TEMATICS | Technology Management and Informatics Research Journals
E-ISSN : 2798-530X

POLITEKNIK IMIGRASI
POLTEKIM

TWITTER SOCIAL NETWORK ANALYSIS AND SENTIMENT IDENTIFICATION OF "*VAKSIN BOOSTER*" KEYWORD

https://10.0.205.137/tematics.v6i2.652
Submitted: 11-12-2024 Reviewed: 11-11-2024 Published: 13-12-2024

Muhammad Fahrury Romdendine
romdendine@poltekim.ac.id
Immigration Polytechnique

Okky Pratama Martadireja
okkypm@poltekim.ac.id
Immigration Polytechnique

Alif Sofa Danutirta
alifsofa@gmail.com
Correctional Sciences Polytechnique

Mitsal Shafiq Sulasno
mitsal@gmailc.om
Correctional Sciences Polytechnique

Abstract. *The low acceptance level and limited coverage of booster vaccines, despite their critical importance for public health, highlight the need for deeper insights into societal perceptions and behaviors. Social networks, as a significant medium for information dissemination, offer a valuable opportunity to understand public discourse and identify influential factors. This study leverages graph topology analysis to map and analyze the dynamics of vaccine-related discussions within social networks. By identifying key individuals who play pivotal roles in spreading booster vaccine information, the analysis reveals the structure and flow of information within the network. Furthermore, sentiment analysis indicates that neutral interactions dominate these discussions, followed by negative and positive sentiments. Notably, the neutral content largely pertains to travel procedures, which aligns with the "mudik" tradition during the data collection period. These findings provide a framework for understanding the sociotechnical landscape of vaccine acceptance and offer actionable insights for designing targeted, effective strategies to enhance booster vaccine uptake.*

Keywords: social network analysis, sentiment identification, booster vaccine, tweet data.

Abstrak. *Tingkat penerimaan dan cakupan vaksin booster yang rendah, meskipun sangat penting untuk kesehatan masyarakat, menunjukkan perlunya pemahaman lebih mendalam mengenai persepsi dan perilaku masyarakat. Jaringan sosial, sebagai media penting untuk penyebaran informasi, menawarkan peluang berharga untuk memahami wacana publik dan mengidentifikasi faktor-faktor yang berpengaruh. Penelitian ini memanfaatkan analisis topologi graf untuk memetakan dan menganalisis dinamika diskusi terkait vaksin dalam jaringan sosial. Dengan mengidentifikasi individu-individu kunci yang berperan penting dalam penyebaran informasi vaksin booster, analisis ini mengungkap struktur dan aliran informasi dalam jaringan tersebut. Selain itu, analisis sentimen menunjukkan bahwa interaksi netral mendominasi diskusi ini, diikuti oleh sentimen negatif dan positif. Konten netral terutama berhubungan dengan prosedur perjalanan, yang relevan dengan tradisi "mudik" yang terjadi selama periode pengumpulan data. Temuan ini memberikan kerangka kerja untuk memahami lanskap sosial-teknis penerimaan*

vaksin dan menawarkan wawasan yang dapat ditindaklanjuti untuk merancang strategi yang lebih efektif guna meningkatkan penerimaan vaksin booster

Keywords: analisis jejaring sosial, identifikasi sentimen, data *tweet*, vaksin *booster*.

1. INTRODUCTIONS

Third dosage of COVID-19 vaccine or widely known as booster vaccine has become an emerging issue since the government issued a regulation regarding travel procedure that endorses passenger to be vaccinated up to third dosage. Government primary reason to endorse people to get third dosage is to maintain an effective level of immunity in the population (Wagner et al., 2022). While the acceptance level of primary dosage is about 65%, the accepatence level of booster vaccine is found to be lower (Benny et al., 2022). However, if the booster vaccine mandated for work, travel, or public activities, the acceptance level increased slightly (Benny et al., 2022). Even the acceptance level could be measured via survey to selected respondent using survey method, the bigger picture of information spreading across country need to be analysed. The need to identify which person plays an important role in information spreading become urgent since the coverage of booster vaccine is only 22.14% of the target population by the first of June 2022 (Indonesian Ministry of Health, 2022). Information that capture pattern of information spreading and interaction happened in social media could be used by stakeholder to plan effective social media campaign regarding booster vaccine (Yum, 2020). Thus, the campaign would push the coverage level of the booster vaccine indirectly.

While the pattern of information spreading could be captured using social network analysis, the public sentiment about the issue could be analysed using sentiment analysis. The previous work from (Yum, 2020) has succeed to identify the key person in COVID-19 related topics among US people using graph analysis. A study by (Rustam et al., 2021) compared various machine learning model to classify people sentiment pertaining COVID-19 using social media text data. The previous work from (Lyu et al., 2021) shows that tweet data from Twitter could be enriched using thorough exploratory data analysis (EDA) and then be used as features to train predictive model and then predict weather someone would use controvercial terms or not in his/her tweet. Though the work was good, there is no graph model that capture the information spreading pattern.

In Indonesia alone, there has been research about social network analysis and sentiment analysis of people in social media Twitter to capture the polarization occured among society during presidential election in 2019 (Nur Habibi & Sunjana, 2019). Using similar method, this research performed to identify key person, information spreading patter, and interaction sentiment among networks of interaction regarding a topic about booster vaccine. The keywords “vaksin booster” was selected because it draws a clear intuition about the topic of interest. The insight from this research could be used by particular stakeholder to plan or make decision about any campaign related to booster vaccine as this kind of work is similar to that in domain of business (Zhou et al., 2019).

2. METHOD

This research was conducted in four main steps. First was data acquisition from Twitter API, second was data preparation, third was sentiment labelling, and the last was graph data preparation and analysis. Each main step could be done in several sub-steps. All programming language used was in Python and written in Jupyter Notebook as the text editor and analysis report. Graph visualizations and analysis were performed using Gephi.

2.1. DATA ACQUISITION

Data was acquired using open access API provided by Twitter by previosly applying for the access via https://developer.twitter.com/. API fetching was performed using Tweepy client in user's side. The parameters of which the acquisition was done are the tweet creation date, the username of the tweet's user, the tweet itself, the mentioned user inside the tweet, the user's followers count, and the mentioned user's followers count. Data fetching was done repeatedly from 11/05/2022 to 22/05/2022 due to 7-days-range limitation from Twitter policy for elevated access user.

The tweet that was fetched is limited to tweet related to "*vaksin booster*" keywords, and the tweet that has mentions of other users inside the tweet. It could be retweet, mention, or reply. This constraint was set to ensure that the retrieved data is suitable for the construction of graph data which rely on the interactions among users.

3.2. DATA PREPARATIONS

Data preparation was done in three steps. data merging, data filtering, and EDA. The purpose of data merging is to merge all datasets that was collected during 12 days tweet fetching.

After all the datasets had been merged, the next step is to filter the tweet redundancy and the uniqueness of the users. The author only proceeds a unique tweet from a unique user. This step was required since there is the possibility that daily tweet fetching was fetching the same tweet data from the same user. This will lead to erroneous interaction data. The EDA was performed from filtered data from previous step to enhance our understandings of the data.

3.3. SENTIMENT LABELLING

The sentiment labelling task was done using IndoBERT pre-trained NLP classification model (Koto et al., 2020). Since the main objective of this experiment was to analyze the social network or interactions among users in Twitter, the sentiment analysis step was not the main focus. This step was performed only to add additional attributes to the edge of the graph which is the sentiment of each interaction.

The text was cleaned first including punctuations removal, links removal, numbers removal, non-alphanumeric characters removal, and hashtags removal. All those removals were performed using regular expression filtering. After that, the cleaned text then labelled using the pre-trained explained above. The complete explanation of the method and architecture of the IndoBERT could be found in (Koto et al., 2020).

3.4. GRAPH PREPARATIONS AND ANALYSIS

To process the analysis and graph visualization using Gephi, the labelled tweet data from previous step was shaped to form node table and edge table data. Node table is the matrix of username from each user in the tweet data including the author of the tweet and the mentioned user inside the tweet. Additional information was the user's followers count. This additional information could later be utilized to enrich the visualization parameter. Edge table is the matrix of interactions among users in the tweet data that contain the source node, the target node, the sentiment of the interaction, and the score of the sentiment. The specification of each node table and edge table is presented in Table 1 and Table 2.

The next step was to import the node table and edge table, which was exported to csv files before, to Gephi using the "import from spreadsheet" menu. The first to import is node table and second is edge table. Both of the table are appended to one worksheet. After that, the visualization parameter was set to enhance the visualization of the networks. The parameter settings is presented in Table 3. The next was to layout the networks using layouting algorithm provided by Gephi. This layout step was trial and error. But, for social networks, suggested combination of the algorithm used was *Fruchterman-Reingold, OpenOrd, and ForceAtlas2* (Hansen et al., 2020; Zhuhadar & Yang, 2012). The last analysis performed in Gephi was graph topology analysis that would yield centrality measures of each nodes, giant component, ego-network of selected or interesting node, and k-core subgraph.

Table 1. Node Table Spesification

Field	**Description**
ID	The username of tweet author and the mentioned user inside the tweet
Followers	The followers count of the respective user

Table 2. Edge Table Specification

Field	**Description**
Source	Source of the tweet (author)
Target	Target of the tweet (mentioned user)
Sentiment	The tweet sentiment which annotate the interaction sentiment
Score	The score of the labelled sentiment

Table 3. Visualization Parameter Settings

Parameter	Setting
Node Size	Based on the degree of the node
Edge Color	Based on the interaction sentiment
Node Label Size	Based on the followers count

3. RESULTS

3.1. DATA ACQUISITIONS

After 12 days of data acquisition, 25,485 tweets had been collected. This tweet separated in 12 different datasets from each day tweet fetching. During the retrieval of tweet data, a scene was discovered in which there are users for whom we are unable to directly retrieve the user's follower count due to Twitter constraints. To handle that, an error handling code was added to the acquistion script.

3.2. DATA PREPARATIONS

All the 12 different datasets from each tweet data fetching merged into one big dataset and yielded a dataset with 25,485 rows and 6 columns. After filtering, the unique tweet data was only 1,682 tweets. The filtering step yielded a dataset with 1,682 rows and 6 columns.

After that, EDA was performed with filtered dataset. The first EDA was to seek the frequency of mentioned user in the tweet data. Thus, the user with highest frequency was the the user with most interaction. Fig. 1 shows the frequency of the mentioned user. “yuno_yejun” user was the most mentioned user followed by “DokterTifa” and “Mi9aWempy”. The next step was to see closer to each tweet where three top mentioned users were mentioned. Unfortunately, the most retweeted tweet which is from “yuno_yejun” was not in the context of COVID-19 booster vaccine event though the tweet contains “vaksin booster” keywords. Thus, we remove all the tweet related with “yuno_yejun” user.

The last EDA was word cloud visualization. The result is shown in Fig. 2. The most highlighted words was *“vaksin”*, *“booster”*, *“info”*, *“udah”*, and *“vaksinasi”* which are common and general words and from here we can deduce that most of the tweet would be a neutral tweet.

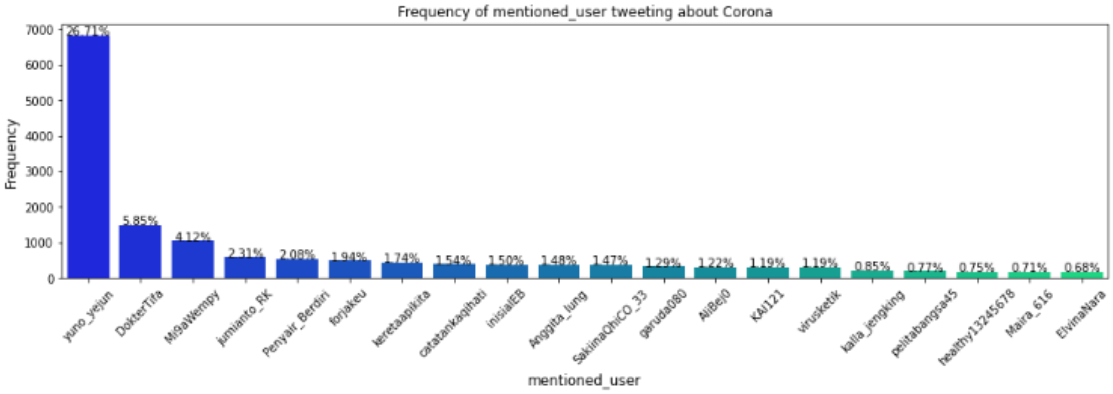


Fig 1. Most mentioned users frequency.

Fig 2. Word cloud of the tweets.

3.3. SENTIMENT LABELLING

The text cleaning step remove all hashtags, punctuations, non-alphanumeric characters, and links completely well. But, the common word *"RT"* which means *retweet* could not be remove due to lack of regular expression knowledge of the the researcher. This unmeaningful word might affect or not affect the sentiment classification results.

Normally, if the research focus was about sentiment analysis and classification, the tweet labelling was done by data annotator who annotate each tweet manually. But, in this research, the IndoBERT pre-trained NLP model by (Koto et al., 2020) was utilized and yielded a result of 66% neutral tweet, 10% positive tweet, and 24% negative tweet as shown in Fig 3. The word cloud visualization of each tweet sentiment was presented in Fig.4. All of the words highlighted in each word cloud visualization could not tell us a specific word that marked the tweet sentiment. Though, some words like *"teruji klinis"*, *"bangsa vaksin"*, and *"infrastruktur Jokowi"* exist in positive tweet. In neutral tweet, one of interesting highlighted word was *"info"*. Lastly, word cloud of negative tweet highlighted no related word.

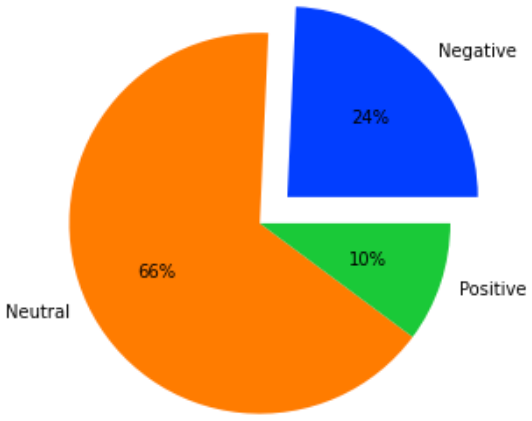


Fig 3. Sentiment labelling pie chart.

(a) (b)

(c)

Fig 4. Word cloud of negative sentiment (a), positive sentiment (b), and neutral sentiment (c).

3.4. GRAPH ANALYSIS

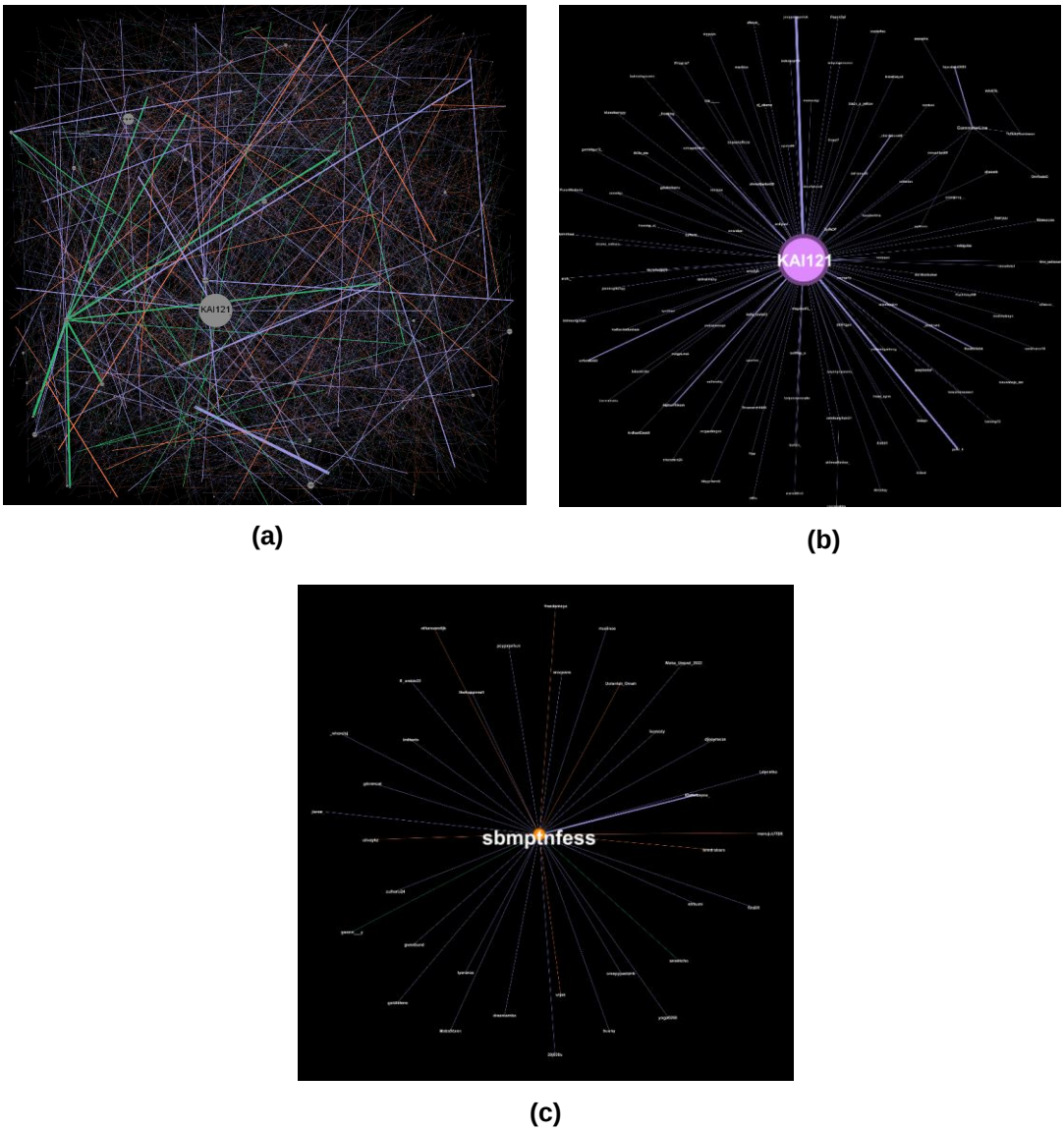
Graph data consists of two main elements which are node table and edge table. Both of that are formed using the labelled tweet data from previous step. The node table consists of 2,144 unique users exist in the tweets while the edge table consists of 1,680 edge from source to target node which later interpretated as interaction among users.

The all networks produced are shown in Fig. 5. The giant component sub-graph analysis yielded a networks with most connected node and most interactions. The giant component shown that user “KAI121” had the most interactions which interpreted as that this user mentioned or retweeted most of the time and could be stated as key person. The giant component networks are shown in Fig. 6 and the topology analysis of top five users in term of centrality measures are presented in Table 4.

The second largets sub-graph was the ego-network of *“sbmptnfess”* user. This user’s ego-network are shown in Fig. 9. Not like the giant component that had

only neutral interactions, this user's ego-network had all three sentiments inside the interactions. The sample tweet from positive and negative interaction is presented in Fig. 7.

Two user that considered as interesting users which are *"jokowi"* and *"detikcom"* are analysed. The ego-network of *"jokowi"* who are the president of Indonesia shown that this user only had small interactions and all the interactions are neutral. The visualization of *"jokowi"* ego-network along with "*detikcom*" is shown in Fig. 8. *"detikcom"* user also had a few number of interactions. But, the interations happened were in positive, neutral, and negative sentiment. The tweet sample from each sentiment is presented in Table 5.



(a) **(b)**

(c)

Fig 5. Networks visualizations of all tweets (a), giant component (b), and @sbmptnfess ego network.

Table 4. Giant Componenet Centrality Measures

Username	Followers	Betweeness Centrality	Closeness Centrality
KAI121	1025108	0.002647	0.888

nottelon	49	0.000136	0.497758
agilesss	12	0.000136	0.497758
sembunyihati21	1136	0.000048	0.476395
cyfranz	559	0.000048	0.476395

Table 5. Tweet Sample From Sbmptnfess Ego Network As Second Largest Sub-Graph

Username	Tweet	Sentiment	Score
olivsykz	OOT gua mau nanya gays emang klo utbk skrng harus bawa sertifikat vaksin ya trus kudu wajib booster apa gimane klo booster wah parah sih gua ga vaksin booster soalnya	negative	0.981185
ethanvandijk	Maaf oot yg pusat utbk UNS apa harus vaksin booster Gue ngecek di Instagram UNS kok ga pernah update kaya ginian Atau gue salah ngecek akun	negative	0.933749
gibrancat	RT ptn utbk di ui wajib vaksin booster kah	neutral	0.977619

(a) **(b)**

Fig 8. Ego network of @jokowi (a), and @detikcom (b).

Table 6. Tweet Sample from @detikcom Ego Network

Username	Tweet	Sentiment	Score
aditya_lasbor	2020 pada ogah2an disuruh pake masker 2022 pada g mau copot masker emang dasarnya rakyat indo susah diatur kyaknya kl dr awal nurut kn enak pk masker nurut diem di rumah nurut vaksin nurut booster nurut lepas masker nurut semua kebijakan kn udh ada pertimbangannya	negative	0.959789
lutfiabizar23	Vaksin booster aman dan halal	positive	0.963949
AdhiStanza	UAS udah vaksin booster ya	neutral	0.843691

4. DISCUSSIONS

There was a limitation in data acquisition process which is the limitation of tweet date range and maximum tweet fetched per day. This limitation happened because of the researher only had *elevated access* type of access to Twitter API. To tackle this issue, researcher did the data fetching in daily fashion. But, this approach had a drawback, namely the lack of diversity in the tweet data fetched. This happened because of the tweet from last day could be fetched again today or next day. This limitation leaded to huge number of fetched tweet data but only small numbers of them are unique. However, only unique tweet data that could be proceed to analysis.

There was also an issue inside the acquisition script which could not handle a tweet if there are more than one user mentioned in that tweet. The script only recorded the first mentioned user. While fetching the mentioned user's folower count, an issue emerged which is there were several users that are blocked or inactive preventing the user's follower count from being fetched. But, researcher could solve this issue by adding some error handling method inside the script.

Using *"vaksin booster"* as the keywords somehow is a tricky keywords. This was proven by the result from user *"yuno_yejun"* who told about booster vaccine but not in term of COVID-19 vaccine. Though tweet from *"yuno_yejun"* was only two tweet, this user was the most mentioned user. In other words, this user would be the giant component. If the manual tweet checking was not performed, the giant component analysis result might lead to a totally wrong way.

Visulization of the most common words found in the text using word cloud told us some hindered information. Before sentiment labelling, some highlighted words are natural word. This information then confirmed that after sentiment labelling the neutral class was the most common class. The word cloud of each sentiment class also told us some information regarding most common words of each sentiment. But, there was no unique word that could be used as the negative sentiment marker.

Furthermore, after the sentiment labelling, there was some concern regarding the negative sentiment class. This concern was based on the word cloud that did not highlighted any widely known words of negative sentiment. So, we pick some samples from negative tweet and read the text carefully. Some of the tweet are not showing direct rejection of booster vaccine. The fact that those tweets contain word that belong to negative class made IndoBERT labelled them as negative. But, using human intuition, there are hindered context that machine could not read. For example, tweet from *"aditya_lasbor"* in Table 6 labelled as negative. If we read the tweet carefully, the intention and the context of the tweet was not a direct rejection of the booster vaccine instead the tweet was a direct rejection of the people wrong attitudes during pandemic. Therefore, using IndoBERT might be a tricky way to automatically label the tweets.

The graph topology analysis could not directly applied to whole networks immediately after the networks are formed. This because of the whole netoworks is not a connected graph. The networks consist of many different sub-graph. Thus, we only perfomed topology analysis on the component we wanted to analyse. The giant component analysis shown us that user *"KAI121"* was the key person booster vaccine information spreading. All interactions in giant component

was neutral one. The tweets mostly talk about information regarding train passenger procedure. This happened because there was an event called *"mudik"* happened during the data acquisition where there are so many users interacted with this user to find information.

The other event occurred during the data acquisition was *"SBMPTN"* or *"UTBK"* which is an entrance test for Indonesian university. Thus, we found user *"sbmptnfess"* as second largest component or sub-graph in the networks. Unlike the giant component interaction, the ego-network of this user shown two different type of interaction which was neutral and negative. Interestingly, the negative tweet was not harmful for the information spreading of vaccine booster, instead a subtle rejection from the tweet writer about some procedures during the *SBMPTN*/*UTBK*.

Lastly, we analyse the user who are widely known in Indonesia which are the president and a news channel. User *"jokowi"* who are President Joko Widodo had only small networks so we could not conclude that first person in this country as important person regarding booster vaccine information spreading. The news channel *"detikcom"* also had few interactions. Some of negative interactions shown that the tweet had similar context with what we have discussed earlier. Thus, we found that in this relatively small tweet data there is no direct rejection or obvious negative sentiment exist regarding booster vaccine information spreading. But, we were succeed to identify who were the key persons and what was the event behind it so that the stakeholders could use this information to effectively plan information campaign.

5. CONCLUSION

Tweet data from twitter could be used to uncover hinder information about how people react to some events, topics, or issues in society. A computational method to model the interactions happened in social networks using graph theory approach might be an effective way to reveal information spreding pattern. Event though the tweet data did a good information capture, the limitation of date range and maximum number of tweet might lead to a poor analysis due to a minimum scope of information. Therefore, our analysis shown that key person in information spreading did exist. The interaction modelling using the sentiment of each interaction might be a good way to reveal how peoples are polarized, segmented, and segregated regarding a topic or issue of interest. The last, due to the limited number of nodes and edge formed and the networks also consist of many unconnected components, the graph topology analysis became less important than it should be. Therefore, for the next experiment it is a must to gather more data and design other related keywords so that the information spreading pattern could be captured finer than what we done.

REFERENCES

Benny, G., Wirawan, S., Putu, N., Harjana, A., Nugrahani, N. W., & Januraga, P. P. (2022). *Health Beliefs and Socioeconomic Determinants of COVID-19 Booster Vaccine Acceptance : An Indonesian Cross-Sectional Study*. 1–14.

Hansen, D. L., Shneiderman, B., Smith, M. A., & Himelboim, I. (2020).

Installation, orientation, and layout. In D. L. Hansen, B. Shneiderman, M. A. Smith, & I. Himelboim (Eds.), *Analyzing Social Media Networks with NodeXL* (Second Edi, pp. 55–66). Elsevier. https://doi.org/10.1016/B978-0-12-817756-3.00004-2

Indonesian Ministry of Health. (2022). *National COVID-19 Vaccination*.

Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). *IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP*. 757–770. https://doi.org/10.18653/v1/2020.coling-main.66

Lyu, H., Chen, L., Wang, Y., & Luo, J. (2021). Sense and Sensibility: Characterizing Social Media Users Regarding the Use of Controversial Terms for COVID-19. *IEEE Transactions on Big Data*, *7*(6), 952–960. https://doi.org/10.1109/TBDATA.2020.2996401

Nur Habibi, M., & Sunjana. (2019). Analysis of Indonesia Politics Polarization before 2019 President Election Using Sentiment Analysis and Social Network Analysis. *International Journal of Modern Education and Computer Science*, *11*(11), 22–30. https://doi.org/10.5815/ijmecs.2019.11.04

Rustam, F., Khalid, M., Aslam, W., Rupapara, V., Mehmood, A., & Choi, G. S. (2021). A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis. *PLOS ONE*, *16*(2), e0245909. https://doi.org/10.1371/journal.pone.0245909

Wagner, C. E., Saad-Roy, C. M., & Grenfell, B. T. (2022). Modelling vaccination strategies for COVID-19. *Nature Reviews Immunology*, *22*(3), 139–141. https://doi.org/10.1038/s41577-022-00687-3

Yum, S. (2020). Social Network Analysis for Coronavirus (COVID-19) in the United States. *Social Science Quarterly*, *101*(4), 1642–1647. https://doi.org/10.1111/ssqu.12808

Zhou, Q., Xu, Z., & Yen, N. Y. (2019). User sentiment analysis based on social network information and its application in consumer reconstruction intention. *Computers in Human Behavior*, *100*, 177–183. https://doi.org/10.1016/j.chb.2018.07.006

Zhuhadar, L., & Yang, R. (2012). Cyberlearners and learning resources. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, 65–68. https://doi.org/10.1145/2330601.2330621