What do Indonesians talk when they talk about COVID-19 Vaccine: A Topic Modeling Approach with LDA

Theresia Ratih Dewi Saputri[1], Caecilia Citra Lestari[2], Salmon Charles Siahaan[3]

[1,2]*School of Information Technology, Universitas Ciputra Surabaya, Indonesia*
[3]*School of Medicine, Universitas Ciputra Surabaya, Indonesia*

[1]theresia.ratih@ciputra.ac.id, [2]cecilia.citra@ciputra.ac.id,
[3]charles.siahaan@ciputra.ac.id

Abstract - To end the COVID-19 pandemics, the government attempted to accelerate the vaccination through various programs and collaboration. Unfortunately, the number is still relatively small compared to the number of populations in Indonesia. There are some reasons attributed to this challenge, one of them being the reluctance of citizens to accept the COVID-19 vaccine due to various factors. Knowing this factor to increase public compliance, the vaccination program can be speed-up. Unfortunately, traditionally acquiring the knowledge related to COVID-19 vaccine rejection can be challenging. One of the ways to capture the knowledge is by conducting a survey or interview related to COVID-19 vaccine acceptance. This method can be inefficient in terms of cost and resources. To address those problem, we propose a novel method for analyzing the topics related to the COVID-19 Indonesians' opinions on Twitter by implementing topic modeling algorithm called Latent Dirichlet Allocation. We gathered more than 22000 tweets related to the COVID-19 vaccine. By applying the algorithm to the collected dataset, we can capture the what is general opinion and topic when people discuss about COVID-19 vaccine. The result was validated using the labeled dataset that have been gathered in the previous research. Once we have the important term, the strategy based on can be determined by the medical professional who are responsible to administer the COVID-19 vaccine.

Keywords: COVID-19, vaccine, topic modeling, LDA

I. INTRODUCTION

The first COVID-19 case in Indonesia was reported in the early March 2020. The report stated in [1] shows that COVID-19 cases increased to 1285 in almost entire provinces in Indonesia. The World Health Organization (WHO) declared the COVID-19 pandemic in March 2020 [2]. Unfortunately, the number of cases has been reported until these days, as of January 2022. The situation aggravated with all the reported new COVID-19 variants. Tregoning et.al. reported the efficacy of existing vaccines based on virus variants [3]. The study shows that the reported vaccines have more than 50% efficacy. As the new variants surfaced, it affected the effort to reduce the number of cases and developed vaccines and treatments [4].

In the attempt to end the pandemic, several approaches have been implemented in Indonesia including the vaccination programs. As of January 2022, 295M vaccine doses were given and 40.9M citizens had been fully vaccinated [5]. However, this number is relatively small compare with Indonesia's population because only 14.9% of population was fully vaccinated. There is an issue related to the lack of available vaccines. Fuady et.al in [6] argued that the increase target allocation and vaccines' availability is necessary to reduce the number of reported COVID-19 cases. Another significant problem in vaccination program is the reluctance of citizens to accept the COVID-19 vaccine [7-8]. Even though the acceptance is relatively high [9], ignoring the citizens who refuse to get vaccination was not a wise strategy. Therefore, it is important to understand the reason for this rejection such as religion beliefs, political interests, health issues. The vaccination program can be accelerated by knowing this factor to increase public compliance.

Unfortunately, traditionally acquiring the knowledge related to COVID-19 vaccine rejection can be challenging. One of the ways to capture the knowledge is by conducting a survey or interview related to COVID-19 vaccine acceptance. This method can be inefficient in terms of cost and resources. Another problem related to direct interviews is that people' openness to the interviewers when they have objection to a certain topic. Therefore, this study utilized the social media dataset, Twitter, to gather public opinion related to COVID-19 vaccines. Various studies show that social media can be an effective source for sharing information, knowledge, and opinion [10]–[12]. Even though there are plentiful benefit of using social media dataset, analyzing those data is not a trivial task. Manually analyzing the captured data can be costly and error prone.

To address those problem, we propose a novel method for analyzing the topics related to COVID-19 Indonesians' opinions on Twitter by implementing topic modeling algorithm called Latent Dirichlet Allocation (LDA). Topic modeling has been studied in the past decades [13-14]. It is one of the analytical methods in text evaluation especially for identifying the topics. The source of topic modeling can come from different forms such as text, image, video, and geospatial data. This research evaluates textual data on public opinions stated through the Twitter platform. With the help of topic modeling, the time-consuming iterations to determine the latent variables from a large dataset can be reduced with the consideration of overcoming the uncertainty [15].

Due to its remarkable benefits, topic modeling has been used to retrieve information in various domain application. In the area of social science, Schmiedel et.al. in [16] used topic modeling to determining the inquiry strategy to understand the organizational culture. Meanwhile, Chen et.al. in [17] utilized topic modeling to analyze the stock market in China based on social media data. This study shows that topic modeling is useful for understanding human behavior based on their posted opinion about the Chinese stock market. Porturas and Tylor in [18] proposed a method for using topic modeling to analyze the trends in emergency medicine research. With the topic modeling approach, they discovered 40 latent topics research abstract over the last 40 years. With the ongoing COVID-19 pandemic, topic modeling was applied for various purposes. Boon-Itt and Skunkan in [19] conducted topic modeling study on the public sentiment related to the COVID-19 pandemic based on collected Twitter Datasets. Melo and Figueiredo [20] attempted to understand the COVID-19 phenomena and the impact by comparing Brazil's news and public opinion.

One of the widely known topic modeling topic algorithms is LDA. This method was proposed by Blei et al. in 2003 [21] to get the combination of a model that can understand the exchangeability between words and documents. LDA outperforms the other information retrieval methods such as tf-idf [22], LSI [23], and pLSI [24]. In LDA, words and documents are assumed to be independent. Therefore, the prior knowledge of those artifacts is unnecessary. LDA generates the model using a probabilistic distribution approach based on the hierarchical Bayesian Model.

Related to our study, we explored some applications of LDA for COVID-19 related analysis, especially COVID-19 vaccines. For example, Lyu et al. in [25] used the LDA algorithm to extract common topics in COVID-19 vaccine-related tweets. This study shows that there is an influx on vaccine discussion when the news about vaccine development appears in the news. Zhuang et al. in [26] proposed the combination of LDA and Autoregressive moving average (ARMA) for analyzing the public opinion evolution of COVD-19. Some studies related to COVID-19 textual analysis were targeted on Indonesian opinions [27-28]. However, most of the works were focused on the sentiment analysis rather than finding the topics. To reduce the research gap, LDA is implemented in this study to discover the topics related to COVID-19 vaccines in Indonesia. Therefore, Bahasa Indonesia was used as the targeted language in analyzing the topic using the LDA algorithm.

The corpus in this research was collected by filtering the COVID-19 vaccines opinions in Twitter that use Bahasa Indonesia. LDA algorithm was chosen in this study due to its ability to link the artifacts, in this case words. Another benefit of LDA is that there are no requirements of prior knowledge related to the topics. With LDA, we are able to generate the model that explore the possible topic formation based on the corpus. By conducting this research, the topics on COVID-19 vaccines discussion can be determined. Those topics can be used to identify the possible strategies that can be implemented to advance the effort for increasing the number of fully vaccinated citizens. Unfortunately, there is limited topic modeling research COVID-19 vaccine conducted on Bahasa Indonesia. Applying a model from different language may result in inaccurate representation of the uncovered topics. Therefore, this study focusses on the tweets that were posted using Bahasa Indonesia.

II. METHOD

We proposed a novel that consists of three methodology phases as seen in Fig.1. The first phase is intended to gather the data source from Twitter accessed through Twitter Application Programming Interface (API) based on specified language and keywords. The extracted data are stored as excel files used as the input for the second phase which is text pre-processing. This phase is conducted to prepare the data to be effectively used for the next phase. Several techniques are applied in the text pre-processing phase including case folding, stemming, and stop word removal. Once the text pre-processing phase was completed, we move to the next phase which is a model generation and visualization. In this phase, the sentences are analyzed using the LDA algorithm. We also perform a model evaluation based on specified metrics such as coherence and perplexity score.

A. Data Collection

The data collection phase is conducted in three steps: (1) defining a filter to collect the COVID-19 Vaccine-related posts in Bahasa Indonesia, (2) extracting twitter posts based on the defined filter using Twitter streaming API, and (3) removing redundant posts to increase the data accuracy in representing the current condition. Twitter data set has been used in various research and proven to be used as the source to gather public opinions [29-30]. Therefore, we also use the Twitter dataset to gather Indonesian public opinion regarding the COVIC-19 vaccine.

Before the data is extracted, the filters such as keywords and language need to be defined. This process is crucial to increasing this proposed work's efficiency and effectiveness. Using the defined filter can exclude the unnecessary tweet that can mislead the model generation process. The filter used in this study is listed in Table I. We use three type of filter including language, keywords, and tweet type. We decided to use retweet based on the common notion that retweeting a tweet means a person agree with another person’s opinion or tweet.

Once the filter is defined, the next process is conducted. This study is implemented using python. Therefore, we utilize Tweepy, a python library to access twitter API, to extract the tweets. To optimize the tweets extraction process, three researchers perform the process simultaneously. This strategy is carried out due to the limited number of tweets that can be extracted for each developer account. The extracted tweets are collected in an Excel sheet created using XLWT, a library for writing data to Excel files.

After the tweets were extracted, the duplicate data should be removed. This process is important because there is abundance of tweets with the same content as a non-retweet. By doing so, we can extract a retweet type that includes additional opinion. Before, the redundant tweets are removes, we remove the URL link to extend the tweets so that the generated link is not considered the content.

TABLE I
DEFINE FILTERS

Type	Filter
Language	Bahasa Indonesia (id)
Keywords	“covid-19 vaksin”, “vaksin”, “covid vaksin”
Tweet Type	Regular Tweet, Retweet

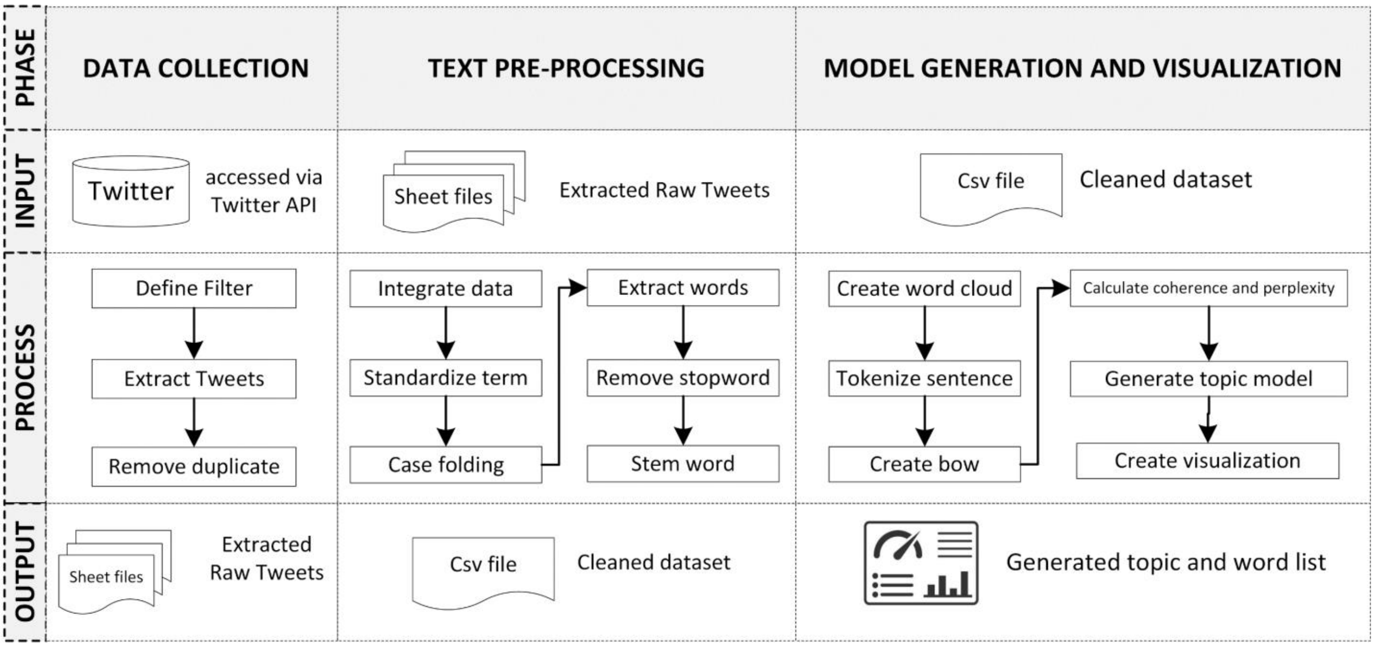


Fig. 1 Proposed methodology for generating topic model

B. Text Pre-processing

Since they are extracted separately based on date, the first process in text pre-processing is integrating the generated excel files as a single csv file. Using CSV file instead of Excel is that CSV file requires less memory and faster processing. Having less memory process is important due to huge number of extracted tweets. Another reason is that by using a CSV file, we can also edit the file using text editor.

Once the data were integrated, the text is standardized. The term standardization refers to the process for changing irregular or informal term in Bahasa Indonesia into the formal term. For example, the terms such as "*nggak*", "*gak*", "*ga*", and "*gk*" are changed into "*tidak*" which means "no" in English. The process aims to make different words that has same meaning is threated in the same way.

The next step is case folding. This process is used to convert the text into a standard form. There are five case-folding techniques applied in this study such as changing to lower case, remove mark and number, and remove whitespace. This process is important to match the words, for example the instance of "*Vaksin*" will be matched with the instance "*vaksin*". Table II shows the case-folding techniques implemented for the standardization process.

After the entire texts have been standardized in term of the case, the next step is word extraction. This process is used to prepare for the next step which is stop word removal. The gathered tweets are mostly in the form of sentences. Therefore, we need to separate each word in the tweets. Then, we remove stop words in the next step. Stop words refers to common words that usually do not provide significant information. For example, the stop words include particles conjunction, and pronounce in English. Since we target language in this study is Indonesia, we perform stop word removal specifically in Bahasa Indonesia such as "*yang*", "*dengan*", "*pada*", and "*dan*". We utilize the NLTK library to conduct this step, specifically in Indonesia.

Following the stop words removal, we perform word stemming in the next step. The purpose of stemming is to reduce the text variant based on the root form without changing the meaning of the text. For example, the word "*vaksinasi*", "*vaksi*", "*divaksin*" can be treated as same word which is *"vaksin"*. Xu and Croft in [31] argued that stemming process could improve the program performance to match the vocabulary and query. Afterward, the pre-processed tweets were gathered into a single csv file. This csv file consists of a cleaned dataset used in the next phase which is a model generation and visualization.

TABLE II
IMPLEMENTED CASE FOLDING TECHNIQUES

Technique	**Task Description**
Lower case	Changing the entire tweets into lower case
Remove mark and number	Deleting the unused character in the tweets such as question mark, acclamation mark, and number
Remove whitespace	Deleting unnecessary space character in the beginning and end of tweets.

C. Model Generation and Visualization

The third phase in this study is a model generation and visualization based on the cleaned dataset resulting in the previous phase. This third phase aims to discover the topics related to covid-19 by generating a topic model using the LDA algorithm. Then, the identified topics will be visualized in a graphical view to ease the user's understanding in analyzing the topics and identifying the possible strategy based on the topics.

The first step of this step is creating word cloud. This step is intended to discover the most discussed terms when people tweeted about covid-19 vaccine. The terms will be visualized with different sizes. The bigger of the size, the more frequent the word appears in the discussion. With this word clouds visualization, we allow the audience to summarize the topics related to covid-19 vaccine. After that, we move the process into tokenization. Webster and Kit [32] defined tokenization as a process to identify the basic units decomposed from a certain sentence. The most common token used in the NLP is a word.

After the tweets were tokenized, the bag-of-words (BoW) creation was conducted. The BoW creation process aims to gather the information on the term document frequency of each token. BoW is represented as a pair that consists of token and its term document frequency. The BoW is generated by creating the dictionary that consists of gathered token. Once the entire term in the dictionary is listed as the corpus, then the term document frequency is calculated for each word.

Afterwards, we calculated the metrics to evaluate the model. This study uses perplexity and coherence scores to determine the number of generated topics. The perplexity score measures the ability of the model to adapt to the unseen data. However, this metric is not enough to evaluate the model due to the low correlation between predictive likelihood which measured with perplexity and human judgment. Therefore, we add another evaluation metric which is the coherence score. Coherence means the human judgment of whether the term consistently and logically belongs to a certain topic.

However, this can be hard to measure statistically. Therefore, we follow the normalized pointwise mutual information (NPMI) and cosine similarity proposed by Syed and Spruit in [33].

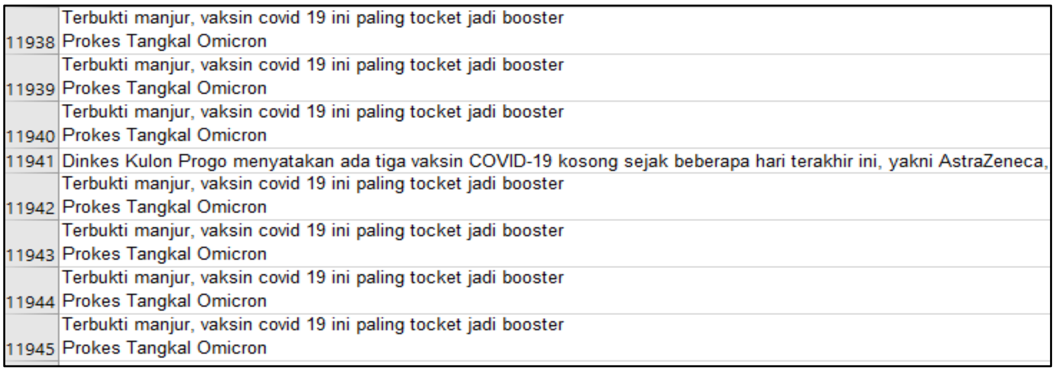
The next step after we determine the number of topics based the optimal combination perplexity and coherence score is generating the model and visualized the generated topics. We use a pure Python library to generate the LDA topic model called Gensim [34]. We adopt this library due to its ability to perform fast and scalable algorithms to create an LDA model. Once the model is generated, we visualize the result as an interactive graph. The visualization is created using Python library called LDAvis developed by Sievert and Shirley in [35].

III. RESULT AND DISCUSSION

The datasets were gathered from October 2021 to January 2022. By using the defined filter, we were able to gather 21498 tweets. Unfortunately, there are huge number of duplicated tweets. This duplication appeared because same tweeted the same content vigorously as seen in Fig. 2. It is important to remove this kind of tweets due to its inorganic nature of the tweets. From our point of view, the massively duplicated tweets cannot represent the public's substantial opinions. As the result, we got 8388 tweets ready to be used in the pre-processing phase.

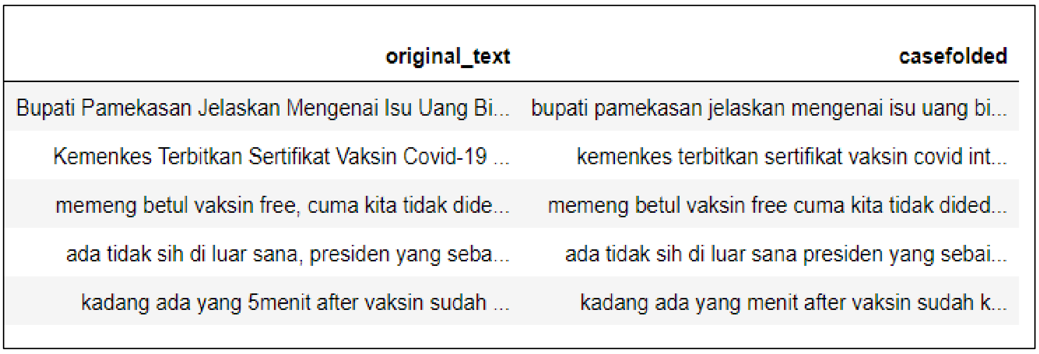
In the second phase, there is no difference in term of the number of tweets. The transformation is focused on the content of each tweet. The result of applying the proposed case folding techniques is shown in Fig.3. As we can see, there are some differences resulting from the case folding step. For example, there is no uppercase in the tweets. There is also a noticeable no number in the case of folded tweets.

Afterwards, the Indonesian stop words removal was conducted. Fig.4 shows the result of the stop words process. Using the NLTK library especially for the Indonesia language, we remove the common and insignificant words from the case folded tweets. For example, as we can see in the first row, the words such as “*harusnya*”, “*adanya*”, “*segera*”, and “*setelah*” was discarded from the tweet. Another example is from second row, the words such as “ketika”, “*sedang*”, and “*dan*” had been removed.



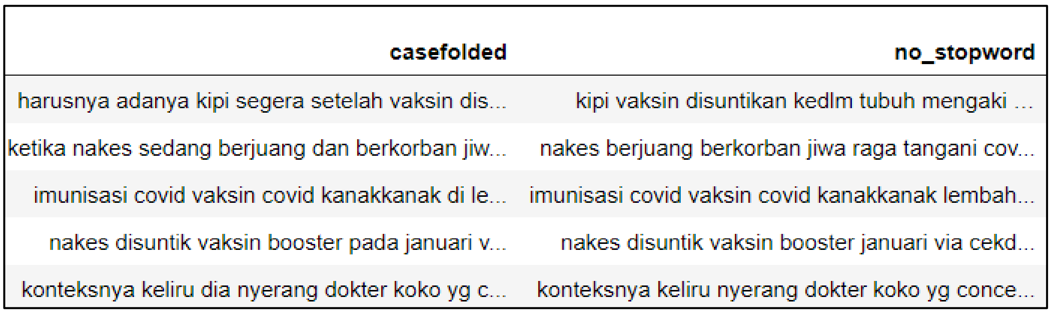
11938	Terbukti manjur, vaksin covid 19 ini paling tocket jadi booster Prokes Tangkal Omicron
11939	Terbukti manjur, vaksin covid 19 ini paling tocket jadi booster Prokes Tangkal Omicron
11940	Terbukti manjur, vaksin covid 19 ini paling tocket jadi booster Prokes Tangkal Omicron
11941	Dinkes Kulon Progo menyatakan ada tiga vaksin COVID-19 kosong sejak beberapa hari terakhir ini, yakni AstraZeneca,
11942	Terbukti manjur, vaksin covid 19 ini paling tocket jadi booster Prokes Tangkal Omicron
11943	Terbukti manjur, vaksin covid 19 ini paling tocket jadi booster Prokes Tangkal Omicron
11944	Terbukti manjur, vaksin covid 19 ini paling tocket jadi booster Prokes Tangkal Omicron
11945	Terbukti manjur, vaksin covid 19 ini paling tocket jadi booster Prokes Tangkal Omicron

Fig. 2 Excerpt of duplicated tweets



original_text	casefolded
Bupati Pamekasan Jelaskan Mengenai Isu Uang Bi...	bupati pamekasan jelaskan mengenai isu uang bi...
Kemenkes Terbitkan Sertifikat Vaksin Covid-19 ...	kemenkes terbitkan sertifikat vaksin covid int...
memeng betul vaksin free, cuma kita tidak dide...	memeng betul vaksin free cuma kita tidak dided...
ada tidak sih di luar sana, presiden yang seba...	ada tidak sih di luar sana presiden yang sebai...
kadang ada yang 5menit after vaksin sudah ...	kadang ada yang menit after vaksin sudah k...

Fig. 3 Excerpt of case folded tweets

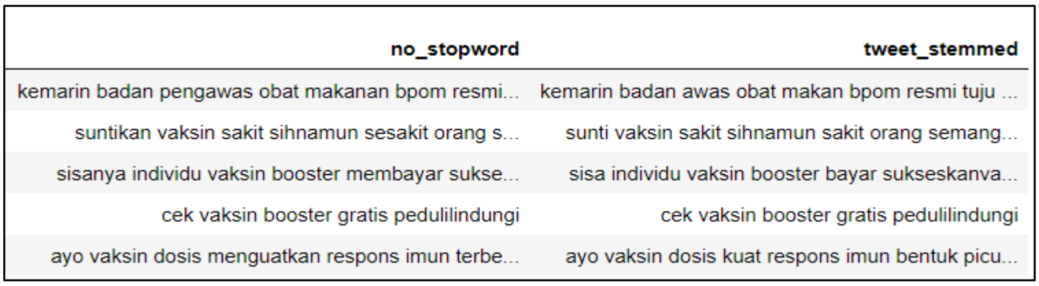


casefolded	no_stopword
harusnya adanya kipi segera setelah vaksin dis...	kipi vaksin disuntikan kedlm tubuh mengaki ...
ketika nakes sedang berjuang dan berkorban jiw...	nakes berjuang berkorban jiwa raga tangani cov...
imunisasi covid vaksin covid kanakkanak di le...	imunisasi covid vaksin covid kanakkanak lembah...
nakes disuntik vaksin booster pada januari v...	nakes disuntik vaksin booster januari via cekd...
konteksnya keliru dia nyerang dokter koko yg c...	konteksnya keliru nyerang dokter koko yg conce...

Fig. 4 Excerpt of stop words removed tweets

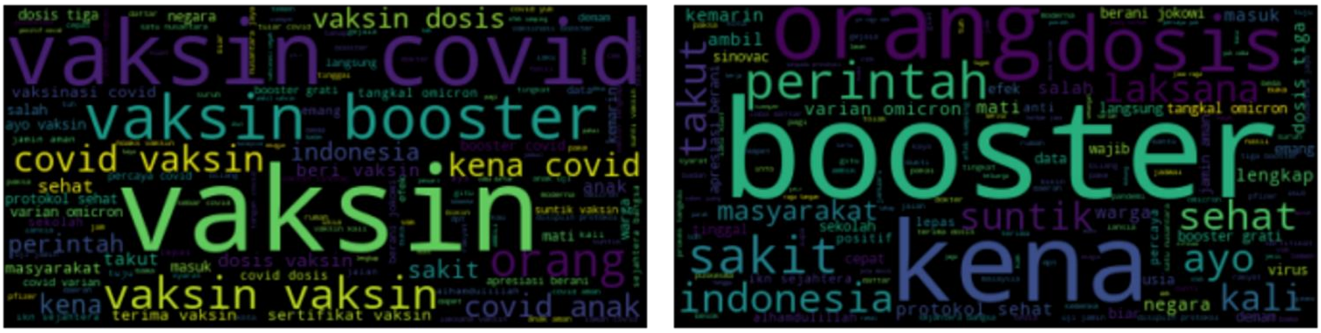
After removing the stop words from the dataset, we stemmed the tweets using the Sastrawi library. Fig. 5 shows the result of stemming process that will be used as the input in the next phase. For example, the word "*pengawas*" in the first row became "*awas*". Meanwhile, the word "*membayar*" in the third row became "*bayar*". Once all the tweets had been stemmed, a csv file that consists of pre-processed tweets was generated so that it can be used in the model generation and visualization phase.

Fig. 6(a) shows that most frequent terms from the corpus of pre-processed tweets. However, this word cloud is not adequate. Knowing the most frequent terms are *"vaksin"* and "covid" does not add new knowledge because those term are the search keywords in the extraction process. Therefore, those words were removed to increase the other words representation in the corpus. The result of this strategy is presented in Fig.6 (b). Based on this information, we understood that people are talking about vaccines booster. There were also large numbers of tweets discussing about the infection. This data is not surprising due to the existence of omicron resulting in exponential increase of the COVID-19 infection cases. Based on Fig.6 (b), we also can see that the government attempt to push the booster vaccination program do reduce the number of cases. This proved that our pre-processed data are able to represent the current condition in Indonesia. Therefore, use used the new data to generate the topic model.



no_stopword	tweet_stemmed
kemarin badan pengawas obat makanan bpom resmi...	kemarin badan awas obat makan bpom resmi tuju ...
suntikan vaksin sakit sihnamun sesakit orang s...	sunti vaksin sakit sihnamun sakit orang semang...
sisanya individu vaksin booster membayar sukse...	sisa individu vaksin booster bayar sukseskanva...
cek vaksin booster gratis pedulilindungi	cek vaksin booster gratis pedulilindungi
ayo vaksin dosis menguatkan respons imun terbe...	ayo vaksin dosis kuat respons imun bentuk picu...

Fig. 5 Excerpt of stemmed tweets



(a)

(b)

Fig. 6 Word cloud visualization for the pre-processed tweets (a) and after removing the "vaksin" and "covid" term

The model evaluation process based on coherence score in terms of NPMI and perplexity score returned the optimal number of topics is five. The optimal number of topics was determined based on the elbow methods that shows in which point the number is started to converge. Fig. 7 shows the score of perplexity from the model. We can see that the score consistently reduced when we increased the number of topics. This is one of the reasons we required additional measures, in this case, coherence score. The calculation result is presented in Fig. 8. Model evaluation based on coherence score calculation returned the optimal score, 0.4239, and perplexity score is -8.3. As we can see in Fig. 7, the score consistently decreases after the number of topics is two. The result from the perplexity score is supported with the coherence score seen in Fig. 8. We can see that the highest perplexity score was achieved from a model that has two topics. Based on those metrics, the optimal number of topics that should be used for this study is two.

Once the number of optimal topics is determined, the next process generates the model with the LDA algorithm. The overall result of the LDA algorithm visualized using LDAvis library [35] can be seen in Fig. 9. We can see that the 30 most salient topics that were discussed when people talk about COVID-19 vaccine in Indonesia include "*sertifikat*" (vaccine certificate), "*booster*", "*aman*" (safe), "*protokol*" (protocol), "*percaya*" (trust), "*apresiasi*" (appreciation). The analysis result shows that many people in Indonesia took the vaccine due to government regulation on the need to have vaccine certificates to enter public spaces. Some amounts of citizens stated their appreciation for the vaccine booster program.

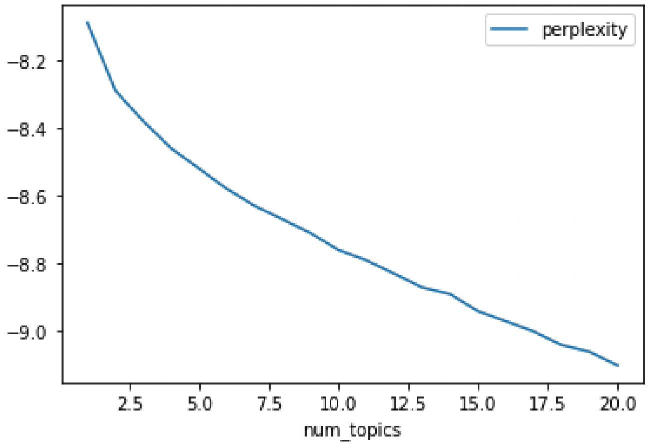


Fig. 7 The perplexity score versus the number of topics

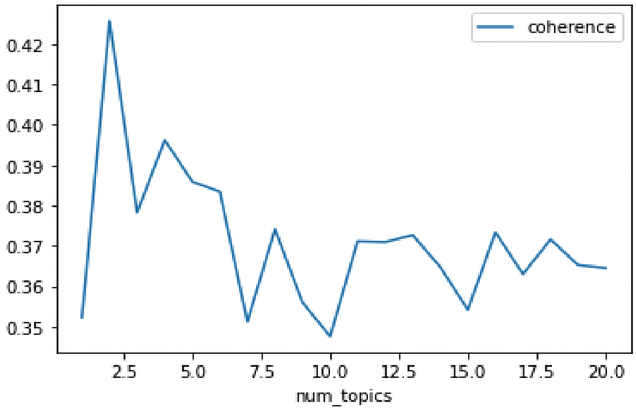


Fig. 8 The coherence score versus the number of topics

Meanwhile, based on the distance metric we can see that the topics were located in different spectrum. It means that there is an opposite view on the COVID-19 vaccines in Indonesia. The number of generated topics align with the previously conducted research on the COVID-19 vaccines sentiment analysis in Indonesia. In that research, we found that there are two sentiments, which are negative and positive sentiment.

We further investigated on the top-30 most salient terms for each topic presented in Table III. Even though we can see that some terms were exists in both topics, the further analysis result show that it has different sentiment. For example, the term "booster" (refers to the third dose of vaccine) appeared in both topics, but in the topic where people were mostly discussed on the negative sentiment related to vaccine booster. Some people question whether the vaccine booster has effectively protected them from the omicron variant. Meanwhile in topic 2, most people have positive perspective on the vaccine booster. They mentioned that vaccine booster is safe to use and beneficial to protecting them from the virus.

TABLE III
TOP-30 MOST SALIENT TERMS FOR TOPIC

Topic	Terms
Topic 1	*booster, dosis, omicron, terima, orang, laksana, masyarakat, kena, gratis, lindungi, tiga, sehat, warga, varian, prokes, tinggal, sekolah, wajib, ayo, sertifikat, sakit, aman, beri, Jokowi, tular, perintah, giat, suntik, sebar, negara*
Topic 2	*booster, aman, kena, orang, dosis, sehat, perintah, prokes, uji, percaya, cegah, daftar, jamim, Jokowi, sakit, laksanan, masuk, beri, ayo, virus, takut, coba, tinggal, omicron, yuk, sinovac, suntik, lengkap, efek, lawan*

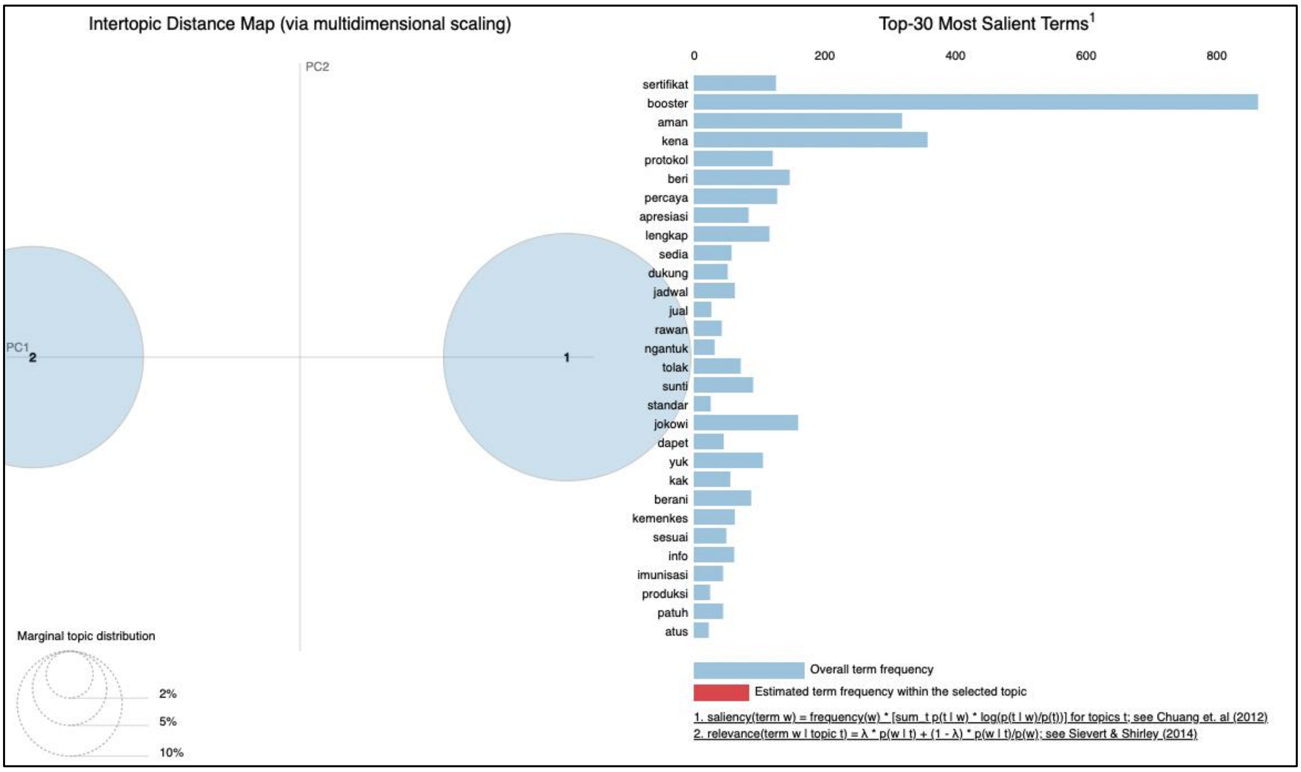


Fig. 9 The overall result of the LDA algorithm

From this analysis, we understand that most of the people that has negative perspective on the COVID-19 vaccine are mostly concerned about its effectiveness. Some people also mentioned the side effect of the vaccines as their concern. We also found that there are some concerns with the difficulty to issue the certificate. It is interesting to see that even though some people against the vaccine, they still decided to get the vaccine so that they can enter public space such as shopping center and theater.

Based on the topic modeling result, there are several approaches can be done to accelerate the COVID-19 vaccination program. The first approach is educating the citizen about the importance of getting a vaccine booster with the aim to increase their understanding and enthusiasm. The Indonesian's citizen should be aware that vaccine is safe and effective to fight the virus. One way to achieve this by spreading the information that the COVID-19 vaccines had been certified by BPOM (National Agency of Drug and Food Control). Moreover, most of the vaccines had been certified as a halal product by MUI (Indonesian Ulama Council).

Another approach that can be done is keep informing the basic information about the COVID-19 vaccines including the procedure to get the vaccine, where to get the vaccine, and correct information about the misconception about the vaccine. The government should be able to accentuate that, based on various study, there is a decrease of antibody six-months after the primary dose of vaccine has been administered to the body. Therefore, there is a need to get another dose or booster to build an extra protection to the body especially for the people who are prone to the COVID-19 virus.

IV. CONCLUSION

By implementing a topic modeling algorithm called Latent Dirichlet Allocation algorithm on more than 8000 tweets that have been cleaned, we are able to determine the most important terms that occurs when people talk about the COVID-19 vaccine. Based on the perplexity and coherence score, we found that the optimal number of topics was two. It means that there were two different perspectives on the COVID-19 vaccine discussion. The first topic has negative sentiment compared to the other topic. This result can further be explored to identify the sentiment of each sentence. Therefore, the further work of this research includes sentiment analysis prediction with machine learning algorithm.

ACKNOWLEDGEMENT

This research was supported by Matching Fund Program, Minister of Education, Culture, Research, and Technology of The Republic of Indonesia, contract number: 362/E1.1/KS.06.02/2022.

REFERENCES

[1] Tosepu, R., Gunawan, J., Effendy, D. S., Lestari, H., Bahar, H., & Asfian, P., "Correlation between weather and Covid-19 pandemic in Jakarta, Indonesia," *Sci. Total Environ.*, vol. 725, 2020, doi: 10.1016/j.scitotenv.2020.138436.

[2] M. Ciotti, M. Ciccozzi, A. Terrinoni, W. C. Jiang, C. Bin Wang, and S. Bernardini, "The COVID-19 pandemic," *Crit. Rev. Clin. Lab. Sci.*, vol. 57, no. 6, pp. 365–388, 2020, doi: 10.1080/10408363.2020.1783198.

[3] J. S. Tregoning, K. E. Flight, S. L. Higham, Z. Wang, and B. F. Pierce, "Progress of the COVID-19 vaccine effort: viruses, vaccines and variants versus efficacy, effectiveness and escape," *Nat. Rev. Immunol.*, vol. 21, no. 10, pp. 626–636, 2021, doi: 10.1038/s41577-021-00592-1.

[4] T. Koyama, D. Weeraratne, J. L. Snowdon, and L. Parida, "Emergence of Drift Variants That May A ff ect COVID-19 Vaccine Development and," *Pathogens*, no. 2020, pp. 1–7, 2020, doi: 10.1109/BIBM52615.2021.9669839

[5] Mathieu, E., Ritchie, H., Ortiz-Ospina, E., Roser, M., Hasell, J., Appel, C., Giattino, C. and Rodés-Guirao, L.., "A global database of COVID-19 vaccinations," *Nat. Hum. Behav.*, 2021, doi: 10.1038/s41562-021-01122-8.

[6] A. Fuady, N. Nuraini, K. K. Sukandar, and B. W. Lestari, "Targeted vaccine allocation could increase the covid-19 vaccine benefits amidst its lack of availability: A mathematical modeling study in indonesia," *Vaccines*, vol. 9, no. 5, 2021, doi: 10.3390/vaccines9050462.

[7] A. Z. Sarnoto and L. Hayatina, "Polarization of the muslim community towards government policies in overcoming the COVID-19 pandemic in Indonesia," *Linguist. Cult. Rev.*, vol. 5, no. April, pp. 642–652, 2021, doi: 10.37028/lingcure.v5nS1.1449.

[8] G. B. S. Wirawan, P. N. T. Y. Mahardani, M. R. K. Cahyani, N. L. P. S. P. Laksmi, and P. P. Januraga, "Conspiracy beliefs and trust as determinants of COVID-19 vaccine acceptance in Bali, Indonesia: Cross-sectional study," *Pers. Individ. Dif.*, vol. 180, no. January, p. 110995, 2021, doi: 10.1016/j.paid.2021.110995.

[9] Harapan, H., Wagner, A.L., Yufika, A., Winardi, W., Anwar, S., Gan, A.K., Setiawan, A.M., Rajamoorthy, Y., Sofyan, H., Vo, T.Q. and Hadisoemarto, P.F. "Willingness-to-pay for a COVID-19 vaccine and its associated determinants in Indonesia," *Hum. Vaccines Immunother.*, vol. 16, no. 12, pp. 3074–3080, 2020, doi: 10.1080/21645515.2020.1819741.

[10] A. N. Mason, J. Narcum, and K. Mason, "Social media marketing gains importance after Covid-19," *Cogent Bus. Manag.*, vol. 8, no. 1, 2021, doi: 10.1080/23311975.2020.1870797.

[11] A. R. Rahmanti, D. N. A. Ningrum, L. Lazuardi, H. C. Yang, and Y. C. Li, "Social Media Data Analytics for Outbreak Risk Communication: Public Attention on the 'New Normal' During the COVID-19 Pandemic in Indonesia," *Comput. Methods Programs Biomed.*, vol. 205, p. 106083, 2021, doi: 10.1016/j.cmpb.2021.106083.

[12] L. J. Peng, X. G. Shao, and W. M. Huang, "Research on the Early-Warning Model of Network Public Opinion of Major Emergencies," *IEEE Access*, vol. 9, pp. 44162–44172, 2021, doi: 10.1109/ACCESS.2021.3066242.

[13] L. Hong and B. D. Davison, "Empirical Study of Topic Modeling in Twitter," in *1st Workshop on Social Media Analytics (SOMA '10)*, 2010, p. 138.

[14] C. H. Papadimitriou, P. Raghavan, H. Tamaki, and S. Vempala, "Latent semantic indexing: A probabilistic analysis," *J. Comput. Syst. Sci.*, vol. 61, no. 2, pp. 217–235, 2000, doi: 10.1006/jcss.2000.1711.

[15] I. Vayansky and S. A. P. Kumar, "A review of topic modeling methods," *Inf. Syst.*, vol. 94, p. 101582, 2020, doi: 10.1016/j.is.2020.101582.

[16] T. Schmiedel, O. Müller, and J. vom Brocke, "Topic Modeling as a Strategy of Inquiry in Organizational Research: A Tutorial With an Application Example on Organizational Culture," *Organ. Res. Methods*, vol. 22, no. 4, pp. 941–968, 2019, doi: 10.1177/1094428118773858.

[17] W. Chen, K. Lai, and Y. Cai, "Topic generation for Chinese stocks: a cognitively motivated topic modelingmethod using social media data," *Quant. Financ. Econ.*, vol. 2, no. 2, pp. 279–293, 2018, doi: 10.3934/qfe.2018.2.279.

[18] T. Porturas and R. A. Taylor, "Forty years of emergency medicine research: Uncovering research themes and trends through topic modeling," *Am. J. Emerg. Med.*, vol. 45, no. xxxx, pp. 213–220, 2021, doi: 10.1016/j.ajem.2020.08.036.

[19] S. Boon-Itt and Y. Skunkan, "Public Perception of the COVID-19 Pandemic on Twitter: Sentiment Analysis and Topic Modeling Study," *JMIR Public Heal. Surveill.*, vol. 6, no. 4, p. e21978, Nov. 2020, doi: 10.2196/21978.

[20] T. De Melo and C. M. S. Figueiredo, "Comparing News and Tweets about COVID-19 in Brazil Table of Contents," *JMIR Public Heal. Surveill.*, vol. 7, no. 2, 2021, doi: 10.2196/24585.

[21] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003, doi: 10.1016/B978-0-12-411519-4.00006-9.

[22] G. Salton, "Some research problems in automatic information retrieval," 1983. doi: 10.1145/1013230.511830

[23] T. Hofmann, "Probabilistic latent semantic indexing," *Proc. 22nd Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retrieval, SIGIR 1999*, pp. 50–57, 1999, doi: 10.1145/312624.312649.

[24] Scott Deerwester, Richard Harshman, Susan T, George W, and Thomas K, "Indexing by Latent Semantic

Analysis," *J. Am. Soc. Inf. Sci.*, vol. 41, no. 6, pp. 391–407, 1990. doi: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

[25] J. C. Lyu, E. Le Han, and G. K. Luli, "Covid-19 vaccine-related discussion on twitter: Topic modeling and sentiment analysis," J. Med. Internet Res., vol. 23, no. 6, pp. 1–12, 2021, doi: 10.2196/24435.

[26] M. Zhuang, Y. Li, X. Tan, L. Xing, and X. Lu, "Analysis of public opinion evolution of COVID-19 based on LDA-ARMA hybrid model," Complex Intell. Syst., vol. 7, no. 6, pp. 3165–3178, 2021, doi: 10.1007/s40747-021-00514-7.

[27] A. M. Tri Sakti, E. Mohamad, and A. A. Azlan, "Mining of opinions on COVID-19 large-scale social restrictions in indonesia: Public sentiment and emotion analysis on online media," J. Med. Internet Res., vol. 23, no. 8, 2021, doi: 10.2196/28249.

[28] D. A. Nurdeni, I. Budi, and A. B. Santoso, "Sentiment Analysis on Covid19 Vaccines in Indonesia: From the Perspective of Sinovac and Pfizer," 3rd 2021 East Indones. Conf. Comput. Inf. Technol. EIConCIT 2021, pp. 122–127, 2021, doi: 10.1109/EIConCIT50028.2021.9431852.

[29] P. Otero, J. Gago, and P. Quintas, "Twitter data analysis to assess the interest of citizens on the impact of marine plastic pollution," Mar. Pollut. Bull., vol. 170, no. June, p. 112620, 2021, doi: 10.1016/j.marpolbul.2021.112620.

[30] U. Naseem, I. Razzak, M. Khushi, P. W. Eklund, and J. Kim, "COVIDSenti: A Large-Scale Benchmark Twitter Data Set for COVID-19 Sentiment Analysis," IEEE Trans. Comput. Soc. Syst., vol. 8, no. 4, pp. 976–988, 2021, doi: 10.1109/TCSS.2021.3051189.

[31] J. Xu and W. B. Croft, "Corpus-based stemming using cooccurrence of word variants," ACM Trans. Inf. Syst., vol. 16, no. 1, pp. 61–81, 1998, doi: 10.1145/267954.267957.

[32] J. J. Webster and C. Kit, "Tokenization as The Initial Phase in NLP," in COLING: The 14th International Conference on Computational Linguistics, 1992, vol. 4. doi: 10.3115/992424.992434

[33] S. Syed and M. Spruit, "Full-Text or abstract? Examining topic coherence scores using latent dirichlet allocation," Proc. - 2017 Int. Conf. Data Sci. Adv. Anal. DSAA 2017, vol. 2018-January, pp. 165–174, 2017, doi: 10.1109/DSAA.2017.61.

[34] R. Řehůřek and P. Sojka, "Gensim — Statistical Semantics in Python," Retrieve from gensim.org, no. May 2010, 2011.

[35] C. Sievert and K. Shirley, "LDAvis: A method for visualizing and interpreting topics," in Workshop on Interactive Language Learning, Visualization, and Interfaces, 2015, pp. 63–70, doi: 10.3115/v1/w14-3110.