RESEARCH PAPER

CROWDSOURCING DATA TO VISUALIZE POTENTIAL HOTSPOTS FOR COVID-19 ACTIVE CASES IN INDONESIA

Noorhadi Rahardjo[1*], Djarot Heru Santosa[2], Hero Marhaento[3]
[1]Faculty of Geography, Universitas Gadjah Mada, Yogyakarta 55281, Indonesia
[2]Faculty of Cultural Sciences, Universitas Gadjah Mada, Yogyakarta 55281, Indonesia
[3]Faculty of Forestry, Universitas Gadjah Mada, Yogyakarta 55281, Indonesia
***Corresponding author:** noorhadi@ugm.ac.id
Received: February 1th, 2021 / Accepted: May 25th, 2021 / Published: July 1st, 2021
https://DOI-10.24057/2071-9388-2021-011

ABSTRACT. As the COVID-19 outbreak spread worldwide, multidisciplinary researches on COVID-19 are vastly developed, not merely focusing on the medical sciences like epidemiology and virology. One of the studies that have developed is to understand the spread of the disease. This study aims to assess the contribution of crowdsourcing-based data from social media in understanding locations and the distribution patterns of COVID-19 in Indonesia. In this study, Twitter was used as the main source to retrieve location-based active cases of COVID-19 in Indonesia. We used Netlytic (www.netlytic.org) and Phyton's script namely GetOldTweets3 to retrieve the relevant online content about COVID-19 cases including audiences' information such as username, time of publication, and locations from January 2020 to August 2020 when COVID-19 active cases significantly increased in Indonesia. Subsequently, the accuracy of resulted data and visualization maps was assessed by comparing the results with the official data from the Ministry of Health of Indonesia. The results show that the number of active cases and locations are only promising during the early period of the disease spread on March – April 2020, while in the subsequent periods from April to August 2020, the error was continuously exaggerated. Although the accuracy of crowdsourcing data remains a challenge, we argue that crowdsourcing platforms can be a potential data source for an early assessment of the disease spread especially for countries lacking the capital and technical knowledge to build a systematic data structure to monitor the disease spread.

KEYWORDS: covid-19, crowdsourcing data, map visualization, netlytic, phyton, Indonesia

CITATION: Noorhadi Rahardjo, Djarot Heru Santosa, Hero Marhaento (2021). Crowdsourcing Data To Visualize Potential Hotspots For Covid-19 Active Cases In Indonesia. Geography, Environment, Sustainability, Vol.14, No 4, p. 125-130
https://DOI-10.24057/2071-9388-2021-011

ACKNOWLEDGMENTS: The authors acknowledge the Institute for Research and Community Services (LPPM) Universitas Gadjah Mada, Indonesia for providing the research grant entitled «Pemanfaatan Hasil Penelitian Dan Penerapan Teknologi Tepat Guna». We greatly appreciate Abghy Aunurrahim and Damar for their help during conducting this research.

Conflict of interests: The authors reported no potential conflict of interest.

INTRODUCTION

A coronavirus disease-19 (COVID-19) caused by a novel coronavirus has been considered the most crucial global health calamity of the century. It was started appeared in China by the end of 2019, and the first reported death from COVID-19 also in China in January 2020 (WHO 2020). The first case outside of China was found in Thailand on 13 January 2020 (Hui 2020). Ever since, the COVID-19 has become a global concern because of its high transmission rate that is from person-to-person via airborne respiratory droplets, direct contact with body fluids or secretions, or through contaminated objects (Xu et al. 2020). As results, on March 11th 2020 the World Health Organization (WHO) declared the COVID-19 outbreak as a global pandemic since it has affected all aspects of human life and has challenged health care systems worldwide (Arora et al. 2020).

With all global attentions are currently on the COVID-19, governments and the scientific community including health professionals are challenged in response to this pandemic e.g., to develop vaccines as well as curative medicines (Iyer et al. 2020). Furthermore, multidisciplinary researches on COVID-19 are then vastly developed not only focusing on the medical sciences like epidemiology and virology but also to the social behaviour issues since COVID-19 has affected much more far-reaching rather than just medical issues, like affecting social and economic of nations (Zhang and Shaw 2020).

One of the particular studies that have been demanded in COVID-19 measures are to understanding the spread of disease (Ponjavic et al. 2020). By understanding the transmission speed which the disease has spread throughout the world and visualizing the data with a clear presentation of the geographical area and time interval, it may help to clarify the extent and impact of the pandemic (Franch-Pardo et al. 2020). However, to develop such a system that integrates data structure analysis and modelling to geocoding and mapping of active cases and visualizing infection spread over time may require a large investment for hardware and software as well as man-hours that is a challenge for low- and middle-income countries like Indonesia.

In order to tackle this limitation, crowdsourcing may offer huge potential to contribute to the modelling and visualizing the spread of coronavirus. Crowdsourcing

basically is a process of outsourcing a business task or activity to a network of individuals (Paniagua & Korzynski 2017). Heipke (2010) defined crowdsourcing is an effort to recruit human workers to perform tasks that are inherently hard for computers to perform, for instances sentiment analysis of text, image or video classification or tagging, and matching data records that belong to the same entity. In the geospatial study, Heike (2010) described crowdsourcing is data acquisition by large and diverse groups of people, who in many cases are not trained surveyors and who do not have special computer knowledge, using web technology. During this COVID-19 pandemic, crowdsourcing and social media have played an unprecedented role which can help understand disease dynamics in space and time when testing is limited (Al-Omoush et al. 2020). This is because social media has been the preferred platforms to communicate, collaborate, and convey a sense of unity during the times of crisis e.g., during pandemic COVID-19 (Gui et al. 2017; Abdulhamid et al. 2020).

Social media platforms such as Facebook, Instagram, and Twitter have provided direct access to account users' preferences and attitudes, algorithms mediate, and facilitate content promotion (Kulshrestha et al. 2017). This has attracted the attention of researchers from various fields including cartography to analysing and synthesizing social media data. However, like no other platforms, Twitter provides a feature called «geo-location» that is open-access information regarding the user's location when uploading information. According to Twitter (developer.twitter.com), there are three metadata sources for geo-referencing tweets that can be used for map visualization: 1) tweet location: tweets that are geo-tagged with an exact location (i.e. a single landmark with longitude and latitude coordinates) or twitter place (i.e. an area with four pairs of longitude and latitude coordinates that define a bounding box), 2) mentioned location: parsing the Tweet message for geospatial location, and 3) profile location: parsing the account-level location for locations of interest. These facilities have helped to carry out a spatial analysis as well as map visualization.

This study aims to review the implementation of crowdsourcing-based data from social media in understanding locations and the distribution patterns of COVID-19 in Indonesia. In this study, we used Twitter as crowd-sourced data to retrieve location-based active cases of COVID-19 in Indonesia. According to the latest report of statistica.com, Twitter users in Indonesia reach 13.2 Million is the seventh-largest Twitter user in the world. From these users, approximately around 80% are active users producing 5 billion tweets a year. By these large number of Twitter users in Indonesia, our hypothesis is that crowdsourcing-based data from Twitter may provide an early assessment of the disease spread in Indonesia.

MATERIALS AND METHODS

Data source

In this study, tweets contain information related to COVID-19 active cases in Indonesia were used and analysed. We used Netlytic (www.netlytic.org) to retrieve the relevant online content about COVID-19 cases including audiences' information such as username, time of publication, and locations. Netlytic is a community-supported text and social networks analyser that can capture publicly available posts from social media sites, discover popular topics, find and explore emerging themes of discussions, analyse online communication networks using social network analysis, and map geo-coded social media data (www.netlytic.org). However, since the accessible data from netlytic.org is only limited to the maximum data acquisition up to the past 7 days, we used another method by using Phyton's script namely GetOldTweets3 and pandas packages to retrieve older tweet data. We determined the range time period from January 2020 to August 2020 when COVID-19 active cases significantly increased.

In order to find related information needed, we searched queries including several keywords (some words are in Bahasa Indonesia, the national language) such as: covid, corona, pasien covid, odp, pdp, otg, virus, virus covid, physical distancing, social distancing, positif corona, psbb, new normal, pandemi, karantina, quarantine, stay at home, bantuan covid, and vaksin covid which were all posted in Indonesia. It should be noted that all data collected were publicly available and obtained legally.

Data Analysis

All tweets retrieved by data-crawling using Netlytic and Phython script was then filtered according to the required criteria such as containing chosen keywords with information of active cases of COVID-19, geo-referenced tweets, and tweeted between January 2020 and August 2020. Subsequently, selected data is transformed into a shapefile format that can be visualized in Geographic Information System (GIS) environment. In this study, we used Quantum GIS (QGIS) to carry out map editing and map visualization of the geo-referenced tweets. QGIS is a cross-platform desktop (open source) software on geographic information systems (GIS) that has been widely used worldwide to analyse spatial data (Jaya & Fajar 2019; Ahmad & Kim 2020).

The resulted data and maps were then compared with the official data from the Ministry of Health of Indonesia in order to validate the results. We used a commonly used t-test to determine whether the results from crowdsourcing data are equal with the official data. The null hypothesis is that the two means are equal, and the alternative is that they are not. In addition, we also visually observed the resulted scatter-plot graph by comparing it relative with the x = y line as well as the resulted map visualizations.

RESULTS

Data and Map visualizations

By using Netlytic and Phyton, we were able to retrieve the suspected COVID-19 active cases within thirty-four provinces in Indonesia. However, both methods worked in different time periods, which Netlytic only covered in the period from July to August 2020, while Phyton was able to cover old tweet data from January 2020 to August 2020. In total, from the Netlytic in the period July – August 2020, we discovered 89 active cases in Indonesia spread over nine provinces (see Figure 1), where Jawa Barat (West Java) province has the largest cases with 35 active cases, followed by Jawa Tengah (Central Java) and Jawa Timur (East Java) provinces. However, it should be noted that these were based on tweeted spots which in each spot often contained more than one active cases.

Different from Netlytic, by using Phyton script to crawl the old tweets, we were able to retrieve monthly information from January 2020 to August 2020. The results show that the top six provinces having the most tweets contain information about the COVID-19 active cases were all on the Java Island. DKI Jakarta, Indonesia's capitol, has consistently been the

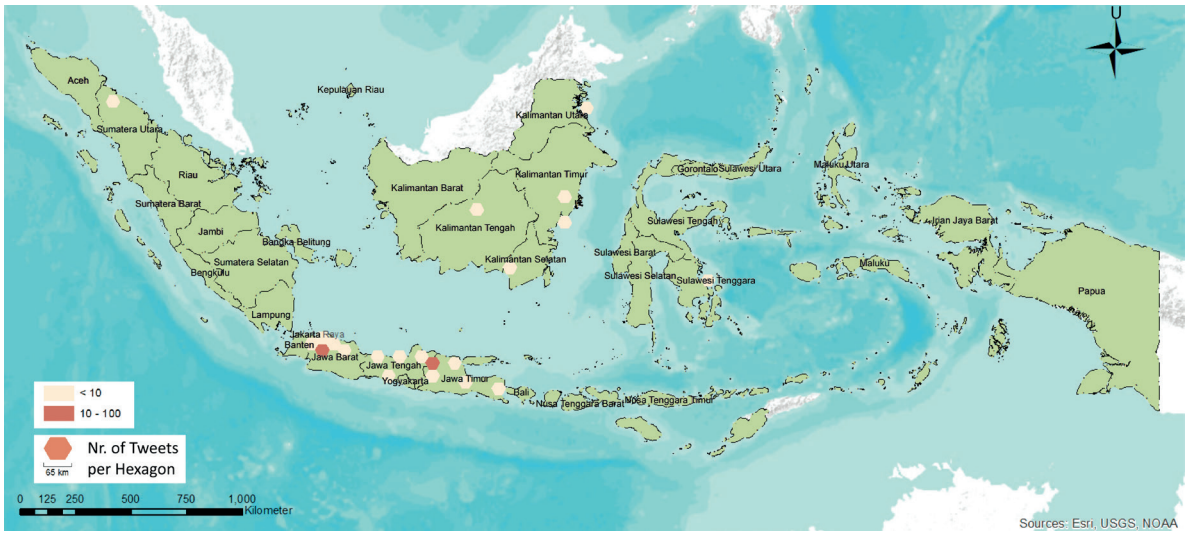

Fig. 1. Map visualization of the tweets contains information on COVID-19 active cases in Indonesia from July to August 2020 based on Netlytic

largest COVID-19 active cases in Indonesia which reach the peak reported cases on March 2020, and slowly decreased in the subsequent months. The following provinces having the largest informed COVID-19 active cases after DKI Jakarta from January 2020 to August 2020 were Banten, Jawa Barat, D.I. Yogyakarta, Jawa Timur, and Jawa Tengah (see Figure 2). Figure 3 shows the map visualization from the tweets contains information on COVID-19 active cases in Indonesia from January to August 2020 where most of the tweeted active cases were located in Java Island. Some provinces outside Java Island i.e. Sumatra Barat (West Sumatra), Papua, Kalimantan Timur (East Kalimantan), and Riau were also reported having quite high number of COVID-19 active cases but only by less than 200 tweets of cases.

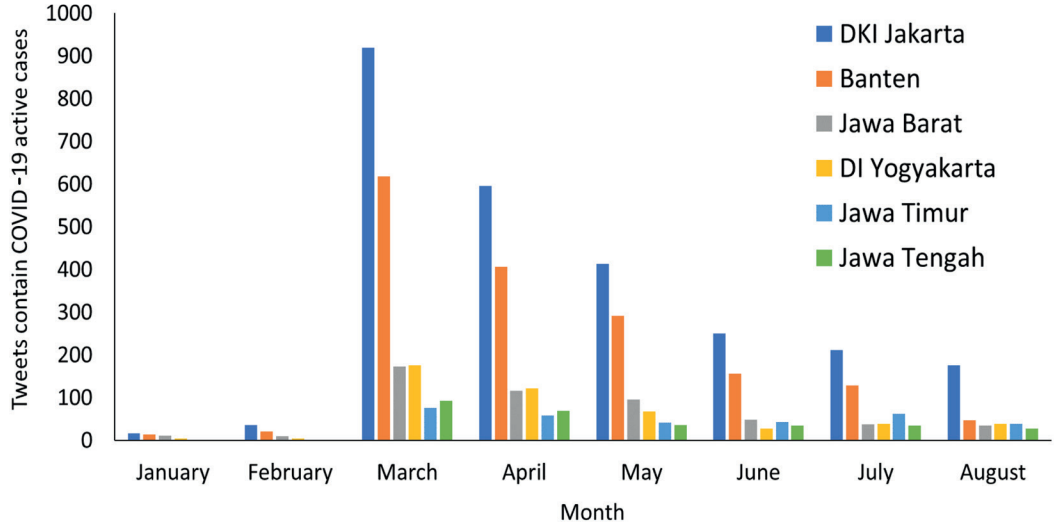

Fig. 2. Number of tweets contain information of COVID-19 active cases in the top six provinces in Indonesia

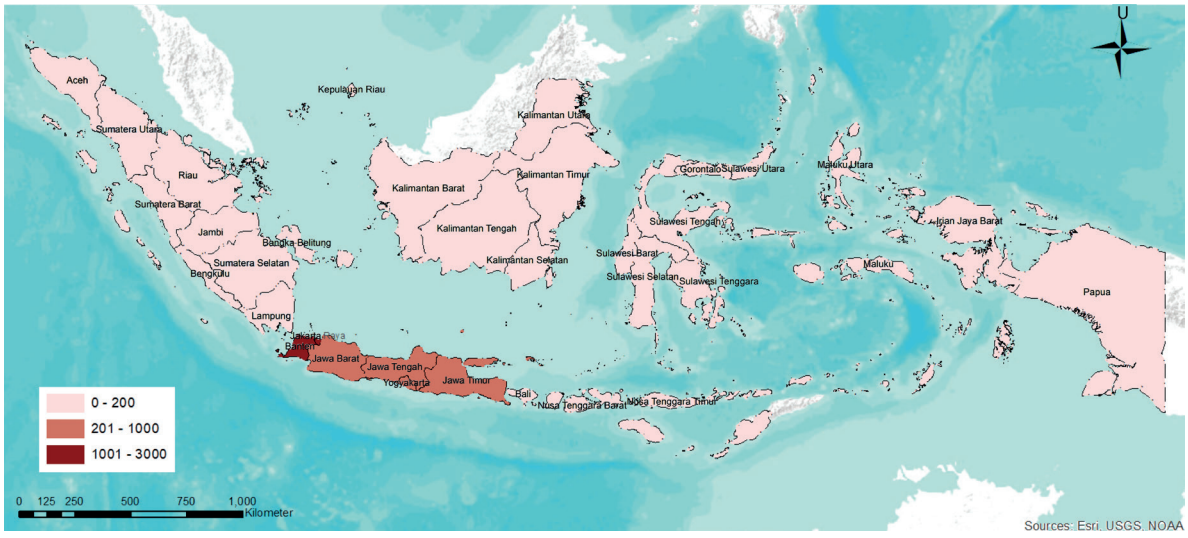

Fig. 3. Map visualization of the tweets contains information on COVID-19 active cases Indonesia from January to August 2020 based on data crawling using Phyton

Accuracy of results

We carried out the accuracy analysis by performing t-test statistic to the two sets data from crowdsource-based data (i.e. Twitter) and official data from the Ministry of Health of Indonesia. For this study, the comparisons were applied for three different periods: March – April, May – June, and July – August. The results show that in the period of March – April, the null hypothesis was accepted with the p-value was 0.736295, greater than 0.05, the applied statistic significant level. While for the subsequent period, May – June and July – August, the alternative hypothesis was accepted with the p-values for both periods were 0.004341 ($p < 0.01$) and 0.000982 ($p < 0.001$). These statistical test results show that the crowdsourcing data was relatively accurate to predict the COVID-19 active cases only for the period of March – April, while the accuracy was getting worst for the subsequent periods. By visual inspections of the scatter-plot graphs, it was observable that during the period March – April, the resulted scatter-plot was relatively closer to the x = y line (see Figure 4a). However, during the May – June period, the resulted scatter-plot was below the x = y line indicating an under-estimation from the crowdsourcing data compared to the official data (see Figure 4b). This bias due to under-estimation was exaggerated in the period of July – August (see Figure 4c).

Similar results were shown by comparing the visualization maps of COVID-19 active cases between the crowdsourcing-based data and the official data as seen in Figure 5. It was observable that in the period of March – April 2020, crowdsourcing data has comparable results with the official data, where the Java Island was the epicentre of the spread disease. However, in the subsequent periods, the crowdsourcing data were not able to match the official data due to under-estimation results. As seen in Figure 5c and 5d, while the results of crowdsourcing data visually pointed Java Island as the most findings COVID-19 active cases, the government data showed that the active cases of COVID-19 have been spread in all over Indonesia (i.e., 27 provinces out of 34 provinces) with a range of 101 – 10,000 cases/province. In the subsequent period, the differences become larger as seen in Figure 5e and 5f, which the crowdsourcing data only resulted in 7 provinces in Indonesia that have 101 – 10,000 cases/province. This is far below the government data which found 33 provinces out of 34 provinces have COVID-19 active cases more than 100 cases/province.

DISCUSSION

Research on social media and its unique communities have now been often studied (Marwick & Boyd 2011; Gaffney & Puschmann 2013). Carley et al. (2015) argue that following patterns on social media e.g. Twitter, Facebook, Instagram can help in making accurate predictions about future trends. Through social media, public involvement in the scientific processes is now openly available; not just in the data collection process, but also in the planning and data visualization (Lamoureux & Fast 2019). However, it should be noted that information spread through social media has been often inaccurate (Thomson et al. 2012), outdated, and contain irrelevant information (Acar & Muraki 2011). For this reason, it is necessary to explore to what extent the crowdsourcing data from social media can be used to provide reliable information.

In this research, we used Twitter as the main source of information to visualize potential hotspots for COVID-19 active cases in Indonesia. We found a promising result of crowdsourcing data visualisation only during the early period of COVID-19 transmission, when it was going viral on the social media of Twitter. In the subsequent periods, a

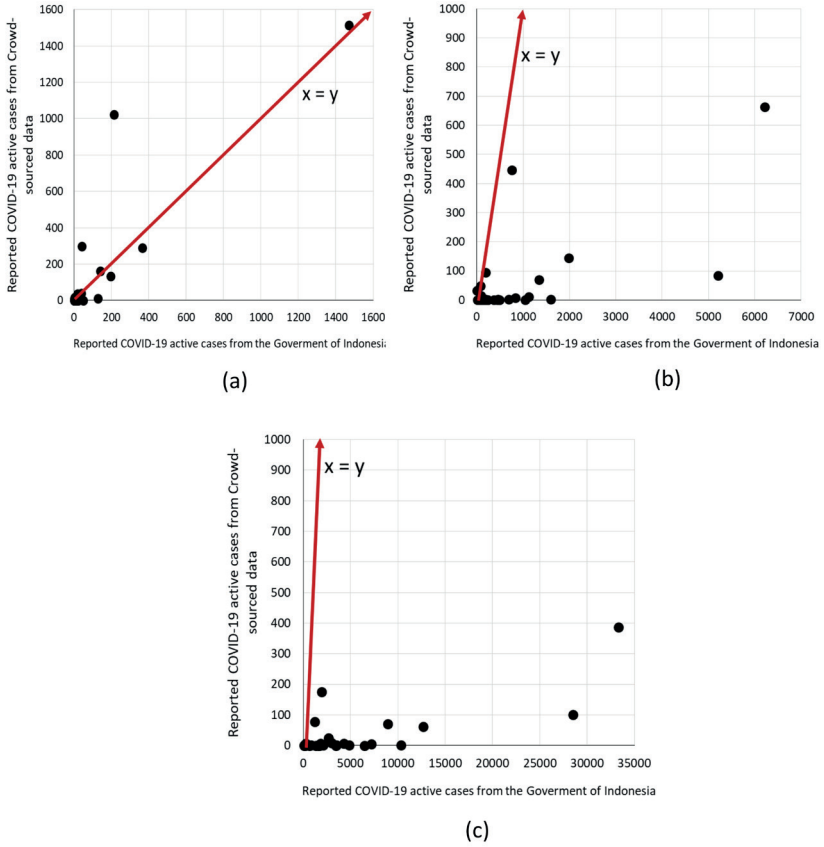

Fig. 4. Scatter-plot graphs between reported COVID-19 active cases from crowdsourcing data and government data relative to the x = y line for the period March – April 2020 (a), May – June 2020 (b), and July – August (c)

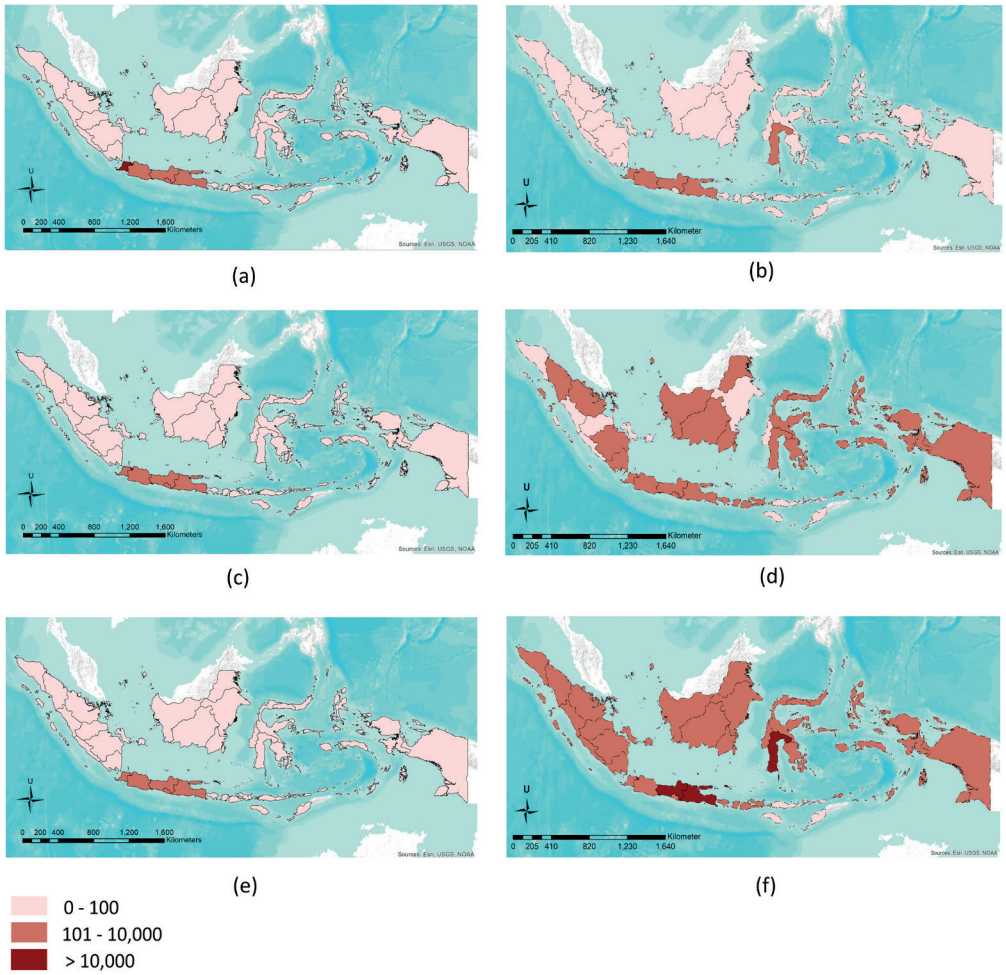

Fig. 5. Maps of COVID-19 active cases in Indonesia from March 2020 to April 2020 according to Crowdsourcing (a) and Government data (b), from May 2020 to June 2020 according to Crowdsourcing (c) and Government data (d), and from July 2020 to August 2020 according to Crowdsourcing (e) and Government data (f)

significant decrease in the accuracy was observable when it was compared to the government data. This phenomenon apparently can be explained by a social media behaviour which during a meaningful event like COVID-19 initial spread, social excitement has influenced social media user in content creation and sharing (Wakefield & Wakefield 2016). However, once the meaningful event was over, in this study the COVID-19 initial cases, the only social excitement was not sufficient to motivate content creation and sharing activities in social media resulting inaccuracies of crowdsourcing data.

Despite revealing the challenge on its accuracy, the crowdsourcing platform used and discussed in this research can be a potential source for those lacking the capital and technical knowledge to build a systematic data structure for data collection, management, and visualization platforms of the COVID-19 spread. Our findings are similar to the results of Larson (2018) and Chakraborty et al. (2020), among others that during the pandemic, crowdsourcing data can support monitoring of social distancing, contact tracing, as well as the disease spread. However, as found in our research that the accuracy of crowdsourcing data has remained questionable. This is similar to the finding of Moturu & Liu (2010) who argue that only a fair portion of social media information is useful and has proven to be a great source of knowledge, which most of the information shared should be taken carefully. One of the reasons is that much of the information shared through social media has been contributed by strangers with little or no apparent reputation to share information. For this reason, Chung et al. (2012) emphasized the importance of the source of information that transmits the news. Our results showed that crowdsourcing data make citizen science-based project more attractive and accessible to everyone. Indeed, the accuracy and information credibility remain the main issues of working with the crowdsourcing data requiring more study focusing on the information credibility and crowdsourcing data verifications.

CONCLUSIONS

In this study, we were able to visualize the distribution patterns of COVID-19 active cases in Indonesia by using crowdsourcing-based data from social media Twitter. However, based on the accuracy-test using independent t-test and visual inspection to the resulted scatter plots against the official data, it was found that the prediction is only promising during the early period of the disease spread on March – April, 2020, where most of people (i.e. netizen) tweeted about COVID-19 active cases. In the subsequent periods from April to August 2020, the prediction error was exaggerated from time to time. Although it has challenges on the data accuracy, we argue that crowdsourcing platform can be a potential data source for an early assessment of the disease spread especially for those (e.g. countries) lacking the capital and technical knowledge to build a systematic data structure.

REFERENCES

Abdulhamid N.G., Ayoung D.A., Kashefi A., & Sigweni B. (2020). A survey of social media use in emergency situations: A literature review. Information Development, DOI: 10.1177/0266666920913894.

Acar A., & Muraki Y. (2011). Twitter for crisis communication: lessons learned from Japan's tsunami disaster. International Journal of Web Based Communities, 7(3), 392-402.

Ahmad S., & Kim D.H. (2020). Quantum GIS based descriptive and predictive data analysis for effective planning of waste management. IEEE Access, 8, 46193-46205.

Al-Omoush K.S., Orero-Blat M., & Ribeiro-Soriano D. (2020). The role of sense of community in harnessing the wisdom of crowds and creating collaborative knowledge during the COVID-19 pandemic. Journal of Business Research.

Arora G., Kroumpouzos G., Kassir M., Jafferany M., Lotti T., Sadoughifar R., Sitkowska Z., Grabbe S. and Goldust M. (2020). Solidarity and transparency against the COVID–19 pandemic. Dermatologic Therapy.

Carley K.M., Malik M.M., Kowalchuck M., Pfeffer J., & Landwehr P. (2015). Twitter usage in Indonesia. Computational Analysis of Social and Organizational Systems (CASOS) Technical Report CMU-ISR-15-109. Available at SSRN 2720332.

Chakraborty K., Bhatia S., Bhattacharyya S., Platos J., Bag R., & Hassanien A.E. (2020). Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers—A study to show how popularity is affecting accuracy in social media. Applied Soft Computing, 97, 106754.

Chung C.J., Nam Y., & Stefanone M.A. (2012). Exploring online news credibility: The relative influence of traditional and technological factors. Journal of Computer–Mediated Communication, 17(2), 171-186.

Franch-Pardo I., Napoletano B.M., Rosete-Verges F., & Billa L. (2020). Spatial analysis and GIS in the study of COVID-19. A review. Science of The Total Environment, 140033.

Gaffney D., & Puschmann C. (2013). Data collection on Twitter. In K. Weller, A. Bruns, J. Burgess, M. Mahrt, & C. Puschmann (Eds.), Twitter and Society, New York: Peter Lang, 55-68.

Gui X., Kou Y., Pine K.H., & Chen Y. (2017, May). Managing uncertainty: using social media for risk assessment during a public health crisis. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, 4520-4533.

Heipke C. (2010). Crowdsourcing geospatial data. ISPRS Journal of Photogrammetry and Remote Sensing, 65(6), 550-557.

Hui D.S., Azhar E.I., Madani T.A., Ntoumi F., Kock R., Dar O., ... & Zumla A. (2020). The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health—The latest 2019 novel coronavirus outbreak in Wuhan, China. International Journal of Infectious Diseases, 91, 264-266.

Iyer M., Jayaramayya K., Subramaniam M.D., Lee S.B., Dayem A.A., Cho S.G., & Vellingiri B. (2020). COVID-19: an update on diagnostic and therapeutic approaches. BMB reports, 53(4), 191.

Jaya M.T.S., & Fajar A.N. (2019). Analysis of The Implementation Quantum GIS: Comparative Effect and User Performance. J. Theor. Appl. Inf. Technol, 97, 2596-2605.

Kulshrestha J., Eslami M., Messias J., Zafar M.B., Ghosh S., Gummadi K.P., & Karahalios K. (2017, February). Quantifying search bias: Investigating sources of bias for political searches in social media. In Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, 417-432.

Lamoureux Z. & Fast V. (2019). The tools of citizen science: An evaluation of map-based crowdsourcing platforms. Spatial Knowl. Inf. Canada, 7(4), 1.

Larson H.J. (2018). The biggest pandemic risk? Viral misinformation. Nature, 562(7726), 309-310.

Marwick A.E. & Boyd D. (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. New Media & Society, 13(1), 114-133, DOI: 10.1177/1461444810365313.

Moturu S.T., & Liu H. (2011). Quantifying the trustworthiness of social media content. Distributed and Parallel Databases, 29(3), 239-260.

Paniagua J. & Korzynski P. (2017). Social Media Crowdsourcing. E.G. Carayannis (ed.), Encyclopedia of Creativity, Invention, Innovation and Entrepreneurship. Springer Science+Business Media LLC, DOI: 10.1007/978-1-4614-6616-1_200009-1.

Ponjavic M., Karabegovic A., Ferhatbegovic E., Tahirovic E., Uzunovic S., Travar M., Pilav A., Mulić M., Karakaš S., Avdić N., Mulabdić Z., Pavić G., Bičo M., Vasilj I., Mamić D., Hukić, M. (2020). Spatio-temporal data visualization for monitoring of control measures in the prevention of the spread of COVID-19 in Bosnia and Herzegovina. Med Glas (Zenica), 17(2), 265-274.

Thomson R., Ito N., Suda H., Lin F., Liu Y., Hayasaka R., ... & Wang Z. (2012, April). Trusting Tweets: The fukushima disaster and information source credibility on Twitter. In 9th ISCRAM conference, 10.

Wakefield R., & Wakefield K. (2016). Social media network behavior: A study of user passion and affect. The Journal of Strategic Information Systems, 25(2), 140-156.

World Health Organization W. (2020). Coronavirus disease (COVID-19) pandemic. Retrieved August 12, 2020, from https://www.who.int/emergencies/diseases/novel-coronavirus-2019.

Xu C., Luo X., Yu C., & Cao S.J. (2020). The 2019-nCoV epidemic control strategies and future challenges of building healthy smart cities.

Zhang H., & Shaw R. (2020). Identifying research trends and gaps in the context of covid-19. International journal of environmental research and public health, 17(10), 3370.