Indonesian Citizens' Health Behavior in a Pandemics: Twitter Conversation Analysis using Latent Dirichlet Allocation

Dian Sa'adillah Maylawati
Department of Informatics
UIN Sunan Gunung Djati Bandung
Bandung, Indonesia
diansm@uinsgd.ac.id
https://orcid.org/0000-0002-1193-3370

Muhammad Ali Ramdhani
Department of Informatics
UIN Sunan Gunung Djati Bandung
Bandung, Indonesia
m_ali_ramdhani@uinsgd.ac.id
https://orcid.org/0000-0002-6492-067X

***Abstract*—Health is an essential thing in carrying out human activities. The COVID-19 pandemic has made people aware of the importance of maintaining health and hygiene for individuals, families, and the surrounding community. All countries, including Indonesia, are impacted by the COVID-19 pandemic, which has undoubtedly changed health behavior in the community. This study aims to reveal changes in health behavior during the pandemic through conversations on social media such as Twitter. The study was conducted using the Latent Dirichlet Allocation (LDA) method to analyze changes in Indonesian citizens' health behavior during the pandemic through social media analysis. The results of social media analysis using LDA on 495,740 tweet data indicate that it is true that there has been a change in public health behavior. At the beginning of the pandemic, many people still did not believe that various hoaxes were spread, and it was difficult to comply with health protocols. Hence, the government massively appealed to make regulations to break the chain of the spread of COVID-19. However, at a critical time with many victims falling, the public became more aware, maintained health protocols, followed the vaccination program, and finally, people got used to coexistence with COVID-19. These results indicate that the Indonesian people are wiser in dealing with the COVID-19 pandemic and following the applicable health protocols.**

Keywords—COVID-19, health behavior, latent dirichlet allocation, social media analysis, twitter

I. INTRODUCTION

The COVID-19 pandemic is a global nightmare that has created fear, social isolation, anger and uncertainty about the future. Based on data from the World Health Organization (WHO), as of June 3, 2022, as many as 528,816,317 people in the world have been confirmed positive for COVID-19 [1]. In Indonesia, as of June 1, 2022, 6,055,341 residents have been confirmed positive for COVID-19 [2]. Of course, this high COVID-19 transmission rate changes the views and lifestyles of the global community, especially in the health aspect. Since the COVID-19 pandemic entered Indonesia, the government, with various regulations, has regulated community activities to comply with health protocols to the policy of implementing restrictions on community activities. This is of course implemented for the benefit of individuals and society to break the chain of the spread of COVID-19. Various opinions are undoubtedly widespread in society, especially through social media.

In the Industry 4.0 era, all human activities cannot be separated from the use of technology, especially during a pandemic where most activities are carried out from home. No less than 4.66 billion people in the world will be connected to the internet by 2021 [3], [4]. Even in Indonesia, by 2021, there will be 212.35 million Internet users in Indonesia or 76.8 percent of the 276.3 million Indonesian population [5]. Of the many Internet users in Indonesia, around 61.8 percent spend their activities on the internet to access social media [6]. This phenomenon cannot be separated from the rapid development of technology. This social media is also a big data source with a huge volume of data, varied data types, and fast flow [7], [8]. This much data from social media is certainly an interesting source that can be processed into meaningful information and knowledge (insight knowledge) using social media analysis technology [9]–[12]. Youtube will become Indonesia's most popular social media in 2021, followed by WhatsApp, Instagram, Facebook, Twitter, and other social media applications [13], [14].

Twitter is one social media that is widely used in social media analysis research. This is due to the easy access and permission to pull data from Twitter compared to other social media. However, despite the more accessible data collection techniques, Twitter is a more open, open-minded, egalitarian place and a place for trending news [15]. Compared to Facebook, which has many fake accounts and hoaxes, and Instagram, which is mainly used for self-promotion and self-exposure, Twitter has far more educated users. The lack of mental judgement shows this and if there is news that is a hoax, has a toxic smell, and is fighting with each other, Twitter users are wiser in dealing with it. In essence, social media analysis technology is currently one of research strengths in the digital era [16].

There are several previous research on social media analysis: (1) Retrospective analysis of the probability of predicting the COVID-19 epidemic in China using Internet searches and social media data [17]; (2) Social media sentiment analysis based on COVID-19 using Recurrent Neural Network [18]; (3) a social media analysis for the Alternative für Deutschland which is the government's third-largest party [19]; (4) The rise of "flightshame" on social media can be traced back to the climate issue [20]; (5) Artificial intelligence was used to analyze social media responses to the Covid19 epidemic [21]; and (6) A quantitative analysis of social media data using Twitter to create COVID-19 stigma by referencing the novel coronavirus as the "Chinese Virus."

In several previous studies, social media analysis technology can help obtain necessary information and knowledge related to public views. Therefore, this study utilizes social media analysis technology to identify changes in public health behavior in Indonesia during the pandemic. From various social media analysis techniques, this research uses a topic modeling approach with Latent Dirichlet Allocation (LDA) algorithms. These two methods produce clusters of topics that are discussed on social media based on

978-1-6654-5122-2/22/$31.00 ©2022 IEEE

the proximity of the conversation. After the clusters were formed, further social media analysis was carried out in depth on the Twitter data obtained.

II. Research Methods

A. Cross-Industry Standard Process for Data Mining

The methodology used in this research is a mixed-method between quantitative and qualitative. Quantitative is done during the development and evaluation of the modeling topic. While qualitative when interpreting and analyzing the results of model development. This research was conducted with a data science approach using the CRISP-DM (Cross-Industry Standard Process for Data Mining) data science methodology [22], [23]. Activities in the CRISP-DM Data Science methodology consist of business understanding, data understanding, data preparation, modeling, model evaluation, deployment, and model management. In this study, no deployment and model management activities were carried out.

B. Latent Dirichlet Allocation (LDA)

LDA is one of the popular topic modeling algorithms with generative probabilistic models for discrete data collections such as text data [24]. LDA is a Bayesian model where each item being modeled consists of a series of topic probabilities. In the context of topic modeling, LDA generates topic possibilities by providing an explicit representation of a document. LDA is the most simple and efficient algorithm for topic modeling [25]. The basic idea is that the document is represented as a random mixture of latent topics, where a distribution of words characterizes each topic. Figure 1 shows an illustration of obtaining topics from documents based on extracted words.

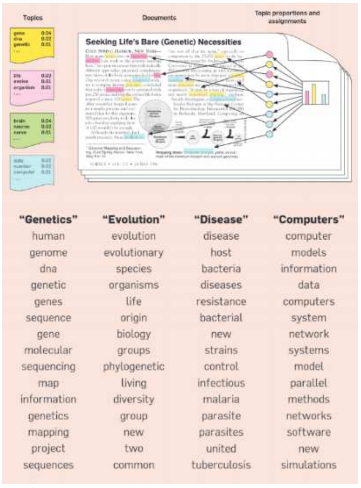


Fig. 1. Illustration of Topic Modeling using LDA [25]

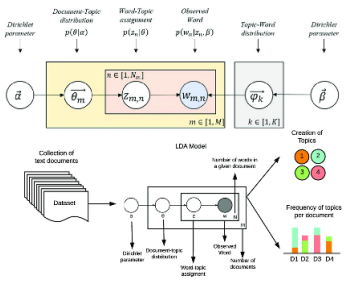


Fig. 2. Illustration of the Vector Space Calculation Process and Topic Formation in LDA [26]

LDA forms two assumptions [26]: first, that a document is a collection of topics, and second, that a topic is a collection of tokens or words. In LDA, this topic uses probability distribution in generating words. In statistical parlance, documents are known as the probability density (or distribution) of topics and topics are the probability density (or distribution) of words. The processes in LDA include: (1) A collection of text documents will be mapped in the form of a document word matrix. The rows indicate the number of documents, while the columns indicate the unique words in the document. Finding these individual words can use Term Frequency and Inverse Document Frequency (TF-IDF) weighting [27]–[29]; and (2) Next, enter the vector space calculation process from the LDA algorithm which is illustrated in Figure 2. The process starts by calculating Dirichlet parameters, the distribution of topics in the document, the distribution of words in topics, the distribution of selected words, the calculation of topic distribution based on words, and the formation of topics and frequency of topics for each document.

C. Coherence Value Evaluation

LDA applies an unsupervised learning technique, which means how many topics will be generated from our document collection before running the model is unknown. So that the coherence value measurement is used to determine how many optimal groups will be formed in the model that is run [30]. The coherence value calculation process starts from document segmentation, calculating the probability of word occurrences in the document. Then there is a confirmation size process that takes a single pair of words or word subsets and the reasonable possibility to calculate the strength of the word set. The last is the aggregation process to produce coherence values.

III. Result and Discussion

This section presents the result of social media analysis to know the Indonesian citizens' health behavior during the COVID-19 pandemic. The explanation begins with business understanding, data preparation, modeling, and evaluation results. Last the interpretation of the clustering result.

A. Business Understanding

Social media is a communication platform rich in sources of information related to public opinion on hot issues being discussed, one of which is Twitter. People tend to be free to voice their opinions through social media. During the COVID-19 pandemic, various opinions, pros and cons, hoaxes, and other important information were scattered through social media. Changes in health behavior during the COVID-19 pandemic are exciting things to study in-depth.

Social media analysis technology, part of Natural Language Processing, can be used to find important information through social media. There needs to be a mapping of the topics most discussed by the public through social media without the need to enumerate the messages conveyed on social media one by one. The view of the community as a whole can be done using topic modeling techniques with a clustering approach. Text data will be grouped based on similarity in content or widely discussed topics. Furthermore, an in-depth analysis was carried out on

the text data based on the topics that emerged from each cluster that was formed. The algorithm used to create these topic clusters is Latent Dirichlet Allocation (LDA).

B. Data Understanding

This study uses data taken from social media Twitter (Indonesian), with several data collection scenarios according to the needs and objectives of the study. Data was collected from Twitter from January 1, 2021, to July 31, 2021. This timeframe was considered when COVID-19 began to enter Indonesia until the period of the *Pemberlakukan Pembatasan Kegiatan Masyarakat*/ Emergency Community Activity Restrictions (PPKM). In addition, the data is grouped according to four-time series: (1) the January to March 2020 period at the beginning of the COVID-19 pandemic; (2) The period from March to May 2020 during the COVID-19 pandemic increases; (3) the period of June-December 2020 during the new normal (new habits), where people begin to adapt to COVID-19; and (4) the period from January to July 2021 during the PPKM period until the COVID-19 emergency. Twitter data retrieved using keywords *COVID-19 dan masker* (COVID-19 and mask)*; COVID-19 dan cuci tangan* (COVID-19 and washing hand)*; COVID-19 dan menjaga jarak* (COVID-19 and physical distancing)*; COVID-19 dan kerumunan* (COVID-19 and crowd)*; COVID-19 dan mobilitas* (COVID-19 and mobility)*; COVID-19 dan interaksi* (COVID-19 and interaction)*; COVID-19 dan berjemur* (COVID-19 and sunbathe)*; COVID-19 dan vitamin* (COVID-19 and vitamin)*; COVID-19 dan obat* (COVID-19 and medicine)*; COVID-19 dan herbal* (COVID-19 and herbal)*; COVID-19 dan madu* (COVID-19 and honey)*; COVID-19 dan vaksin* (COVID-19 and vaccine)*; COVID-19 dan rumah sakit* (COVID-19 and hospital)*; COVID-19 dan puskesmas* (COVID-19 and public health center)*;* and *COVID-19 dan olah raga* (COVID-19 and exercise).

C. Data Preparation

The intermediate data preparation activity process is conducting a review and pre-processing of the text data. Pre-processing text data is essential because, at this stage, the data is prepared, cleaned, and selected based on the need to maintain the quality of the input data [31]–[33]. Activities carried out in the pre-processing of this text include: (1) labeling the data as a sentiment analysis requirement, the data is labeled with positive, negative, and neutral labels; (2) case folding, namely the uniform process the size of the text into lowercase letters, because in computing using the Python language, capital letters and lowercase letters are distinguished (case sensitive); (3) cleaning the text of unused characters (regular expressions) and redundant characters, such as the use of excessive punctuation; (4) changing Indonesian slang words into their formal words; (5) changing emoticons that will affect the results of sentiment analysis; (6) change the abbreviation into the standard word; (7) eliminating stop-words or unimportant words that have no meaning, for example the conjunctions "which", "at", "is", "that is", and so on; (8) stemming, namely the process of changing words that have affixes into their stem word, such as "*memakan*" to "*makan*" (eat); and (9) tokenizing is the process of cutting words which will then be represented in a structured form such as bag of words and 2-grams, and the frequency of occurrence of words is calculated using Term Frequency and Inverse Document Frequency (TF-IDF) [28], [29].

D. Modeling

This process implements the LDA algorithms to form clusters of topics that are widely discussed by the public through Twitter social media. These clusters were formed based on coherence value evaluation (CV). The best clusters formed were 5 clusters with the best CV value of 0.434. Figure 3 shows a CV graph that determines the best number of clusters that can be formed. Figure 4 shows five clusters formed from the LDA algorithm with two far apart clusters, which means that each cluster has unrelated similarities between clusters. Each of these clusters has a different topic of conversation. Meanwhile, the other three clusters overlap, which means that there are similar topics of conversation between the two clusters.

Table I shows a word cloud that visualizes the most discussed topics in each cluster in the form of words and word pairs. Table II shows a graph of CV evaluation for each time of the COVID-19 and Health data groups. In the first time period, the clustering process resulted in the 11 best clusters with the best CV of 0.479. In the second time series, the two best clusters were produced with a CV value of 0.619, the third time series had 8 clusters with a CV value of 0.493, while in the third time series, the best three clusters were produced with a CV value of 0.65. Table III shows the LDA results applied to each time series in this data group. Meanwhile, Table 12 presents a collection of word cloud visualizations for each cluster formed in each time series.

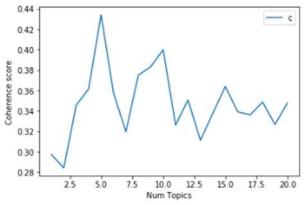


Fig. 3. CV for COVID-19 and health behavior

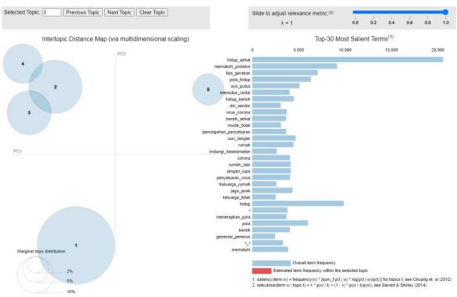


Fig. 4. LDA results for COVID-19 and health behavior

TABLE I. VISUALIZATION OF THE WORD CLOUD FOR COVID-19 AND HEALTH BEHAVIOR IN GENERAL

TABLE II. CV GRAPH FOR COVID-19 AND HEALTH BEHAVIOR BASED ON TIME SERIES

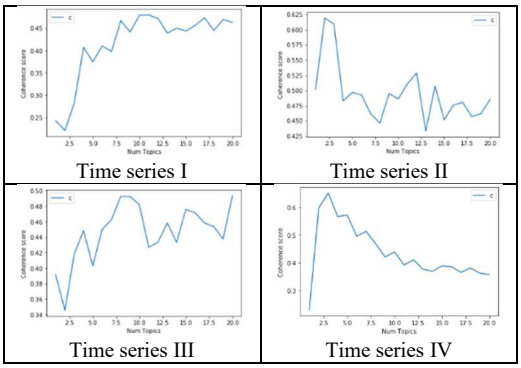


TABLE III. LDA RESULT OF COVID-19 AND HEALTH BEHAVIOR BASED ON TIME SERIES

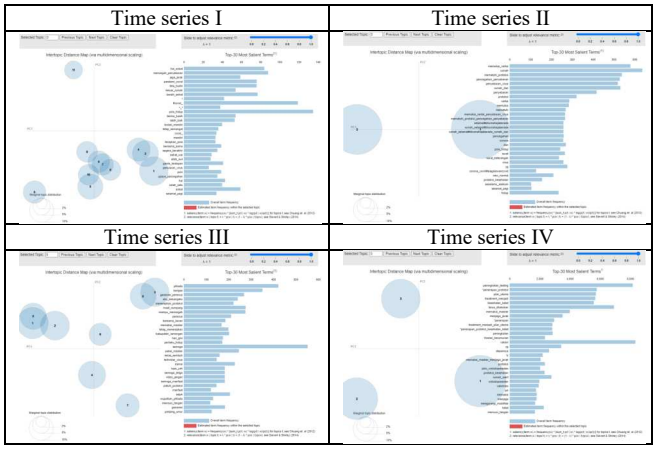
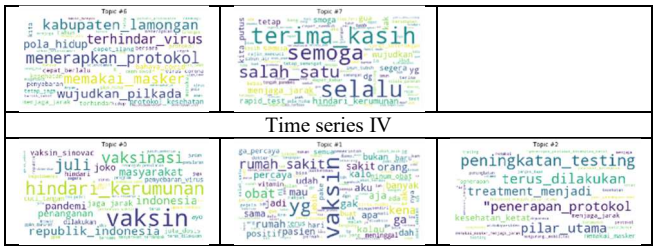


TABLE IV. WORD CLOUD VISUALIZATION OF COVID-19 AND HEALTH BEHAVIOR BASED ON TIME SERIES



E. *Interpretation Result*

Generally, Five clusters were formed related to COVID-19 and Indonesian public health behavior. Clusters 2, 3, and 4 intersect each other. From all clusters, it can be seen that the Indonesian people actually understand a healthy and clean lifestyle, such as washing hands, obeying health protocols, keeping a distance, and complying with the recommendation not to go home to break the chain of the spread of COVID-19. This effort is to protect themselves and their families from COVID-19.

In cluster 1, the words "Indonesia" and "*sehat*" (healthy) appeared during the PPKM period, the government and community leaders invited them to maintain health protocols and carry out vaccinations during the PPKM period to lead to a healthy Indonesia, extraordinary Indonesia, and a rising economy. In the 2nd, 3rd, and 4th clusters that intersect, the word "*pencegahan penyebaran*" (preventing the spread) is an effort to prevent the spread of COVID-19 which is apparently mostly voiced by the police. It means, the police exist enough to appeal to the public through social media (especially Twitter). A clean and healthy lifestyle here also appears as an effort to prevent the COVID-19 virus. Many educations are delivered via Twitter, such as how to wash hands properly to break the chain of spreading COVID-19. While cluster 5 emerged from an appeal to protect yourself and your family by not going home and having picnics during the COVID-19 pandemic, especially during the PPKM period. This appeal not to go home came during the Eid moment as one of the steps to prevent the spread of COVID-19.

The process of clustering by time shows several views. For the first time series, 11 clusters were formed, where clusters 1, 2, and 4 intersect. Clusters 3, 6-10 also have similarities, while clusters 5 and 11 have different characteristics. Clusters 1, 2, and 4 showed a lot of public concern at the beginning of the COVID-19 pandemic, including the public's view of health workers who were deemed not good at handling COVID-19 patients, such as: "*Ini bikin ngeri sihh masuk rumah sakit bukan sehat malah nambah penyakit klo begini semoga segera teredukasi staf medisnya kayaknya mereka takut ngerawat pasien covid*" (It's scary to go to the hospital, it's not healthy, it just adds to the disease. Hopefully, if this is the case, the medical staff will be educated soon. It seems like they are afraid to take care of Covid patients). In addition to people who hope that the pandemic will end soon, such as: "*Sehat-sehat nggihBuat kalian gausah risau akan virus COVID-19 ini yaa..kalian punya Allah, kalian punya tuhan kan? Mintalah pertolongan pengampunan sama Allah azza wajjala. Ingat Allah gaakan ngasih berupa penyakit kalau tidak ada obatnya. Yang terpenting kita selalu tetap hidup bersih sehat.*" (Stay healthy. Don't worry about this COVID-19 virus, okay? You have God, and you have God, right? Ask for forgiveness from

Allah *azza wajjala*. Remember God will not give you a disease if there is no cure. The most important thing is that we always live clean and healthy). Appeals to maintain a clean and healthy lifestyle, as well as to maintain distance have also begun to be spread, such as: "*Yuk kita sebarkan informasi pencegahan covid di medsos, WA grup, line grup, telegram, dan sebagainya, Saling mengingatkan untuk menerapkan pola hidup bersih sehat, terapkan sosial distancing*." (Let's spread covid prevention information on social media, WA groups, line groups, telegrams, and so on, remind each other to apply a clean and healthy lifestyle, apply social distancing). In clusters 3 and 6 to10, many appeals and invitations to maintain a healthy, clean lifestyle and health protocols were conveyed. In addition, in this cluster, there are many mutually encouraging tweets in the face of the COVID-19 pandemic. Meanwhile, in other clusters, appeals not to panic in the face of COVID-19 and self-isolation education appeared.

In the second time series, 2 clusters were formed, which showed that to prevent and break the chain of the spread of COVID-19, complying with the health protocol was something that had to be taken. During this period, irresponsible persons emerged who traded COVID-19 free letters, some examples of discussions in the community such as: "*Surat bebas covid dijual seharga 70 ribu hingga 39 juta. Senyumin aja. Urusan bayar surat keterangan sehat/sakit dokter udah dari jaman dulu. Doain semoga semoga yang manipulasi hisabnya nanti ga berat, kasihan sama yang begitu :)*" (Covid-free letters are sold for 70 thousand to 39 million. Just smile. The business of paying for a doctor's health/ill certificate has been around for a long time. Pray that the manipulation of the reckoning will not be too heavy, sorry for those who are like that :)), "*Baru-baru ini beredar di media sosial surat keterangan sehat bebas COVID dari sebuah rumah sakit swasta. Surat ini dijual seharga Rp 70 ribu*." (Recently circulating on social media a health certificate free of COVID from a private hospital. This letter is sold for Rp. 70 thousand), "Tokopedia membenarkan ada penjual nakal yang menjual surat keterangan sehat bebas COVID-19" (Tokopedia marketplace confirms that there are rogue sellers who sell COVID-19-free health certificates). When the positive number of COVID-19 increases, some people are not responsible, which is a very unfortunate thing and can impact people's mental health while facing COVID-19.

At the third time, namely during the new normal, 8 clusters were formed. At this time, calls for continuing to implement a healthy lifestyle and comply with health protocols are still emerging, both from the government, community leaders, and the community itself. Interestingly, in clusters 3 and 6, which overlap each other, "*pilkada*" (regional head selection) appears, which is true. At this time, simultaneous elections are being held. The mechanism for selecting regional heads still pays attention to health protocols for both voters and voter officers when voting is carried out so that the elections run smoothly. However, it turns out that the public's response to the elections during this pandemic has also reaped negative sentiments, such as: "*klaster pilkada jangan sampai ada*" (*Pilkada* cluster should not exist)," *Pilih sehat atau pilkada? Pandemi COVID-19 makin ganas lho ini.*" (Choose healthy or *pilkada*? The COVID-19 pandemic is getting fiercer), "*Kenapa hak dasar rakyat untuk hidup sehat sulit dipenuhi? Sementara di sisi lain, beberapa oknum terlihat ngotot selenggarakan pilkada di masa genting wabah covid?*" (Why is the basic right of the people to live a healthy life difficult to fulfill? Meanwhile, on the other hand, some individuals are seen as insistent on holding local elections during the critical period of the COVID-19 outbreak), and many more. It can be seen that the Indonesian people are more alert and worried that a new COVID-19 cluster would emerge from the *pilkada* cluster. However, in this election, the tagline that was carried was "*Pilkada Sehat*" (healthy *pilkada*) and became an event to invite the public to continue to comply with health protocols.

At the fourth time, namely, during the PPKM period, two ideal clusters were formed where there were not many calls for a clean and healthy lifestyle and maintaining health protocols. Little still appears about the importance of hand washing. The main point here is the appeal to keep your distance and avoid crowds. This shows that it is essential to maintain a healthy lifestyle and health protocols from the beginning of the COVID-19 pandemic until the PPKM period is still emerging, although not as massive as at the beginning of the pandemic. Different from the other three times, at this fourth time,s the two clusters showed that "*vaksin*" (vaccine) was an important step to protect public health from the COVID-19 virus. Most Tweets at this point are talking about the COVID-19 vaccination program. This vaccination program is the government's effort to solve the COVID-19 problem in line with health protocols that are continuously maintained. However, there are pros and cons in the community regarding this vaccination program, including: "*Vaksin bukan solusi, Indonesia butuh segera tarik rem. Ayo gaungkan petisi*" (Vaccines are not a solution, Indonesia needs to immediately pull the brakes. Let's echo the petition), "*Vaksin bisa dijual mahal, vitamin nggak bisa dijual mahal. Jadi permainan kapitalis.*" (Vaccines can be expensive, vitamins can't be expensive. So, it's a capitalist game.), "*Ragu pake vaksin*" (Do not hesitate to use vaccines), and so on. However, not a few also welcomed the presence of the COVID-19 vaccine and hoped that vaccination could protect against COVID-19.

CONCLUSION

Social media analysis technology can be one way to reveal phenomena that occur in society. This study utilizes social media analysis to determine Indonesian public health behavior changes during the COVID-19 pandemic. Using the Latent Dirichlet Allocation method, it was found that advice and education on maintaining health, increasing immunity, complying with health protocols, wearing masks, avoiding crowds and maintaining distance emerged from various parties, both from the government and community leaders. This is a form of public awareness in tackling and breaking the chain of the spread of COVID-19. At the beginning of the pandemic, activities to maintain health protocols were still in the form of appeals and education, which later became the rules set by the government in tackling COVID-19. Even during the new normal, calls for maintaining health protocols are still being voiced. Even the Pilkada is an event to invite the public to continue to comply with health protocols. However, as time goes by and adaptation to COVID-19, there are not many calls for a clean and healthy lifestyle and maintaining health protocols. Little still appears about the

importance of washing hands and wearing masks. The main point here is the appeal to keep your distance and avoid crowds. This fact shows the importance of maintaining a healthy lifestyle and health protocols from the beginning of the Covid-19 pandemic until the Emergency Community Activity Restrictions (PPKM) continue to emerge, although not as massively as at the beginning the pandemic. It also shows that maintaining health protocols and keeping yourself healthy has become a habit of the community. Further research can use social media analysis technology with other methods such as sentiment analysis, topic modeling, segmentation of social media users, opinion mining and others to uncover various phenomena in society.

ACKNOWLEDGMENT

We would like to thank the Center for Research and Publications, Institute for Research and Community Service at UIN Sunan Gunung Djati Bandung for supporting and financing this publication.

REFERENCES

[1] World Health Organization, "WHO Coronavirus (COVID-19) Dashboard," 2022. https://covid19.who.int/ (accessed June 6, 2022).

[2] Satuan Tugas Penanganan COVID-19, "Data Sebaran COVID-19," *covid19.go.id*, 2022. https://covid19.go.id/ (accessed Nov. 08, 2021).

[3] A. S. Wardani, "Pengguna Internet Dunia Tembus 4,66 Miliar, Rata-Rata Online di Smartphone," *liputan6.com*, 2021. https://www.liputan6.com/tekno/read/4469008/pengguna-internet-dunia-tembus-466-miliar-rata-rata-online-di-smartphone (accessed Oct. 03, 2021).

[4] M. I. Marsyaf, "Jumlah Pengguna Internet Sedunia Mencapai 4,66 Miliar," *sindonews.com*, 2021. https://tekno.sindonews.com/read/316920/207/jumlah-pengguna-internet-sedunia-mencapai-466-miliar-1611820860 (accessed Oct. 03, 2021).

[5] V. B. Kusnandar, "Penetrasi Internet Indonesia Urutan ke-15 di Asia pada 2021," *Databoks*, 2021. https://databoks.katadata.co.id/datapublish/2021/07/12/penetrasi-internet-indonesia-urutan-ke-15-di-asia-pada-2021 (accessed Oct. 03, 2021).

[6] R. K. Nistanto, "Berapa Lama Orang Indonesia Akses Internet dan Medsos Setiap Hari?," *Kompas.com*, 2021. https://tekno.kompas.com/read/2021/02/23/11320087/berapa-lama-orang-indonesia-akses-internet-dan-medsos-setiap-hari-?page=all#:~:text=Dari total populasi Indonesia sebanyak,3 persen dibandingkan tahun lalu.

[7] K. Borne, "Top 10 List – The V's of Big Data," *Data Science Central*, 2014. https://www.datasciencecentral.com/profiles/blogs/top-10-list-the-v-s-of-big-data.

[8] S. Sagiroglu and D. Sinanc, "Big data: A review," in *Proceedings of the 2013 International Conference on Collaboration Technologies and Systems, CTS 2013*, 2013, doi: 10.1109/CTS.2013.6567202.

[9] S. Stieglitz, L. Dang-Xuan, A. Bruns, and C. Neuberger, "Socialmedia analytics," *Bus. Inf. Syst. Eng.*, 2014, doi: 10.1007/s12599-014-0315-7.

[10] P. Brooker, J. Barnett, and T. Cribbin, "Doing social media analytics," *Big Data Soc.*, 2016, doi: 10.1177/2053951716658060.

[11] I. Lee, "Social media analytics for enterprises: Typology, methods, and processes," *Bus. Horiz.*, 2018, doi: 10.1016/j.bushor.2017.11.002.

[12] S. Stieglitz, M. Mirbabaie, B. Ross, and C. Neuberger, "Social media analytics: Challenges in topic discovery, data collection, and data preparation," *Int. J. Inf. Manage.*, vol. 39, pp. 156–168, 2018.

[13] Ahmad, "10 Sosial Media Paling Populer di Indonesia," *gramedia.com*, 2021. https://www.gramedia.com/best-seller/sosial-media-paling-populer/ (accessed Oct. 03, 2021).

[14] Y. Dahono, "Data: Ini Media Sosial Paling Populer di Indonesia 2020-2021," *Kompas.com*, 2021. https://www.beritasatu.com/digital/733355/data-ini-media-sosial-paling-populer-di-indonesia-20202021 (accessed Oct. 03, 2021).

[15] "Ciri Khas Yang Membedakan Pengguna Twitter, Instagram dan Facebook," *dipa14.web.id*, 2020. https://dipa14.web.id/2020/12/16/ciri-khas-yang-membedakan-pengguna-twitter-instagram-dan-facebook/ (accessed Aug. 11, 2021).

[16] A. I. Kabir, R. Karim, S. Newaz, and M. I. Hossain, "The Power of Social Media Analytics: Text Analytics Based on Sentiment Analysis and Word Clouds on R.," *Inform. Econ.*, vol. 22, no. 1, 2018.

[17] C. Li, L. J. Chen, X. Chen, M. Zhang, C. P. Pang, and H. Chen, "Retrospective analysis of the possibility of predicting the COVID-19 outbreak from Internet searches and social media data, China, 2020," *Eurosurveillance*, vol. 25, no. 10, p. 2000199, 2020.

[18] L. Nemes and A. Kiss, "Social media sentiment analysis based on COVID-19," *J. Inf. Telecommun.*, vol. 5, no. 1, pp. 1–15, 2021.

[19] J. C. M. Serrano, M. Shahrezaye, O. Papakyriakopoulos, and S. Hegelich, "The rise of Germany's AfD: A social media analysis," in *Proceedings of the 10th international conference on social media and society*, 2019, pp. 214–223.

[20] S. Becken, H. Friedl, B. Stantic, R. M. Connolly, and J. Chen, "Climate crisis and flying: social media analysis traces the rise of 'flightshame,'" *J. Sustain. Tour.*, vol. 29, no. 9, pp. 1450–1469, 2021.

[21] M. Bhat, M. Qadri, M. K. Noor-ul-Asrar Beg, N. Ahanger, and B. Agarwal, "Sentiment analysis of social media response on the Covid19 outbreak," *Brain. Behav. Immun.*, vol. 87, p. 136, 2020.

[22] P. Chapman *et al.*, "Cross Industry Standard Process for Data Mining 1.0," *Step-by-step Data Min. Guid.*, 2000.

[23] P. Chapman *et al.*, "The CRISP-DM user guide," in *4th CRISP-DM SIG Workshop in Brussels in March*, 1999, vol. 1999.

[24] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[25] D. M. Blei, "Probabilistic topic models," *Commun. ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[26] N. Seth, "Part 2: Topic Modeling and Latent Dirichlet Allocation (LDA) using Gensim and Sklearn," *Analytics Vidhya*, 2021. https://www.analyticsvidhya.com/blog/2021/06/part-2-topic-modeling-and-latent-dirichlet-allocation-lda-using-gensim-and-sklearn/ (accessed October 9, 2021).

[27] S. Andayani and A. Ryansyah, "Implementasi Algoritma TF-IDF Pada Pengukuran Kesamaan Dokumen," *JuSiTik J. Sist. dan Teknol. Inf. Komun.*, vol. 1, no. 1, pp. 53–62, 2017.

[28] M. Chahal, "Measuring Similarity between Documents Using TF-IDF Cosine Similarity Function," *Int. J. Res. Publ. Semin.*, vol. 9, no. 1, pp. 53–57, 2018.

[29] P. Bafna, D. Pramod, and A. Vaidya, "Document clustering: TF-IDF approach," in *International Conference on Electrical, Electronics, and Optimization Techniques, ICEEOT 2016*, 2016, doi: 10.1109/ICEEOT.2016.7754750.

[30] M. Roder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proceedings of the eighth ACM international conference on Web search and data mining*, 2015, pp. 399–408.

[31] D. S. Maylawati, H. Aulawi, and M. A. Ramdhani, "Flexibility of Indonesian text pre-processing library," *Indones. J. Electr. Eng. Comput. Sci.*, 2019, doi: 10.11591/ijeecs.v13.i1.pp420-426.

[32] S. Kannan *et al.*, "Preprocessing Techniques for Text Mining," *Int. J. Comput. Sci. Commun. Networks*, 2015.

[33] S. Vijayarani, J. Ilamathi, and M. Nithya, "Preprocessing Techniques for Text Mining - An Overview," *Int. J. Comput. Sci. Commun. Networks*, vol. 5, no. 1, pp. 7–16, 2015, [Online]. Available: http://www.ijcscn.com/Documents/Volumes/vol5issue1/ijcscn2015050102.pdf.