Automatic Smart Crawling on Twitter for Weather Information in Indonesia

Kartika Purwandari, Reza Perdana, Join Sigalingging, Reza Rahutomo, Bens Pardamean

Automatic Smart Crawling on Twitter for Weather Information in Indonesia

Kartika Purwandari
Bioinformatics & Data Science Research Center
Bina Nusantara University
Jakarta, Indonesia
kartika.purwandari@binus.edu

Reza Bayu Perdana
Database Center Division of BMKG Meteorological, Climatological, and Geophysical Agency
Jakarta, Indonesia 10720
reza.perdana@bmkg.go.id

Join W. C. Sigalingging
Database Center Division of BMKG Meteorological, Climatological, and Geophysical Agency
Jakarta, Indonesia 10720
join.wan.chanlyn@bmkg.go.id

Reza Rahutomo
Information Systems Department, School of Information Systems
Bina Nusantara University
Jakarta, Indonesia 11480
reza.rahutomo@binus.edu

Bens Pardamean
Computer Science Department, BINUS Graduate Program – Master of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
bpardamean@binus.edu

***Abstract*—As a popular resource for analyzing social interactions and text data mining, Twitter utilization is facing an automation problem in collecting Twitter users' geolocation. To surpass this problem, the research proposes Support Vector Machine (SVM) model that can be used to automatically design a smart crawling system on Twitter. Twint, a Python-based Twitter scraping program is utilized to perform data crawling based on keywords related to the weather in Indonesia. Null-geolocations are filled toward using aliases generated based on Indonesians' behavior of reporting about Indonesia's location in Twitter tweets. The accuracy of the outcomes of automated smart crawling using the SVM model is 85%.**

Keywords—Social media, Twitter, SVM, smart crawling, weather

I. INTRODUCTION

The importance of social media in disseminating information is getting higher since the number of social media users keeps increasing, and more discussions in the scope of politics, economy, social life, and other fields are opened [1]–[3]. Facebook and Twitter are social media platforms with the most number of users with wide age range [4]. It is believed that the number of users does not necessarily disseminate information through social media, but the presence of users of this information can be used to determine the level of trustworthiness of information [5], [6].

On July 2021, there are more than 500 million Twitter users from all over the world, and Indonesia is ranked in sixth place with 15.7 million users, while the United States is in first place with 73 million users [7]. The achievement symbolizes that the growing number of users in Indonesia keeps increasing and reliable to be a source for big data engineering.

The phenomenon of big data generated by social media can be in the form of public opinion, social behavior, and geospatial points (locations) of characteristics of social media users [8]. Rahutomo et al. conduct research on the people movement called #SaveKPK from 2009 to 2019. This research presents an exploratory data analysis that shows the most frequently used words, most active users, most active account dates, most mentioned accounts, most hashtags, trending topics, and the most relevant keywords [9].

Getting geospatial data is another benefit that can be obtained in data collection through social media. Covering geographical level dimensions and objects' natural characteristics, geospatial data is able to locate the source of public perception and social behavior on the issue. It is considered that the information can be utilized by various interested parties, namely government, company, and research area [10], [11]. As geolocation is embedded in social media posts, a special method is required to extract the information. For certain purposes, this information can be selected so that significantly useful information is obtained [12].

The research focuses on the use of the Twitter platform for data collection in the form of weather conditions in Indonesia. The automatic smart crawling technique is applied to this data collection to fill in the empty geolocation from the collected data based on the aliases of the names of regions in Indonesia that have been created.

II. PREVIOUS WORK

A. Twitter for Data Collection

In infodemiology studies, Twitter is known as an important data source for monitoring public response [13]. The important things that can be collected from Twitter are tweet full text, the numbers of favorites, followers and friends, user' geolocation, user' description/self-created profile, etc. Many researchers use hashtags as a convenient way to collect data, but there is limitation arising from hashtag-based data collection [14], [15]. In specific topics such as political communication during an election period, there is various size dataset for data collection using hashtags. However, the limitation of the specific hashtags has become common practice even though the size of acquired data and users vary from several to thousands of objects [16]. Accordingly,

researchers use different synonyms for keywords that could have been more useful entry points for data collection. When setting up a data collection approach based on keywords or other full-text searches, it is important to summarize the choice of search terms and consider potential alternatives [17].

B. Tools for Crawling Twitter Data

Some Python-based tools are able to collect Twitter data. Twitter Application Programming Interface (API) is a program or application provided by Twitter to make it easier for other developers to access information on the Twitter website [18], [19]. Twitter API allows us to collect updated Twitter posts from any keyword inserted i.e. 'COVID-19' [20]. Several Python-based tools also allow users to obtain data from Twitter servers. For example, Rahutomo et al. utilize Getoldtweets2, a Python programming language library, to collect Twitter posts with hashtag '#SaveKPK'.

Another tool for crawling Twitter data is Tweepy Streaming API that can be utilized to collect a random sample of real-time on community. Despite its capability, the API is limited to track the number of keywords at once. The API must be run separately to another installation to track several keywords [9].

C. Method for Automatic Crawl on Twitter Data

A Java-based data collection application developed by Byun et al. aims to crawl Twitter using keywords, user accounts, number of tweets, and radio buttons for saving the data into a database by inputting keywords, user accounts, and tweets [21], [22]. Twitter Application Programming Interface (API) [23] has been implemented inside crawling system using the PHP programming language and MySQL database.

D. SVM as Binary Classification

In recent times, the task of automatic text classification is being extensively studied, and rapid progress is being recorded in this area, including the deep learning [24]–[26] and machine learning approaches such as Support Vector Machine (SVM) [27]. The SVM problem is to find the decision surface that maximizes the margin between the data points of the two classes. Positive and negative training sets, which are not common for other classification [28] methods, are needed for SVM model. It is proposed to seek the decision surface that best separates the positive from the negative data in the n-dimensional space called the hyperplane [29], [30].

III. Research Methodology

Figure 1 illustrates the Twitter data mining approach. It consists of four stages: data collection, preprocessing, finalizing, and storing.

Data collection, as the first process, starts with crawled Twitter using specific keywords automatically applied by crontab job. The second process is data pre-processing implemented with Indonesian slang words [31] file and Sastrawi stemming tool. Data finalization using SVM model has been implemented in the third process. The last step is data storing into a database called weather report.

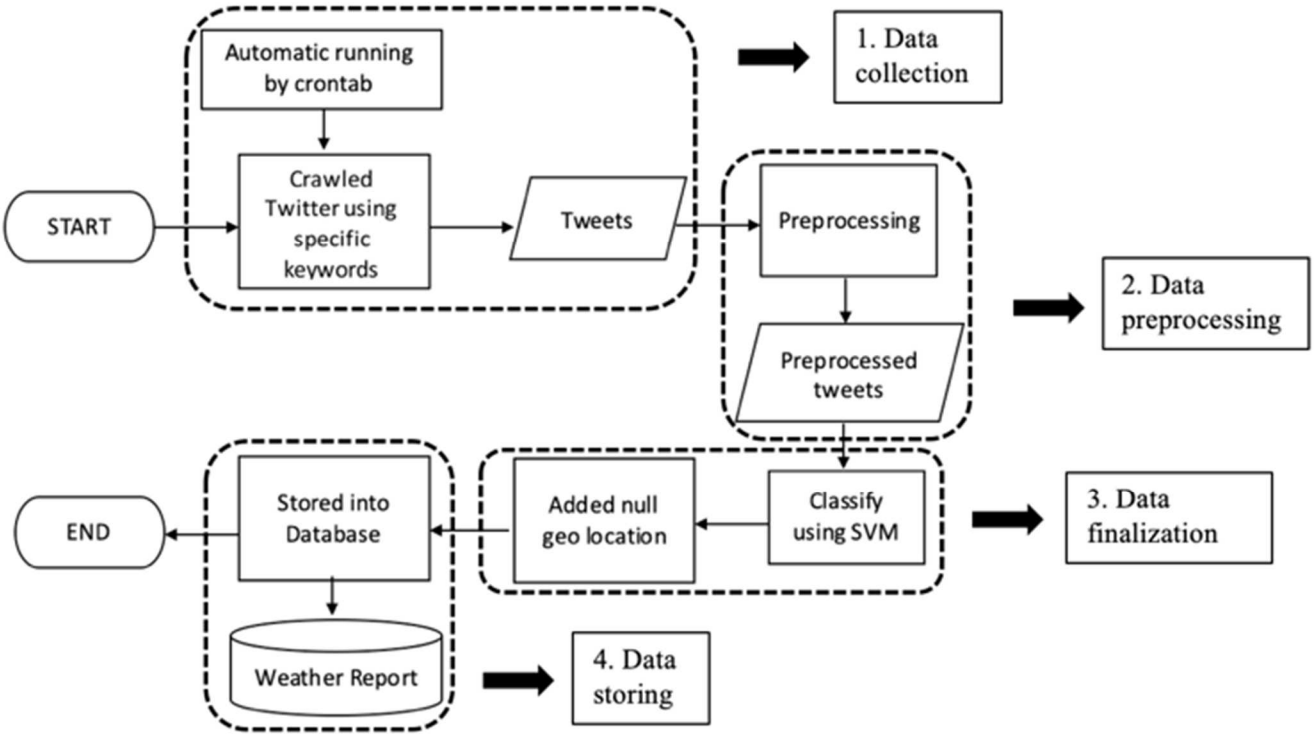

Fig. 1. Automatic Smart Crawling Process.

A. Data Collection

The Twitter scrapers used to collect the data were written in Python, called Twint, which generated a collection of text data based on the posts uploaded by Twitter users. In addition, this process also collects some embedded geospatial data in the form of coordinates that are sourced from the location of Twitter users when checking in while posting on Twitter. The vector data and the resulting coordinates are the requirements of the application that is built-in in determining the location of Twitter users. Data results show the source of the two-dimensional coordinates (longitude and latitude).

Data crawling is influenced by several factors, such as the network connection, the retrieval time, and the latest news updates that are to be crawled. First, a stable internet network connection will facilitate the data crawling process. On the other hand, the unstable internet connection will cause connection errors, or data retrieval process will be slow and intermittent. A second factor is the length of the data crawling process; the longer the crawling process, the more information is obtained.

As the third factor, the latest news about weather information must be crawled in real-time every last hour. Thus, when real-time data will be crawled, the use of the selected keywords running using the current words namely "*currently*" and "*today*" will be processed. Consequently, the process of deleting future and past words is carried out such as "*last time*", "*future*", "*forecasting*", "*tomorrow's weather*", "*yesterday*", "*the day after tomorrow*", "*last week*", etc. Indonesia language code "*id*" was used to crawl tweets in Indonesian. Weather-specific keywords applied in crawling commands are "*weather*" OR "*light*" OR "*sunny*" OR "*cloudy*" OR "*raining*" OR "*drizzle*" OR "*thunder*" OR "*lightning*" OR "*flash*".

Using the cronjob tool, the weather tweets can be automatically scheduled on the server [32]. Cron is a tool on UNIX-based operating systems (Linux, Ubuntu, etc.) that functions to run tasks or scripts automatically. Therefore, cronjob is a term for using cron to schedule a task repeatedly at a predetermined time. Each cronjob is represented by a crontab file. Each crontab file consists of two components namely time and command. Cron syntax crawled between 500 tweets a time using the Twint tool which has stored in a Python file (.py). It run every "00" minute for 24 times and stored in a logging file (.log). An example of the syntax we have used is 00 * * * * /usr/bin/python3.8 /home/ crawling/main.py > /home/crawling/scrap.log 2>&1.

B. Data Pre-processing

The public government accounts, namely the official BMKG account and the BMKG Technical Implementation Unit (UPT) do not qualify for further processing since they do not come from public information. In addition, some accounts detected as bots are also not released.

The tweet filtering and cleaning process performed using Python tools include converting tweets to lowercase, cleaning external links, removing symbols and numbers, and removing mentions. Replacing slang words with appropriate words by following the writing rules is also carried out in the pre-processing process. Indonesian words with affixes have changed into their basic forms using a Python library called Sastrawi [33]. The results carried out by the crawling filtering and cleaning processes are divided into training, validation, and testing data. As shown in table 1, data training and validation are carried out manually using a binary concept that develops tweets into a real-time weather tweet class (1) or a non-real-time weather tweet class (0).

TABLE I. SOME EXAMPLE OF TWEETS BEFORE TEXT PROCESSING

Weather class label	Tweet
1	light drizzle... it's really good to take pictures of the camera's lights dripping with rain, the result is something like this https://t.co/Yz3DlFB40B
1	heavy rain tonight, quite perfect when compared to what the heart says
0	Tomorrow's weather forecast, Jakarta will be cloudy at noon
0	Yesterday it rained heavily so I couldn't run in GBK

C. Data Finalization

The support vector machine (SVM) model can be applied after manually labeled data has been successfully categorized. SVM is a supervised learning method considered familiar for text classification to find out the linear separator hyperplane that maximizes the margin, i.e. the Optimal Separating Hyperplane (OSH), and maximizes the margin between two data sets [34]. Figure 2 illustrates the optimal hyperplane and its support-vector, strengthened by the linear equation shown in equation (1).

$$y = ax + b \tag{1}$$

Performance was calculated by using accuracy, which is presented as equation (2).

$$Accuracy = \frac{correctly\ predicted}{total\ of\ data} \tag{2}$$

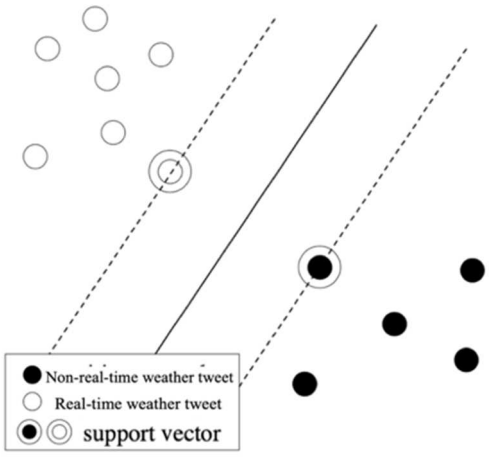

Fig. 2. Illustrated SVM for Weather Tweet Classification.

D. Data Storing

MySQL is known as an open-source relational database management system that uses Structured Query Language (SQL). SQL is the standard language used to access server databases. A database is a collection of structured data. Four basic operations that can be implemented on any database insert, select (query), update, and delete. Data insertion begins with weather report databases with an average of 500 tweets. The id of weather report is generated automatically and filling of the username, created at, tweets, latitude, and longitude have used MySQL command called "INSERT INTO".

IV. Result and Discussion

A. Result

Wordcloud visualization is utilized to describe 100 words that often appear in weather tweets from a total of 4,201 tweets on September 2, 2021, within 24 hours. The larger the font size in the wordcloud, the more often the word appears in the text. Figure 3 shows that the word "*rain*" often appears in the wordcloud (976 times). An example of the tweet is "*how do I go home, I'm caught in heavy rain*".

Fig. 3. Wordcloud Result.

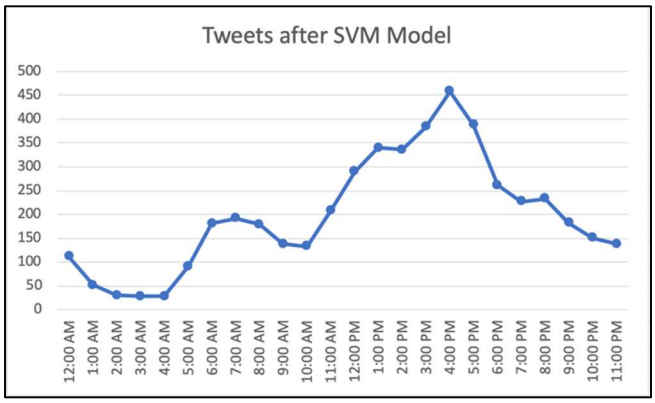

Fig. 4. Tweet Volume using SVM Model.

GMT (Greenwich Mean Time) is the standard time in London. Indonesia uses GMT+7, which is 7 hours ahead of GMT, which was the time zone we used for this study. For example, if London time is 12.00, it is 19.00 in the GMT+7 time zone.

Figure 4 shows the trend of tweets about weather conditions in Indonesia on September 2, 2021, after the SVM model was applied. The volume of tweets per hour continues to increase from 10 AM to the peak point at 4 PM. The data shows that the average number of conversations about the weather has increased starting at 4 PM and has decreased again starting at 5 PM. However, despite the decreased amount, the number of tweets keeps coming at 5 PM. The peak of the conversation occurs at 4 PM, in which there are 458 tweets due to the large number of people starting to return home. At 11 AM, it starts to increase as workers begin to enter break time, causing a decrease in work effectiveness to open social media, especially Twitter.

According to the results of the SVM model, it shows that several tweets still have null values in the geolocation column. The filling of this geolocation is used to help place tweets in locations that correspond to longitude and latitude. Aliases in .csv form are created based on people's tweeting habits. This .csv file contains city and county aliases with two-column definitions. Furthermore, city and district geolocation point data information from the Central Statistics Agency (BPS) is used to help fill in the zero geolocation. Indonesia has collected up to 543 latitude and longitude information points from each district and city, according to the latest data from BPS. Latitude is the line that determines the location north or south of the equator. Latitude is measured from 0° from the equator to 90° at the poles and longitude is measured from 0° in the Greenwich area to 180° on the International Date Line. An example of changing aliases to a complete district or city name along with latitude and longitude is shown in table 2.

TABLE II. Complete Tweet Using Attribute Latitude and Longitude

Aliases	Name of district/city	Latitude	Longitude
jaksel, jkt selatan	Kota Jakarta Selatan	-6.33297	106.807915
mlg, ngalam	Kota Malang	-7.98189	112.626503
jakpus	Kota Jakarta Pusat	-6.21154	106.845172
bgr	Kota Bogor	-6.6	106.8

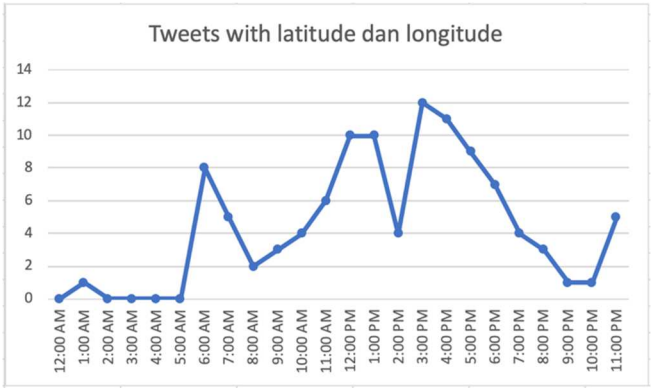

Fig. 5. Tweet Colume After Latitude and Longitude Filling.

Figure 5 shows the total number of tweets per hour after filling out the latitude and longitude. A majority of the tweets were posted at 3 pm (12 tweets). According to the data, only a few Twitter users use geolocation and mention their location when posting a tweet. 106 tweets can be pre-processed as an example according to table 2 by transforming aliases into district/city names.

Twitter is collected by using a SVM algorithm that is trained to detect tweets pertaining to real-time weather conditions in Indonesia. In the first step, a collection of text documents is converted to numerical feature vectors. The most popular method to convert text data into vectors which a model can comprehend is called TF-IDF. As a result of assigning the metrics to each term, this acronym stands for "Term Frequency - Inverse Document" Frequency. TF-IDF is a word frequency scoring system that highlights words from a document with greater interest, e.g. words that are less common across documents [35].

Finally, data will be transformed to Train_X and Test_X to be vectorized to Train_X_Tfidf and Test_X_Tfidf respectively. The rows will now contain a list of unique integer numbers and their importance, as calculated by TF-IDF. SVM classifying weather tweets was performed with 85% accuracy. This was done by using linear kernels.

As a model evaluation, confusion matrix was used to determine the comparison between real-time weather and non-real-time weather. Comparing the classification results reported by the model with the actual results can be accomplished using the confusion matrix. This confusion matrix is displayed as a matrix table that summarizes the SVM model's performance on tweet test data related to weather conditions. The binary classification used as the output value for the class is real-time weather tweet class "1" or non-real-time weather tweet class "0". As shown in figure 6, the combination of the predicted value and the actual value shows the results to predict which tweets belong to the current weather conditions. From the figure of 2001 tweets, the SVM model predicts 532 tweets to be negative real-time weather tweets (FN), and from 2099 tweets, the model predicts 98 tweets to be positive real-time weather tweets (FP).

The correct prediction is located in a diagonal table (bold font), so visually it is easy to see the prediction error since it is outside the area.

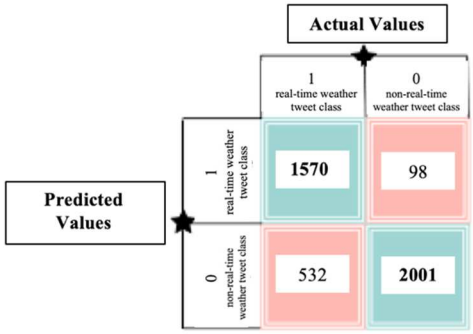

Fig. 6. Confusion Matrix for Weather Condition Classification.

B. Discussion

The increase in anxiety vents experienced by individuals on social media platforms in expressing heart situations greatly affects the results of crawling the resulting tweets. The habit of excessive online activities, dependence on spreading information through social media, the involvement of hateful content, and social media that presents news about glaring differences of opinion can accelerate the emergence of this spiral of anxiety, which is a phenomenon that appears in the era of cyber society and new media at the moment.

It is significant that the SVM model yields low results with the effect of the outpouring factor containing several weather keywords. The following example sentence, "*The mood is gray as cloudy as the sky this afternoon*" contains the word "*cloudy*" as one of the words representing "*weather*" and the word "*this afternoon*" that represents the word "*present*". Poetry-type sentences can also create ambiguity in the SVM model in classifying. The definition of poetry is a branch of literature that uses words as a medium of delivery to produce illusions and imaginations, like a painting that uses lines and colors to describe the artist's ideas.

Approaching language diversity, Indonesia has a great variety of languages, ranging from Javanese, Sundanese, Malay, Batak, Madurese, Betawi and many other regional languages. Indonesia is a country located on the continent of Southeast Asia. Indonesia is a country that has the largest island in the world, and the largest ethnic group and nation on earth, consisting of approximately 1120 ethnic groups spread out over 34 provinces. This is what makes Indonesia a country that has a distinctive ethnic and cultural diversity in the world, including more than 652 regional languages that have been identified and validated. This causes some collections of tweets to give ambiguity in the meaning of words using regional languages.

Diverse names of areas could be incorporated into the form of the name of food in Indonesia. The charging geolocation by name and city district in Indonesia depends on the presence of the word in a sentence. For example, in writing the sentence, *"I was just about to go out, I bought rice Padang, even though it was raining,"* the geolocation provided by the system will point to an area called "*Padang City*" with a latitude of -0.95 and a longitude of 100.3530556.

V. Conclusion

This research develops Indonesian tweets crawling automatically based on implementation of text classification using SVM model. The experiment result shows that SVM achieves 85% performance on Indonesian text. The preprocessed data will be utilized as supplementary data in weather information portal (bmkg.go.id) and Automatic Rain Gauge (ARG). The increasing of the datasets and exploring more on Indonesia language structure will become our future works. Using substituted of stemming algorithm to specify the words and distinguish the label class will make the prediction more accurate.

References

[1] M. J. Magro, "A Review of Social Media Use in E-Government," *Adm. Sci.*, vol. 2, no. 2, pp. 148–161, 2012.

[2] R. N. Bolton *et al.*, "Understanding Generation Y and their use of social media: A review and research agenda," *J. Serv. Manag.*, vol. 24, no. 3, pp. 245–267, 2013.

[3] Z. Tufekci and C. Wilson, "Social Media and the Decision to Participate in Political Protest: Observations From Tahrir Square," *J. Commun.*, vol. 62, no. 2, pp. 363–379, 2012.

[4] B. Sago, "The Influence of Social Media Message Sources on Millennial Generation Consumers," *Int. J. Integr. Mark. Commun.*, vol. 2, no. 2, pp. 7–18, 2010.

[5] T. Kuo, I. Y. Lu, C. H. Huang, and G. C. Wu, "Measuring users' perceived portal service quality: An empirical study," *Total Qual. Manag. Bus. Excell.*, vol. 16, no. 3, pp. 309–320, 2005.

[6] K. Purwandari, J. W. C. Sigalingging, M. Fhadli, S. N. Arizky, and B. Pardamean, "Data mining for predicting customer satisfaction using clustering techniques," in *Proceedings of 2020 International Conference on Information Management and Technology, ICIMTech 2020*, 2020.

[7] Statista, "Leading countries based on number of Twitter users as of July 2021," 2021. [Online]. Available: https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/. [Accessed: 08-Sep-2021].

[8] D. Sui and S. L. Shaw, *Human dynamics in smart and connected communities*, vol. 72, no. February. 2018.

[9] R. Rahutomo, A. Budiarto, K. Purwandari, A. S. Perbangsa, T. W. Cenggoro, and B. Pardamean, "Ten-year compilation of #savekpk twitter dataset," in *Proceedings of 2020 International Conference on Information Management and Technology, ICIMTech 2020*, 2020.

[10] T. H. Rashidi, A. Abbasi, M. Maghrebi, S. Hasan, and T. S. Waller, "Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges," *Transp. Res. Part C Emerg. Technol.*, vol. 75, no. February, pp. 197–211, 2017.

[11] A. Stefanidis, A. Crooks, and J. Radzikowski, "Harvesting ambient geospatial information from social media feeds," *GeoJournal*, vol. 78, no. 2, pp. 319–338, 2013.

[12] P. Panagiotopoulos, A. Z. Bigdeli, and S. Sams, "Citizen-government collaboration on social media: The case of Twitter in the 2011 riots in England.," *Gov. Inf. Q.*, vol. 31, no. 3, pp. 349–357, 2014.

[13] H. W. Park, S. Park, and M. Chong, "Conversations and medical news frames on twitter: Infodemiological study on COVID-19 in South Korea," *J. Med. Internet Res.*, vol. 22, no. 5, 2020.

[14] B. Han, P. Cook, and T. Baldwin, "Text-based twitter user geolocation prediction," *J. Artif. Intell. Res.*, vol. 49, pp. 451–500, 2014.

[15] I. Nurlaila, R. Rahutomo, K. Purwandari, and B. Pardamean, "Provoking tweets by indonesia media twitter in the initial month of coronavirus disease hit," *Proc. 2020 Int. Conf. Inf. Manag. Technol. ICIMTech 2020*, no. August, pp. 409–414, 2020.

[16] S. Stieglitz and L. Dang-Xuan, "Political communication and influence through microblogging - An empirical analysis of sentiment in Twitter messages and retweet behavior," *Proc. Annu. Hawaii Int. Conf. Syst. Sci.*, no. August, pp. 3500–3509, 2012.

[17] A. Madani, O. Boussaid, and Djamel Eddine Zegour, "Real-time trending topics detection and description from Twitter content," *Soc. Netw. Anal. Min.*, vol. 5, no. 59, 2015.

[18] I. Dongo *et al.*, "A qualitative and quantitative comparison between Web scraping and API methods for Twitter credibility analysis," *Int. J. Web Inf. Syst.*, no. August, 2021.

[19] A. Rahimi, T. Cohn, and T. Baldwin, "Twitter user geolocation using a unified text and network prediction model," *ACL-IJCNLP 2015 - 53rd Annu. Meet. Assoc. Comput. Linguist. 7th Int. Jt. Conf. Nat. Lang. Process. Asian Fed. Nat. Lang. Process. Proc. Conf.*, vol. 2, no. 2014, pp. 630–636, 2015.

[20] M. Pobiruchin, R. Zowalla, and M. Wiesner, "Temporal and Location Variations, and Link Categories for the Dissemination of COVID-19–Related Information on Twitter During the SARS-CoV-2 Outbreak in Europe: Infoveillance Study," *J. Med. Internet Res.*, vol. 22, no. 8, 2020.

[21] C. Byun, Y. Kim, H. Lee, and K. K. Kim, "Automated Twitter data collecting tool and case study with rule-based analysis," *Proc. 14th Int. Conf. Inf. Integr. Web-based Appl. Serv.*, pp. 196–204, 2012.

[22] Z. Shahbazi and Y. C. Byun, "Toward social media content recommendation integrated with data science and machine learning approach for e-learners," *Symmetry (Basel).*, vol. 12, no. 11, pp. 1–22, 2020.

[23] Kevin Makice, *Twitter API Up and Running*, First Edit. United States of America, 2009.

[24] T. W. Cenggoro, F. Tanzil, A. H. Aslamiah, E. K. Karuppiah, and B. Pardamean, "Crowdsourcing annotation system of object counting dataset for deep learning algorithm," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 195, no. 1, 2018.

[25] H. Prabowo, T. W. Cenggoro, A. Budiarto, A. S. Perbangsa, H. H. Muljo, and B. Pardamean, "Utilizing mobile-based deep learning model for managing video in knowledge management system," *Int. J. Interact. Mob. Technol.*, vol. 12, no. 6, pp. 62–73, 2018.

[26] T. W. Cenggoro, A. Budiarto, R. Rahutomo, and B. Pardamean, "Information System Design for Deep Learning Based Plant Counting Automation," *1st 2018 Indones. Assoc. Pattern Recognit. Int. Conf. Ina. 2018 - Proc.*, no. September, pp. 329–332, 2019.

[27] E. O. Abiodun, A. Alabdulatif, O. I. Abiodun, M. Alawida, A. Alabdulatif, and R. S. Alkhawaldeh, "A systematic review of emerging feature selection optimization methods for optimal text classification: the present state and prospective opportunities," *Neural Comput. Appl.*, vol. 8, 2021.

[28] R. Rahutomo, F. Lubis, H. H. Muljo, and B. Pardamean, "Preprocessing Methods and Tools in Modelling Japanese for Text Classification," *Proc. 2019 Int. Conf. Inf. Manag. Technol. ICIMTech 2019*, no. August, pp. 472–476, 2019.

[29] S. Rapacz, P. Chołda, and M. Natkaniec, "A Method for Fast Selection of Machine-Learning Classifiers for Spam Filtering," *Electronics*, vol. 10, no. 2083, pp. 1–23, 2021.

[30] K. Purwandari, J. W. C. Sigalingging, T. W. Cenggoro, and B. Pardamean, "Multi-class Weather Forecasting from Twitter Using Machine Learning Aprroaches," *Procedia Comput. Sci.*, vol. 179, no. 2019, pp. 47–54, 2021.

[31] A. Notanto, R. Rahutomo, and B. Pardamean, "Finetunning IndoBERT to Understand Indonesian Stock Trader Slang Language," *2021 Int. Conf. Comput. Sci. Artif. Intell. Oct. 2021.*, no. August, 2021.

[32] O. Omolola, R. Roberts, M. I. Ashiq, T. Chung, D. Levin, and A. Mislove, "Measurement and Analysis of Automated Certificate Reissuance," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12671 LNCS, pp. 161–174, 2021.

[33] M. A. Rosid, A. S. Fitrani, I. R. I. Astutik, N. I. Mulloh, and H. A. Gozali, "Improving Text Preprocessing for Student Complaint Document Classification Using Sastrawi," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 874, no. 1, 2020.

[34] S. Fatima and D. B. Srinivasu, "Text Document categorization using support vector machine," *(IRJET), Int. Res. J. Eng. Technol.*, vol. 4, no. 2, pp. 141–147, 2017.

[35] M. Razzaghnoori, H. Sajedi, and I. K. Jazani, "Question classification in Persian using word vectors and frequencies," *Cogn. Syst. Res.*, vol. 47, pp. 16–27, 2018.