Journal of Sisfotek Global

ISSN (Online): 2721 - 3161, ISSN (Print): 2088 – 1762
DOI: http://dx.doi.org/10.38101/sisfotek.v12i1.448
Vol. 12, No. 1, March 2022, pp. 18-23

Analysis of COVID-19 Information Based On Social Media Big Data Classification Using the K-Means Data Mining Method

Lukas Umbu Zogara[1], Ahmad Sururi[2] & Leny Tritanto Ningrum[3]
19116013651[1], 19116008622[2],19116013323[3]

[1,2,3] Budi Luhur University, Jakarta, Indonesia, 12260
E-mail: [1]lukasumbuzogara68@gmail.com, [2]ahmadsururi71@gmail.com, [3]leny.online@gmail.

ARTICLE HISTORY

Received : 5 February, 2022
Revised : 2 March, 2022
Accepted : 22 March, 2022

KEYWORDS

Covid-19
Data Mining
Social Media
K-Means

ABSTRACT

This Covid-19 began to infect almost all countries in early 2020, including in Indonesia, Covid-19 spread widely throughout the world and was declared as a global pandemic by the World Health Organization (WHO). In the current era of Big Data, large amounts of data have been generated and collected from a variety of rich data sources. Big Data is useful information and valuable knowledge. In this study, the method that will be used for data analysis is the K-Means algorithm with orange tools as a tool to display the results of data classification. One of the information that can be generated is Sentiment Analysis. The purpose of this research was to determining the information such as public sentiment on social media towards government policies in handling COVID-19. In this research, 2000 tweets were used. The keyword used related to government policies are sourced from several online media. The tools used to analyze this twitter data is using Orange Application. The selected keywords are covid19, lockdown, PSBB, and isolation. This keyword is used as a reference to retrieve tweet data from twitter. From each of these keywords, a sentiment classification process will be carried out automatically so that data or tweets are obtained and grouped into positive, negative and neutral sentiment classes. From the result of research conducted, public sentiment on social media towards government policies in handling this virus outbreak tends to be positive.

1. Introduction

Corona or Covid-19 has become a pandemic in early 2020. This virus first came from China, precisely in the city of Wuhan, has spread rapidly almost throughout the world including Indonesia. Indonesia, the entire National Disaster Management Agency (BNPB) has declared a disaster emergency status related to this virus. The Indonesian government finally issued a policy, namely implementing a lockdown or known as Large-Scale Social Restrictions (PSBB). However, the community responded to the policy in various ways, some were pro and some were con. These comments can be found on social media which is for anyone to give their opinion freely. The most widely used social media to express appreciation or comments is Twitter. Data from tweets can be used as a big data source. Of course these comments are used to get good information if they can be processed properly [1].

To process large data used data mining process [2]. There are processes in data mining ranging from pre-processing of data sets to selecting methods to generate valuable information [3]. The results obtained are information that can be generated, namely sentiment analysis. Sentiment evaluation is used to show public opinion on issues, operator satisfaction, policies based on textual data. Referring to the data found on the www.internetworldstats.com page, in 2012, the number of internet users in the world was 2.4 billion. Based on the population of internet users, this results in an increase in the level of internet access and is able to create an increasing data population as well. The very large number of data populations on the internet today is called big data [4]. Big data is able to provide more in-depth information than traditional data analysis [5]. Big data refers to the development of new technologies designed to extract value from data that has four characteristics, namely volume, variation, speed, and correctness [6]. Analysis on big data also aims to extract information that has abundant sources of knowledge and can be processed with data mining techniques to be able to generate predictions, identify trends, explore hidden information, and make decisions [7]. Examples of most of the use of big data are on social media data, network data, disease reports, statistical data [8]. Social media that are widely used today include Facebook, Twitter, LinkedIn, WhatsApp, Instagram [9].

Based on the description described above, the purpose of this research is to analyze Covid-19 information based on the Big Data classification of social media using the K-Means Data Mining method. The analysis of information trends that is the focus of this research is the analysis of sentiment towards Covid-19.

2. Research Method

This research is divided into four parts, namely: data collection, selection of the application of data mining methods, and generating information & knowledge. Process data. At the assembling stage, it is determined the keyword parameters for crawling data from social media, preprocessing the text that has been successfully obtained, and the output data that is clean and ready to be processed. In the process of applying and selecting data mining methods, several tools are used to process data using certain algorithms. The next stage, the results from the previous stage are processed into useful information. The results of the previous stage are table data, statistics, recapitulation, and so on. To become information, it is necessary to make analysis and study that is visualized in a form that is easily understood by humans as users.

2.1 Data Method

In this study, we used the K-Means algorithm for data analysis and used data from twitter. Tweet data is retrieved based on certain keywords in accordance with Covid-19 policies. The tweet data used is 2000 tweets. Keywords are determined from several words. The sources used are trusted online media. Tweet data that has been downloaded will be processed using the K-Means algorithm by first grouping positive, negative, or neutral data and turning it into an attribute with a numerical value so that it can be analyzed by K-Means. Tweet data with a positive value will be assigned a value of 1, negative -1, and neutral with a value of 0. After that, the distance from each data will be calculated to find the closest distance for each row. The closest data distance will be grouped into the same cluster member to form a cluster group. The results of the visualization of the cluster will be displayed using the Orange tools which can be seen in the results section.

Every online media contains articles that discuss terms that emerged during the Covid-19 pandemic. From these several sources, several terms related to government policies were chosen that appeared or intersect in various sources. Determination of the keywords used in the study can be seen in Table 1.

Table 1. List Of Government Policies Keyword Related to COVID-19 from Twitter Online Media

Source / Media	Keyword	Government Policies
Liputan 6	Masker N95, WFH, Suspect, Positif, Lockdown, Social Distancing, Isolasi, Karantina, ODP,ODP, Hand Sanitiser, Fasaynakes	Lockdown, Social Distancing, Isolasi, Karantina, WFH
Detik	Pandemi, PDP, OTG, Suspect, PSBB, ODP, Physical Distancing, WFH, Karantina, Isolasi, Lockdown, Rapid test, Swab Test, PCR, Positif,	PSBB, Social Distancing, WFH, Karantina, Isolasi, Lockdown, Rapid Test, SWAB Test
Alodokter	Social Distancing, Isolasi, Karantina, Lockdown, Flattening Curve, PDP, ODP, OTG, OTG, Herd Immunity, PSBB	Social Distancing, Isolasi, Karantina, Lockdown, PSBB
Tribun	Rapid Test, Antispetik, PDP, Suspect, Positif, Lockdown, Karantina, Social Distancing, Isolasi,, WFH, Wabah, epidemi, Pandemi, ODP, Disinfektan	Lockdown, Social Distancing, Isolasi, Karantina, WFH, Rapid Test

From Table 1 above, it can be seen some keywords related to government policies that will be used in data collection keywords on Twitter. Keywords that appear only once are not included. The keywords “isolation” and “quarantine” were changed to “self-isolation” and “area quarantine” based on available twitter data. The keywords used in this research are: covid19, lockdown, PSBB and isolation. There are 4 keywords to use in retrieving tweet data on Twitter.

2.2 Data Mining Process

Data mining or also known as knowledge discovery in database (KDD) is the discovery of knowledge from big data by utilizing data in a database to be processed in order to produce new information that is more useful [10].

Broadly speaking, the stages of data processing with KDD techniques in data mining are as follows:

1. Data selection, meaning that not all data in the database is used, therefore only the appropriate data is taken from the database [11].
2. Data cleaning is the process of removing noise and inconsistent or irrelevant data.
3. Data integration is merging data from several databases into one new database.
4. Data transformation. Data is converted or combined into a format suitable for processing in data mining.
5. The mining process is the main process before applying the method to find the required knowledge from the data [12].
6. Knowledge presentation is the visualization and presentation of knowledge about the method used, to obtain the knowledge required by the user.
7. Interpretation or evaluation is the resulting pattern of information displayed in an easy-to-understand form [13].

The system architecture used in this research can be seen in Figure 1.

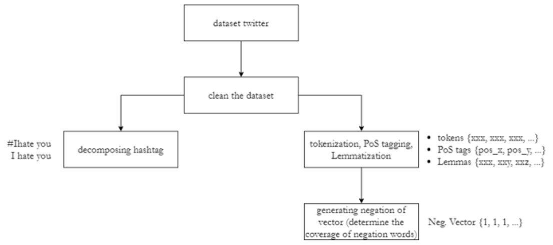

dataset twitter
clean the dataset
#Ihate you
I hate you
decomposing hashtag
tokenization, PoS tagging, Lemmatization
• tokens {xxx, xxx, xxx, ...}
• PoS tags {pos_x, pos_y, ...}
• Lemmas {xxx, xxy, xxz, ...}
generating negation of vector (determine the coverage of negation words)
Neg. Vector {1, 1, 1, ...}

Figure 1. Data Mining Process System Architecture

The stages carried out in this study are [14]:

1. Initial research, studying things related to research.
2. Data collection, the data used is tweet data about covid from Twitter.
3. Data analysis, the data that has been collected is analyzed based on research needs.

The data that used in this research is sourced from Twitter social media data. Some of the social media that are widely used provide features to be able to pull data with an API (Application Programming Interface) access method. The data that can be retrieved is in the form of posts or writings from each user using the licensing rules determined by the social media [15].

Data retrieved from social media will be stored in a database location. At the time of crawling data, keyword parameters can be obtained from the user where these keywords are generated from the process of literature study and analysis of knowledge from researchers. Furthermore, the data that has been stored in the local database will be processed by data mining.

The closing level is processing the outcomes of the previous tiers into a precious data. The output of the results of the preceding degree are table facts, information, recapitulation, and so forth. To end up information, it is essential to make analysis and have a look at this is visualized in a shape this is effortlessly understood with the aid of human beings as customers. The form of visualization may be within the form of photographs, or written narration.

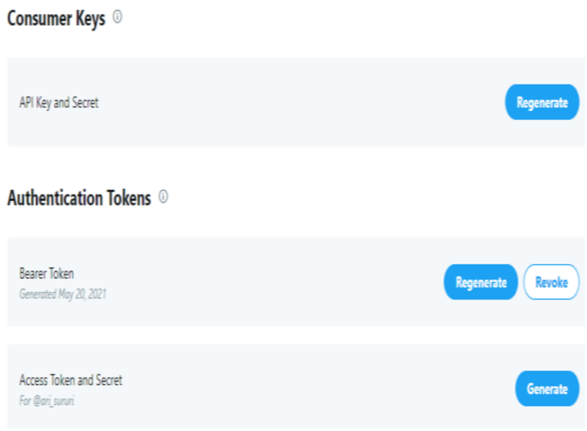
Headings: type and label phase and subsection headings inside the fashion proven on these pages. Use numbered sections, for you to facilitate go references.

3. Result And Review

In this research, the Orange application was used to carry out the sentiment classification process. Some of the libraries used are text mining. Text Mining works as a library to connect to Twitter with standard verification rules that have been determined by Twitter called Twitter developer.

The process of collecting data from twitter is done with orange tools. In order to be able to connect to Twitter, the application created must have some kind of key or key generated from Twitter. There are 4 types of keys that are required to be placed in the program code, those are : access token, access secret, consumer API keys and consumer secret. The page to get the key is: https://developer.twitter.com/en/apps.

The user performs the registration process on the page by providing some required information. These data are useful for Twitter in verifying our needs in making applications. If the specified conditions are met, the verification process will be accepted. After the verification process is accepted, the access key will be obtained as shown in Figure 2.


Consumer Keys

API Key and Secret

Authentication Tokens

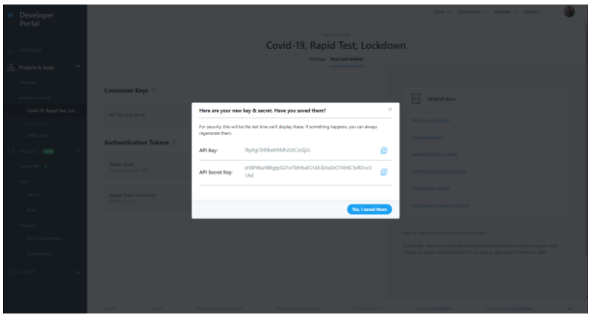
Bearer Token
Generated May 20, 2021

Access Token and Secret
For @ori_sururi

Figure 2. Keys and Tokens Information Page

On that page, there is some information we need regarding to consumer access keys and access tokens.

API keys are unique for each application that we create. While access tokens will always be generated when we need to see it. The display to see the API key is as shown in Figure 3.



Developer Portal

Projects & Apps

Covid-19, Rapid Test, Lockdown

Consumer Keys

Authentication Tokens

Here are your new key & secret. Have you saved them?

For security, this will be the last time we'll display these. If something happens, you can always regenerate them.

API Key: [illegible]

API Secret Key: [illegible]

Yes, I saved them

Figure 3. API Keys and Tokens Page View

After being verified by twitter, the next process is collecting tweet data using orange tools. Before the data is collected, the main step is to create a schema or data processing flow in orange starting from importing data to the final result in cluster form. The schematic display of tweet data collection is shown in Figure 4.

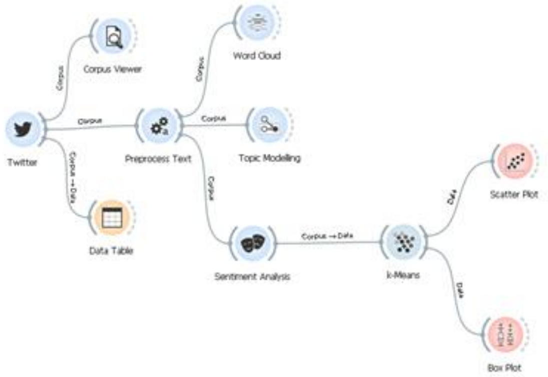
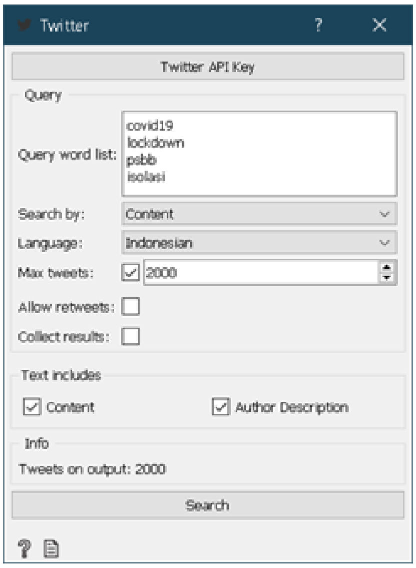


Figure 4. Tweet Data Collection Scheme using Orange Tools

The Orange application can crawl data with certain parameters. Several parameters that need to be entered are keywords and the number of tweets, as shown in Figure 5.



Twitter

Twitter API Key

Query

Query word list:
covid19
lockdown
psbb
isolasi

Search by: Content

Language: Indonesian

Max tweets: ☑ 2000

Allow retweets: □

Collect results: □

Text includes

☑ Content ☑ Author Description

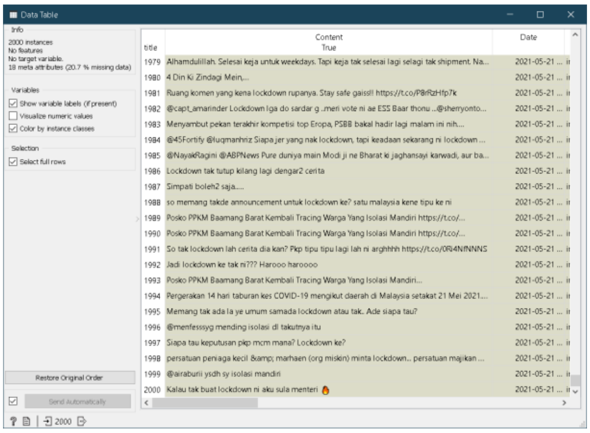
Info

Tweets on output: 2000

Search

Figure 5. Keyword Input Process and Target Number of Tweets Using Orange Tools

Twitter data that can be retrieved is data that is within a maximum period of the last one month. The display of the data collection process is shown in Figure 6.



Data Table

Info
2000 instances
No features
No target variable.
18 meta attributes (20.7 % missing data)

Variables
☑ Show variable labels (if present)
□ Visualize numeric values
☑ Color by instance classes

Selection
☑ Select full rows

Restore Original Order

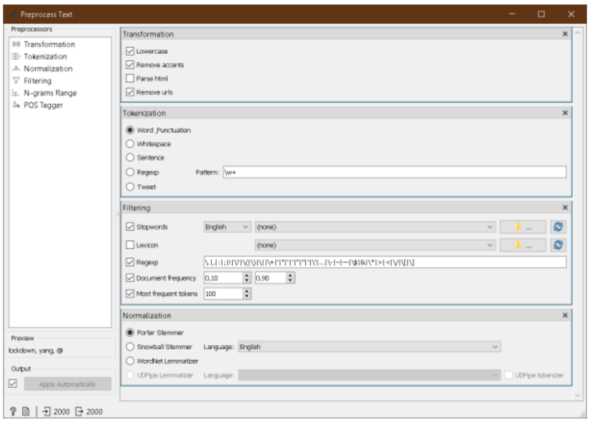
☑ Send Automatically

2000

title	Content True	Date
1979	Alhamdulillah. Selesai keja untuk weekdays. Tapi keja tak selesai lagi selagi tak shipment. Na...	2021-05-21 ...
1980	4 Din Ki Zindagi Mein,...	2021-05-21 ...
1981	Ruang komen yang kena lockdown rupanya. Stay safe gaiss!! https://t.co/P8rRzHfp7k	2021-05-21 ...
1982	@capt_amarinder Lockdown lga do sardar g ..meri vote ni ae ESS Baar thonu ...@sherryonto...	2021-05-21 ...
1983	Menyambut pekan terakhir kompetisi top Eropa, PSBB bakal hadir lagi malam ini nih....	2021-05-21 ...
1984	@45Fortify @luqmanhriz Siapa jer yang nak lockdown, tapi keadaan sekarang ni lockdown	2021-05-21 ...
1985	@NayakRagini @ABPNews Pure duniya main Modi ji ne Bharat ki jaghansayi karwadi, aur ba...	2021-05-21 ...
1986	Lockdown tak tutup kilang lagi dengar2 cerita	2021-05-21 ...
1987	Simpati boleh2 saja.....	2021-05-21 ...
1988	so memang takde announcement untuk lockdown ke? satu malaysia kene tipu ke ni	2021-05-21 ...
1989	Posko PPKM Baamang Barat Kembali Tracing Warga Yang Isolasi Mandiri https://t.co/...	2021-05-21 ...
1990	Posko PPKM Baamang Barat Kembali Tracing Warga Yang Isolasi Mandiri https://t.co/...	2021-05-21 ...
1991	So tak lockdown lah cerita dia kan? Pkp tipu tipu lagi lah ni arghhhh https://t.co/0Ri4NfNNNS	2021-05-21 ...
1992	Jadi lockdown ke tak ni??? Harooo haroooo	2021-05-21 ...
1993	Posko PPKM Baamang Barat Kembali Tracing Warga Yang Isolasi Mandiri...	2021-05-21 ...
1994	Pergerakan 14 hari taburan kes COVID-19 mengikut daerah di Malaysia setakat 21 Mei 2021....	2021-05-21 ...
1995	Memang tak ada la ye umum samada lockdown atau tak.. Ade siapa tau?	2021-05-21 ...
1996	@menfesssyg mending isolasi dl takutnya itu	2021-05-21 ...
1997	Siapa tau keputusan pkp mcm mana? Lockdown ke?	2021-05-21 ...
1998	persatuan peniaga kecil &amp; marhaen (org miskin) minta lockdown... persatuan majikan	2021-05-21 ...
1999	@airaburii ysdh sy isolasi mandiri	2021-05-21 ...
2000	Kalau tak buat lockdown ni aku sula menteri 🔥	2021-05-21 ...

Figure 6. Tweet Data Collected from Twitter Using Orange Tools

Furthermore, after the data from twitter is obtained, the next process is the data mining process. The first stage is the pre-processing of the data. The process begins by decapitating each word. Then the filtering process is to remove words, symbols, and various other characters that have no meaning. As seen in Figure 7.



Preprocess Text

Preprocessors
- Transformation
- Tokenization
- Normalization
- Filtering
- N-grams Range
- POS Tagger

Transformation
- ☑ Lowercase
- ☑ Remove accents
- □ Parse html
- ☑ Remove urls

Tokenization
- ● Word_Punctuation
- ○ Whitespace
- ○ Sentence
- ○ Regexp   Pattern: \w+
- ○ Tweet

Filtering
- ☑ Stopwords   English   (none)
- □ Lexicon   (none)
- ☑ Regexp   [illegible]
- ☑ Document frequency   0,10   0,90
- ☑ Most frequent tokens   100

Normalization
- ● Porter Stemmer
- ○ Snowball Stemmer   Language: English
- ○ WordNet Lemmatizer
- ○ UDPipe Lemmatizer   Language: □ UDPipe tokenizer

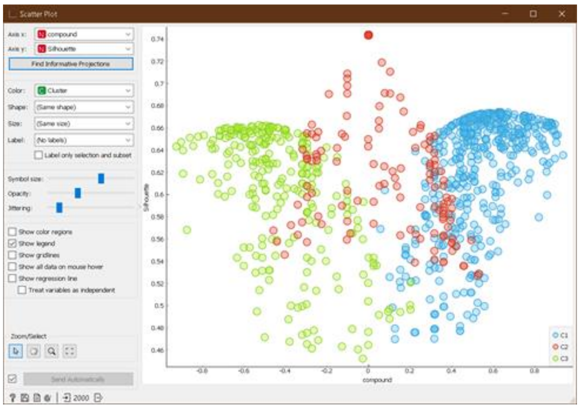
Preview
lockdown, yang, @

Output
☑ Apply Automatically

2000   2000

Figure 7. Pre-Processing Stage of Tweet Data Using Orange Tools

After completing the pre-processing step, the next step is the classification process using the K-Means algorithm. This classification process is carried out by generating a model which can then be used to classify new data that does not yet have a sentiment class. To facilitate the visualization of the output, the classification results are displayed in data grouping such as Scatter Plot and Box Plot. Scatter Plot displays data grouping based on K-Means clusters, while Box Plot displays cluster values in the form of numbers so that positive, negative, and neutral values can be known. The cluster display with Scatterplot can be seen in Figure 8 and Box Plot in Figure 9.



Scatter Plot

- Axis x: compound
- Axis y: Silhouette
- Find Informative Projections
- Color: Cluster
- Shape: (Same shape)
- Size: (Same size)
- Label: (No labels)
- □ Label only selection and subset
- Symbol size:
- Opacity:
- Jittering:
- □ Show color regions
- ☑ Show legend
- □ Show gridlines
- □ Show all data on mouse hover
- □ Show regression line
- □ Treat variables as independent
- Zoom/Select
- ☑ Send Automatically

2000

Figure 8. Tweet Data Cluster Result with Scatter Plot Visualization Using Orange Tools

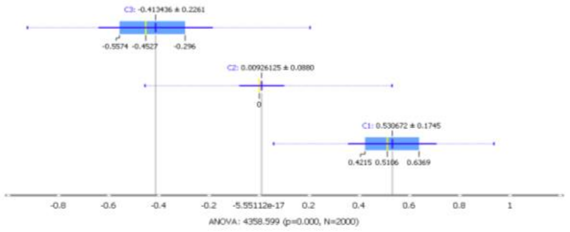


Figure 8. Tweet Data Cluster Result with Box Plot Visualization Showing Positive, Negative and Neutral Values Using Orange Tools

Figure 8 shows the results of the cluster pattern formed from the results of tweet data processed with the K-Means algorithm using the orange tool. The blue scatter plot shows the results of a positive response to the COVID-19 policy. The red scatter plot shows the results of a neutral response, and the green color indicates a negative response. While figure 9 is the result of visualization on the orange tool based on a box plot that shows the overall results of netizens' responses to the policies implemented by the government during the pandemic. Figure 9 shows that negative responses are higher than neutral and positive, thus in this period it can be said that most of the users' responses to tweets lead to negative results.

4. Conclusion

Social media as a place for people to provide comments can be a useful source of data if it is processed properly and correctly. Twitter is one of the social media that can be used to collect data related to public opinion about something. Various features have been provided by Twitter to be used as needed. In this research using data from twitter for 4 keywords (keywords) the total data used is 2000 tweets.

In this Research , one of the methods used in data mining is classification using the K-Means algorithm. The classification used in this research is sentiment classification with three classes, such as : positive, negative, neutral. From the results of research conducted, public sentiment on social media related to government policies tends to be positive. Further research can be done regarding the development of algorithms or the selection of other tasks in data mining. For example, related to the analysis of estimates or policy predictions that can be associated with other policies. In addition, the selection of keywords for extracting datasets can be further expanded.

References

[1] Enda Esyudha Pratama, H. Sastypratiwi, and Yulianti, “Analisis Kecenderungan Informasi Terkait Covid-10 Berdasarkan Big Data

Sosial Media dengan Menggunakan Metode Data Mining," *J. Inform. Polinema*, vol. 7, no. 2, pp. 1–6, 2021, doi: 10.33795/jip.v7i2.453.

[2] N. Elgendy and A. Elragal, "Big Data Analytics in Support of the Decision Making Process," *Procedia Comput. Sci.*, vol. 100, pp. 1071–1084, 2016, doi: 10.1016/j.procs.2016.09.251.

[3] E. R. E. Sirait, "Implementasi Teknologi Big Data Di Lembaga Pemerintahan Indonesia," *J. Penelit. Pos dan Inform.*, vol. 6, no. 2, p. 113, 2016, doi: 10.17933/jppi.2016.060201.

[4] C. E. wahyudi Utomo, "Implementasi Bussiness Intelligent dalam e-Tourism Berbasis Big Data," *J. Tour. Creat.*, vol. 3, no. 2, p. 163, 2019, doi: 10.19184/jtc.v3i2.14065.

[5] Y. He, F. R. Yu, N. Zhao, H. Yin, H. Yao, and R. C. Qiu, "Big Data Analytics in Mobile Cellular Networks," *IEEE Access*, vol. 4, pp. 1985–1996, 2016, doi: 10.1109/ACCESS.2016.2540520.

[6] L. A. Tawalbeh, R. Mehmood, E. Benkhlifa, and H. Song, "Mobile Cloud Computing Model and Big Data Analysis for Healthcare Applications," *IEEE Access*, vol. 4, pp. 6171–6180, 2016, doi: 10.1109/ACCESS.2016.2613278.

[7] M. Marjani *et al.*, "Big IoT Data Analytics: Architecture, Opportunities, and Open Research Challenges," *IEEE Access*, vol. 5, pp. 5247–5261, 2017, doi: 10.1109/ACCESS.2017.2689040.

[8] C. K. Leung, Y. Chen, C. S. H. Hoi, S. Shang, Y. Wen, and A. Cuzzocrea, "Big Data Visualization and Visual Analytics of COVID-19 Data," *Proc. Int. Conf. Inf. Vis.*, vol. 2020-September, no. Iv, pp. 415–420, 2020, doi: 10.1109/IV51561.2020.00073.

[9] Q. Qi and F. Tao, "Digital Twin and Big Data Towards Smart Manufacturing and Industry 4.0: 360 Degree Comparison," *IEEE Access*, vol. 6, pp. 3585–3593, 2018, doi: 10.1109/ACCESS.2018.2793265.

[10] F. Nur, M. Zarlis, and B. B. Nasution, "Penerapan Algoritma K-Means Pada Siswa Baru Sekolahmenengah Kejuruan Untuk Clustering Jurusan," *InfoTekJar (Jurnal Nas. Inform. dan Teknol. Jaringan)*, vol. 1, no. 2, pp. 100–105, 2017, doi: 10.30743/infotekjar.v1i2.70.

[11] F. Yunita, "Penerapan Data Mining Menggunkan Algoritma K-Means Clustring Pada Penerimaan Mahasiswa Baru," *Sistemasi*, vol. 7, no. 3, p. 238, 2018, doi: 10.32520/stmsi.v7i3.388.

[12] R. Helilintar and I. N. Farida, "Penerapan Algoritma K-Means Clustering Untuk Prediksi Prestasi Nilai Akademik Mahasiwa," *J. Sains dan Inform.*, vol. 4, no. 2, pp. 80–87, 2018, doi: 10.34128/jsi.v4i2.140.

[13] A. Asroni, H. Fitri, and E. Prasetyo, "Penerapan Metode Clustering dengan Algoritma K-Means pada Pengelompokkan Data Calon Mahasiswa Baru di Universitas Muhammadiyah Yogyakarta (Studi Kasus: Fakultas Kedokteran dan Ilmu Kesehatan, dan Fakultas Ilmu Sosial dan Ilmu Politik)," *Semesta Tek.*, vol. 21, no. 1, pp. 60–64, 2018, doi: 10.18196/st.211211.

[14] M. A. Rheza and F. Metandi, "Implementasi Metode K-Means Clustering Untuk Penentuan Jenis Komentar Pada Tweet Pssi," *Just TI (Jurnal Sains Terap. Teknol. Informasi)*, vol. 12, no. 2, p. 73, 2020, doi: 10.46964/justti.v12i2.363.

[15] X. Chen, M. Vorvoreanu, and K. P. C. Madhavan, "Mining social media data for understanding students' learning experiences," *IEEE Trans. Learn. Technol.*, vol. 7, no. 3, pp. 246–259, 2014, doi: 10.1109/TLT.2013.2296520.