Social media data crowdsourcing as a new stream for environmental planning & monitoring: A review

B L Lawu¹, F Lim², A Susilo³, and N Surantha⁴*

¹-⁴ Computer Science Department, BINUS Graduate Program-Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia, 11480

Email: nico.surantha@binus.ac.id

Abstract. This research focused on social media applications that had been used by large-scale users. Data might be in the form of text, image, video, each with its own data processing complexity. In this study, the researchers had performed a systematic review of literature on the development and usage of crowdsourcing data collection in recent days, with a range of techniques/methods/approaches to present meaningful data results/conclusions for environmental management. From the collected references, it is found that most papers discuss the monitoring and planning the environment through the use of social media data. An in-depth review of the literature was performed on the state-of-the-art environmental monitoring and planning and the general architecture of the data crowdsourcing system. Finally, based on the literature review, the challenge of data crowdsourcing for environmental monitoring and planning is identified and addressed.

Keywords: social media, data crowdsourcing, environmental planning

1. Introduction

Environmental problems related to energy, food security, global warming, and global population growth are some of the key issues that need to be resolved effectively to prevent global crises. Information technology is only one of the many technologies that can be used to mitigate or prevent environmental issues. Geospatial information systems (GIS) and Internet-of-things are two of the information technologies that are commonly used for environmental management. Interestingly, however, social media recently also appears to be one of the possible solutions for environmental management.

Social media is home to a vast amount of data, both valuable and invaluable. The ability to filter/sort a specific data set can lead to new information and situational awareness. For example, a certain flood situation user post in his/her home area can be used in real-time for canvas/map of affected flood areas. This data mining/analysis is called crowdsourcing, a compilation of data that uses mass data available on the Internet. Data can also be used to forecast certain incidents that may occur in the future, i.e., people's health history conditions may be used for people with similar health issues. Even night city light may be used to map people's density rates in certain areas.

This study focuses on social media applications that have been used by a large scale of users. Data may be in the form of text, image, and video, each with its own difficulty processing data. This study emphasizes the development and usage of crowdsourcing data collection in recent days, with a range technique/methods/approach to present meaningful data results/conclusions. The identification of social media is also part of this paper study. Research on related topic papers is conducted to identify further
aspects/patterns in crowdsourced data used, i.e., data type developed application, and data processing techniques, thus classified as purpose similarities. The research outcome is designed to model crowdsources from raw data to ready-to-use data.

This paper is structured as follows. The approach for the literature review is discussed in Section 2. Section 3 addresses an in-depth analysis of the literature review. It demonstrates the state-of-the-art of data crowdsourcing approach for environmental management. The challenge of data crowdsourcing for environmental management is addressed in Section 4. A conclusion with the findings of the literature review will be addressed at the end of the article.

2. Review methodology

The reviewed papers were mainly collected from online academic directories such as Google Scholar, IEEE Xplore, and ScienceDirect, with keyword crowdsourced social media, crowdsourced environmental management, crowdsourced environmental science, crowdsourced natural resource, crowdsourced ecosystem service, crowdsourced tourism, crowdsourced nature, and crowdsourced social network. The period of publication of the papers is from 2012 to recent. Literature reviews on Social Media Crowdsourcing have increased from year to year, as shown in Figure 1.

![Figure 1. Paper about Crowdsourcing on Social Media](image)

Most studies use data from social media, as shown in Figure 2. In China, studies specifically use Weibo Network (3%), and other common social media such as Facebook (3%), Foursquare (3%), Instagram (6%), Panoramio (9%), Flickr (26%), and Twitter (44%). The data provided was in the form of an image and a text. Unfortunately, popular social media such as Instagram and Facebook have certain limitations due to their policy on data access restriction. The general use for accessing and collecting social media data is through the application programming interface (API), i.e., The Power Track API is provided by Twitter. Some applications use the Client-Server Architecture to upload data directly, such as Waze and VGI (Volunteered Geographic Information).

From various Social Media and information resulting from multiple studies, the data collection purpose categories are generally classified into 2 (two) categories, as shown in Figure 3, for monitoring and planning. Therefore, in this study, we focus on reviewing the state-of-the-art of research in both categories.
3. In-depth review and analysis
In this chapter, an in-depth review and analysis are conducted. Firstly, we review the state-of-the-art of research in environmental monitoring and planning. Then, based on the literature review, the general architecture of the system and the issue within the system are discussed.

3.1. Monitoring
In this category (Table 1), the main focus was to monitor the current situation/condition and potential adjustments. Many of social media data are used for population distribution mapping [1], tourism and recreation [2][3][4][5], to identify the current condition of a certain area, environmental monitoring [6], biodiversity including the ecosystem support [7][8][9][10][11], and climate change [12]. Social media also used for social and environmental awareness [13][14][15], certain disaster distribution pattern using social media, i.e., flood [16][17], forest fire, environment pollution [18][19], hurricane, land deposits [20], land used [21], and heatwave [22].

| Paper | Summary | Media | Social |
|-------|---------|-------|--------|
| [13]  | Social Media as tools for building public awareness and communication on Environmental Issues. | Twitter |
| [2]   | Tourist Favorite Destination Identification using Uploaded Photo in Flickr Social Media | Flickr |
| [4]   | Tourist Distribution Analysis using GIS and Geo-tag Photo to predict most visited Tourist Destination, and Tourist Density in the location. | Panoramio |
| [7]   | Forest and The Ecosystem Monitoring using Social Media | Twitter |
| [1]   | Data Mining for Identifying Location Point of Interest using Geo-tag Photo | Flickr |
3.2. Planning

In this phase (Table 2), the focus of the study will be on planning preparation, either for disaster mitigation, to reduce damage impact, or to conduct better city infrastructure planning. The data provided in social media may provide leads/hints for a future decision, i.e., to cut an illegal rare orchid trading [24], disaster mitigation [25], and better city planning or environment [26][27][28]. Social media data can also be synergized with the decision making process [29].

| Paper | Summary | Media Social |
|-------|---------|--------------|
| 24    | Horticultural Orchids Trade Identification using Social Media | website |
| 25    | Disaster Mitigation using Social Media | twitter |
| 29    | Social Media as Decision Support System on Hazard Events Risk Perception | twitter |
| 26    | Cycling and Air Pollution Exposure Analysis using Social Media Strava | strava |
| 27    | Domestic Water Charges Analysis using Social Media | Twitter |
| 28    | Cultural Ecosystem Service and Landscape Features Mapping using Social Media Photos | Flickr, Panoramio |

**Table 2. Planning Papers**
3.3. Generic Modelling

Based on the literature that is mapped in sections 3.1 and 3.2, the generic model of the data crowdsourcing system is defined in this section. Generic modelling for Crowdsourcing Data illustrated in Figure 4. These are the core process of data crowdsourcing started from data collection until results visualization. The detail of the model is described as follow:

3.3.1. Data collection
Data is collected through social media API or directly from a specified mobile application such as Waze, Google Maps, and GIS. The example of the collected data is shown in Figure 5.

3.3.2. Processing
In this part, if data collection cannot directly lead to a presentable outcome, there is a need for a further processing method with the main objective of producing a more valid data result. The data validation process can be made simpler for a particular application that retrieves data from user input, i.e., the water level in flood application input columns, or accident location in Waze Apps.

In other cases, data is processed through natural language processing (NLP) [30][31] and text mining. Any methods/function/parameter calculations will be needed, such as clustering data [1], model training system [30][31], regular expression patterns based on data mining are combined to provide better analysis of the result. The machine learning method may be useful in some forecasting environment. However, it does not yet deliver on the promise for unvaried time series forecasting.

Figure 4. Generic Model of Data Crowdsourcing System

Figure 5. The Highlight Text Are Data to be Identified in Tweet [17]
Table 3. Example of Flooding Tweets Processed by Natural Language Processing [31]

| Tweet ID           | Posted time | Tweet                                                                 |
|-------------------|-------------|----------------------------------------------------------------------|
| 648973656161394688 | 2015-09-29 14:32 | Roanoke River over 12 feet at Walnut St bridge in Roanoke now expected to top 13. Flood stage is 10. #swvawx |
| 649935518038405120 | 2015-10-02 06:14 | Flood currently on 3 NE Wrightsboro. in New Haven, NC. 1-2 ft. of water on Tandem CT. #newx #flood #flooded #flooding #rain #HurricaneJoaquin |
| 650019989353852928 | 2015-10-02 11:50 | Chicod creek at flood stage in Pitt County, Likely to 11 ft by tonight. If it hits 12, it's running over the road. |
| 650393844115243008 | 2015-10-03 12:35 | Helped nearby drivers by reporting a flood on PR-1, Ponce on @username - Drive Social. |
| 650403125698891776 | 2015-10-03 13:12 | The Market Street portion of Water Street in Downtown Wilmington is under 9 inches of water |

As shown by Table 3, it uses named entity recognition (NER) recognition tool and model training provided by Stanford, [31] to identify the location of flood resulted as follows (Table 4).

Table 4. Results of Location Detection in ALTA 14 Dataset [31]

| Models                        | Precision | Recall | F-score |
|-------------------------------|-----------|--------|---------|
| Stanford NER                  | 94.51%    | 32.07% | 47.88%  |
| Stanford NER retrained        | 86.68%    | 69.72% | 77.20%  |

This study [31] also employs Convolutional Neural Networks (CNN) technique to automatically classify the crowdsourcing photos through Clarifai API [31]. The use of Geocoding for locating geographical coordinates in this data processing section offers a better result. The next final step is to integrate a data validation to assure data “cleanliness” quality [22].

3.3.3. Visualization result
Crowdsourced data can be presented in a range of information systems, such as Google map, Waze, DRM-based app, SIG-based app, to provide an interactive result for users.

4. Challenge discussion
Qualitative Literature Studies indicate that the mass data collection (Crowdsourcing) is an additional need and even become a new stream of data collection in this modern era. Many literature studies using this approach have shown that this new data collection medium is a more efficient way to collect data and affect the creation of other information system applications. The growing use of social media is becoming a key motivator for this technique of collecting data for the collection and processing of different forms of text and image/video information. Every other type of information accumulates in a large data unit called Big-Data. Defining Big Data in social media, a 3V approach is used: Volume, Variety, and Speed. The terms of ethics and data privacy will also be discussed.

4.1. Volume: big-data
Social media consists of huge amounts of data that can be analyzed for a variety of needs. Data is very easily obtained through various social media applications. These data can help the development of
studies in various ways, for example, studies of environmental planning and environmental monitoring. Social media data can help the government in terms of disaster mitigation and governance systems-related to social and culture. Through social media, communities can voluntarily participate in providing information about floods, landslide-prone areas, accident-prone areas, robber-prone areas, forest fires, illegal actions, and so on. Social media can also contribute to the economic-tourism sector. Even new tourist destinations can be exposed to economic growth in frequently visited areas.

4.2. Velocity: real-time data
Crowdsourcing on social media allows almost real-time availability of data. Data can be quickly obtained for analysis. This is necessary in some cases that require a rapid response. In disaster mitigation, the mapping of the latest flood conditions, landslides, forest fires, new epidemics, and air pollution can be retrieved almost immediately. Fast emergency response is also needed in the event of a crime, a traffic accident, etc. Traffic control can also be improved by the availability of traffic data, road delays, damaged roads, dangerous objects, and others due to their real-time advantages.

However, social media can provide incorrect information for an investigation that takes a longer period of time. An event may not occur continuously, i.e., land subsidence/displacement at a certain time, very high rainfall at present does not lead to direct flooding, areas, where the virus presents in some areas, are not necessarily a frequent tourist destination, and climate change requires data analysis over a longer period of time. In the analysis of data from social media such as Twitter, special handling of retweet features is often necessary, since the likelihood of retweet information may not represent the current situation, but information that has already occurred in the past.

4.3. Variety: informations diversity
Some misleading information can occur in various studies describing the information diversity data collected from social media. Meta-data information (geotags and timestamps) may not be relevant to the subject of the photo in the photo media. The type of social media application used may also affect the quality of the results of the analysis. In the study [4], Panoramio social media had a high percentage of photo relevance compared to Flickr. Photo diversity domination on Flickr is far more than Instagram; thus, photos with human objects appear more often on Instagram than Flickr. There are fewer Geotags errors in Panoramio compared to Flickr; this may be due to differences in photo qualification standards that can be uploaded. High-resolution photos can bias the analysis when compared to low-resolution photos. Information on photo titles and tags cannot always be assumed to be the location of the photo. The diversity of data can also be affected by variations in areas, cultures, the majority of populations, and languages.

Diversity of information is one of the weak advantages of using crowdsourcing as a data mining tool in social media because the information may be biased/irrelevant. Therefore, a further quality control, filtering, and data validation are required to deliver a better analysis result.

4.4. Ethics and data privacy
Several studies suggest the need for more discussion on privacy and ethics in collecting data through social media. Although the data collected are based on publicly available data, users do not willingly provide information, and they may be unaware of the purpose of the data used. This may be a problem in the areas of intellectual property law and online privacy law.

Crowdsourcing on social media may produce inappropriate data if taken from users unaware that their personal data may be used for analysis. This contradicts traditional data collection methods such as surveys, where respondents are fully aware of the purpose of the survey. Therefore respondents may provide more relevant and accurate information.

4.5. Data validation
Unfortunately, data validation is rarely a topic to be discussed, as in some cases, the crowdsourcing data image [31] showed a flooded street. The information was not correct, and vice versa, due to inaccuracy
and irrelevant data tweeting at a certain location. When the user is not aware that their data may be used for any analysis, there may be a lack of validity. Even if the user is aware of a potential analysis of the data, the user may intentionally provide incorrect details.

Moreover, piling up invalid information, crowdsourcing data mining processes may lead to the inefficient analysis result. This number decreased significantly for data image. Due to CNN technology for machine learning, detecting image is still far from accurate. Dataset training is required to increase the validity of the data image. It can also be combined with digital validation using physical sensors, manual validation by relevant users (government/organization), and accurate data itself.

5. Conclusions
The need and development of Big Data Analytics will grow stronger due to the increased of crowdsourcing data mining. The availability of API in social media will ease the data collection process for any required data, as shown by the significantly increased number of published research and developed applications related to this technology, i.e., mapping and distribution applications for population, disaster, and trends.

The crowdsourcing method can be used for real-time data collection; therefore, results can quickly obtain more than any other means of data collection. The development of current data analysis techniques, such as deep learning and natural language processing, greatly affects the analysis of data obtained using crowdsourcing methods. Data filtering can be done optimally with the training model, and results generated can be more qualified. Besides, crowdsourcing can synergize with other data collection techniques, such as physical sensors, the internet-of-things (IoT), and surveys. On the other hand, standardization of data crowdsourcing is required in relation to data recovery ethics. This is due to the potential infringement of privacy.

The advantage of data analysis using social media is size, with the availability of more data samples and combine with traditional data collection, such as surveys. This makes data collection highly effective and efficient, does not consume too many resources, relatively inexpensive, and the time spent is short, especially if it can carry out automatic data retrieval. However, the generated data would be more reliable in large populations and or in areas with well-known tourist attractions, rather than remote locations and or areas with small tourist attractions.

The identified consideration key points of crowdsourcing data using social media are data integration, the integration methods, data mining quality, the ethics aspect of data, and data availability assurance. Data integration and integration methods require a certain level of procedure development to fully integrate various data from multiple sources of social media and combine it with existing data. Data quality also needs to be considered due to its nature in dealing with a large scale of data size and data complexity. Meanwhile, the ethics of data retrieved is related to the ethic of obtaining sensitive information. Data Availability Assurance also needs to be considered because the data source is still related to the existence of Social Media itself.

Data Validation will still be one of the key challenges topics in crowdsourcing data mining; hence the future of computation technology will open a new chapter on driving a significant role in data validation to give better accuracy and speed.

Acknowledgement
This research publication is fully supported by Bina Nusantara University.

References
[1] Lee I, Cai G and Lee K 2014 Exploration of geo-tagged photos through data mining approaches Expert Systems with Applications 41 397–405 doi: 10.1016/j.eswa.2013.07.065
[2] Wood S A, Guerry A D, Silver J M and Lacayo M 2013 Using social media to quantify nature-based tourism and recreation Sci Rep 3 2976 doi: 10.1038/srep02976
[3] Heikinheimo V, Minin E D, Tenkanen H, Hausmann A, Erkkonen J and Toivonen T 2017 User-Generated Geographic Information for Visitor Monitoring in a National Park: A Comparison of Social Media Data
[4] Orsi F and Geneletti D 2013 Using geotagged photographs and GIS analysis to estimate visitor flows in natural areas *Journal of Nature Conservation* **21** 359–368 doi: 10.1016/j.jnc.2013.03.001

[5] García-Palomares J C, Gutiérrez J and Mínguez C 2015 Identification of tourist hot spots based on social networks: A comparative analysis of European metropolises using photo-sharing services and GIS *Applied Geography* **63** 408–417 doi: 10.1016/j.apgeog.2015.08.002

[6] Giuliani M, Castelletti A, Fedorov R and Fraternali P 2016 Using crowdsourced web content for informing water systems operations in snow-dominated catchments *Hydrol. Earth Syst. Sci.* **20** 5049–5062 doi: 10.5194/hess-20-5049-2016

[7] Daume S, Albert M and von Gadow K 2014 Forest monitoring and social media – Complementary data sources for ecosystem surveillance? *Forest Ecology and Management* **316** 9–20 doi: 10.1016/j.foreco.2013.09.004

[8] Daume S and Galaz V 2016 Anyone Know What Species This Is? Twitter Conversations as Embryonic Citizen Science Communities *PLoS ONE* **11** e0151387 doi: 10.1371/journal.pone.0151387

[9] Gliozzo G, Pettorelli N and Haklay M 2016 Using crowdsourced imagery to detect cultural ecosystem services: a case study in South Wales, UK *E&S* **21** ar6 doi: 10.5751/ES-08436-210306

[10] Atsumi K and Koizumi K 2017 Web image search revealed large-scale variations in breeding season and nuptial coloration in a mutually ornamented fish, Tribolodon hakonensis *Ecol Res* **32** 567–578 doi: 10.1007/s11284-017-1466-z

[11] ElQadi M M, Dorin A, Dyer A, Burd M, Bukovac Z and Shrestha M 2017 Mapping species distributions with social media geo-tagged images: Case studies of bees and flowering plants in Australia *Ecological Informatics* **39** 23–31 doi: 10.1016/j.ecoinf.2017.02.006

[12] Kirilenko A K, Molodtsova T and Stepchenkova S O 2015 People as sensors: Mass media and local temperature influence climate change discussion on Twitter *Global Environmental Change* **30** 92–100 doi: 10.1016/j.gloenvcha.2014.11.003.

[13] Autry M K and Kelly A R 2012 Merging Duke Energy and Progress Energy: Online Public Discourse, Post-Fukushima Reactions, and the Absence of Environmental Communication *Environmental Communication* **6** 278–284 doi: 10.1080/17524032.2012.672444

[14] Hutchins B 2016 The Many Modalities of Social Networking: The Role of Twitter in Greens Politics *Environmental Communication* **10** 25–42 doi: 10.1080/17524032.2014.966853

[15] Zhou X and Zhang L 2016 Crowdsourcing functions of the living city from Twitter and Foursquare data *Cartography and Geographic Information Science* **43** 393–404 doi: 10.1080/15230406.2015.1128852

[16] Fohringer J, Dransch D, Kreibich H and Schröter K 2015 Social media as an information source for rapid flood inundation mapping *Nat. Hazards Earth Syst. Sci.* **15** 2725–2738 doi: 10.5194/nhess-15-2725-2015

[17] Smith L., Liang Q, James P and Lin W 2017 Assessing the utility of social media as a data source for flood risk management using a real-time modelling framework: Assessing the utility of social media for flood risk management *J. Flood Risk Manage* **10** 370–380 doi: 10.1111/jfr3.12154

[18] Chong Z, Qin C and Ye X 2017 Environmental Regulation and Industrial Structure Change in China: Integrating Spatial and Social Network Analysis *Sustainability* **9** 1465 doi: 10.3390/su9081465

[19] Wang Y, Fu X, Jiang W, Wang T, Tsou M H and Ye X 2017 Inferring urban air quality based on social media *Sustainability* **9** 110–116 doi: 10.1016/j.compenvurbsys.2017.07.002

[20] Zhou X and Xu C 2017 Tracing the Spatial-Temporal Evolution of Events Based on Social Media Data *IJGI* **6** 88 doi: 10.3390/ijgi6030088

[21] Jendryke M, Balz T, McClure S C and Liao M 2017 Putting people in the picture: Combining big location-based social media data and remote sensing imagery for enhanced contextual urban information in Shanghai *Computers, Environment and Urban Systems* **62** 99–112 doi: 10.1016/j.compenvurbsys.2016.10.004

[22] Jung J and Uejo C K 2017 Social media responses to heat waves *Int J Biometeorol* **61** 1247–1260 doi: 10.1007/s00484-016-1302-0

[23] Mancilla-Garcia M 2015 Does Social Media Benefit Dominant or Alternative Water Discourses? *Water Alternatives* **8** 125-146

[24] Hinsley A, Lee T E, Harrison J R and Roberts D L 2016 Estimating the extent and structure of trade in horticultural orchids via social media: Social-Media Orchid-Trade Networks *Conservation Biology* **30** 1038–1047 doi: 10.1111/cobi.12721.

[25] Kryvasheyeu Y 2016 Rapid assessment of disaster damage using social media activity *Sci. Adv.* **2** e1500779 doi: 10.1126/sciadv.1500779
[26] Sun Y and Mobasheri A 2017 Utilizing Crowdsourced Data for Studies of Cycling and Air Pollution Exposure: A Case Study Using Strava Data *IJERPH* 14 274 doi: 10.3390/ijerph14030274

[27] Quinn M, Lynn T, Jollands S and Nair B 2016 Domestic Water Charges in Ireland - Issues and Challenges Conveyed through Social Media *Water Resour Manage* 30 3577–3591 doi: 10.1007/s11269-016-1374-y

[28] Oteros-Rozas E, Martín-López B, Fagerholm N, Bieling C and Plieninger T 2018 Using social media photos to explore the relation between cultural ecosystem services and landscape features across five European sites *Ecological Indicators* 94 74–86 doi: 10.1016/j.ecolind.2017.02.009

[29] Shook E and Turner V K 2016 The socio-environmental data explorer (SEDE): a social media–enhanced decision support system to explore risk perception to hazard events *Cartography and Geographic Information Science* 43 427–441 doi: 10.1080/15230406.2015.1131627

[30] Athuraliya C D, Gunasekara M K H, Perera S and Suhothayan S 2015 Real-time natural language processing for crowdsourced road traffic alerts 2015 Fifteenth International Conference on Advances in ICT for Emerging Regions (ICTer), Colombo, Sri Lanka 58–62 doi: 10.1109/ICTER.2015.7377667

[31] Wang R Q, Mao H, Wang Y, Rae C and Shaw W 2018 Hyper-resolution monitoring of urban flooding with social media and crowdsourcing data *Computers & Geosciences* 111 139–147 doi: 10.1016/j.cageo.2017.11.008

[32] Levin N, Kark S and Crandall D 2015 Where have all the people gone? Enhancing global conservation using night lights and social media *Ecol Appl* 25 2153-2167