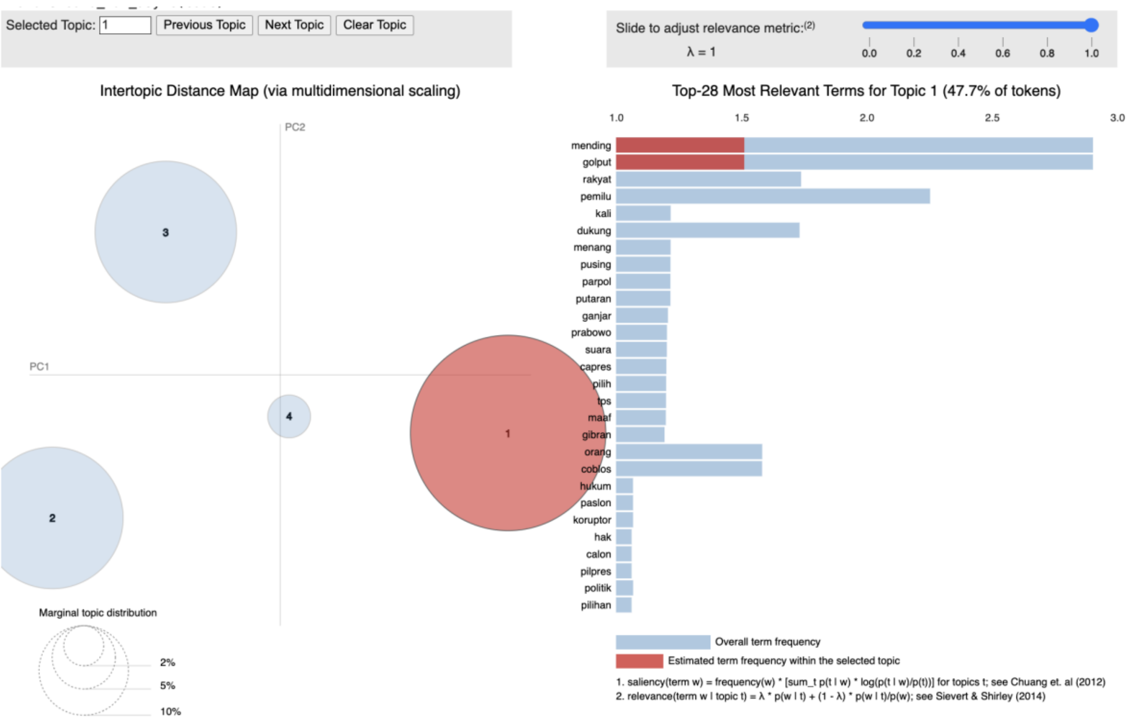

Selected Topic: 1
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0 0.2 0.4 0.6 0.8 1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
Marginal topic distribution
2%
5%
10%
Top-28 Most Relevant Terms for Topic 1 (47.7% of tokens)
1.0 1.5 2.0 2.5 3.0
mending
golput
rakyat
pemilu
kali
dukung
menang
pusing
parpol
putaran
ganjar
prabowo
suara
capres
pilih
tps
maaf
gibran
orang
coblos
hukum
paslon
koruptor
hak
calon
pilpres
politik
pilihan
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 8. LDA Visualization for Pro Stance

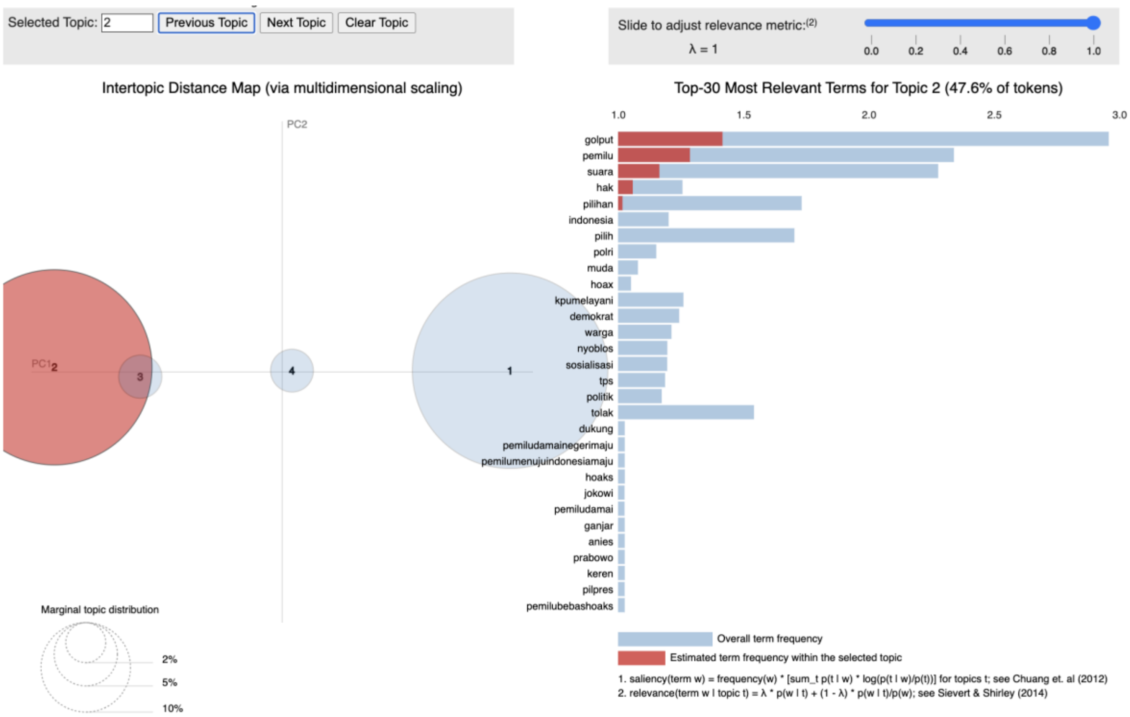

Selected Topic: 2
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0 0.2 0.4 0.6 0.8 1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 2 (47.6% of tokens)
1.0 1.5 2.0 2.5 3.0
golput
pemilu
suara
hak
pilihan
indonesia
pilih
polri
muda
hoax
kpumelayani
demokrat
warga
nyoblos
sosialisasi
tps
politik
tolak
dukung
pemiludamainegerimaju
pemilumenujuindonesiamaju
hoaks
jokowi
pemiludamai
ganjar
anies
prabowo
keren
pilpres
pemilubebashoaks
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 9. LDA Visualization for Contra Stance

After conducting stance analysis and topic modelling on the collected tweets, several insights can be gleaned. Random Forest and SVM are the best classifiers for performing stance analysis, as evidenced by their high evaluation results. In this study, the classifier with the highest level of accuracy was selected with the aim of obtaining the most accurate stance prediction results. Therefore, Random Forest was used as the classification model to achieve maximal results with an accuracy rate of 84%.

If neutral responses are disregarded, a higher proportion of people in the community are pro-abstention (golput). After analyzing 10,000 data points, tweets categorized as neutral were the majority, accounting for 54.5%. However, it is challenging to draw a conclusion, as neutral can imply tweets that are non-partisan, or irrelevant to golput. Alternatively, neutral could also indicate that someone might shift to being pro or against in the future. Therefore, the researchers chose to ignore the neutral category and focus on the pro and contra stances. Analysis in Table 2 reveals that the majority are more inclined to be pro-abstention, accounting for 30.17%, which is significantly higher than the anti-abstention at 15.33%. This suggests that in the 2024 election, those who voice opinions about golput are likely to choose not to vote for the president and vice president.

There are four main topics in the pro and anti-abstention stance. In the pro-abstention category, the first topic discusses presidential candidates, the second topic discusses voting rights, the third topic discusses election dynamics, and the fourth topic discusses corruption. These four topics indicate that the public is reluctant to vote or chooses to abstain due to low trust in the presidential and vice-presidential candidates. This is caused by numerous indications of corruption or legal issues. The tweets also show that the public feels dissatisfied with the pairs of candidates, leaving the public confused about making their choice.

The first topic in the anti-abstention stance tends to emphasize rejecting abstention and creating a fair and integrity-filled election atmosphere for national development. In line with the first topic, the second topic discusses using the right to vote and the importance of socialization from the General Election Commission (KPU) to increase public participation. The third topic emphasizes using the right to vote, especially among young people, and avoiding hoax news that influences the election. The fourth topic discusses support for candidate nominees. Therefore, the entire society is encouraged to vote and use their voice to elect national leaders in Indonesia.

D. Conclusion

To answer the research question, the researchers undertook several steps, starting from stance analysis and topic analysis of tweets expressing abstention from voting (golput). The majority of the public seemed neutral regarding golput, but 30.17% expressed a pro stance, signifying a tendency for some individuals to opt for golput, which accounted for 15.33% of the population. The primary discussion topic among those supporting abstention is the lack of trust in the government, particularly concerning corruption issues. Other topics mentioned the public's distrust in all candidates, be it Anies, Prabowo, or Ganjar, leading people to choose pro-golput. On the contrary, tweets opposing golput emphasized that elections are the right of every citizen and should be conducted properly.

The limitation of this research lies in the fact that the tweets were collected based on the last three months' data. Therefore, future improvements in research could extend the timeframe to gather a more extensive collection of tweets.

Further exploration into various topics via modelling could significantly enhance the overall quality of the study.

E. References

[1] Pemerintah Indonesia, *Undang Undang Republik Indonesia Nomor 7 Tahun 2017 tentang Pemilihan Umum*. Indonesia: Lembaran Negara Republik Indonesia Tahun 2017 Nomor 182, 2017.

[2] Badan Pusat Statistik, “Statistik Politik 2019,” Jakarta, Dec. 2019.

[3] B. Arianto and R. Ali Haji, “ANALISIS PENYEBAB MASYARAKAT TIDAK MEMILIH DALAM PEMILU,” *Jurnal Ilmu Politik dan Ilmu Pemerintahan*, vol. Vol 1, pp. 51–60, 2011.

[4] E. Rosenberg *et al.*, “Sentiment analysis on Twitter data towards climate action,” *Results in Engineering*, vol. 19, Sep. 2023, doi: 10.1016/j.rineng.2023.101287.

[5] P. Gunhal *et al.*, “Stance Detection of Political Tweets with Transformer Architectures,” in *International Conference on ICT Convergence*, IEEE Computer Society, 2022, pp. 658–663. doi: 10.1109/ICTC55196.2022.9952951.

[6] S. Hisamitsu, S. Cho, H. Jin, M. Toyoda, and N. Yoshinaga, “Diachronic Analysis of Users' Stances on COVID-19 Vaccination in Japan using Twitter,” in *Proceedings of the 2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 237–241. doi: 10.1109/ASONAM55673.2022.10068695.

[7] T. Alqurashi, “Stance Analysis of Distance Education in the Kingdom of Saudi Arabia during the COVID-19 Pandemic Using Arabic Twitter Data,” *Sensors*, vol. 22, no. 3, Feb. 2022, doi: 10.3390/s22031006.

[8] Q. G. To *et al.*, “Anti-vaccination attitude trends during the COVID-19 pandemic: A machine learning-based analysis of tweets,” *Digit Health*, vol. 9, Jan. 2023, doi: 10.1177/20552076231158033.

[9] H. Hassani, C. Beneki, S. Unger, M. T. Mazinani, and M. R. Yeganegi, “Text mining in big data analytics,” *Big Data and Cognitive Computing*, vol. 4, no. 1, pp. 1–34, Mar. 2020, doi: 10.3390/bdcc4010001.

[10] M. R. Shamsuddin, S. Abdul-Rahman, and A. Mohamed, “Exploratory analysis of MNIST handwritten digit for machine learning modelling,” in *Communications in Computer and Information Science*, Springer Verlag, 2019, pp. 134–145. doi: 10.1007/978-981-13-3441-2_11.

[11] Y. Kang, Z. Cai, C. W. Tan, Q. Huang, and H. Liu, “Natural language processing (NLP) in management research: A literature review,” *Journal of Management Analytics*, vol. 7, no. 2. Taylor and Francis Ltd., pp. 139–172, Apr. 02, 2020. doi: 10.1080/23270012.2020.1756939.

[12] S. Fareri, G. Fantoni, F. Chiarello, E. Coli, and A. Binda, “Estimating Industry 4.0 impact on job profiles and skills using text mining,” *Comput Ind*, vol. 118, p. 103222, 2020, doi: https://doi.org/10.1016/j.compind.2020.103222.

[13] A. Bechini, P. Ducange, F. Marcelloni, and A. Renda, “Stance Analysis of Twitter Users: The Case of the Vaccination Topic in Italy,” *IEEE Intell Syst*, vol. 36, no. 5, pp. 131–139, 2021, doi: 10.1109/MIS.2020.3044968.

[14] J. Sainz-Santamaria, D. Moctezuma, A. L. Martinez-Cruz, E. S. Téllez, M. Graff, and S. Miranda-Jiménez, "Contesting views on mobility restrictions in urban green spaces amid COVID-19—Insights from Twitter in Latin America and Spain," *Cities*, vol. 132, p. 104094, 2023, doi: https://doi.org/10.1016/j.cities.2022.104094.

[15] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The Author-Topic Model for Authors and Documents," 2004.

[16] D. M. Blei, "Probabilistic topic models," in *Communications of the ACM*, Apr. 2012, pp. 77–84. doi: 10.1145/2133806.2133826.

[17] R. Alghamdi and K. Alfalqi, "A Survey of Topic Modelling in Text Mining," *International Journal of Advanced Computer Science and Applications*, vol. 6, Mar. 2015, doi: 10.14569/IJACSA.2015.060121.

[18] J. J. Davis and A. J. Clark, "Data preprocessing for anomaly based network intrusion detection: A review," *Computers and Security*, vol. 30, no. 6–7. pp. 353–375, Sep. 2011. doi: 10.1016/j.cose.2011.05.008.