Journal of Computers for Society
Journal homepage: https://ejournal.upi.edu/index.php/JCS
Vol. 1 No. 1, June-2020, pp. 91-100

A Text Mining Implementation Based on Twitter Data to Analyse Information Regarding Corona Virus in Indonesia

Enda Esyudha Pratama[a], Rizqia Lestika Atmi[b]

[a] *Informatic Study, Universitas Tanjungpura, Indonesia*
[a]enda@informatika.untan.ac.id
[b] *Politeknik Negeri Ketapang, Indonesia*
[b]rizqia.lestika@gmail.com

Abstract

CORONA virus outbreak (COVID-19) began to infect almost all countries in early 2020 including Indonesia. Since its distribution, various information has been spread in the community from various sources, one of them is social media. Various terms also appear on social media related to the corona virus. This study analyzes related terms that emerge from social media-based. The data used was sourced from Twitter in the past month where the data processed was text data. The method used is text mining. Text Mining is a method used to extract important information from a group of texts. From the results of the research conducted, there are several terms or information that tend to appear frequently on social media, namely "*PSBB*", "*new normal*", "*karantina*", and "*juru bicara Dr. Reisa*".

Keywords: Corona Virus, Terms, Twitter, Text Mining

First draft received: Juni 17, 2020
Date Accepted: June 23, 2020
Final proof received: June 29, 2020

1. Introduction

The CORONA virus outbreak (COVID-19) began infecting almost all countries in early 2020. The World Health Organization (WHO) has established a global status in the global emergency related to the COVID-19 virus since January 2020. National Disaster Management Authority (Badan Nasional Penanggulangan Bencana - BNPB) has established a state of disaster emergency from 29 February 2020 to 29 May 2020. Various countries have taken policies to anticipate this widespread outbreak including

in Indonesia. The Indonesian government itself has also issued several policies related to this virus pandemic.

The policy has drawn various responses from the people of Indonesia. One of the means or media for the public in voicing their responses is social media. Instagram and Twitter are some of the most popular social media choices used by the public to comment and issue opinions on Covid-19 [1]. In this study, data from twitter will use to crawl data.

Various terms also appear based on tweets generated from Twitter. The large and varied amount of data makes it difficult to obtain information regarding the terms most widely discussed if the process is done manually. An automation mechanism is needed to carry out the analysis process and produce information from this large amount of data. One method that can be used is text mining.

Text mining method is a technique or automation method to get important information contained in a collection of texts [2]. This approach is generally used to analyze data that is semi-structured such as text.

Based on the description above, an analysis of information trends related to terms related to corona virus will be carried out based on tweets from Twitter using text mining methods.

2. Methods

2.1 Corona Virus (COVID-19)

Corona virus is an RNA virus with a particle size of 120-160 nm. This virus mainly infects animals, including bats and camels. Before the COVID-19 outbreak, there were 6 types of coronaviruses that could infect humans, such as alphacoronavirus 229E, alphacoronavirus NL63, betacoronavirus OC43, betacoronavirus HKU1, Severe Acute Respiratory Illness Coronavirus (SARS-CoV), and Middle East Respiratory Syndrome Coronavirus (MERS-CoV).

Coronavirus which is the etiology of COVID-19 belongs to the genus Betacoronavirus. The results of phylogenetic analysis showed that this virus entered the

same subgenus as the coronavirus that caused the outbreak of Severe Acute Respiratory Illness (SARS) in 2002-2004, namely Sarbecovirus. On this basis, the International Committee on Taxonomy of Viruses submitted the name SARS-CoV-2. [3].

2.2 Text Mining

Text mining is one of the techniques that can be used to do classification where, text mining is a variation of data mining that tries to find interesting patterns from a large collection of textual data. [4].

According to [5], text mining is similar with data mining, except for data mining techniques that are designed to work on structured data in a database, but text mining can work on unstructured or semi-structured data such as complete text documents, code/script web pages, and others.

In general, the major stages in text mining consist from three main parts namely text pre-processing, feature selection, and text analytics [6]. On the stages of text pre-processing in generally are tokenizing, filtering, stemming, tagging, and analyzing. Tokenizing is a step to separate each word (token) in an input document. Filtering is a selection process for words that are generated from the tokenizing process, can be done with a stop list or word list algorithm. The stop list algorithm will discard words that are not important such as pronouns, adverbs, conjunctions, prepositions, and clothing. Instead, the word list algorithm will store important words.

The working mechanism of text mining algorithms generally has similarities with data mining algorithms. The utilization of text mining can be used to solve problems such as analysis, classification, clustering, or prediction and information retrieval [7].

In this study, the approach used is Term Frequency (TF) to calculate the frequency of occurrence of words. The TF approach does not heed the terms contained in other documents. The TF method simply calculates the appearance of terms in a document. Terms that have a high frequency of occurrence will be a feature of a document where the term is located [8]. Moreover, some implementations of text mining can be found at the studies [9, 10, 11, 12, 13].

2.3 Research Flow

Generally, this research stage is divided into three main parts, namely: data collection, text mining process, output visualization. At the data collection stage, various processes are carried out starting from determining the keyword parameters to crawling data from social media. Determination of keywords is done by conducting studies from various trusted online sources on terms related to corona virus.

Furthermore, for the text mining process there are two main processes namely the pre-process stage and the application of the algorithm. At the pre-process stage will be tokenize, filtering using a black list, ie discarding words that do not have meaning. Then in the application phase algorithm the Term Frequency (TF) algorithm is used to count the number of words.

The last stage is processing the results of the previous stage into valuable information. The output from the results of the previous stage, generally in the form of data tables, statistics, recapitulation, etc. To become an information, it is necessary to create a visualization that explains the analysis and study so that it is easily understood by humans as users. Forms of visualization can be in the form of graphics, narratives, or writing. The detailed stages of the research can be seen in **Figure 1**.

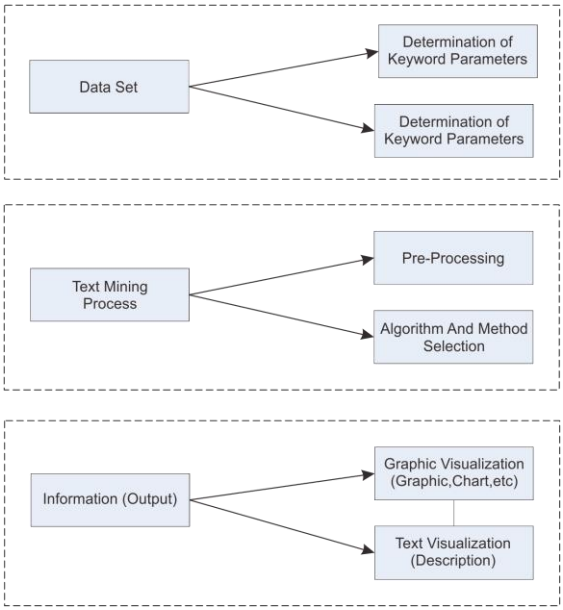

***Figure 1**. The stage of research flow.*

3. Results and Discussion

In this research, Pyhton programming language based Spyder tools are used to process text mining. Some of the libraries used are tweepy, nltk, and wordcloud. Tweepy functions as a library to connect to twiter with standard verification rules that have been determined by Twitter, namely Oauth and REST API architecture. After being verified by Twitter, the application can process data crawling with certain parameters. Some parameters that need to be entered are keywords, number of tweets, date limits. Twitter data that can be retrieved within a maximum period of one month. This process can be seen in **Figure 2**.

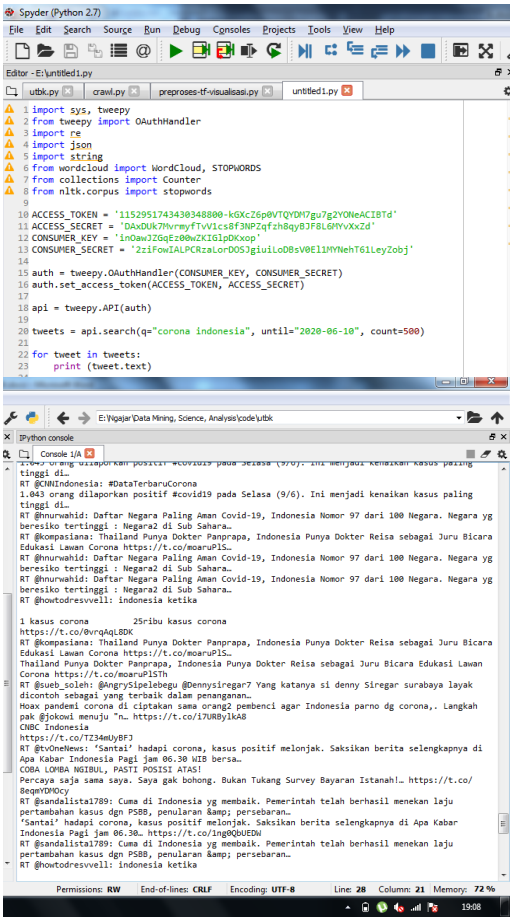

***Figure 2**. The display of the data crawling program from twitter using Spyder.*

Furthermore, after the data from Twitter is obtained, the next process is the text preprocessing. At this stage, the process of fragmenting each word is carried out. Then the filtering process discards words, symbols, and various other characters that have no meaning. Python has also provided special stopword data for the Indonesian language. Stopword is a collection of words that have no meaning and in the process will be deleted or discarded.

After preprocessing, the next step is to count the number of occurrences of words with the term frequency (TF) algorithm. To facilitate output visualization, the results of the term frequency will be displayed in tabular and WordCloud form. Tables and WordCloud has seen using a browser. The results of the program code can be seen in **Figure 3**.

***Figure 3**. The display of the program for the process of text mining and output visualization.*

The analysis process is done by comparing the term frequency results to the parameters of search keywords. The crawling process is repeated with different keywords. Next will be seen in which words intersect between these keywords.

On keyword "*corona indonesia*", the list of 20 words that most often appears is "*Indonesia*", "*kasus*", "*positif*", "*corona*", "*covid19*", "*kenaikan*", "*paling*", "*tinggi*", "*dokter*", "*beresiko*", "*tertinggi*", "*juru"*, "*reisa*", "*membaik*", "*menekan*", "*pemerintah*", "*penularan*", "*persebaran*", "*pertambahan*", and "*PSBB*". If word related to keywords are not included, the word list becomes: "*kasus*", "*positif*", "*kenaikan*", "*paling*", "*tinggi*", "*dokter*", "*beresiko*", "*tertinggi*", "*juru*", "*reisa*", "*membaik*", "*menekan*", "*pemerintah*", "*penularan*", "*persebaran*", "*pertambahan*", and "*PSBB*". From the word list, some of the terms that are identical to the corona virus case are "*juru bicara dokter reisa*", "*kasus positif dengan kenaikan palingg tinggi*", "*upaya pemerintah dalam menekan penularan/ persebaran/ pertambahan kasus corona*" and "*PSBB*".

Furthermore, from some information obtained from the keyword "*corona Indonesia*", several related keywords were developed again such as "*juru bicara corona*", "*positif corona*", "*pemerintah cegah corona*", and "*psbb*".

In the keyword "*juru bicara corona*", several terms appear, such as: "*satgas*", "*informasi*", "*mengedukasi*", "*gugus tugas*", and "*achmad yurianto*". Then if each of these terms is reprocessed, the information produced will intersect with "*juru bicara corona*". For the keyword "*pemerintah cegah corona*", several information or terms that appear are: "*program*", "*penanganan*", "*karantina*", "*kebijakan*", "*psbb*", and "*new normal*". Meanwhile, for the keyword "*positif corona*", the information and terms that appear are: "*pesawat*", "*penerbangan*", "*penumpang*", "*tembus1000*", and "*kasus*". The result of information and terms for the keyword "*PSBB*" are: "*new normal*", "*karantina*", "*pelonggaran*", "*perekonomian*", and "*tarif listrik*".

The other keyword parameters used are "*istilah corona*" and "*informasi corona*". From the keyword "*istilah corona*", information that appears are: "*new normal*", "*pandemic*", "*virus*", "*second wave*". Meanwhile, for the keyword "*informasi corona*", the results that most appear are: "*juru bicara*", "*dr reisa*", "*update*", "*satgas*", and "*data*". From the description above, there are several terms or information from several

keywords that have similarities or intersecting. The terminology or information contained: "*PSBB*", "*new normal*", "*karantina*", and "*juru bicara Dr. Reisa*".

The analysis of this research can be compared with other applications of text mining, such as in [9, 14]. So, it can be seen that this research provides different strategy to obtain the results. Moreover, some methods in machine learning, for example: Rough Sets [15], Fuzzy Sets [16], and Gradient Descent [17] can be embedded to improve and enhance results. To reduce the computational cost, approaches using parallel computing [18] and data streaming [19] can be performed.

4. Conclusion

Based on the discussion above, it can be concluded several things as follows: (i) Social media, especially Twitter can be a source of data for analyzing public information related to the corona virus; (ii) The keywords used are "*corona Indonesia*", "*juru bicara corona*", "*positif corona*", "*pemerintah cegah corona*", "*psbb*", "*istilah corona*," and "*informasi corona*"; (iii) Trends in information or terms produced and analyzed, i.e.: "*PSBB*", "*new normal*", "*karantina*", and "*juru bicara Dr. Reisa*". The analysis process is done by using the text mining method and the term frequency (TF) algorithm to see the number of words that appear most often and intersect between one search keyword with another search keyword.

References

[1] Chen, E., Lerman, K., & Ferrara, E. (2020). Tracking social media discourse about the COVID-19 pandemic: development of a public coronavirus Twitter data set. *JMIR Public Health and Surveillance*, *6*(2), 1-9.

[2] Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of emerging technologies in web intelligence*, *1*(1), 60-76.

[3] Lai, C. C., Shih, T. P., Ko, W. C., Tang, H. J., & Hsueh, P. R. (2020). Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and corona virus disease-2019 (COVID-19): the epidemic and the challenges. *International journal of antimicrobial agents, 55*, 1-9.

[4] Feldman, R., & Sanger, J. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press.

[5] Liao, S. H., Chu, P. H., & Hsiao, P. Y. (2012). Data mining techniques and applications–A decade review from 2000 to 2011. *Expert systems with applications*, *39*(12), 11303-11311.

[6] Gaikwad, S. V., Chaugule, A., & Patil, P. (2014). Text mining methods and techniques. *International Journal of Computer Applications*, *85*(17), 42-45.

[7] Berry Michael, W. (2004). Automatic discovery of similar words. *Survey of Text Mining: Clustering, Classification and Retrieval", Springer Verlag, New York, LLC*, 24-43.

[8] Hakim, A. A., Erwin, A., Eng, K. I., Galinium, M., & Muliady, W. (2014). Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (TF-IDF) approach. In *2014 6th International Conference on Information Technology and Electrical Engineering (ICITEE)* (pp. 1-4). IEEE.

[9] Mediayani, M., Wibisono, Y., Riza, L. S., & Pérez, A. R. (2019). Determining trending topics in twitter with a data-streaming method in R. *Indonesian Journal of Science and Technology*, *4*(1), 148-157.

[10] Riza, L. S., Putra, B., Wihardi, Y., & Paramita, B. (2019). Data to text for generating information of weather and air quality in the R programming language. *Journal of Engineering Science and Technology*, *14*(1), 498-508.

[11] Riza, L. S., Pertiwi, A. D., Rahman, E. F., Munir, M., & Abdullah, C. U. (2019). Question Generator System of Sentence Completion in TOEFL Using NLP and K-Nearest Neighbor. *Indonesian Journal of Science and Technology*, *4*(2), 294-311.

[12] Eslami, B., Rezaei, Z., Habibzadeh, M., Fouladian, M., & Ebrahimpour-Komleh, H. (2020). Using deep learning methods for discovering associations between drugs and side effects based on topic modeling in social network. *Social Network Analysis and Mining*, *10*, 1-17.

[13] Liu, Y., Peng, H., Li, J., Song, Y., & Li, X. (2020). Event detection and evolution in multi-lingual social streams. *Frontiers of Computer Science*, *14*(5), 1-15.

[14] Salloum, S. A., Al-Emran, M., Monem, A. A., & Shaalan, K. (2017). A survey of text mining in social media: facebook and twitter perspectives. *Adv. Sci. Technol. Eng. Syst. J*, *2*(1), 127-133.

[15] Riza, L. S., Janusz, A., Bergmeir, C., Cornelis, C., Herrera, F., Ślezak, D., & Benítez, J. M. (2014). Implementing algorithms of rough set theory and fuzzy rough set theory in the R package "RoughSets". *Information Sciences*, *287*, 68-89.

[16] Riza, L.S., Bergmeir, C., Herrera, F., Benítez, J.M. (2015). frbs: Fuzzy rule-based systems for classication and regression in R. *Journal of Statistical Software, 65*(6), 1-30.

[17] Riza, L. S., Nasrulloh, I. F., Junaeti, E., Zain, R., & Nandiyanto, A. B. D. (2016, August). gradDescentR: An R package implementing gradient descent and its variants for regression tasks. In *2016 1st International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)* (pp. 125-129). IEEE.

[18] Riza, L. S., Rachmat, A. B., Munir, T. H., & Nazir, S. (2019). Genomic Repeat Detection Using the Knuth-Morris-Pratt Algorithm on R High-Performance-Computing Package. *Int. J. Advance Soft Compu. Appl*, *11*(1), 94-111.

[19] Riza, L. S., Pratama, F. D., Piantari, E., & Fashi, M. (2020). Genomic repeats detection using Boyer-Moore algorithm on Apache Spark Streaming. *Telkomnika*, *18*(2), 783-791.