UTILIZING THE GEOGRAPHICAL CONTEXT OF TWITTER FOR FLOOD DISASTER RESPONSE

WIDA WIDIASTUTI
March, 2016

SUPERVISORS:
Dr. F.O. Ostermann
Dr. ir. R.L.G. Lemmens

UTILIZING THE GEOGRAPHICAL CONTEXT OF TWITTER FOR FLOOD DISASTER RESPONSE

WIDA WIDIASTUTI
Enschede, The Netherlands, March, 2016

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfillment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.
Specialization: Geoinformatics

SUPERVISORS:
Dr. F.O. Ostermann
Dr. ir. R.L.G. Lemmens

THESIS ASSESSMENT BOARD:
Prof.dr. A.A. Voinov (Chairman, University of Twente)
Drs. C.W. Quak (External Examiner, Delft University of Technology, OTB Research Institute for the Built Environment)

DISCLAIMER

This document describes work undertaken as part of a programme of study at the Faculty of Geo-Information Science and Earth Observation of the University of Twente. All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the Faculty.

ABSTRACT

During a natural disaster, an up-to-date information is required for providing emergency response. People in the affected areas can provide valuable real-time information through social media since they have better knowledge of the real situation. Social media plays a role in distributing up-to-date information in real-time. Twitter as one of the social media platforms that allowing a user to share their geographic location through their broadcasted message. Twitter has been used for communication in several disaster cases. One of the utilization of Twitter in flood disaster response developed by Jakarta city is "Peta Jakarta". This system collects information from citizens via Twitter to monitor flood situation. However, there are several issues of the use of Tweets for the disaster response system. Firstly, deploying help based only on the number of the reports is insufficient because the places with lack of reports will be deprived of any assistance. Secondly, the unstructured text in Twitter usually has noise that can lead to false information. Thirdly, the location name mentioned in the Tweet can vary from the location of the Tweet itself causing ambiguity of the true location.

Using the case study of "Peta Jakarta" the research developed a conceptual workflow to improve the role of Twitter for communication during a flood disaster response. Based on the conceptual workflow, the prototype system was developed which processes one tweet at a time. The system filtered the detailed information about flood location and flood level to identify relevant Tweet using modified Named Entity Recognition and Part Of Speech (POS) tagging. The location mentioned in the Tweet was then geocoded using two geocoding services namely Open Street Map nominatim and Google geocoding. The distance from the geocoded location and the coordinates of the Tweet location was calculated. If the two coordinates are close to each other, it is likely that flooding occurred in the Tweet location.

Another added value to improve the role of Twitter in this research is the identification of a direction to the nearest shelter from a Tweet location as a feedback information. Tweet location with relevant information and close to the geocoded location was used for identification of the nearest shelter. The nearest shelter was identified by calculating the optimum path from the Tweet location to each possible shelters using Dijkstra algorithm of "pg-Routing". Shelters that were possibly inundated were eliminated from the selection process. The inundated shelters were identified using a scoring system based on the closeness to the Tweet location.

In the evaluation, the F-measure test of filtering shows the system is reliable to extract the detailed information of location content, flood level, and to identify a relevant Tweet. The F-measure test of geocoding services shows that the services are credible to geocode a location name as long as the location name is valid. OSM nominatim provides a better result compared to Google geocoding. Google might return a different location name from the input since it finds a similar name even partially in the case of failure to find the exact match. Moreover, the comparison between the spatial distribution of Tweets after filtering process and flood area obtained from the authoritative data shows a similar pattern. Thus, Twitter is a promising source of reliable flood information as long as its information is filtered properly.

Keywords: Flood disaster, Twitter, Volunteered Geographic Information, Natural Language Processing, Geocoding, Geotagging, Shelter selection

ACKNOWLEDGEMENTS

Alhamdulillahi rabbil alamiin… "All the praises be to God, the Lord of the worlds", I thank and gratitude to Allah, the most beneficent, the most merciful.

I am very grateful to my supervisors: Dr. F.O. Ostermann and Dr.ir. R.L.G. Lemmens for all the help and guidance in the whole period of my research.

I am utmost appreciate all the supports from my colleagues: Yoehannes Kinskij B, John Ray Bergado, and Ipshit Dash who are willing to spend their time for giving me feedbacks on my thesis document.

Also a very special thank you for all of my colleagues in GFM class of 2014-2016 and all of my fellow Indonesians. I would like to thank you for our precious companionship and beautiful memories during our study in ITC.

The last, I would like to express deeply my gratitude to my mother, husband, lovely son and my family for your endless love, pray, support and blessing.

TABLE OF CONTENTS

Abstract....i
Acknowledgements....ii
1. Introduction....7
   1.1. Background....7
   1.2. Case Study....8
   1.3. Research Identification....9
   1.4. Innovation Aimed at....10
   1.5. Research Methodology....10
2. The use of Social Media in Disaster response....11
   2.1. Social Media and Volunteered Geographic Information....11
   2.2. Social Media in Disaster response....11
   2.3. Related Work in Term of Methodology....12
   2.4. "Peta Jakarta" as the Current System....14
3. Conceptual Design....16
   3.1. Requirement Analysis....16
   3.2. The Role of the Users....18
   3.3. Conceptual Workflow....19
4. Methodology....23
   4.1. Data....23
   4.2. Tools Used....24
   4.3. General Workflow....25
   4.4. Data Collection....26
   4.5. Filtering of Twitter Content....27
   4.6. Geocoding of Location Content....33
   4.7. Evacuation Shelter Selection....38
5. Results....41
   5.1. Data Collection....41
   5.2. Authoritative Data for Comparison....42
   5.3. Filtering of Twitter Content....43
   5.4. Geocoding of Location Content....48
   5.5. Evacuation Shelter Selection....50
6. Evaluation And Discussion....52
   6.1. Filtering of Twitter Content....52
   6.2. Geocoding of Location Content....58
   6.3. Evacuation Shelter Selection....64
   6.4. System Performance....64
7. Conclusions and Recommendation....65
   7.1. Conclusions....65
   7.2. Limitation and Recommendations....67
List of references....69
Appendix....74

LIST OF FIGURES

Figure 1.1: Methodological workflow 10
Figure 2.1: Tweet invitation at petajakarta.org (source: http://petajkt.org) 15
Figure 2.2: The workflow to obtain information at petajakarta.org (source: http://petajkt.org) 15
Figure 3.1: Tweet of flood report category 16
Figure 3.2: Tweet of help and evacuation category 16
Figure 3.3: Tweet of review category 17
Figure 3.4: Tweet of information shelter 17
Figure 3.5: The users as seeker 18
Figure 3.6: The users as contributor 18
Figure 3.7: The conceptual workflow of utilizing geographical context of Twitter in flood disaster system 22
Figure 4.1: The flow of data retrieval 23
Figure 4.2: The connection of the tools 24
Figure 4.3: The general workflow of data collection, filtering, geocoding, and shelter selection 25
Figure 4.4: The workflow of data collection 26
Figure 4.5: The workflow of filtering of Twitter content 27
Figure 4.6: The output of first geocoding experiment 34
Figure 4.7: The output of second geocoding experiment 34
Figure 4.8: The output of third geocoding experiment 35
Figure 4.9: The workflow of geocoding 37
Figure 4.10: The example of identification the nearest shelter 39
Figure 4.11: The compass degree 39
Figure 4.12: The workflow of evacuation shelter selection 40
Figure 4.13: The example of identification a direction to an un-named path 40
Figure 5.1: The output of data collection 41
Figure 5.2: The line chart of Tweets activity per week 41
Figure 5.3: The flood area in 2015 42
Figure 5.4: The comparison of relevant Tweets and off-topic Tweets based on filtering category 43
Figure 5.5: The comparison between flood level from Tweet content and flood level from the authoritative data 44
Figure 5.6: The distribution of Relevant Tweets 45
Figure 5.7: The distribution of off-topic Tweets 46
Figure 5.8: Relevant Tweets and Off-topic Tweets per month 47
Figure 5.9: The example of a reply to the invitation 47
Figure 5.10: The example of a noise content 47
Figure 5.11: The results of Geocoding content 48
Figure 5.12: The distribution of Tweets based on the geocoding services 48
Figure 5.13: The distribution of Tweets that close to its location content 49
Figure 5.14: The distribution of selected shelters and the route from Tweets to the shelters 50
Figure 5.15: The direction to the shelter from the Tweet displayed in Cartodb 51
Figure 6.1: The first sample of the fail identification of location name (there is no keyword precede Pulogebang Jakarta timur) 53
Figure 6.2: The second sample of the fail identification of location name ("di" or "in" is a keyword to identify the location name but it written without space) 53

Figure 6.3: The third sample of the fail identification of location name (the identified location name is 'mosque palmerah' instead of 'failaka mosque') 54
Figure 6.4: The forth sample of the fail identification of location name ("dpn" is abbreviation of "depan" or *in front of* in English) 54
Figure 6.5: The first sample of the fail identification of flood level 55
Figure 6.6: The second sample of the fail identification of flood level 55
Figure 6.7: The third sample of the fail identification of flood level 55
Figure 6.8: The forth sample of the fail identification of flood level 55
Figure 6.9 : The fifth sample of the fail identification of flood level 55
Figure 6.10: The sixth sample of the fail identification of flood level 56
Figure 6.11: The first example of error identification of relevant Tweets ("banjir" is flood keyword and "kelapa gading nias" is location name) 56
Figure 6.12: The second example of error identification of relevant Tweets ("banjir" is flood keyword and "cempaka putih, cempaka mas, and arta gading" are location name) 57
Figure 6.13: The third example of error identification of relevant Tweets ("banjir" is flood keyword and "arta gading, matraman" are location name) 57
Figure 6.14: The forth example of error identification of relevant Tweets ("air" is flood keyword and "mata" is flood level keyword) 57
Figure 6.15: The fifth example of error identification of relevant Tweets 57
Figure 6.16: The sixth example of error identification of relevant Tweets (BHP is the location name but the system fails identify it) 57
Figure 6.17: The proportion of the Geocoding output 58
Figure 6.18: The Tweet that produce incorrect location name 59
Figure 6.19: The distance between primary geocoding and secondary geocoding obtained from Google geocoding 59
Figure 6.20: The distance between primary geocoding and secondary geocoding obtained from OSM 59
Figure 6.21: The sample of geocoding output (secondary geocoding) that very close to the geocoding input (primary geocoding) 61
Figure 6.22: The example of primary geocoding and secondary geocoding in the same location but the distance is far 62
Figure 6.23: The example of failure in identification the location name by Google geocoding 62
Figure 6.24: The first example of flood location is not covered in the authoritative data 63
Figure 6.25: The second example of flood location is not covered in the authoritative data 63
Figure 6.26: The Tweet that failed to identify a shelter because of the unconnected path 64

LIST OF TABLES

Table 3.1: The workflow symbol......19
Table 4.1: The list of tools that used in the system......24
Table 4.2: The conversion of flood depth level......30
Table 4.3: The Indonesian tag-set (source: Dinakaramani et al., (2014))......31
Table 4.4: The first comparison of geocoding output between OSM nominatim and Google Geocoding.34
Table 4.5: The second comparison of geocoding output between OSM nominatim and Google Geocoding......35
Table 4.6: The third comparison of geocoding output between OSM nominatim and Google Geocoding......35
Table 4.7: The rule assignment to identify the direction based on the azimuth......39
Table 5.1: Flood level results and comparison......43
Table 5.2: Relevant Tweets, Off-topic Tweets and Comparison......45
Table 5.3: Tweets that close to its content location, relevant Tweets that close to its content location and Comparison......49
Table 5.4: Shelters selection of Tweets that relevant and close to its content location......50
Table 6.1: The contingency table......52
Table 6.2: The contingency table of the identification of location name......53
Table 6.3: The contingency table of the flood level Tweets......54
Table 6.4: The contingency table of the relevant and off-topic Tweets......56
Table 6.5: The comparison output between OSM and Google Geocoding......58
Table 6.6: The quality checking of secondary geocoding......60
Table 6.7: The contingency table of geocoding service......60
Table 6.8: The average time consuming of each process......64

1. INTRODUCTION

1.1. Background

During a disaster, up-to-date information is required to identify areas that need immediate rescue. This information can be obtained from the affected citizens, because they have a better understanding about the areas. As an example, the 2010 Haiti earthquake shows how citizen volunteers play a role in providing information about their situation (Oliver et al., 2014). As the result, it was very useful for relief agencies to allocate the supplies during emergency response. The information is also useful for another disaster case, for example in flood disaster. The in-situ information helps relief agencies to monitor the situation (Poser & Dransch, 2010).

Poser & Dransch (2010) states that during a flood disaster, internet technology provides the possibility to collect in-situ information such as water level and location. The convergence of internet technology and social media provides a high potential for collecting up-to-date information from internet users, since this media is mobile and ubiquitous (Bruns & Liang, 2012).

During a natural disaster, a relief agency uses this social media to get a better understanding of the situation in real-time (Yin et al., 2012), and to find information continuously from the citizens about the on-going disaster (Landwehr & Carley, 2014). One of the social media that can be used to retrieve the information is Twitter. Twitter allows its users to broadcast a short message within 160 characters called Tweet. When the users broadcast a Tweet, they can share the geographic position, known as geolocation. This functionality is useful for those who wants to share the geographic position when reporting the incident location.

The user who share the geographic position of incident location acts as a volunteer. As referred by Elwood, Goodchild, & Sui (2012), the individual or group voluntary activity that reports geographic information results in Volunteered Geographic Information (VGI). The convergence of social media and geographic information develops a new perspective of VGI (Jiang, 2012).

Rojas & Muñoz (2014) show how VGI from geolocation Tweets promises a great opportunity to map a distribution of crisis events in a web application. A similar system has been developed by the relief agency of Jakarta, called petajakarta (DKI Jakarta Capital Government, 2014). This application gathered and shared the flood information derived from geolocation Tweet that were sent by the citizen. This way, the relief agency monitors the distribution of the Twitter reports in order to identify hot spots problem and to respond accordingly.

Monitoring only the number of reports may not be sufficient, because oversimplifying the relationship between geospatial distribution of Twitter activity and the areas that are hit hardest provides a high risk of the insufficient help for the areas that lack much report activities (Shelton et al., 2014). In addition, the geolocation Tweet do not necessarily imply the incident location, because the location name referred in the Tweet message can be different from the location from where the Tweet is sent. Furthermore, the noisy nature of Twitter content can impair the credibility of the information (Flanagin & Metzger, 2008) since it is collected from everyone indiscriminate of motivation or expertise.

To avoid false reports during the disaster response, the filtering of Twitter reports is necessary. The extraction and filtering information in a Twitter message are possible using a natural language processing (Klein & Castanedo, 2013). Moreover, this also allows extracting a location name mentioned in a Twitter message using Named Entity Recognition (NER) (Liu et al., 2011).

The identified location name can be used to check the reliability of the Twitter report by comparing the location from where the Twitter was sent and the location name mentioned in the Tweet. The location name can be converted into geographic coordinates using a geocoding service such as Google geocoding (Davis & de Alencar, 2011). Therefore, the distance between the two locations can be calculated.

During a disaster, people use Twitter not only to report an incident location but also to request a help and rescue as stated by Adam and Muraki (2011). They also state that Twitter is the more reliable tool to communicate than other devices in an emergency condition. Therefore, it raises a challenge for the relief agency to optimize the use of Twitter for providing help and rescue operations. One of the information that can be provided by relief agency is about evacuation shelter. It is because evacuation shelter is a basic need in disaster condition (CCCMCluster, 2014; Rashid, Haider, & McneilL, 2007)

The example of the platform for communicating between citizen and relief agency is a disaster portal that developed by Lickfett et al. (2008). The portal providing various information during the disaster. For instance, the citizens allow accessing information about shelter through the portal. However, when Twitter is a platform for communication, a relief agency can use the same platform to provide the information directly to the citizen.

The exchanged information between the affected citizens and the relief agency during the disaster response is optimizing the role of Twitter as a platform for communication. The citizens act not only as the information seeker but also as the sensor who contribute the information in the field (Goodchild, 2007). Meanwhile, the relief agency obtains the periodically updated information.

1.2. Case Study

The use of Twitter to obtain information during flood disaster has been widely used in various application. One of them is "Peta Jakarta", a system developed by relief agency of Jakarta to gather and to disseminate flood information obtained from Twitter. The system utilizing geolocation Tweet which contains the information about flood in real-time (DKI Jakarta Capital Government, 2014).

Flooding in Jakarta occurs almost every year. The flood is influenced by several factors such as disorganized building construction, increasing urbanization, economic growth, and global climate change. Another factor is the changing of natural condition. It occurs due to population increase and the spread of residential/Industry that covered the catchment area (Team Mirah Sakerti, 2010). This issue is worsened by the population wastes and sediment that cause the silting of the rivers (Steinberg, 2007).

The worst flood disaster in the last ten years happened in 2007, because of the high rainfall for days. The affected areas are estimated around 4.370 km2 with the height reached 4 meters, rendered 340.000 homeless, and the total cost of the disaster is skyrocketing at 8.8 trillion Indonesian Rupiah (Dartmouth Flood Observatory, 2008).

The existence of "Peta Jakarta" helped Jakarta's relief agency to monitor the situation during the disaster. The citizens of Jakarta reported the situation in their neighborhood to the system. The role of the system is to be a platform for communication between citizen and relief agency to obtain information and to provide help and rescue. Therefore, by considering these explanations, this research used "Peta Jakarta" as the case study by offering benefits to improve the role of Twitter in such system.

1.3. Research Identification

The section of research of identification gives more explanation about research objectives (1.3.1) and research questions (1.3.2) that are taken for this research. They were built based on issues that are addressed in the previous sections.

1.3.1. Research Objectives

1. To develop a conceptual design for utilizing geographical context of Twitter in flood disaster response system.
2. To develop an automated process for filtering the Twitter content of flood disaster using natural language processing.
3. To develop an automated process for assessing the geocoding of Twitter and the geocoding of its location content in flood disaster.
4. To develop an automated process for providing on the nearest safe evacuation shelter as a feedback information.

1.3.2. Research Questions

For objective 1:

Research Questions:

a. What is the characteristic of Twitter content in flood disaster?

b. What is the use case to be implemented?

For objective 2:

Research questions:

a. Which natural language processing method to use or modify in filtering the Twitter content?

b. What is the evaluation result of filtering the Twitter content?

For objective 3:

Research questions:

a. How to assess the geocoding of Twitter and the geocoding of its location content?

b. What is the evaluation result of assessing the geocoding of Twitter and the geocoding of its location content?

For objective 4:

Research questions:

a. What is the method to use or modify for finding a route to the nearest shelter under flood disaster conditions?

b. How can these directions be communicated back to the information seeker?

1.4. Innovation Aimed at

The innovation of this research aims to utilize Twitter in flood disaster beyond from the data collection, which is the automated extraction of detail information about the content, the automated geocoding based on the content, and the identification of direction to evacuation shelter as a feedback information.

1.5. Research Methodology

The research methodology were generated to obtain the research objectives. The research methodology are showed as follows:

1. A literature review regarding the VGI and social media in natural disaster, natural language processing, geocoding, and the evacuation shelters selection.
2. Requirement analysis to identify the benefits for the current system where to focus on the automatic processing.
3. Design the conceptual of the automatic processing of citizen volunteer in flood disaster, from Twitter. Then determining the design that doable for the research.
4. Implementation of the conceptual design into a system.
5. Evaluation the result.

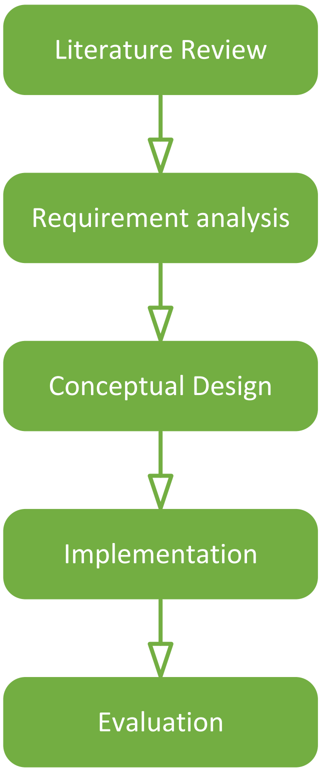


Figure 1.1: Methodological workflow

2. THE USE OF SOCIAL MEDIA IN DISASTER RESPONSE

2.1. Social Media and Volunteered Geographic Information

As referred by Goodchild (2007), volunteered geographic information (VGI) emphasizes the role of volunteers in contributing geographic information. The emergence of Web 2.0 technology (O'Reilly, 2005) offer non-professionals as volunteers to participate in updating process an information (Roche et al., 2011). The web community as citizens act as sensors in sharing information via social web (Web 2.0) (Sheth, 2009). The technology allows users to insert latitude-longitude coordinates called 'geotag' to allocate a place-mark (Crampton et al., 2013). The ability in sharing the geographic position has been adapted by several social media platforms such as Flickr, Instagram, and Twitter.

Twitter is a social media platform that is widely used by the internet community. Its users can broadcast a 140 characters message, called Tweet, via the internet. One of the features that Twitter provide is the location-based feature. This feature enables users to share their personal position in anywhere and anytime. The convergence of social media and geographic information transforms the perspective of VGI (Sui & Goodchild, 2011) from what Goodchild (2007) referred.

Twitter provides an API (Application Program Interface) that allows developers to access Twitter data (Twitter, 2016). The emergence of API and support location-based feature by Twitter offer a high potential to gather a large number of geographical information. Therefore, it gives an opportunity to analyze the spatial distribution of this information for various purposes.

2.2. Social Media in Disaster response

Nowadays, social media has changed the way of people to interact and to communicate in daily life. People can broadcast information easily using this platform. Thus, propagating news can be faster and broader. During a crisis event, people use social media to get up-to-date information about the current situation. This opportunity is used by relief agency to collect data on social media to allocate relief needed and to broadcast the situation using their application software (Landwehr & Carley, 2014).

Since some social media allow users to share their personal position in geographic coordinates, it gives a positive impact for relief agency to analyze a spatial distribution in real-time. As referred by Poser & Dransch (2010), up-to-date information is required in all phases of disaster management. In this case, users of social media act as volunteers who share their spatial information or called VGI. In other words, they act as "sensors" that observe the situation in-site. For instance, in flood disaster VGI could be utilized to detect flood locations and to propagate flood alerts from the local authority (Schade et al., 2011).

Vieweg et al., (2010) revealed how Twitter contributed to develop situational awareness during the Red River floods and the Oklahoma grassfire. Using an automatic system, they extracted the place name that was mentioned in a Twitter message to identify a disaster location. The result shows that the automatic method requires a better understanding how an affected citizen is communicating. Thus, the extraction strategy is necessary for this kind of system. However, the positive benefits of the social media in crisis event are not without some drawbacks. There is a possibility that the information given is purposely misleading. Based on the possibility, a further filtering of every information is required (Madry, 2015).

According to Reza Fazeli et al. (2015), there are several approaches to assess the credibility of VGI in flood hazard mapping. First approach is analyzing the spatial pattern by classifying data. Second approach is analyzing the metadata to extract the information using natural language processing. Last approach is, comparing the authoritative data to assess the quality of VGI.

In the conceptual workflow of the automatic quality assessing of VGI by Ostermann & Spinsanti (2011), several pointers need to be considered in the data collection phase; a keyword could have several meanings, and within a single message could have some information of locations. Furthermore, to enrich the content, the information of places or location name mentioned in the text could be transformed into geographic coordinates. Therefore, it can be used to analyze the distance between the location name and where the message is sent.

From the conceptual model of the automatic credibility assessment of VGI (Idris et al., 2014), there are two main components of the indicator in assessing VGI. The first component is metadata and the second component is the data itself. The recommendation to assessing these components are using machine-based learning to develop a digital metadata and using other government data to assess the credibility via the Linked Data Web infrastructure. However, it is mentioned that currently to link the government data using Semantic Web is still be argued since the lack of research in this technology in supporting the large-scale data.

According to Adam & Muraki (2011), another recommendation to improve the reliability of Twitter in communicating during a crisis is using an official hashtag and using the system that could extract the information. The use of the hashtag and prescribed syntax will make it easier for the machine to read the data (Starbird & Stamberger, 2010). This way, the simultaneous communication between two sides could give benefit to both local authority and citizen.

2.3. Related Work in Term of Methodology

2.3.1. Filtering of Twitter Content

The information written in a Tweet is a kind of human language (English, Indonesia). The structure of most languages are different. The methodology to understand human language computationally is called Natural Language Processing (NLP) (Robin, 2009a). An NLP method allows extracting information in a written text such as a Twitter message.

An example of the use of NLP in Twitter is a system that has developed by Cameron et al. (2012). The system is capable of detecting the incidents from Twitter during an emergency. This system has several abilities, such as (1) Detecting the incident using a burst detection method from historical data and statistic model to count how many times a certain word emerges (2) Summarizing information that is the result of burst detection by clustering similar event. (3) Classifying relevant Tweets using support vector machines. (4) Managing issues by separating the clusters which do not contain similar contents, and (5) Analyzing historical alerts which have been obtained from the burst detector.

The most significant phase in analyzing Tweet content is the process of identification of words using the Part Of Speech (POS) tagging (Gimpel et al., 2011). Each word inside a Tweet is tagged according to its lexical category. For example, A for the adjective, N for the noun, P for preposition, and so on. The result of this tagging process is in a form of annotated corpus that could be used for other linguistic analyzes.

Dinakaramani et al. (2014) designed an Indonesian Part of Speech (POS) tag set to assign a tag or a descriptor into a word. The process consists of two main steps which are assigning the initial tag set and revising the tag set by manually add into the Indonesian corpus consisting of 10,000 sentences. As a result, there are 23 Indonesian tag sets such as; coordinating conjunction (CC), cardinal number (CD), preposition (IN), adjective (JJ) and so on.

Another method in the NLP is Named Entity Recognition (NER) (Ritter et al., 2011) that aims to identify named entities such as persons, organizations, and locations. POS tagging is the prerequisite to identify each word in NER. Considering the content of Twitter is an unstructured text, the different style of vocabulary is needed to overcome out of vocabulary (OOV) issues.

The NER method for the Indonesian language is conducted by Budi et al. (2005). They developed a set of rules to identify named entities such as persons, organizations, and locations. It is derived from a combination of contextual, morphological and part of speech features. To check the accuracy of this method, they conducted an empirical evaluation using the F-measure approach. The result is 63.43% recall and 71.8% precision. The low results occur because of a conjunction that is mostly used in a part of organization name. The following rule is a sample to recognize a location entity.

"**If** a proper noun is preceded by 'di' **then** the proper noun is the name of a location"

Middleton et al. (2014) utilized the information obtained from Twitter for crisis mapping during a natural disaster. To increase the number of Tweets, they used NER method detect 'named entities' or 'place name' within the Tweet which are then geocoded using a gazetteer. They conducted several steps to extract a place name and to get the coordinates of each place. Firstly, during tokenization, the message split into separate words. Then, a noise such as URLs and email address is omitted to avoid false tokens. Secondly, using the POS tagging, each word is classified based on its lexical category. The pattern of its lexical then used to identify a place name that exists inside the Tweet. Lastly, the identified place name is given the geographic coordinate using the geocoding method.

2.3.2. Geocoding of Location Content

Geocoding process is also known as Geotagging, which is a process to identify one part of a text that describes spatial aspect into a geographic coordinate location (Ghahremanlou et al., 2015). The emergence of web services followed by the emergence of online geocoding services (Google geocoding, Yahoo PlaceFinder, Open Street Map nominatim) help the developers to choose a service that suitable to their needs.

Watanabe et al., (2011) detects an event by identifying a non-geotagged location from Twitter using Foursquare's service. They extracted a place name from a Tweet, and then sending them to the Foursquare service. Then, the service calculates a variance of the place name and the physical location. The fewer variation means, the more correct one place name is described in the physical location.

In contrast, with Yin et al. (2012), developed a system that implements the geotagging process for a location in the user profile if a Tweet has no location. The content of the location is sent to the Yahoo Geocoding Service to obtain top five matching location in the world.

Each online geocoding service has the difference of strategy and database that affect to the output quality. Roongpiboonsopit & Karimi (2010) compare and evaluate the output of several online geocoding services; Geocoder.us, Google, Microsoft, MapQuest and Yahoo. Google, MapPoint and Yahoo have the high match rates. MapQuest is of middle quality, and Geocoder.us has high-unmatched rates.

The geocoding services provide a metadata as an output. One of the attributes in the metadata is a granularity of the output level. In Teske (2014), they compared several services as follow: Yahoo Placefinder has a different rank in every level of point, street, and area. Bing Maps has four output levels, which are "rooftop", "parcel", "offset", and "interpolation". Similar to Bing Maps, Google Geocoding Version 3.0 also has four level output: "rooftop", "range_interpolated", "geometric center", and "approximate". Even though OSM does not have the granularity of the output level, the output is the most precise compared to the others, and OSM has an attribute type to distinguish whether the output is point, line, or area.

According to the document of Google Geocoding service (Google Developers, n.d.), the output metadata indicates a different level of the precisions:

- "ROOFTOP", indicates the output has an address precision or a Point of Interest (POI) address.
- "RANGE_INTERPOLATED", indicates the output is the interpolation between two precise points.
- "GEOMETRIC_CENTRE", indicates the output is a polyline or a polygon.
- "APPROXIMATE", indicates the output is approximate.

Even though, OSM does not provide coordinate accuracy metadata, the attribute of OSM *type* from the metadata could be used to distinguish whether the output is point, line, or area.

2.3.3. Evacuation Shelter Selection

With the simultaneous communication between relief agency and citizen through the social media in real time during a disaster, it is possible for the relief agency to give information about the nearest shelter location to an affected citizen. by considering the accessibility, the shelter is better to be located within walking distance, which is maximum 1 km to reach by walking (Sanyal & Lu, 2009).

Regarding the evacuation route, distance is often used in many studies such as Euclidian distance or network path distance as a parameter to calculate a travel cost. Others used time of the cost (Ye et al., 2012). For the shortest path algorithm, the Dijkstra algorithm (Dijkstra, 1959) performs better in an evacuation system (Wang et al., 2011). This algorithm uses a simple yet efficient coding, To generate 19,701 paths in a CPU with a 512 MB memory card, it takes less than 3 seconds (Alçada-Almeida, et al., 2009).

Fortunately, Dijkstra algorithm can be used in the PostGIS/PostgreSQL extension called "pgrouting". "pg-Routing" is a library with various tools to find the optimum paths based on cost parameter. It can be calculated dynamically using the Structured Query Language (SQL) within a very short time execution (Zhang & He, 2012).

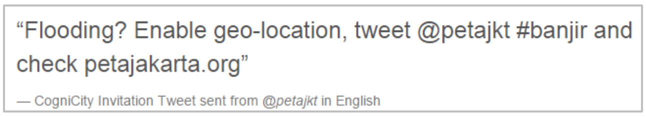
Singh et al., (2015) developed a system using pgrouting to obtain the shortest and the alternative path in different disaster scenarios. Here, its users could give information about the affected road segments through the system that will be stored in a database. Hence, when another user tries to find a route, those segments might not be calculated into the query.

Regarding this research, Dijkstra algorithm on "pg-Routing' is used to determine the shortest path from several potential shelters provided by the system. Ultimately, users will obtain feedback from the system, which shelters are the most potential towards their location.

2.4. "Peta Jakarta" as the Current System

The annual flooding that occurs in Jakarta attracts the local government to collect flood information direct from citizen by deploying a real-time application called "Peta Jakarta" (DKI Jakarta Capital Government, 2014). Using this system, the relief agency of Jakarta city called BPBD gathered and disseminated flood information obtained from Twitter. The effected citizen can report the condition of their neighborhood to the system via Twitter. This way, the relief agency can monitor the situation during the disaster. The system is active every monsoon season, which is from December to March.

In a white paper (Holderness & Turpin, 2014), it is explained that the system retrieved the data in two phases: The first phase is actively persuading any Twitter users who insert the word 'banjir' (flood) inside their Tweet by sending a Tweet invitation (figure 2.2). This invitation is to confirm whether the users are affected by the flood or not. If they are affected, then they can send back a Tweet to the system's official account (@petajkt). In the second phase, the confirmation Tweet is automatically mapped in the system. The Tweet can be inserted to the map if it contains a geolocation. If there is none, then the system will send back another Tweet that asks the user to enable their geolocation. Figure 2.2 describe the workflow of "Peta Jakarta" in obtaining the information.



"Flooding? Enable geo-location, tweet @petajkt #banjir and check petajakarta.org"

— CogniCity Invitation Tweet sent from @*petajkt* in English

Figure 2.1: Tweet invitation at petajakarta.org (source: http://petajkt.org)

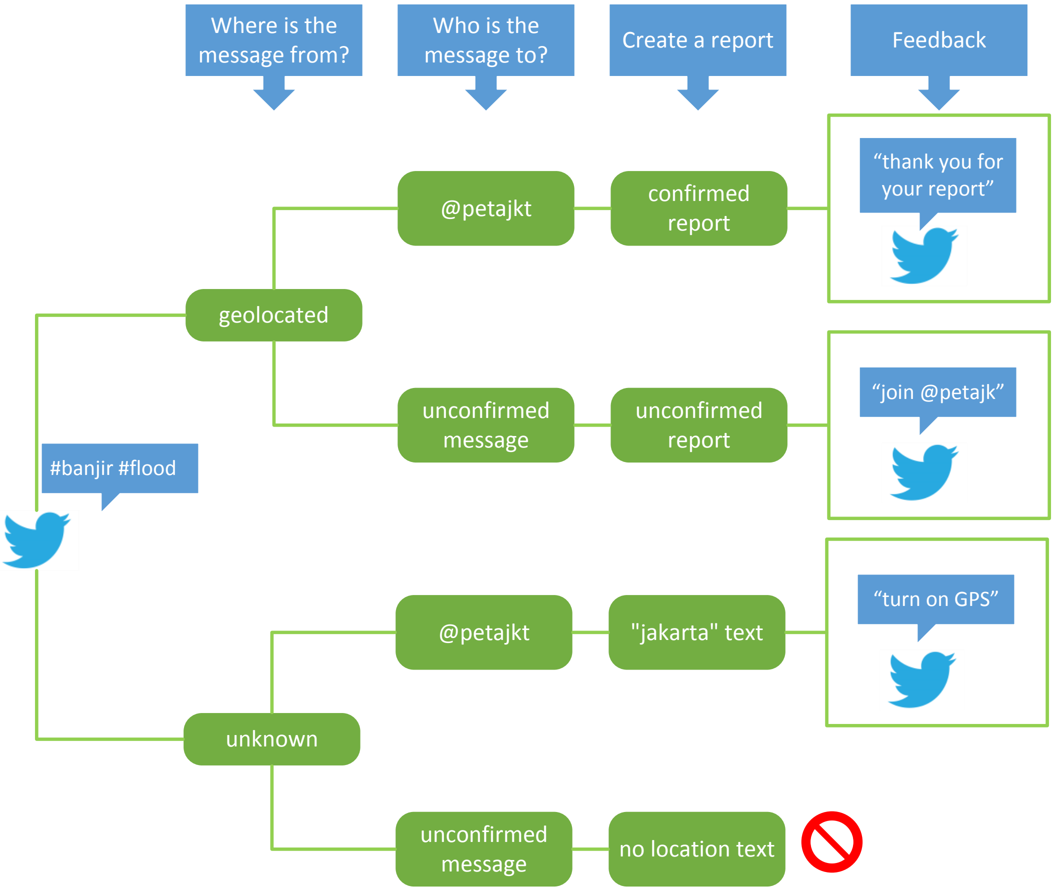


Figure 2.2: The workflow to obtain information at petajakarta.org (source: http://petajkt.org)

3. CONCEPTUAL DESIGN

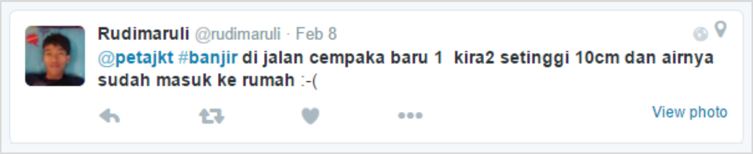
This chapter addressed the first objective of the research; the objective is to develop a conceptual design for utilizing geographical context of Twitter in flood disaster system. There are three steps to obtain the objective. First, identifying the problem based on requirement analysis. Second, identifying the user who involved in the system. Third, designing a concept that can be implemented to the system. Flood disaster system that used as the case study is a system developed by Jakarta city called "Peta Jakarta". Hence, the concept was determined based on the system.

3.1. Requirement Analysis

In a research documentation of "Peta Jakarta" (DKI Jakarta Capital Government, 2014), a frequently asked question is about the accuracy and the validity of the report. Currently, the validity process that has been conducted, which are:

1. Observing the concentrated areas of Tweets, then manually validating it by crosschecking with other Tweets on the same areas.
2. Monitoring the other official Twitter accounts that give information about flood as the addition cross checking. Those are @BPBDJakarta as the official account of the Jakarta relief agency, @TMCPoldaMetro as the official police account for traffic management center in Jakarta, and @RadioElshinta as the official account of one radio news in Jakarta.
3. Monitoring electronic media such as radio, television and other internet sources.

The documentation explained, there are three categories of Tweet that received by @petajkt. The first is "flood report" category that includes details about location name and flood depth level. The second is "help and evacuation" category that asks for help, support, and evacuation. The third is "review" category that sends a feedback or review related to problematic reports. The examples of flood report category, help and evacuation category, and review category are shown in Figures 3.1, 3.2, and 3.3 respectively.



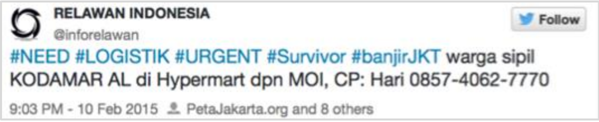
Rudimaruli @rudimaruli · Feb 8

@petajkt #banjir di jalan cempaka baru 1 kira2 setinggi 10cm dan airnya sudah masuk ke rumah :-(

View photo

@petajkt #flood in cempaka baru 1 street is around 10cm and the water has already got inside the house

Figure 3.1: Tweet of flood report category



RELAWAN INDONESIA
@inforelawan

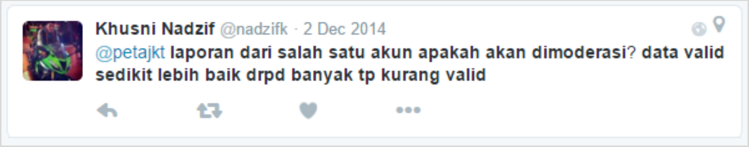
Follow

#NEED #LOGISTIK #URGENT #Survivor #banjirJKT warga sipil KODAMAR AL di Hypermart dpn MOI, CP: Hari 0857-4062-7770

9:03 PM - 10 Feb 2015 PetaJakarta.org and 8 others

#need #logistic #urgent #Survivor #banjirJKT citizen KODAMAR AL in Hypermart in front of MOI, CP: Hari 0857-4062-7770

Figure 3.2: Tweet of help and evacuation category



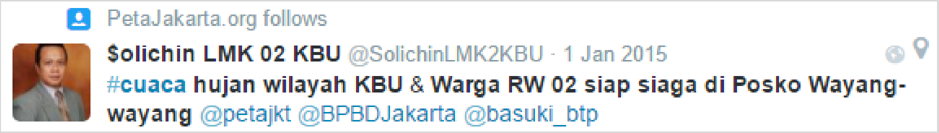
Khusni Nadzif @nadzifk · 2 Dec 2014

@petajkt laporan dari salah satu akun apakah akan dimoderasi? data valid sedikit lebih baik drpd banyak tp kurang valid

@petajkt Does the report from one of the account will be moderated? Less data but more valid is better than more data but less valid.

Figure 3.3: Tweet of review category

In a Twitter message that sent to "Peta Jakarta" account, there is a message that provides information related to available shelter. Figure 3.4 shown a user sent a report about the weather condition from a shelter.



PetaJakarta.org follows

Solichin LMK 02 KBU @SolichinLMK2KBU · 1 Jan 2015

#cuaca hujan wilayah KBU & Warga RW 02 siap siaga di Posko Wayang-wayang @petajkt @BPBDJakarta @basuki_btp

#weather raining at KBU area & citizen of RW 02 ready at Wayang-wayang shelter @petajkt @BPBDJakarta @basuki_btp

Figure 3.4: Tweet of information shelter

These Tweet categories show the different interest of the people in their report. Therefore, a classification of Tweets is required before there are shown on a map. Otherwise, there will be misunderstandings when looking at the spatial distribution. The high concentration of Tweets in one area, does not always mean that the area suffers the most (Shelton et al., 2014).

Tweets of the "flood report" category, contains a location name where the flood occurs. The "Peta Jakarta" mapped the flood reports based on the geolocation Tweet (location of the mobile phone when the Tweet is sent). However, the geolocation Tweet might be different from the location mentioned in the Tweet. For instance, users can send a Tweet from their workplace to report the flood condition of their house. Due to this, any application that focuses on the location content at the base of the disaster decision making should be aware whether the geolocation Tweet is suitable for the system (Hahmann et., 2014).

Disaster is always closely connected to the shelter and evacuation, as can be seen in the category "help and evacuation" Tweet where the users request help in logistic or evacuation. As a prone area, Jakarta city has evacuation shelters plan that can be used during disasters. If there is a quick feedback from the relief agency to the users about nearby shelters, then they will able to save themselves without having to wait for the arrival of the evacuation team.

Therefore, the added value that can be implemented in such system (Peta Jakarta) are:

1. Automatic filtering to obtained a detail information about flood location and flood depth level to identify relevant Tweet and off-topic Tweet;
2. Disambiguate the flood location between geolocation content (location mentioned in the Tweet) and geolocation Tweet (location where the Tweet sent);
3. Management of the feedback regarding the useful information for the user.

The requirement analysis was identified the characteristic of Twitter content in flood disaster. Therefore, the section addressed the research question about what is the characteristic of Twitter content in flood disaster.

3.2. The Role of the Users

The role of the users is explaining the type of users that involved in the current system. Based on the explanation in the research background (chapter 1) and the requirement analysis, the conceptual design proposed the role of the users could be divided into the following categories:

1. As a seeker, users consists of two categories as seen in Figure 3.5: first, the relief agency who seek the information about the flood. Second, the affected citizen who seek a help and rescue.

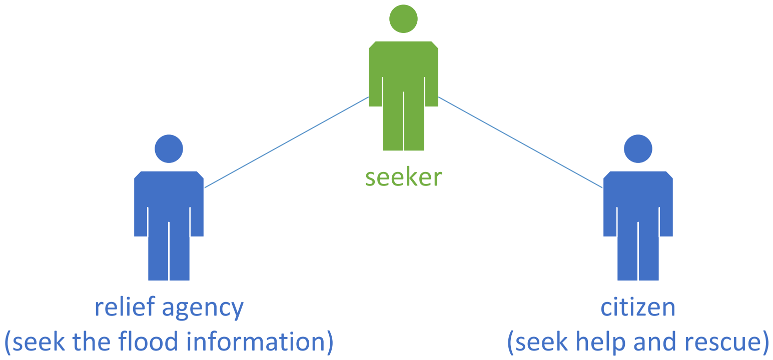


Figure 3.5: The users as seeker

2. As a contributor, users consists of two categories as seen in Figure 3.6: first, a citizen who contribute the information about the flood. Second, the citizen who contribute to update the information about a shelter.

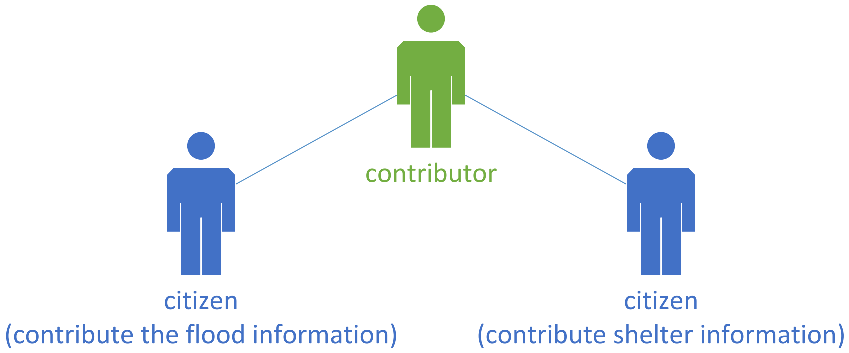

contributor
citizen
(contribute the flood information)
citizen
(contribute shelter information)

Figure 3.6: The users as contributor

Even though the role of users is divided into several categories, it does not mean that a citizen cannot be acted as a seeker and a contributor simultaneously. Because, when they were sending a Tweet, the message might be contained both information and request. Therefore, this research proposed to focus on one group of the users to limit the problem. The group is the users who act as contributor of flood information.

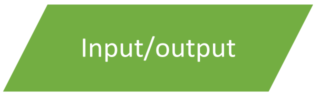
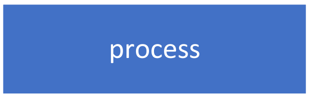
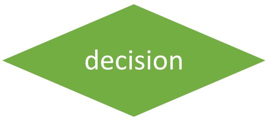
3.3. Conceptual Workflow

The conceptual workflow describes the flow of process in utilizing geographical context of Twitter in flood disaster system. This section addressed the research question about what is the use case to be implemented.

As explained in the role of the users, this research focused on the users who act as contributor of flood information. Considering in a disaster, a citizen might be needed a help and rescue, a feedback information should be informed to the users. The research proposed a direction to the nearest shelter as the feedback. However, only users who provide a relevant information that get the feedback. Therefore, filtering the content is implemented before identification the shelter. After filtering, another step is assessing the geolocation of the Tweet (primary geocoding) and the geolocation of the location content (secondary geocoding) mentioned in the Tweet. The assessing is to calculate a distance between the two geolocations. The close distance means, the Twitter users report a location in the same area where they were sending the Tweet.

Before explaining more detail about each process, the following Table 3.1 is describing briefly the list of symbol that used in every workflow in this research.

Table 3.1: The workflow symbol

Symbol	Name	Functionality
	Input/Output	A parallelogram that represents an input or an output
	Process	A rectangle that represents a process
	Decision	A diamond that represents a decision
	Arrow	An arrow that represents a flow of process

As already explained there are three main steps in the conceptual workflow; filtering the content, assessing the geolocation, and identifying a direction to the nearest shelter. The following points are the use case scenario of the concept. The workflow of the concept is portrayed in Figure 3.7.

1. The announcement and invitation

   In Twitter, it is a common using a specific hashtag to highlight a topic. Currently, “Peta Jakarta” use a hashtag *banjir* #banjir (flood) in their invitation. According to Starbird & Stamberger (2010), the hashtag is a prescribed syntax can help the system to extract the information for different purposes in an emergency situation. For examples to share flood location (e.g. #info #location), to seek a help and rescue (e.g. #help #evacuation), and to share shelter information (e.g. #info #shelter). These hashtags can be introduced when the relief agency send the invitation to persuade the citizen.

The invitation is conducted with two approaches, first the relief agency broadcast the invitation with a regular interval, and second the system send the invitation automatically to Twitter users if a specific keyword (e.g. banjir/flood) identified in their tweet.

Before a Tweet is processed in the system, checking the correct format is necessary. In the current system, the checking means, if the Tweet contains the geolocation and the official account (@petajkt), then it will be used in the system.

2. Filtering

   The filtering process allows classifying a tweet into different categories. Because this research is focused on the group of users who contribute information about flood, then the filtering aims to identify a relevant Tweet that related to flood infomation. As mentioned in the requirement analysis, a flood report Tweet contains information about flood location or flood depth level. This mean, the relevant Tweet is a Tweet that contains one of the information.

   The extraction of flood location and flood depth level is using a Natural Language Processing (NLP) approach. In the NLP, the use of a set predefined keyword can help the process to distinguish the group message (Klein & Castanedo, 2013). The content of flood location and flood depth level then used to determine whether the information is relevant or not.

3. Geocoding

   Location name (location content) mentioned in a Tweet is used by citizen to refer a flood location. The closeness between the Tweet location (primary geocoding) and the location content (secondary geocoding) could be varied. A large distance between these locations can be misleading. It is related to the uncertainty about the true location. Therefore, checking the distance is important to be considered before it added into a system such "Peta Jakarta".

   Checking the distance needs information about the coordinates of the two location. The first coordinates obtained from the geolocation of the Tweet (where the Tweet is sent) and the second coordinates obtained from the geolocation of the location name (location content). For this purpose, the location name should be converted into coordinates. A geocoding service has the ability to find coordinates from a given location name. The coordinates from geocoding content (location name) are represent the center of point, line, or area. For example, if a given location name is a street, the coordinates from geocoding content is the center of the street. This mean, the distance is calculated from the primary geocoding to the center of secondary geocoding.

   After identifying the secondary geocoding (location content), the distance between the two locations can be known. If the distance is too far, then a feedback can be sent to the user for request confirmation of the true location. In this concept, a Tweet that close to its location content has a maximum distance of one kilometer.

   As mentioned in Chapter 2.3, Google geocoding allows returning an approximation output in its metadata. The approximation indicates the quality of geocoding is low. Therefore, a Tweet that has an approximation output can be marked, and a request confirmation about the true location can be sent to the user.

4. Shelter selection

Shelters dataset obtained from the authoritative data. The dataset provides a large number of shelters plan. However, during the disaster, it does not ensure that all shelters are available. Therefore, the conceptual workflow is proposed the identification of the shelters using Tweets that categorized as credible Tweets. The credible means, the Tweets have relevant information and have close distance from its location content. The close distance implies is likely if flooding occurs in the location where the Tweet was sent.

In selection of the nearest shelter, shelters that has a possibility being inundated should be omitted from the selection. Since the Tweet is representing a flood location, then the possibility-inundated shelters can be identified based on the following assumption:

- If the shelter is close to the Tweet, then the more possible of the shelter is inundated.
- If there is another Tweet in the same location, then the more possible of the shelter is inundated.

After elimination of the possibility-inundated shelters, the nearest shelter can be determined by calculating the optimum path. Therefore, the direction to the shelter can be identified as well. Ultimately, it used as feedback information to the user.

The communication between citizens and relief agency in this conceptual workflow can be implemented in a real-time system. However, the concern of this research is more of data processing (filtering content, geocoding of location content, and identifying the nearest shelter). Therefore, a sufficient number of Tweets is required in this research. For this purpose, this research was used the Twitter data from the monsoon season 2013-2014.

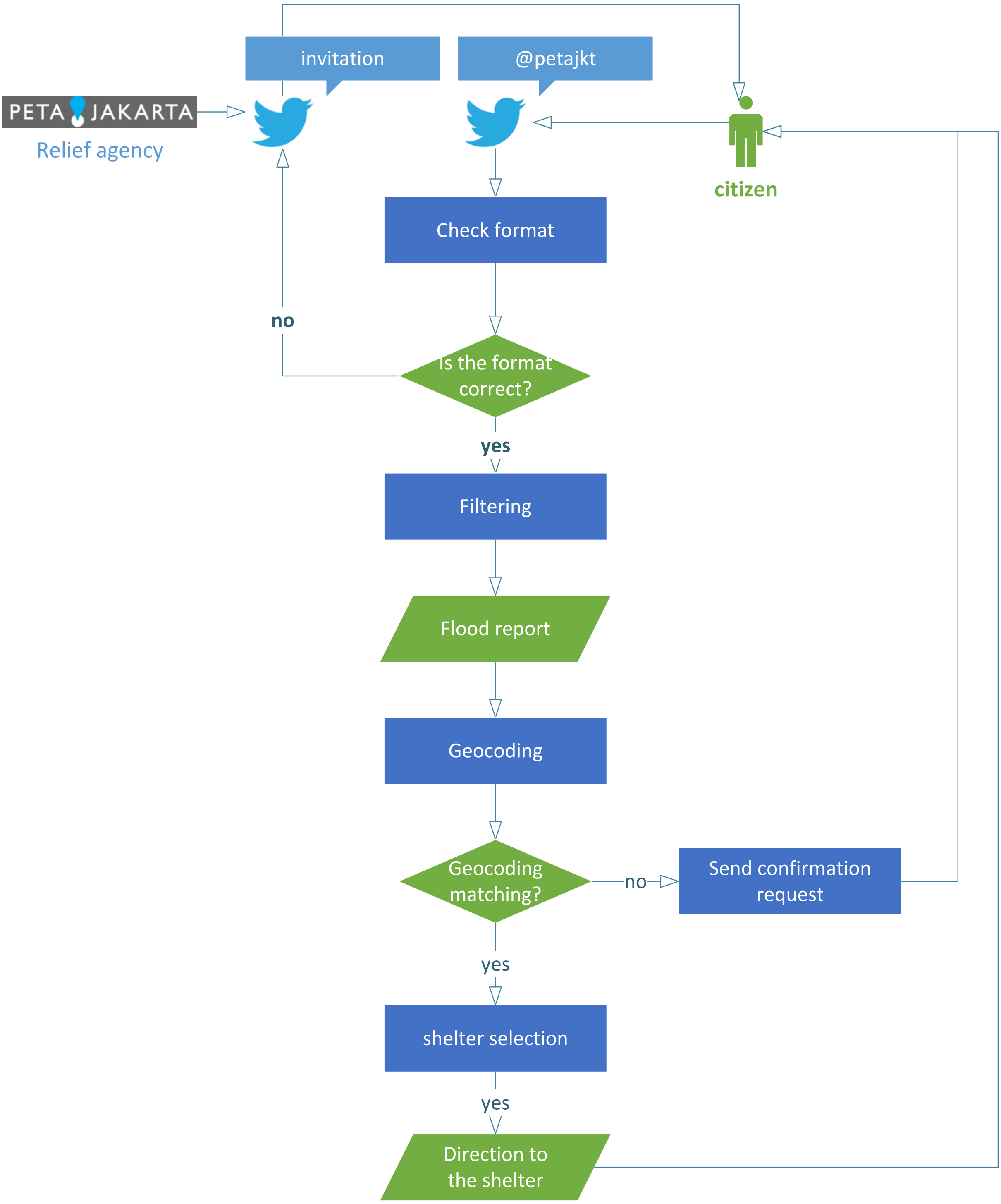


Figure 3.7: The conceptual workflow of utilizing geographical context of Twitter in flood disaster system

4. METHODOLOGY

Based on the conceptual workflow in Chapter 3, the filtering content, the geocoding of location content, and the identifying shelters are steps that proposed in the workflow. This chapter is explaining the methodology of each step. The three steps addressed a research question from the three last research objectives of the research.

4.1. Data

This research used several sets of data from various sources. First, Tweets from the monsoon season December 2014 to March 2015. The strategy for retrieving the Tweets was used a keyword of the official account of "Peta Jakarta" (@petajkt) and blocked the Tweets account of the official account. Second, Indonesian tagged corpus obtained from "https://github.com/famrashel/idn-tagged-corpus" as the research output by Dinakaramani et al., (2014). Third, administrative name of Jakarta city obtained from authoritative data. Fourth, street map obtained from OpenStreetMap. Fifth, Shelter map obtained from authoritative data. Jakarta capital government provides the shelter for spatial Plan 2014-2030, there are about 2563 shelters plan in the dataset.

Figure 4.1 shows the flow of the data retrieval. In the Figure, both filtering and selection used the street dataset. However, the filtering used only of the attribute name for the list of toponym where the shelter selection used both of the attribute and the spatial data.

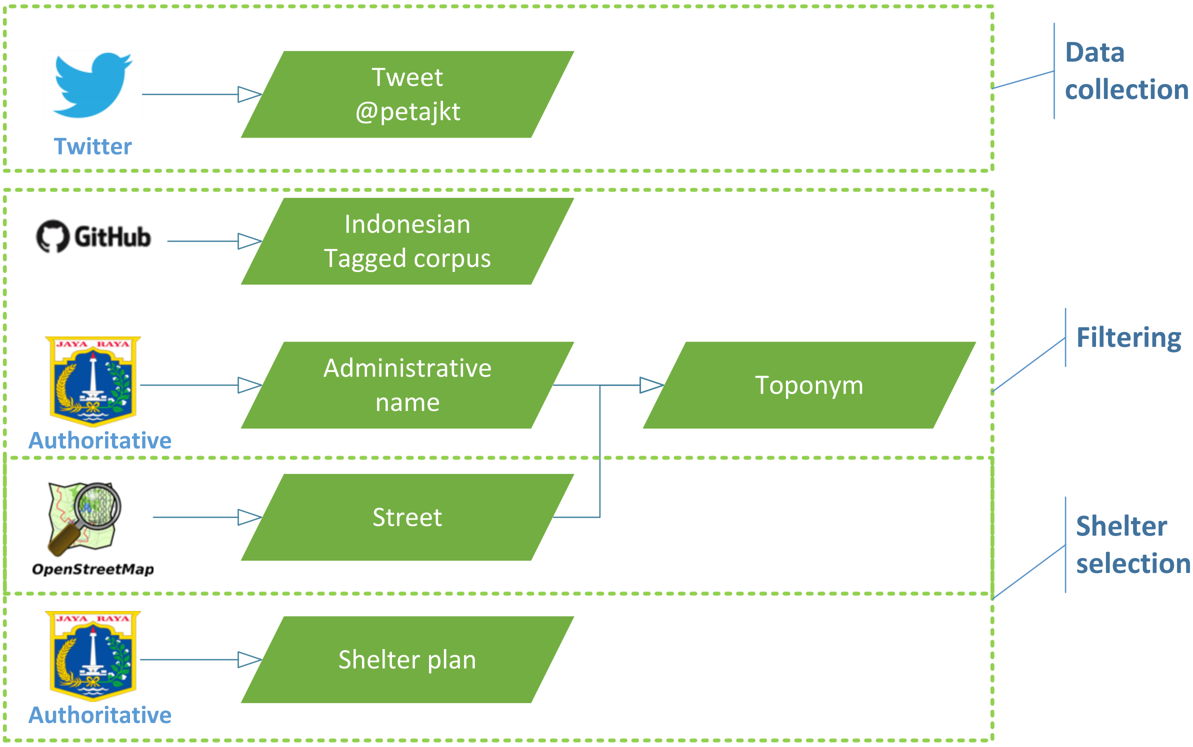


Figure 4.1: The flow of data retrieval

4.2. Tools Used

The system was developed using three mains platform. First, Python as the programming language, PostgreSQL as the database, and Application Programming Interface (API) as the services. Several python libraries were connecting to these platforms, and two extensions were embed in the database. Each process in the system was used a different tool as seen in Table 4.1. The connection of the tools shown in Figure 4.2. The output of the research was stored in CartoDb, an online visualization platform that allowed its users to store a spatial data in its platform and allowed its users to design the visualization.

Table 4.1: The list of tools that used in the system

Purpose	Tool Name	Explanation	Platform
Database	PostGIS	An extension for PostgreSQL to support a spatial data (PostGIS, n.d.)	PostgreSQL
	Psycopg2	A database adapter of PostgreSQL for the Python programming language (Python Software Foundation, 2015)	Python
Data collection	Twitter Search API	Part of the REST API (application programming interface) that authorized to find or to query a recent Tweet (Twitter.inc, 2015). The REST (Representational State Transfer) is a software architectural style for communicating trough HTTP (Hypertext Transfer Protocol) to request data (Rouse, 2014).	API
	Tweepy	Python library that allows accessing the Twitter API (Roesslein, 2009)	Python
Filtering	Nltk (Natural Language Toolkit)	A platform that works for human language data in python programming language (NLTK Project, 2015)	Python
Geocoding	Google Maps Geocoding	An API service for converting an address into geographic coordinates (geocoding) or vice versa (Google Developers, n.d.)	API
	OSM Nominatim	A tool to find an address from OpenStreetMap data. Currently, there is an API that used nominatim for geocoding (OpenStreetMap Wiki, 2016)	API
	Geopy	A client in Python programming language for geocoding using third-party geocoding services such as OpenStreetMap nominatim, Google Geocoding, ESRI ArcGis, Bing Maps, and so on (geopy, 2016)	Python
Shelter selection	pgRouting	A PostGIS/PostgreSQL extension that is providing functionality for geospatial routing (pgRouting Community, n.d.)	PostgreSQL

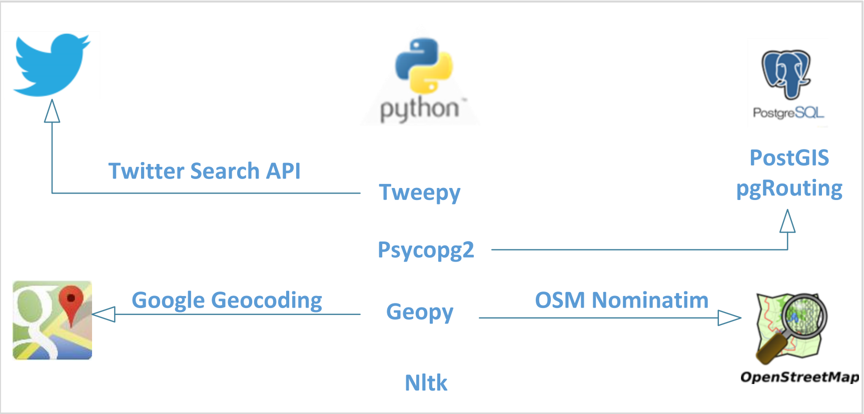


Figure 4.2: The connection of the tools

4.3. General Workflow

In general, the workflow consists of four processes; data collection, filtering, geocoding, and shelter selection. Data collection was designed to retrieve the Tweets and to store the result into a table in the database. Each Tweet in the table was processed by filtering to obtain the content (flood level and location content). Based on these, the Tweet can be identified whether it is relevant or off-topic. The geocoding process used to check the closeness between the primary geocoding and the secondary. Thus, the Tweet can be identified whether it is close to its location content or not.

The Tweet after the two processing stored into a new table called Tweets result. Afterward, for the Tweet that categorized as relevant Tweet and close to its location content was processed in the shelter selection. The output was a location of the nearest shelter and its direction to the shelter. The two output were used to update of the attributes of the Tweet. This way, the user of the Tweet obtained the information about the shelter because the user id is includes in the attributes. The general workflow can be seen in Figure 4.3.

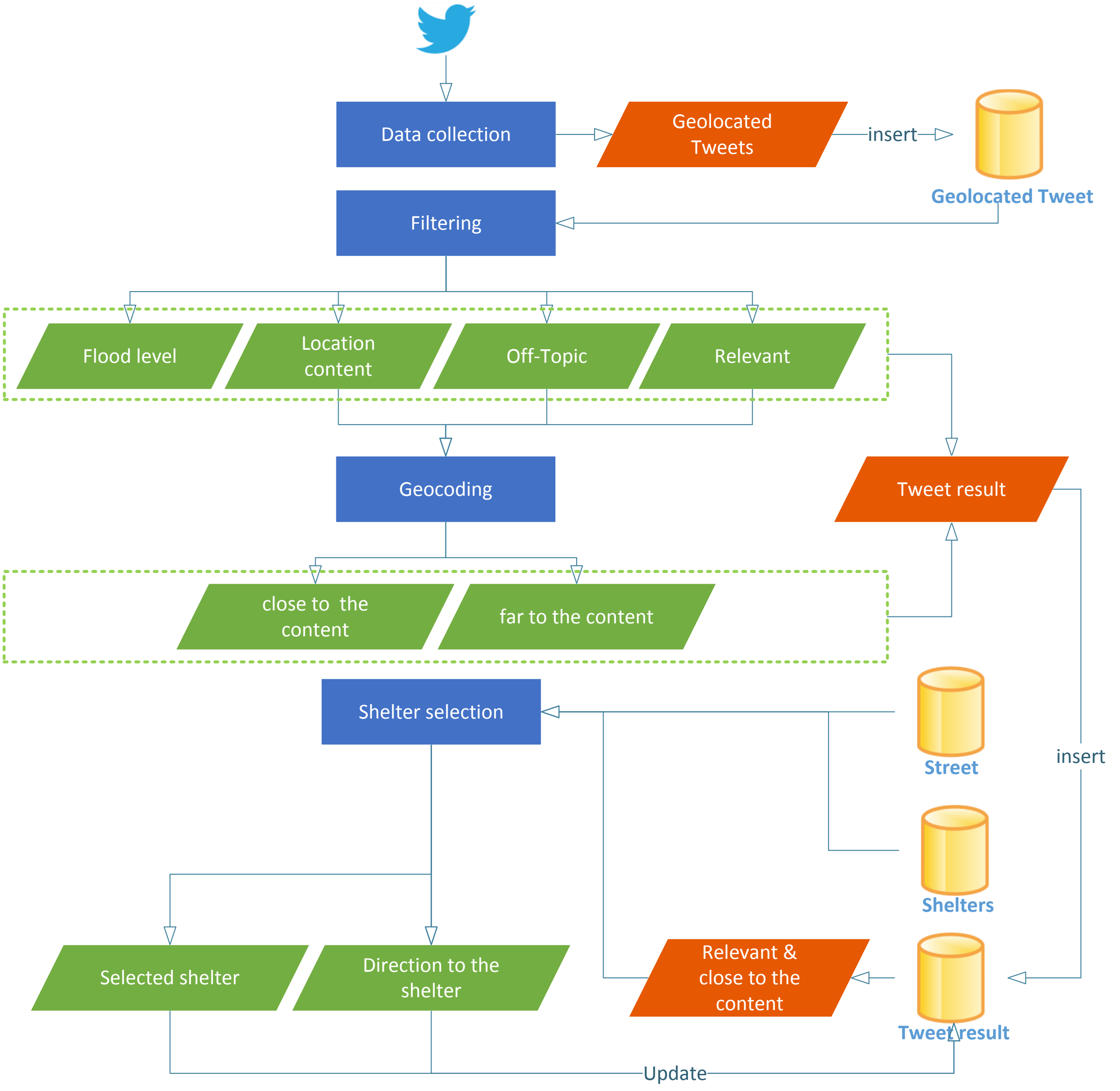


Figure 4.3: The general workflow of data collection, filtering, geocoding, and shelter selection

4.4. Data Collection

There is a time constraint of Twitter Search API that does not retrieve Tweets older than 7 days (Twitter.inc, 2016) whereas the required data is from last year's monsoon season. However, there is a possibility to search a Tweet without the API by using URL "https://twitter.com/i/search/timeline?" or "https://twitter.com/search?" these URLs allow searching a Tweet by adding a query after the question mark (?). This strategy makes it possible to retrieve a Tweet at the period. The different between both URLs are the first URL returning an HTML (Hyper Text Markup Language) page whereas the second URL returning the HTML page in a JSON file. The following URL is the strategy to retrieve the Tweets:

"https://twitter.com/search?f=realtime&q="since:2014-12-01 until:2015-03-31 @petajkt"&src=typd"

The HTML consists of tag elements inside an angle bracket (< >). It is enclosing Tweet attributes such as status id, text message, and date. The status id is stored in a tag element "data-tweet-id". To retrieve the status id in the HTML, a regex function was used as seen in the following code:

```
…
html = page.read().decode("utf-8")
pat = re.compile('data-tweet-id="(.+?)"')
status_id = pat.findall(html)
…
```

Unfortunately, the HTML does not have information about the geolocation of Tweet. Therefore, the next process was querying the geolocation and other attributes (text, date, username) via Tweepy library.

In general, the process to collect the Tweets consist third steps. First, retrieving an HTML tag element from URL's Twitter search to obtain the status id. Second, retrieving geolocation and remaining attributes using Twitter API. Third, inserting the Tweet into the database for the Tweet with geolocation. The flow of the process is in Figure 4.4.

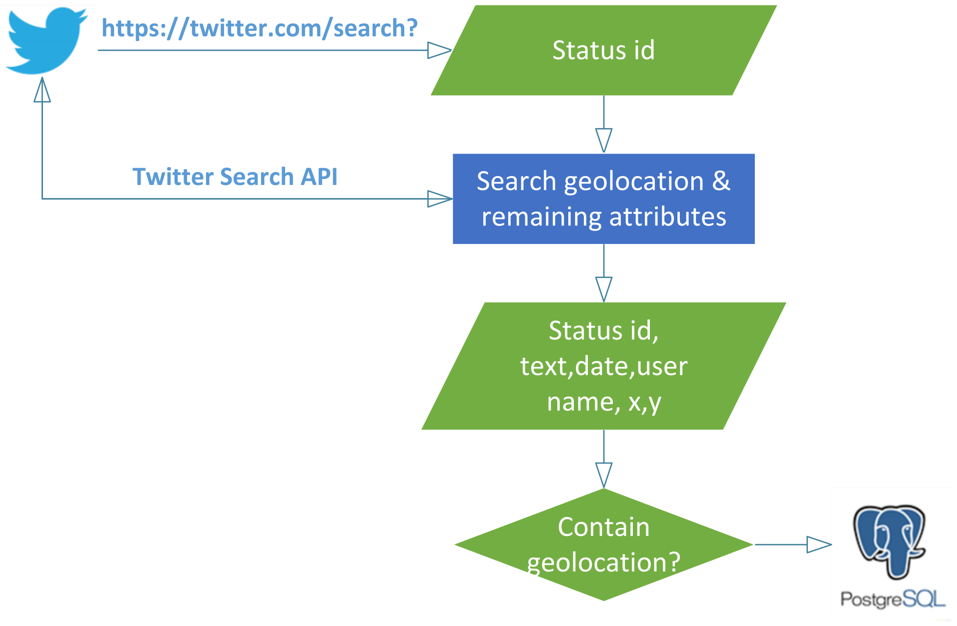

https://twitter.com/search?
Status id
Twitter Search API
Search geolocation & remaining attributes
Status id, text,date,user name, x,y
Contain geolocation?
PostgreSQL

Figure 4.4: The workflow of data collection

4.5. Filtering of Twitter Content

The first objective of the conceptual workflow is to filter the detail information of Twitter content. The information is location content and flood depth level. These contents are determining whether the Tweet is relevant or off-topic. As mentioned in the conceptual design, NLP approach can be implemented in this process. The explanation of the method in this section addressed the research question of the second research objective about which natural language processing method to use or modify in filtering the Twitter content

In NLP, a method that allows to extract entity in a set of text from a given query called Name Entity Recognition (NER) (Jung, 2012). The most common entities extracted are persons, organizations, and locations. In this case, location content is the entity that required being identified.

The strategy to acquire the entity could be different in every system. The main step in NER method is tokenization and Part Of Speech Tagging (POS). The objective of tokenization is to break the sentence into separate word where POS is to tag of each word based on its category. The tokenization and POS are part of the NLTK library that used in the system.

The filtering is divided into seven steps as seen in general workflow Figure 4.5, the explanation of each step is in the following sections.

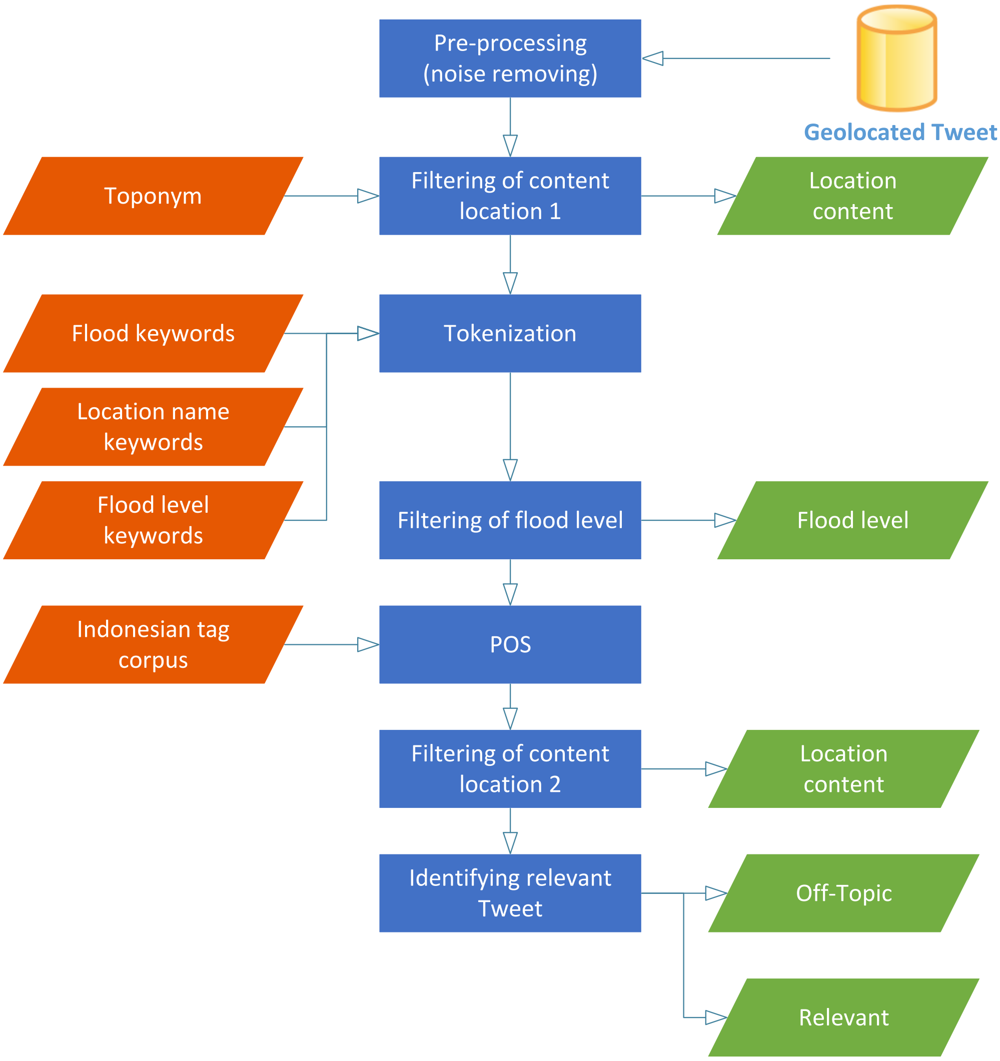


Figure 4.5: The workflow of filtering of Twitter content

4.5.1. Pre-processing

The objective of pre-processing is for normalization of Twitter content. In this system, the normalization has several steps as follow:

- Convert the Tweets into lower case
- Convert an invitation message into a tag "INVITATION."
- Convert a URL into a tag "URL."
- Convert a username with prefix @ into a tag "USER."
- Removing a hashtag
- Remove all punctuations

The invitation is occasionally mentioned in a Tweet if the user replies directly or retweet the invitation. The invitation is not the original message from the user hence it is necessary to be removed. The system used the regex expression in the process. The pre-processing is shown in the following code:

```
tweet = tweet.lower()
tweet = re.sub('((kena banjir?.*))','INVITATION',tweet)
tweet = re.sub('((www\.[^\s]+)|(https?://[^\s]+)|(pic\.[^\s]+))','URL',tweet)
tweet = re.sub('((rt@[^\s]+)|(@[^\s]+))','ATUSER',tweet)
tweet = re.sub(r'rt AT_USER', 'ATUSER', tweet)
         replace_punctuation = tweet.maketrans(string.punctuation, '
         '*len(string.punctuation))
tweet = tweet.translate(replace_punctuation)
tweet = re.sub('[\s]+', ' ', tweet)
tweet = re.sub(r'#([^\s]+)', r'\1', tweet)
```

The following Tweet is a sample of pre-processing:
Original Tweet:

@petajkt #banjir banjir sepaha orang dewasa, Jl. Satria 2 Jelambar JakBar pic.twitter.com/CmdYPd4A7l

The Tweet after noise removing:

ATUSER banjir banjir sepaha orang dewasa Jl Satria 2 Jelambar JakBar URL

4.5.2. Filtering of Location Content Using Toponym

To identify a location content the most straightforward way by using a list of toponym. As mentioned in Chapter 4.1 about Data, the list of toponym obtained from merging between a street name of OSM and administrative name. Then, it sorted per line from the longest name to the shortest name and it saved as a txt file.

To find a match string name in the list, the code read the text line by line using regex expression. If a string in a Tweet matched with a string in the list, then it converted into tag "LOCNAME" and the string name stored into a location name list. The following code is the filtering of location content using the list of toponym.

```
…
fileloc = open('location_name.txt')
cortloc = fileloc.read().splitlines()
…
result = ''
for address in cortloc:
    result = re.findall('\\b'+address+'\\b', tweet, flags=re.IGNORECASE)
    if len(result) > 0:
        tweet = tweet.replace(address, "LOCNAME")
        tweetrem = tweetrem.replace(address, "") #remaining words
        resnameloc.append(''.join(result)) #location name
result = None
```

The following Tweet is the sample of filtering of location content:
Before processing:

ATUSER banjir banjir sepaha orang dewasa Jl Satria 2 Jelambar JakBar URL

After processing:

ATUSER banjir banjir sepaha orang dewasa Jl Satria 2 LOCNAME JakBar URL

Since the location name could be contained multiple words, then this process needs to be done before the tokenization. It is important because tokenization aims to break the sentence (text) into separated word called tokens (Robin, 2009c).

4.5.3. Tokenization

The separated words after tokenization were used for several purposes. First, to identify a keyword related to flood. Second, to identify a keyword related to location name. Third, to identify a keyword related to flood depth level.

The keywords related to flood are *banjir* (flood), *air* (water), *genangan* (inundate), *meluap* (overflow), *terendam* (submerged). The keywords related to location name is a set of words that usually precede a location name, for examples; *jalan* (street), *wilayah* (area), *kampung* (village) and so on. The keywords of location name included also a set of list preposition which are; *di* (at), *depan* (in front of), *dekat* (near), *belakang* (back), *dekat* (next to). The abbreviation of keywords was included in the list as well.

A word related to a part of human body is commonly used to describe a flood depth level. For example; waist-high, thigh-high, and so on. These words were used as a set of keywords to identify a flood depth level. Later, these words were converted into a numeric number to obtain the value of water level.

Every string/word that intersection with the list was converted into a tag. The tag are; KEYFLOOD for keyword related flood, PREP for a keyword related to the location name, and DEEPVAL for a keyword related to flood depth level. The regex expression was used to search these words as seen in the following code:

```
…
#replace keyflood
keyflood = set(tweetlist).intersection(keyfloodlist)
for i in keyflood:
    tweet = tweet.replace(i, 'KEYFLOOD')
    tweetrem = tweetrem.replace(i,'') # remaining words
keyflood = None

#replace prepolist
keyprep = set(tweetlist).intersection(preploclist)
for i in keyprep:
    tweet = re.sub(r'\b'+i+r'\b', "PREP", tweet, 1)
    tweetrem = re.sub(r'\b'+i+r'\b', "PREP", tweetrem, 1)
keyprep = None
…
```

The following Tweet is the sample of identifying the keywords:
Before processing:

ATUSER banjir banjir sepaha orang dewasa Jl Satria 2 LOCNAME JakBar URL

After processing:

ATUSER banjir KEYFLOOD sepaha orang dewasa PREP Satria 2 LOCNAME JakBar URL

4.5.4. Filtering of Flood depth Level

There are two ways of a user sends an information about flood level. First, using a part of human body as explained before. For example:

@petajkt #banjir banjir ***sepaha*** *orang dewasa, Jl. Satria 2 Jelambar JakBar*

@petajkt #flood **thigh-high**, Jl. Satria 2 Jelambar JakBar

Second, using a metric unit of length (centimeter/cm). For example:

@BPBDJakarta @TMCPoldaMetro @petajkt jam 06:48 ketinggin air ***50cm*** *di jln danau sunter barat*

@BPBDJakarta @TMCPoldaMetro @petajkt at 06:48 the water level is **50cm** at jln danau sunter barat

In the tokenization, a word that intersection with the set of keywords related to flood level was tagged into DEEPVAL tag. Afterward, it converted into a numeric value. Since the keywords are a list of parts of human body words, then the flood level is an approximation value from the high of the parts of human body. The following Table 4.2 shows the conversion of the word to an approximate value.

Table 4.2: The conversion of flood depth level

Flood level (Indonesia)	Flood level (English)	Approximate value in cm
Sepinggang	waist-high	100
Sepaha	thigh-high	70
Selutut	knee-high	30
Semata kaki	ankle-high	10

The following code is the code for identifying flood depth level based on the conversion list

```
…
keydeep = set(tweetlist).intersection(deeplist)
deepval = None
deepkey = ''
for i in keydeep:
    tweet = re.sub(r'\b'+i+r'\b', "DEEPVAL", tweet, 1)
    tweetrem = re.sub(r'\b'+i+r'\b', "", tweetrem, 1)
    deepkey = i
        #translate deepvalue
    if deepkey == 'sepinggang':
        deepval = 100
    if deepkey == 'sepaha':
        deepval = 70
    if (deepkey == 'selutut') or (deepkey == 'lutut'):
        deepval = 30
    if (deepkey == 'semata') or (deepkey == 'mata') :
        deepval = 10
…
```

The regex expression was applied to convert the second type of flood level into a numeric value. The regex identified the numeric number after a metric unit of length (centimeter or meter) using a set of rule. Therefore, a list of metric units was required in this step included its abbreviation. If a metric unit of meter was used in a Tweet, then it converted into centimeter unit (multiplied by 100). The following code is the code for identifying flood level using a metric unit:

```
…
if word in deepunitlist:
    x = tweetremsplit[(tweetremsplit.index(word)) - 1]
    xx = re.findall(r'\d+', x)
    if len(xx) != 0:
        xx = xx[len(xx)-1]
        i = int(xx[0])
        if i != 0:
            deepval = xx
            tweet = tweet.replace(deepval + ' ' + word, 'DEEPVAL')
            if word == 'm' or word == 'meter':
                deepval = int(xx) * 100
            tweetrem = re.sub(r'\b'+word+r'\b', "", tweetrem, 1) # remaining words
    x = None
    xx = None
…
```

4.5.5. Part of Speech Tagging

A Tag is a descriptor that assigned to a given token (word). The process aiming to assign this descriptor to a part of speech (sentence) that called Part of Speech Tagging (POS) (Robin, 2009b). In this research, the collection of tag-set was obtained from the result of research by Dinakaramani et al. (2014). The corpus was used to tag (label) a word during the POS processing. The result of this process was a part of a sentence in a Tweet that converted into a tag. Table 4.3 shows the list of the Indonesian tag-set that exists in the corpus:

Table 4.3: The Indonesian tag-set (source: Dinakaramani et al., (2014))

Tag	Description
CC	Coordinate conjunction
CD	Cardinal Number
OD	Ordinal Number
DT	Determiner/Article
FW	Foreign Word
IN	Preposition
JJ	Adjective
MD	Modal and auxiliary verb
NEG	Negation
NN	Noun
NNP	Proper Noun
NND	Classifier, partitive, and measurement noun
PR	Demonstrative Pronoun
PRP	Personal Pronoun
RB	Adverb
RP	Particle
SC	Subordinating conjunction
SYM	Symbol
UH	Interjection
VB	Verb
WH	Question
X	Unknown
Z	Punctuation

Before POS processing, the remaining words were filtered using a list of stop words. The stop word referred to a list of common words that considered having no meaning. A word that intersection with the list was converted into a tag STOPWORD. The following code is the code to identify a stop word:

```
…
tweetremsplit = tweetrem.split() #split the remaining words
i = 0
for word in tweetremsplit:
    if word in stopwords.words('indonesia'):
        tweet = re.sub(r'\b'+word+r'\b', "STOPWORD", tweet, 1)
        tweetrem = re.sub(r'\b'+word+r'\b', "", tweetrem, 1)
…
```

Afterward, the POS is implemented as seen in the following code:

```
…
wordlists = PlaintextCorpusReader(corpus_root, '.*')
cortagg = wordlists.words(fileids='indonesia_tagg_new.txt')
…
if (word in cortagg) == True:
    inword = (cortagg).index(word)
    tagg = cortagg[inword + 2]
…
```

4.5.6. Filtering of Location Content Using a Rule Assignment

The first approach of identification a location content is using a list of toponym. However, there is a possibility a location name cannot be found in the list. Therefore, the location content can be identified using a rule assignment approach to overcome the problem.

In the tokenization step, there is a predefined keyword usually precede a location name. The keyword was converted as PREP tag. Then, a word after PREP tag can be identified as a location name and then it inserted into the list of location name.

There is a challenge for a location name consisting of more than one word because the second word and after might be failed to be identified. As seen in the following Tweet:

banjir sepaha orang dewasa di jl. budi mulia gunung sahari @petajkt #banjir

Thigh-high flood at jl. budi mulia gunung sahari @petajkt #banjir

"jl. budi mulia gunung sahari" is a set of the location name. The predefined keyword identified "budi" as a location name since it preceded by "jl" which is a member of the predefined keyword. Then, how to identify the rest of the words (mulia gunung sahari)?

The POS identified the remaining words as a Noun, the output was; PREP NN NN NN NN. Furthermore, Noun after the PREP tag was identified as a location name using a rule assignment. In another word, if there was a set of Noun after a PREP tag, then those words was identified as a set of the location name. Therefore, "jl. budi mulia gunung sahari" identified as a location name.

The used of unstructured text such as abbreviation caused a failure in the identification. A word caused by the failure identification called an unknown word. Unfortunately, the unknown word was often found as a location name. Therefore, if an unknown word was preceded by PREP tag, then it identified as a location name as seen in the following code:

```
…
tweetsplit = tweet.split() #remaining words
i = 0
for word in tweetsplit:
    if tweetsplit[i-1] == 'PREP' :
        resnameloc.append(''.join(word))
        tweet = tweet.replace(word, 'LOCNAME')
   i += 1
```

4.5.7. Identifying Relevant Tweet

The remaining step in the filtering is to determine whether the Tweet is relevant or off-topic. As explained in the conceptual design, a relevant Tweet is if it mentioned a location content or a flood depth level.

The output of previous steps is a set of tagging words resulted from the identification of keyword related flood, flood depth level, and location content. The following Tweet is an example of the steps:

Original Tweet:

@petajkt #banjir banjir sepaha orang dewasa, Jl. Satria 2 Jelambar JakBar pic.twitter.com/CmdYPd4A71

The Tweet after the rule assignment process:

ATUSER KEYFLOOD KEYFLOOD DEEPVAL NN RB LOCNAME 2 LOCNAME LOCNAME URL

The rule assignment to identify a relevant Tweet is "if the output contains a floods keyword and a location name, or if the output contains a floods keyword and a flood depth level, then it is classified as a relevant Tweet". Since the sample contains KEYFLOOD tag and DEEPVAL tag, then the Tweet was classified as a relevant Tweet. The rule assignment can be seen in the following code:

```
…
tweetsplit = tweet.split() #remaining words
resrelevant = None
if (('KEYFLOOD' in tweetsplit) and (('LOCNAME' in tweetsplit) or ('KEYHEIGHT' in
tweetsplit )) or ('DEEPVAL' in tweetsplit)):
    resrelevant = 'y'
else:
    resrelevant = 'n'
```

4.6. Geocoding of Location Content

The second objective of the conceptual workflow is to check the closeness between primary geocoding (geolocation of the Tweet) and secondary geocoding (geolocation of the location content). If the distance between both coordinates is less than one kilometer, then it considered as close to its location content. The explanation of the method in this section addressed the research question of the third research objective about how to assess the geocoding of Twitter and the geocoding of the location content.

The geocoding aims to convert location content obtained from the filtering process. The location content was sent to OSM nominatim and Google geocoding as the online geocoding services. As explained in the literature review, the output from OSM provides the best fitting compare to other services.

Before the services was used in the system, the experiment was conducted to compare between both services. The experiment was geocoded of three location names using the online geocoding comparator on "http://www.gisgraphy.com/compare/".

The first location name is "kampung baru , pondok pinang", the result is shown in Figure 4.6 and Table 4.4. The output shows that OSM is the best fitting to the street name "kampung baru". The second location name is "Plaza semanggi", the result is shown in Figure 4.7 and Table 4.5. The geocoding output of OSM is located at the correct location "plaza semanggi" but Google geocoding is located at "jalan jendral gatot subroto". The third location name is "pondok indah". The result is shown in Figure 4.8 and Table 4.6. It can be seen, the distance between the two outputs is quite far. Google geocoding returns an address "pd. indah" instead of "pondok indah" whereas OSM returns the correct address.

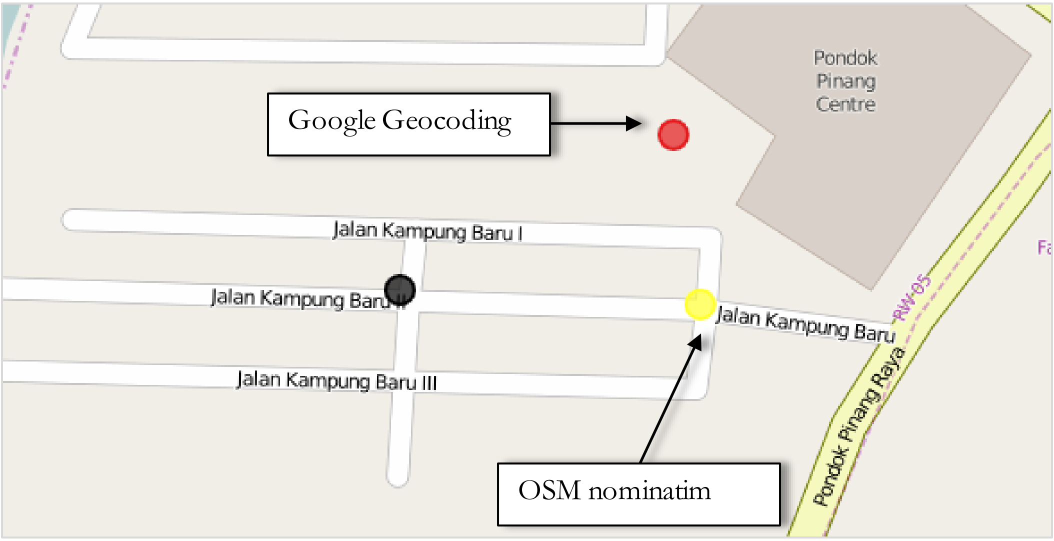


Figure 4.6: The output of first geocoding experiment

Table 4.4: The first comparison of geocoding output between OSM nominatim and Google Geocoding

Attribute	OSM nominatim	Google Geocoding
Address	Jalan Kampung Baru, Bintaro, Jakarta Special Capital Region, 12310, Indonesia	Jl. Kp. Baru, Kby. Lama, Kota Jakarta Selatan, Daerah Khusus Ibukota Jakarta 12310, Indonesia
Longitude	106.7730941	106.77303440000003
Latitude	-6.2744991	-6.2740124
Quality	residential	GEOMETRIC_CENTER

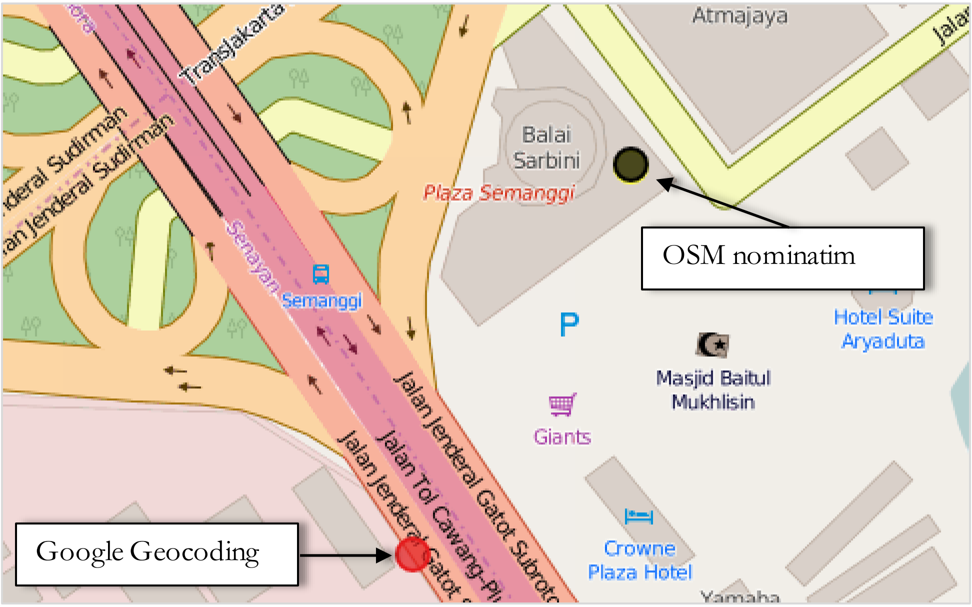


Figure 4.7: The output of second geocoding experiment

Table 4.5: The second comparison of geocoding output between OSM nominatim and Google Geocoding

Attribute	OSM nominatim	Google Geocoding
Address	Plaza Semanggi, Kav. 50, Jalan Jend. Sudirman, Gelora, Jakarta Special Capital Region, 12930, Indonesia	Plaza Semanggi, Senayan, Kebayoran Baru, South Jakarta City 12190, Indonesia
Longitude	106.814890563593	106.81374000000005
Latitude	-6.21980365	-6.22189
Quality	mall	APPROXIMATE

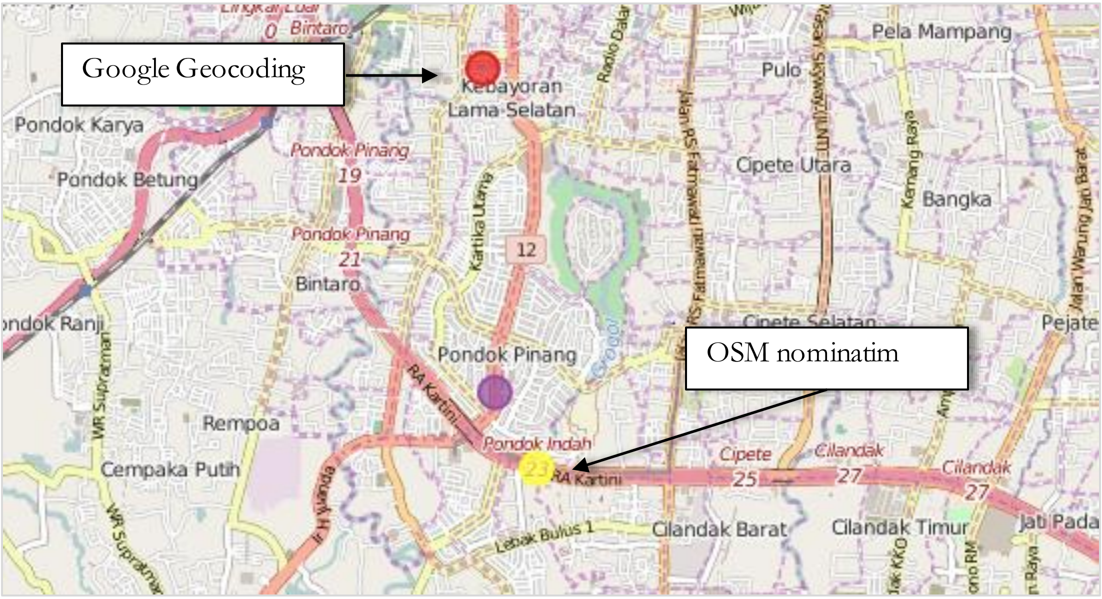


Figure 4.8: The output of third geocoding experiment

Table 4.6: The third comparison of geocoding output between OSM nominatim and Google Geocoding

Attribute	OSM	Google Geocoding
Address	Pondok Indah, Jalan Tol Lingkar Luar Jakarta, Cilandak Barat, Jakarta Special Capital Region, 12430, Indonesia	Pd. Indah, Jl. Tanah Kusir II, Kby. Lama, Kota Jakarta Selatan, Daerah Khusus Ibukota Jakarta 12240, Indonesia
Longitude	106.7838466	106.77919020000002
Latitude	-6.2915604	-6.2574396
Quality	motorway_junction	APPROXIMATE

Based on these experiments, OSM nominatim was used as the primary geocoding. If OSM failed to geocode a location name, then the location name was sent to Google geocoding as the secondary service.

The location name that sent to the geocoding services was the original name as mentioned in the Tweet. The several factors can fail the geocoding are the following points;

- the location name is not found in the database of the services
- the typo in writing the location name
- the ambiguity because more than one location is sent to the services

Google Geocoding provides the approximate attribute if the output is an approximation. Therefore, it was used to mark the Tweet that indicating the ambiguous location. The process of sending a location name to the services is shown in the following code:

```
geolocator_g = GoogleV3(api_key=g_api_key,domain='maps.googleapis.com',timeout=None)
#geocoding google
geolocator_y = Nominatim(timeout=None)#geocoding osm


…
if len(resnamelocnew) != 0: #location name
…
    location_geo = geolocator_y.geocode(resnamelocnew)

if location_geo is not None:
    resgeo_geo = location_geo.raw
…
    resgeoapi = 'osm'
    x_geo = resgeo_geo['lon']
    y_geo = resgeo_geo['lat']
else:
    resgeoapi = 'Google'

if resgeoapi == 'Google':
…
    location_geo = geolocator_g.geocode(query=resnamelocnew,bounds=(-5.90,107.10,-
6.37,106.53))

    if location_geo is not None:
        resgeo_geo = location_geo.raw
        …
            if resgeo_geo['geometry']['location_type'] != 'APPROXIMATE':
                resgeoapi = 'google'
                x_geo = resgeo_geo['geometry']['location']['lng']
                y_geo = resgeo_geo['geometry']['location']['lat']
            else:
                resgeoapi = 'APPROXIMATE'
        else:
            resgeoapi = 'fail'
    else:
        resgeoapi = 'fail'
```

Afterward, the distance between both geocoding can be calculated. If the distance is not far from one kilometer, then the Tweet is considered in the same area with its location content or close to its location content. The process is shown in the following code:

```
…
if resgeoapi != 'fail' and x geo is not None:
     xytweet = 'POINT(%s %s)' %(x_tweet, y_tweet)
     xygeo = 'POINT(%s %s)' %(x_geo,y_geo)
     cur = conn.cursor()
     cur.execute("""SELECT ST_Distance(ST_Transform(ST_GeomFromText(%s,4326),3857),\
ST_Transform(ST_GeomFromText(%s, 4326),3857));""", (xytweet,xygeo ))
     queryresults = cur.fetchall()

     if (len(queryresults) != 0):
         geodist = queryresults[0][0]
         if geodist > 1000: #1000 m
               resnearby = 'n'
         else:
               resnearby = 'y'
         cur.close()
     else:
         resnearby = None
```

The overall process of geocoding can be seen in the following Figure 5.9.

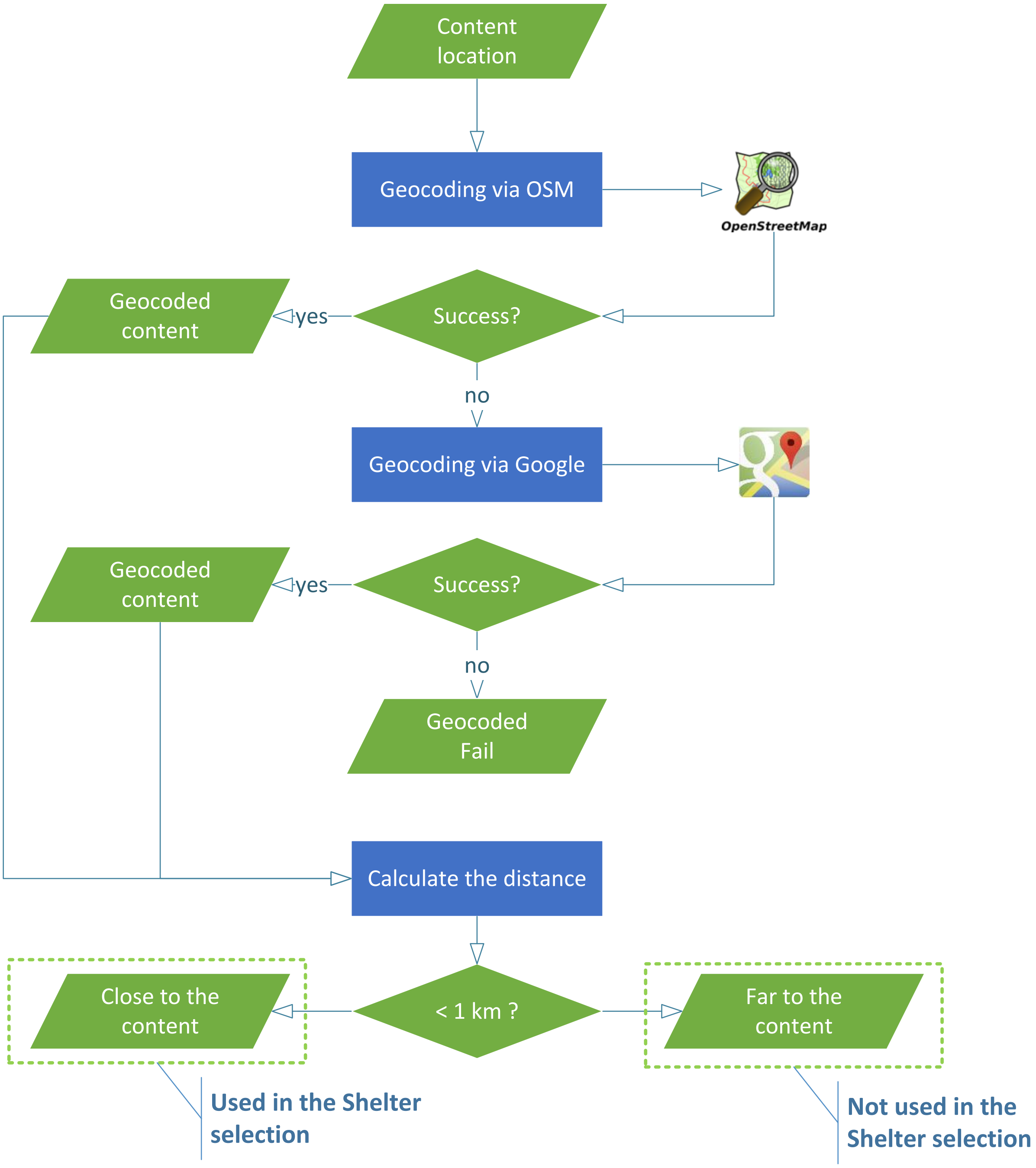


Figure 4.9: The workflow of geocoding

4.7. Evacuation Shelter Selection

The third objective of the conceptual workflow is to identify the nearest shelter from a Tweet that identified as relevant Tweet and close to its location content. The explanation of the method in this section addressed the research questions of the third research objective about what is the method for finding a route to the nearest shelter under flood disaster conditions and how can these directions be communicated back to the information seeker?.

The relevant Tweet was identified using the filtering approach and the close to its location content was identified using geocoding approach. Before the identification of shelter, a Tweet after passing the two steps (filtering and geocoding) was inserted into a new table. The Tweet has information whether it was contained location name, flood depth, close to its location content, as a relevant topic or not. Only relevant Tweet and close to its location content will be used in the selection process.

The first step of the shelter selection is eliminating shelter that has a possibility inundated. Based on the conceptual workflow, two conditions are proposed in the system. First, if the shelter is close to the Tweet, then the more possible of the shelter is inundated. Second, if there is another Tweet in the same location, then the more possible of the shelter is inundated.

Currently, this system used distance and scoring as the parameters. Based on the two parameters, the score was added to a shelter if it located close to the Tweet and if another Tweet found in the same area with the shelter. The scoring was calculated using the following rules:

- If there a Tweet within 100 meter radius from the shelters, the score is 0.5
- If there is a Tweet within 1000 meter radius from the shelters, the score is 0.25
- If there is another Tweet within 1000 meter radius from the shelters, the score is 0.25

The shelter dataset was updated every time a new Tweet came to the system. Therefore, the number of relevant Tweets affect to the number of the inundated shelters

The second step is identifying shelters that its location is located close to the Tweet (nearby shelters). The distance to the shelter was determined as two kilometers because a shelter in one kilometer considered as inundated. The distance will be extended to three kilometers if it failed to find a shelter within two kilometers.

The third step is identifying the optimum path from the Tweet to each nearby shelter. The identification used Dijkstra algorithm (Dijkstra, 1959) approach. The algorithm is available in the "pgrouting" extension of PostGIS. The algorithm determines the optimum path based on cost value of each path. The system used length path as the cost value. Therefore, the nearest shelter is a shelter that has the shortest path from the Tweet. The following points are the steps to identify the shortest path:

1. Assign the length of the path (road) as the weight cost.
2. Assign the nearest path from the Tweet as the source path (the road next to the Tweet)
3. Assign the nearest path from the shelters as the target path (the road next to each nearby shelter)
4. Find the least cost of possible path from the source to the target path

The output of the query using Dijkstra algorithm is the cost of the optimum path of each nearby shelter. Hence, the nearest shelter is determined based on the shelter that has the least cost compare to other shelters.

Figure 4.10 shows the example of a Tweet to obtain the nearest shelter. The blue dots are the shelters with a score more than zero or have a possibility inundated. The green dots are the shelter with a zero score or have not a possibility inundated. After calculating the optimum path to each green shelter, the least cost path identified as shown by the red route. This way, the nearest shelter is identified.

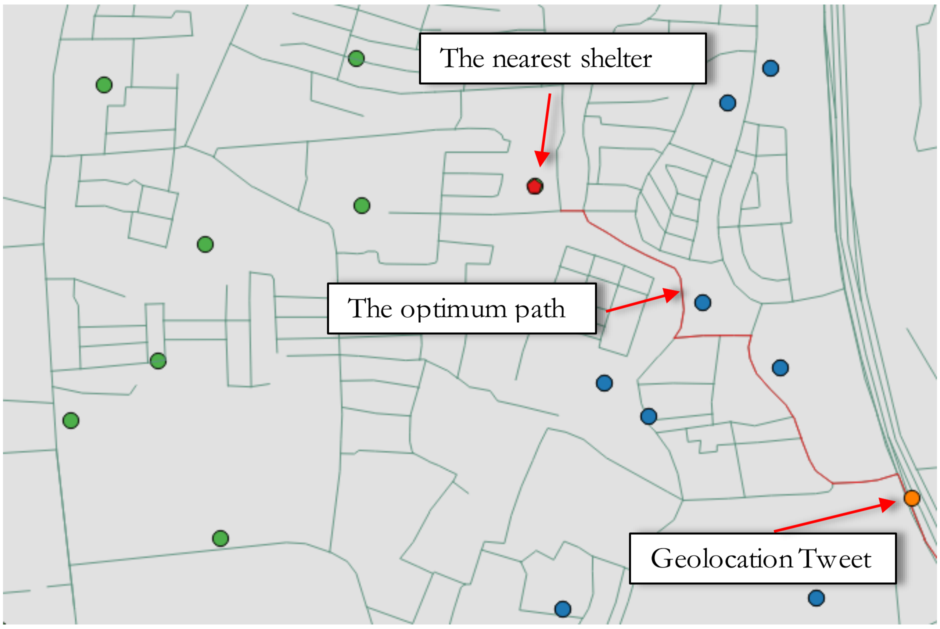


Figure 4.10: The example of identification the nearest shelter

The last step is identifying the direction to the nearest shelter. The query carries not only the cost of the optimum path but also the route of the optimum path. Since each path has an attribute name, then each name's path was inserted into a list of path name to obtain the direction.

Unfortunately, not all paths have an attribute name. Therefore, a query using the azimuth was calculated to overcome the problem. The azimuth is returning a clockwise north-based angle between two points (PostGis, n.d.). The points are the first second coordinates of the un-named path and the last second coordinates the previous path of the un-named path. The angle represents the compass degree as seen in Figure 4.11. Based on the degree, a point of the compass was identified based on a set of rule in Table 4.7.

Table 4.7: The rule assignment to identify the direction based on the azimuth

direction	Rule assignment
North	0^0 <= azimuth <= 20^0 or azimuth > 340^0
North East	20^0 < azimuth <= 70^0
East	70^0 < azimuth <= 110^0
South East	110^0 < azimuth <= 160^0
South	160^0 < azimuth <= 200^0
South West	200^0 < azimuth <= 250^0
West	250^0 < azimuth <= 290^0
North West	290^0 < azimuth <= 340^0

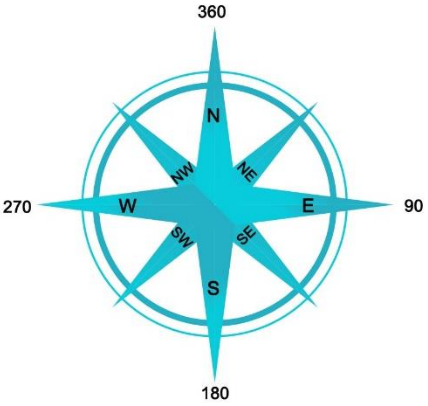


Figure 4.11: The compass degree

Figure 4.12 shows the example of identification a direction from a path named "jalan pecenongan" to an un-named path. The azimuth between the two coordinates is 279⁰. Based on the table, it is in the range of the west direction. Therefore, the direction is "…,go to jalan pecenongan, go to west,…". The information of the direction was used to update the table of Tweet result. Because the table contains attribute of the Twitter user then the information of the direction can be sent to the user.

It should be noted that the identification of shelters was conducted based on Twitter location. Therefore, the shelters was updated every time a new Tweet came to the system. Consequentially, if a user already received a direction to the shelter but later the shelter is not available anymore the updated direction should be sent to the user. The shelter is not available if another Tweet identified it as an inundated shelter. The workflow of evacuation shelter selection can be seen in Figure 4.13

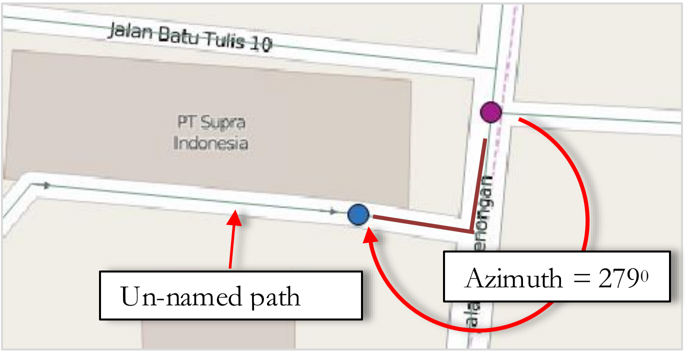


Figure 4.13: The example of identification a direction to an un-named path

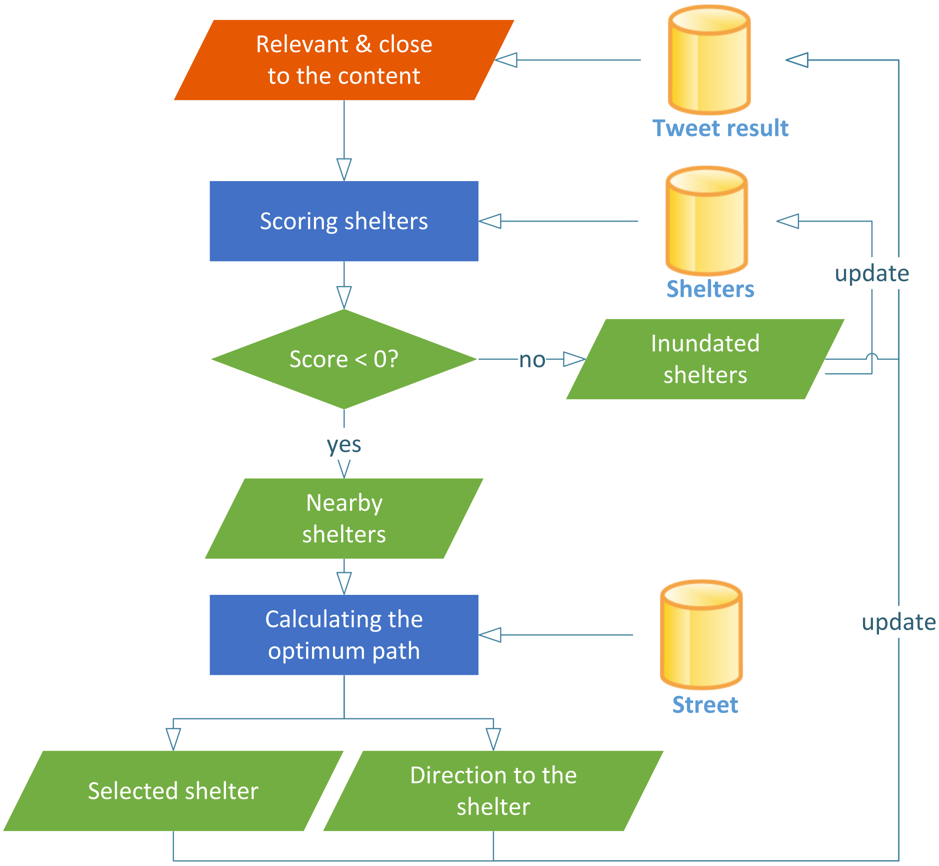


Figure 4.12: The workflow of evacuation shelter selection

5. RESULTS

This chapter addressed the research question of research objective number two and number three. The research question is about; what is the evaluation result of filtering the Twitter content and what is the evaluation result of assessing the geocoding of Twitter and the geocoding of the location content. The first research question was answered in the filtering of twitter content section and the second one was answered in the geocoding section.

5.1. Data Collection

The system was collected 5281 Tweets from December 2014 to March 2015 that contain a keyword @petajt. It was included Tweets with geolocation and without geolocation. Since this research considers only Tweets with geolocation and Tweets with located in Jakarta, the number of Tweets was decreased based on those criteria. As seen in Figure 5.1, from the total Tweets, only 26% Tweets that contain geolocation and 22% Tweets that contain geolocation and within Jakarta boundary. This mean, the number of participants that understand of the flood report format is relatively small.

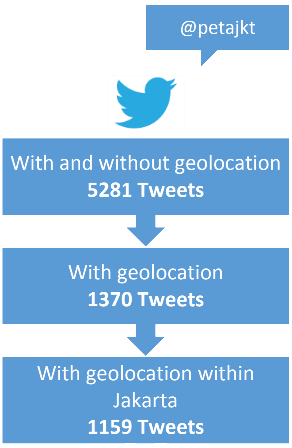


Figure 5.1: The output of data collection

Figure 5.2 shows the activity of Tweets during corresponding period per week. The peak time of the activity occurred in February 2015. The rise of the activity is coherence with the peak of monsoon season which were represented at the great number of flood refugees about 41,201 people in the same month in Jakarta (BPBD Jakarta, 2015).

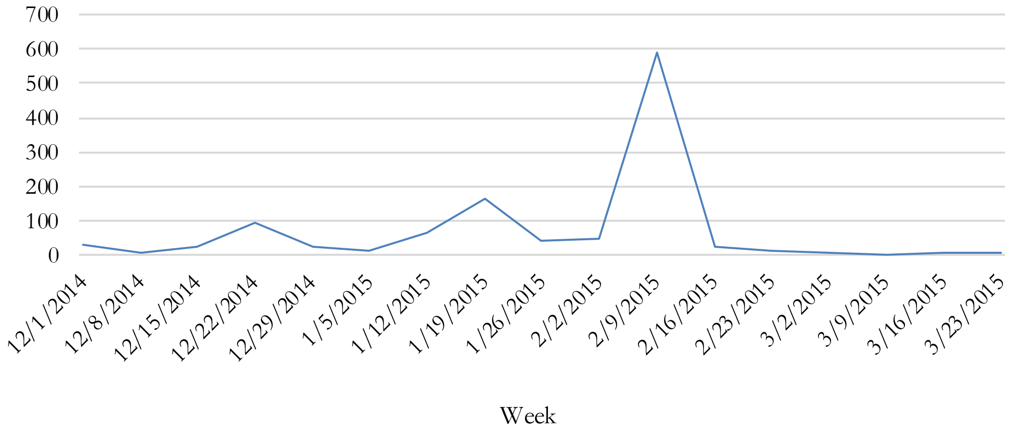


Figure 5.2: The line chart of Tweets activity per week

5.2. Authoritative Data for Comparison

To see the spatial pattern, the comparison between flood reports derived from this system and flood area from the authoritative data was conducted. The data is a flood area map in 2015 derived from SDI (Spatial data infrastructure) of Jakarta relief agency (GIS BPBD DKI Jakarta, 2015) that can be seen in Figure 5.3.

The flood area is determined based on the report from a leader of a neighborhood. The neighborhood is the lowest hierarchy of administrative area in Jakarta. The average area of the neighborhood is around 2 Km. If a flood occurs at somewhere in a neighborhood and if it reported by the leader, then that neighborhood is marked in the authoritative data as flood area. Since it obtained from the aggregation of neighborhood boundary, it leads to the uncertainty of the real flood boundary.

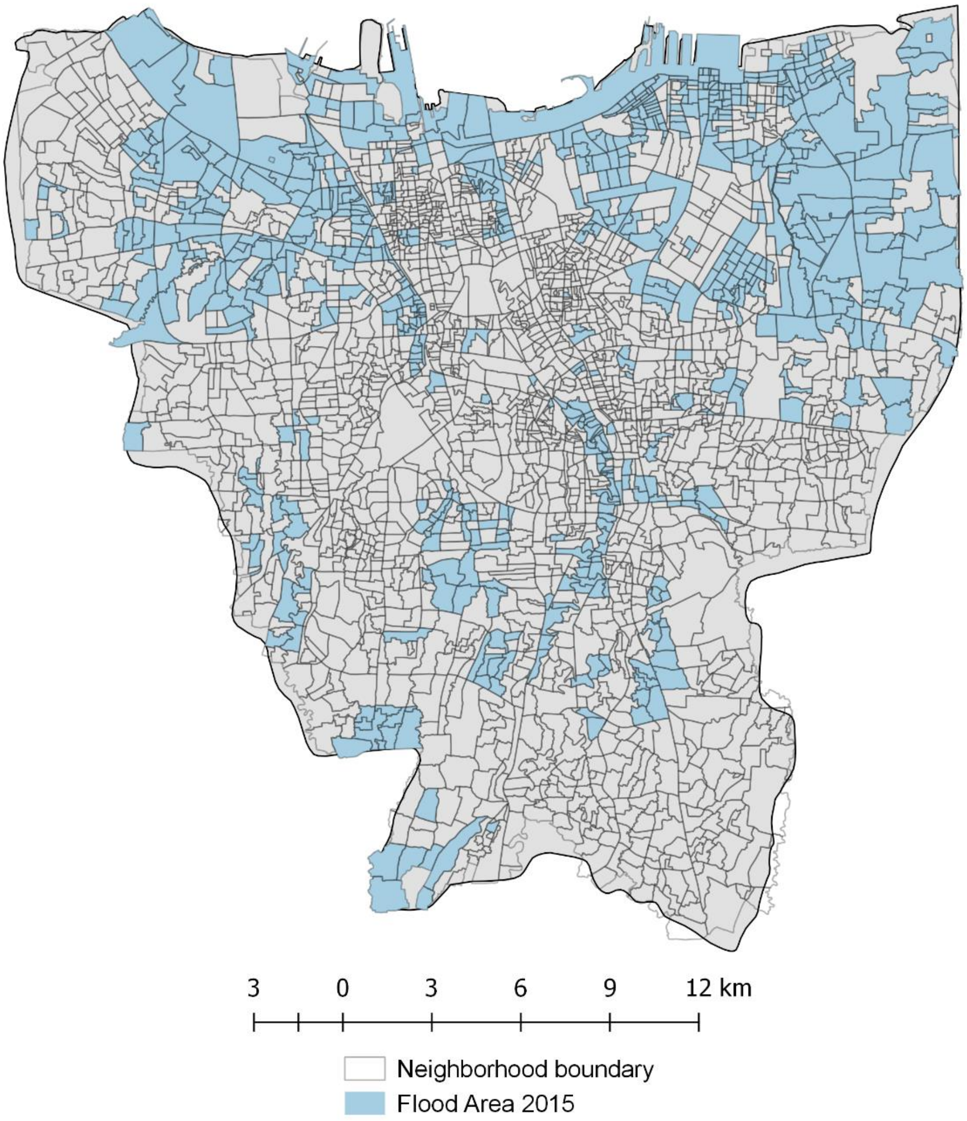


Figure 5.3: The flood area in 2015

5.3. Filtering of Twitter Content

The filtering aims to extract the detail information of location content and flood level information in every Tweet. If a location content or a flood depth level mentioned in a Tweet together with a flood keyword, then it classified as a relevant Tweet.

As shown in Figure 5.4 (Appendix Table 1), the total number of Tweets between relevant and off-topic is slightly difference. About 57% Tweets from the total Tweets contain a location content and about 73% of these Tweets categorized as relevant Tweets. The location content was possibly mentioned in the off-topic Tweets even it was only a fairly a small number compare with relevant Tweets, it was categorized as an off-topic Tweets since the system cannot recognize a flood keywords in the Tweets.

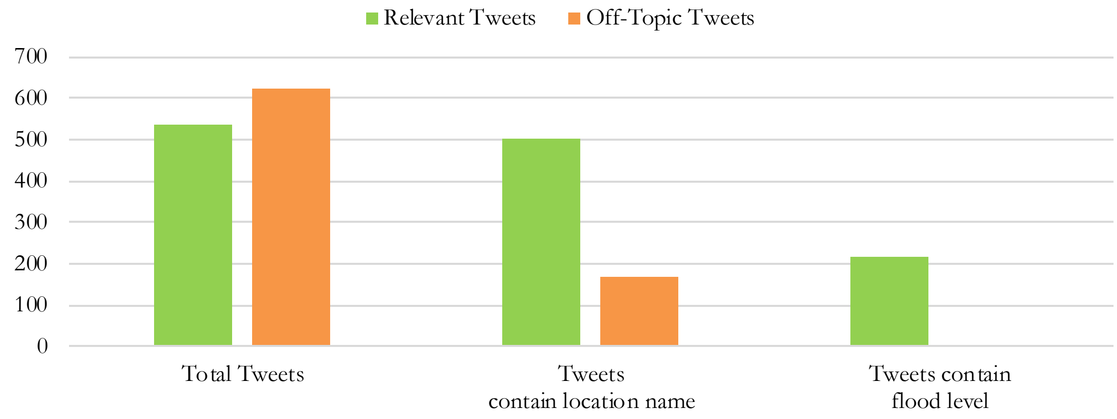


Figure 5.4: The comparison of relevant Tweets and off-topic Tweets based on filtering category

On the other hand, a few number of Tweets contain flood level information and all the Tweets were recognized as relevant Tweets. Figure 5.5 shows the comparison between flood level from the Tweet content and flood level from the authoritative data. It can be seen, the Tweets are clustered in where the flood has occurred. Then, when it looked in more detail as seen in Table 5.1, most of the Tweets are located in the flood area but only 30% the number of Tweets that matched with each category. 50 cm used as the break value considering the average number of flood level in both datasets.

It should be noted that the flood area does not portray flood boundary but represents flooded neighborhood. Therefore, if there is a difference of flood level at several locations in a neighborhood, it cannot be shown in that dataset. It can be said that flood level of the neighborhood is only represented flood level of one location.

Table 5.1: Flood level results and comparison

Category	Flood Level Tweets	Inside Flood Area <50cm	Inside Flood Area >50cm	Outside Flood Area
Tweets < 50 cm	**150**	**45**	55	50
Tweets > 50 cm	**96**	28	**35**	33

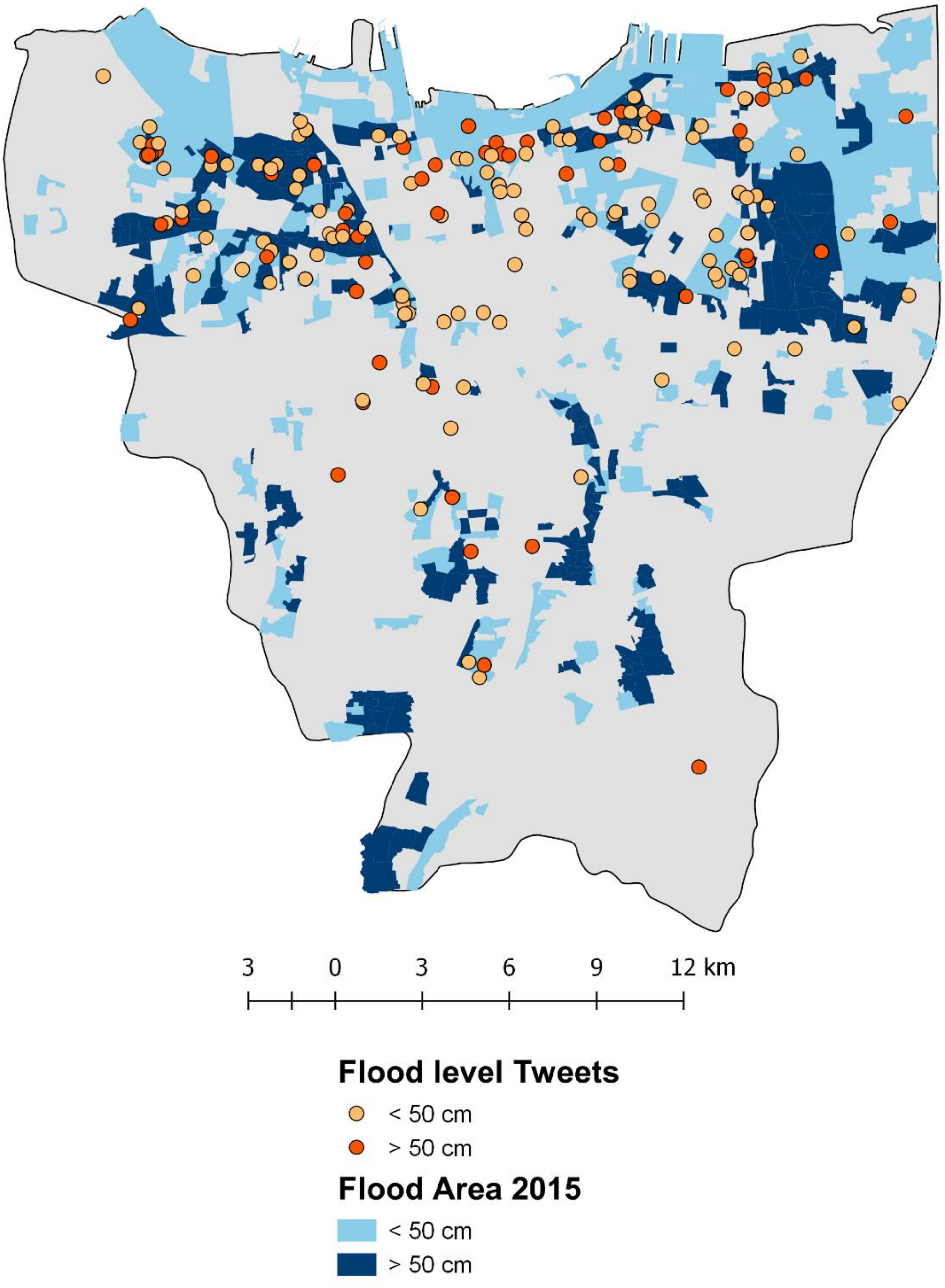


Figure 5.5: The comparison between flood level from Tweet content and flood level from the authoritative data

The same experiment was conducted to compare the relevant Tweets towards the authoritative data, as shown in Table 5.2. Overall, the number of relevant Tweets are more located in the flood area. Whereas, the number of off-topic Tweets are more located outside of flood area. As seen in Figure 5.6, the spatial pattern of relevant Tweets is apparently clustered around the flood area. On the other hand, the off-topic Tweets are more dispersed that can be seen in Figure 5.7.

Table 5.2: Relevant Tweets, Off-topic Tweets and Comparison

Category	Total Tweets	Flood Area	
		Inside	Outside
Relevant Tweets	534	305	229
Off-Topic Tweets	625	241	384

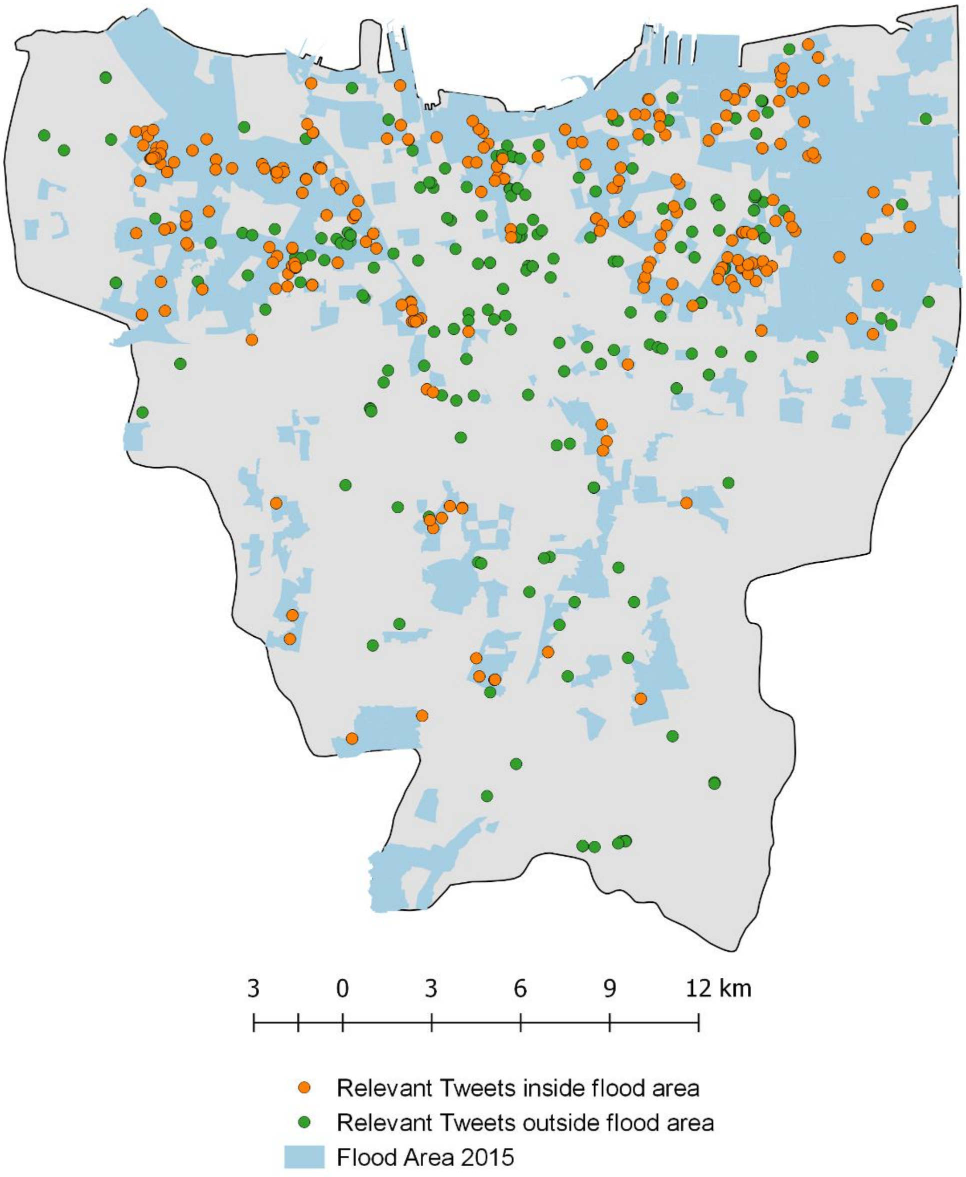

3
0
3
6
9
12 km
● Relevant Tweets inside flood area
● Relevant Tweets outside flood area
▨ Flood Area 2015

Figure 5.6: The distribution of Relevant Tweets

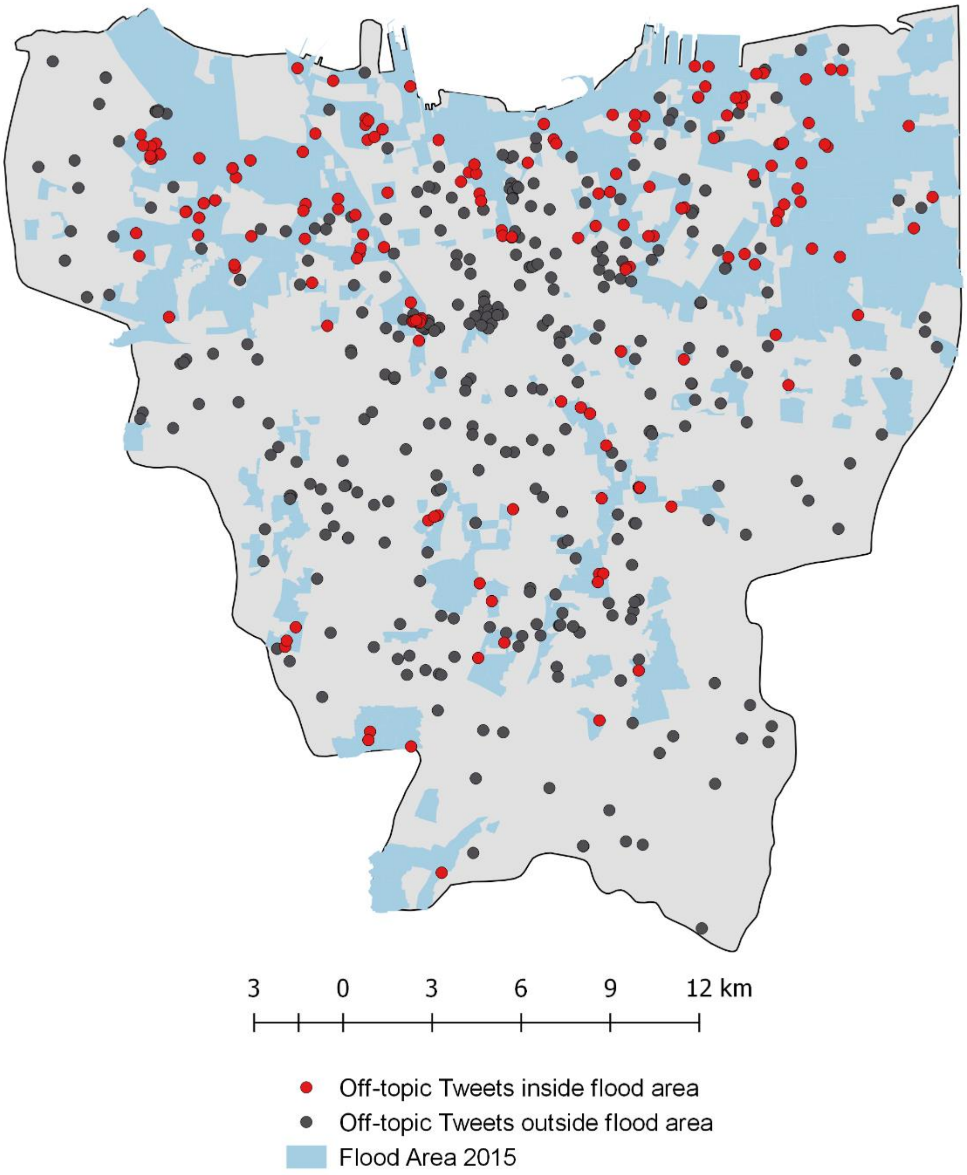

3
0
3
6
9
12 km
● Off-topic Tweets inside flood area
● Off-topic Tweets outside flood area
Flood Area 2015

Figure 5.7: The distribution of off-topic Tweets

Furthermore, the Tweets are grouped by month to analyze the trend of Tweet activity over time (Figure 5.8 and Appendix Table 2). In every month, there are more off-topic Tweets than relevant Tweets, except in the peak monsoon season (February 2015). As known before, the Tweet invitation sent by "Peta Jakarta" (@petajkt) from December 2014. Most of the Tweets during December until January are about reply the invitation or noise content. The examples are shown in Figure 5.9 and 5.10 respectively

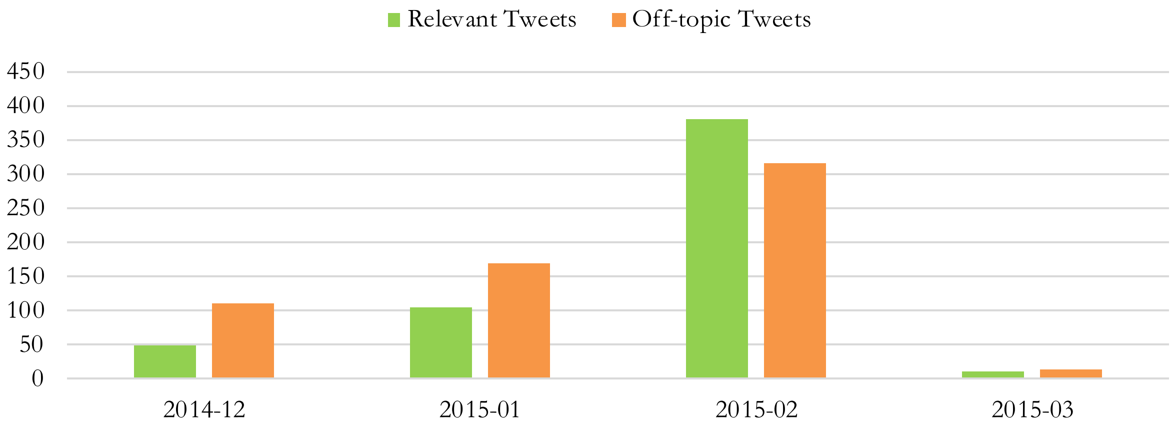
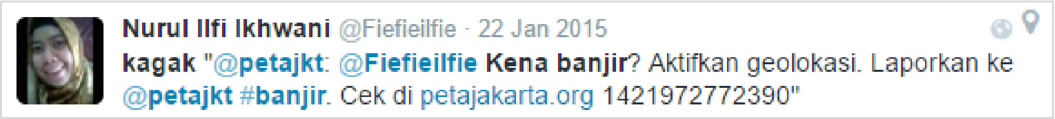


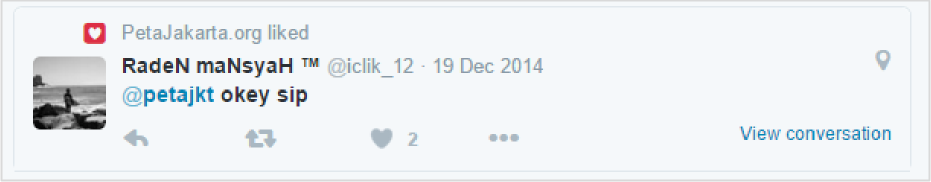
Figure 5.8: Relevant Tweets and Off-topic Tweets per month


Nurul Ilfi Ikhwani @Fiefieilfie · 22 Jan 2015

kagak "@petajkt: @Fiefieilfie Kena banjir? Aktifkan geolokasi. Laporkan ke @petajkt #banjir. Cek di petajakarta.org 1421972772390"

No @petajkt; @Fiefieilfie hit by flood? Activate the geolocation. Send a report to @petajkt #banjir. Check at petajakarta.go.id

Figure 5.9: The example of a reply to the invitation


PetaJakarta.org liked

RadeN maNsyaH ™ @iclik_12 · 19 Dec 2014

@petajkt okey sip

@petajkt okay sip

Figure 5.10: The example of a noise content

5.4. Geocoding of Location Content

The purpose of geocoding in this system is to check the closeness between primary geocoding (coordinates of the Tweet) and secondary geocoding (coordinates of the content-location). This process runs just if the location content recognized by the system.

Tweets mentioned a content-location are about 670 Tweets from the total 1159 Tweets. As seen in Figure 5.11 (Appendix Table 3), from the 670 Tweets, there are about 82% location content Tweets that successfully geocoded and 17% failed in the process. Out of all the successfully geocoded, only 40 % Tweets which the primary geocoding and the secondary geocoding are close to each other.

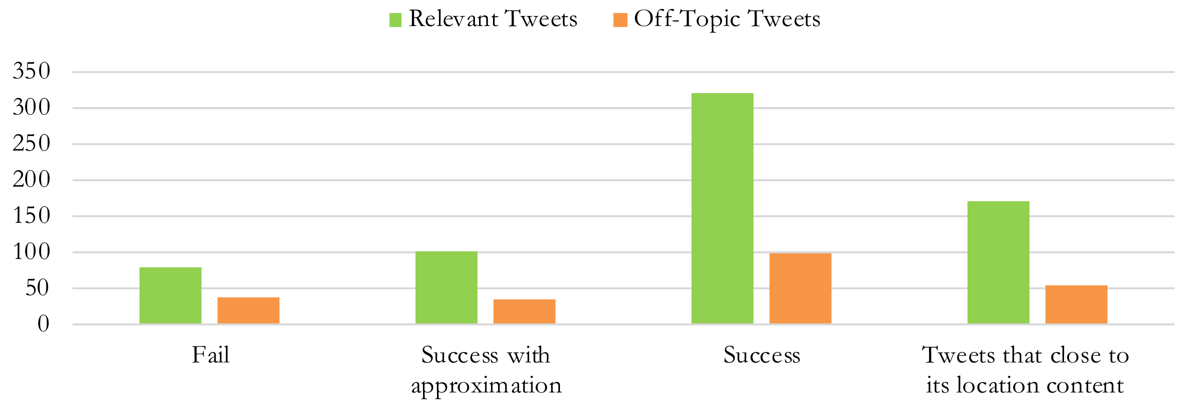


Figure 5.11: The results of Geocoding content

Google geocoding has the ability to return an approximation result of the location name. Out of all the successfully geocoded, there are about 12% content-location Tweets that return the approximation output. Figure 5.12 portray the distribution of Tweets based on geocoding services. The spatial pattern of the Tweets is clustered around the flood area.

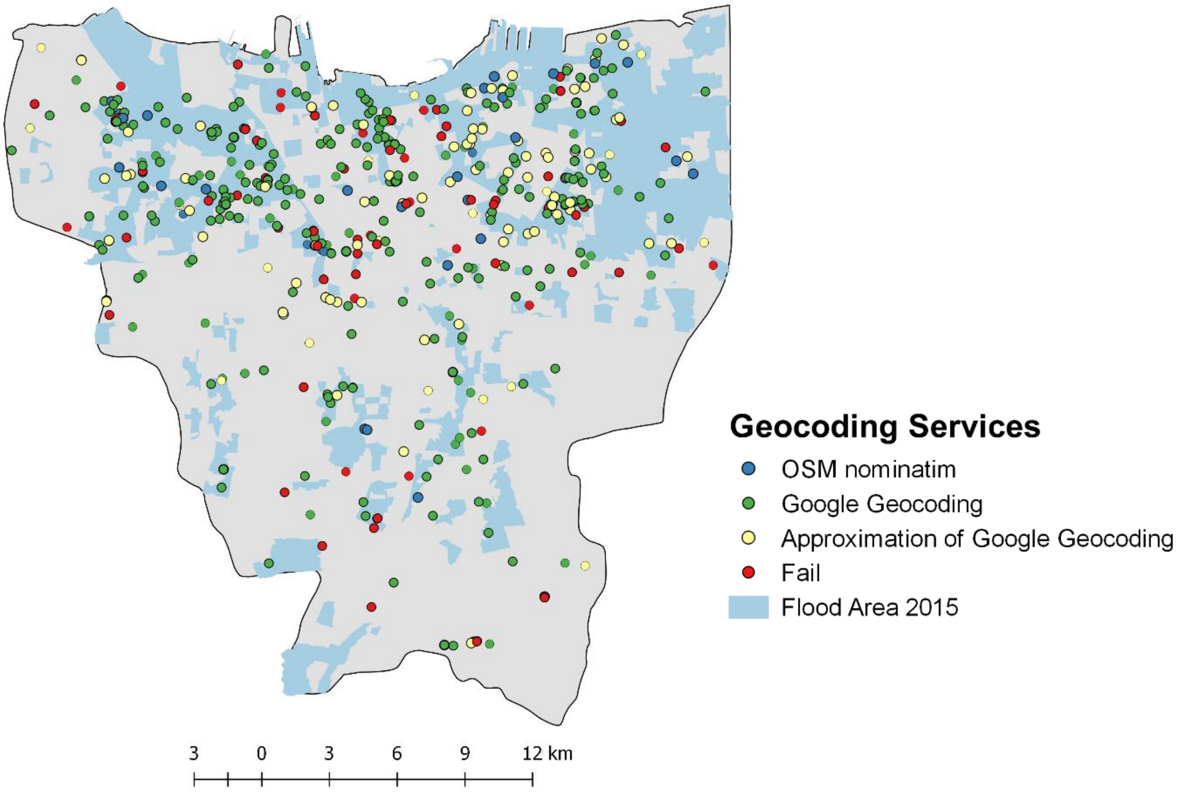


Figure 5.12: The distribution of Tweets based on the geocoding services

Furthermore, to see the spatial pattern, Tweets that have the closeness between primary geocoding and secondary geocoding was compared with the authoritative data (Table 5.3). The Tweets were grouped into two categories; Tweets that geocoding of primary and secondary is close and relevant Tweets that geocoding of primary and secondary is close. The distribution of Tweets in both categories apparently more clustered inside of the flood area.

Even though there are some Tweets located on the outside of flood area, Figure 5.13 shows the distribution of the Tweets is clustered around the flood area. There are about 68% of relevant Tweets located in the flood area. Apparently, adding the closeness of the location content does not have any significant impact. However, when the flood boundary is expanded up to 200 m, the number of the Tweets inside the flood area is increased to 148 Tweets or 87% are clustered around the flood area.

Table 5.3: Tweets that dose to its location content, relevant Tweets that dose to its location content and Comparison

Category	Total Tweets	Flood Area	
		Inside	Outside
Geocoding of primary and secondary is close	224	153	71
Geocoding of primary and secondary is close & the content is relevant	169	109	60

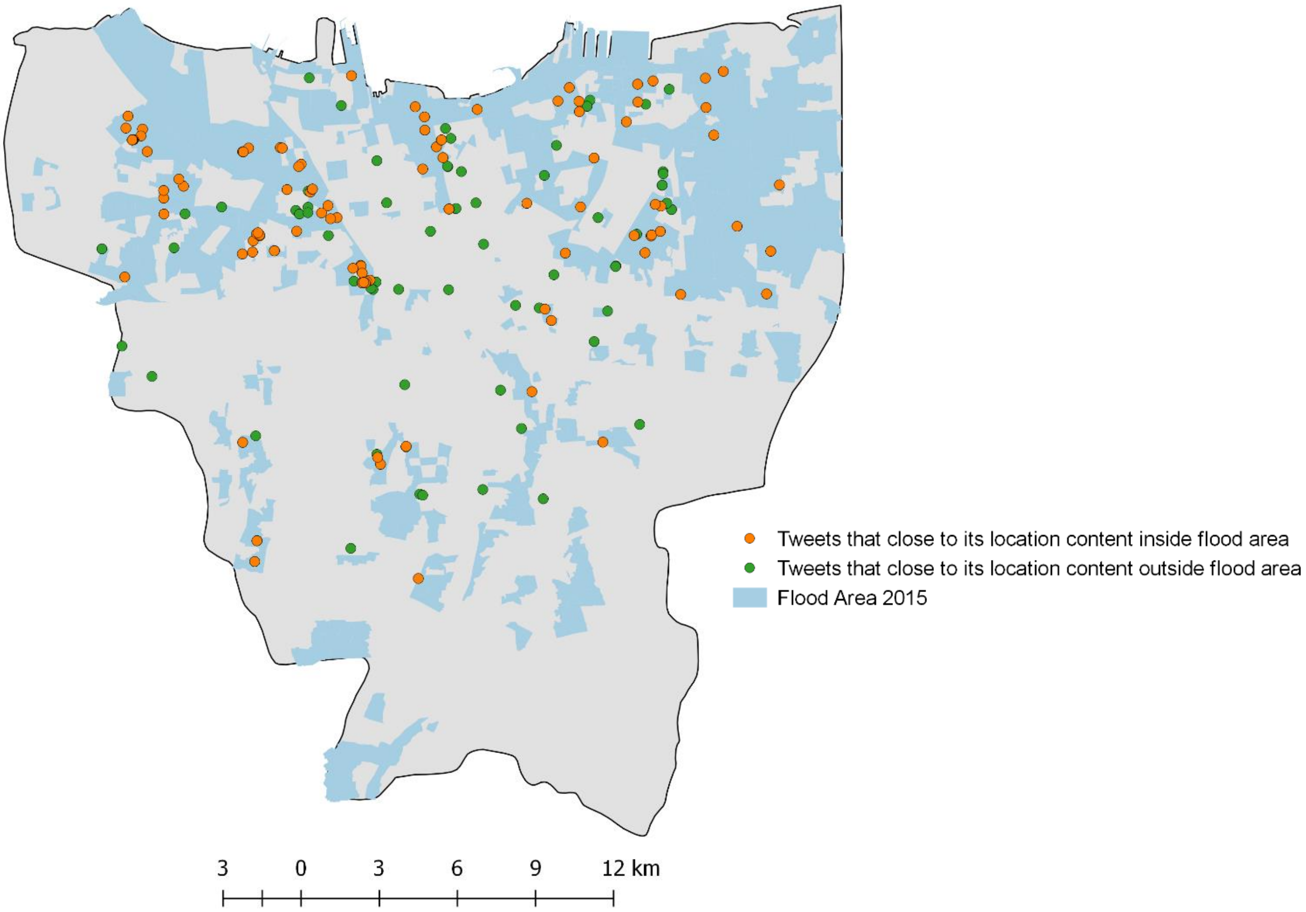

● Tweets that close to its location content inside flood area
● Tweets that close to its location content outside flood area
▨ Flood Area 2015
3
0
3
6
9
12 km

Figure 5.13: The distribution of Tweets that dose to its location content

5.5. Evacuation Shelter Selection

The last process of the system is to identify the potential shelter from the location of Tweet. As mentioned in the Chapter 4.1, the shelters dataset was obtained from the authoritative data. There are about 2563 shelters available in the dataset. The system allows identifying the shelters that might be inundated using a scoring system. If a shelter was close to a relevant Tweet, this increase the probability of being flooded. The system identified 1520 shelters that have no possibility inundated, and 81 shelters that was selected for the nearest shelter.

The shelter was identified from the Tweets that relevant and close to its location content. In the other word, if users mentioned a location name in their Tweet and its location in the same area where they were sending the Tweet, then the users get an information about the closest shelter, as long as the location is not too far or still reachable by walking distance. Table 5.4 shows only 1 Tweet from all Tweets that relevant and close to its location content failed to being matched to the closest shelter.

Table 5.4: Shelters selection of Tweets that relevant and close to its location content

Category	Number of Tweet
Relevant & close to its location content	169
Success	168
Fail	1

The spatial pattern of the selected shelters is shown in Figure 5.14. Most of the shelters are dispersed on the outside of flood area. When it was calculated by a spatial query, from 81 of selected shelters there are 24 shelters or about 29 % shelters inside of flood area. Since the flood area is the aggregation of neighborhood boundary and accumulated flood area in 2015, there is a possibility the number of shelters that located in flood area is smaller than 24 shelters. The selected shelters were determined based on the optimum path from the Tweet location. Therefore, the route to the shelter can be identified.

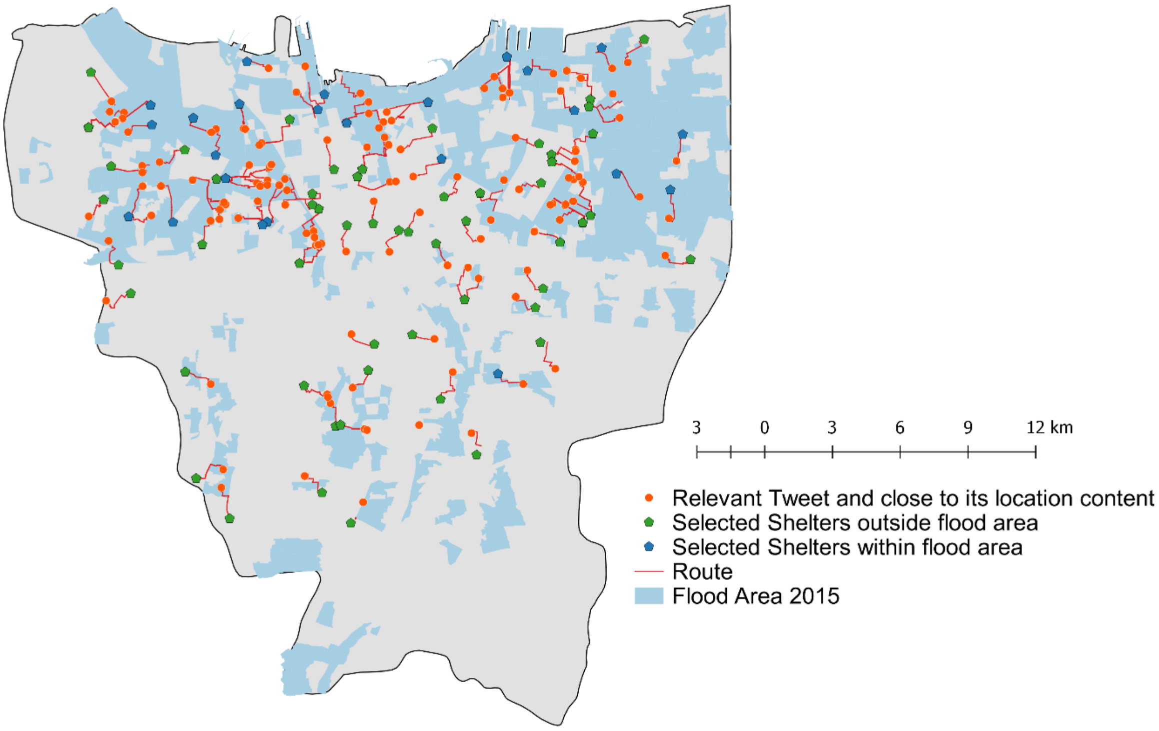
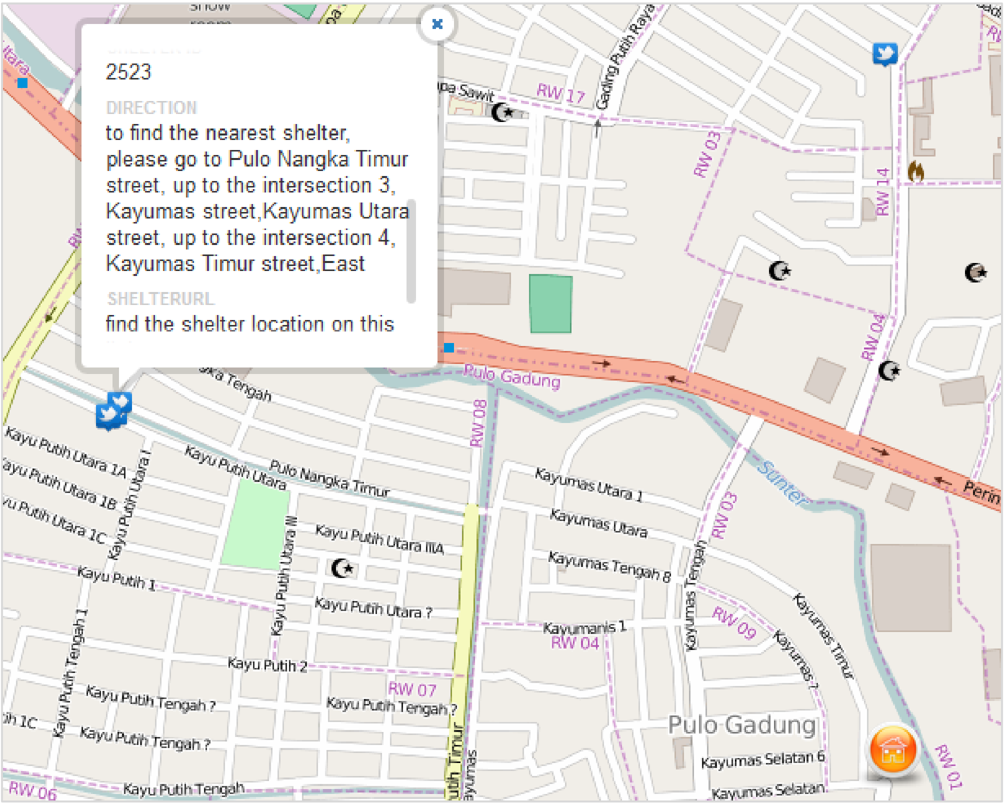

3
0
3
6
9
12 km
Relevant Tweet and close to its location content
Selected Shelters outside flood area
Selected Shelters within flood area
Route
Flood Area 2015

Figure 5.14: The distribution of selected shelters and the route from Tweets to the shelters

As a feedback to the Twitter user, this research proposes a text format as the direction to the shelter. The system allows translating the route into a text format. Therefore, it is possible sending the result to the user using the same platform (Twitter). To assess the output, the Tweets and shelters dataset were stored into CartoDb. Using CartoDb, the toponym of the street can be seen clearly. Figure 5.15 represents the sample of Tweet and the direction to the shelter. It can be seen; the direction calculates the number of the intersection and the points of the compass for the un-named road.



2523

DIRECTION

to find the nearest shelter, please go to Pulo Nangka Timur street, up to the intersection 3, Kayumas street,Kayumas Utara street, up to the intersection 4, Kayumas Timur street,East

SHELTERURL

find the shelter location on this

Figure 5.15: The direction to the shelter from the Tweet displayed in Cartodb

6. EVALUATION AND DISCUSSION

Similar to Chapter 5, this chapter is related to the research question from research objective number two and number three. The research question is about; what is the evaluation result of filtering the Twitter content and what is the evaluation result of assessing the geocoding of Twitter and the geocoding of the location content.

6.1. Filtering of Twitter Content

As adapted from Powers (2011) F-measure (F1 score) is a method to measure the accuracy of prediction and condition. In this case, the prediction is a result of the system, and the condition is a result of checking manually. The F-measure is considered precision and recall that retrieved from a binary classification of contingency as shown in Table 6.1. The cells of the contingency table point out the correct or incorrect of the prediction. The green cells refer the correct prediction and the red cells refer the incorrect prediction.

Table 6.1: The contingency table

	Condition positive	**Condition negative**
Predicted positive	True positive	False positive (Type I error)
Predictive negative	False negative (Type II error)	True negative

Precision (p) or called Positive Predictive Value is the number of True Positive results divided by the number of all positive predictive. Recall (r) or called True Positive Rate is the number of True Positive results divided by the number of all condition positive. In the other word, the precision is to measure the quality of the system and the recall is to measure the completeness of the result.

The evaluation aims to measure the true average between precision and recall. Hence, based on these parameters, the F-measure is performed to calculate the approximation of the average of both. The formula of the F-measure is shown in the following equation:

$$\text{F-measure} = 2 \text{ x } \frac{\text{precision x recall}}{\text{precision} + \text{recall}}$$

In this case, the F-measure approach is to measure the accuracy of the identification of the location content, flood level, and relevant Tweet derived from the filtering.

The first evaluation using F-measure is the identification of location content. As explained in the methodology, there are two approaches to identify a location content; first, using toponym and rule assignment. Since the number of toponyms are limited, then there is a possibility of failure in the identification. All the Tweets were checked manually and categorized into the contingency Table 6.2.

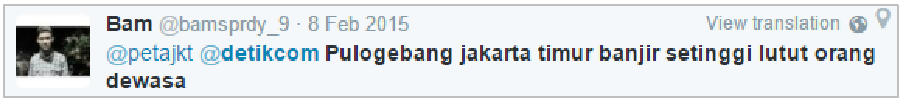
Table 6.2: The contingency table of the identification of location name

	identified has a valid location name	identified has no valid location name
The Tweet has a valid location name	537	89
The Tweet has invalid location name	54	479

The output shows, both the correct cells are far higher than the incorrect cells. Based on this contingency table, precision, recall, and F-measure is computed as follows.

Precision	= 537/(537+89) = 86%
Recall	= 537/(537+54) = 90%
F-measure	= 88%

The result of F-measure is relatively high. This mean, that the system is quite successful in identifying a Tweet with location content or without location content. The system failed to determine the location content because of several factors. For examples; in Figure 6.1, the location name of "Pulogebang Jakarta timur" was not recognized in the gazetteer and there was no keyword precede it. In Figure 6.2, the use of "didaerah" (atarea) instead of "di daerah" (at the area), it caused the failure of the identification of "Islamic centre Jakarta utara" as the location name. In Figure 6.3, the incomplete location name of "masjid palmerah" (palmerah mosque) instead of "masjid failaka palmerah" (failaka palmerah mosque) because the system was not recognized the unknown word (failaka) as a location name without a keyword that preceded it. In Figure 6.4, the use of abbreviation "dpn" as a keyword precedes location name "pt wings" failed to identify it as the location name.

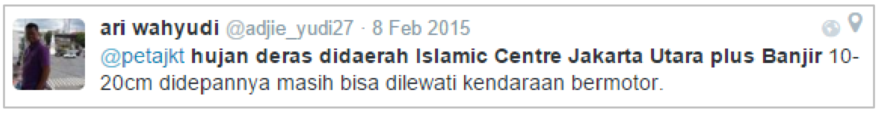


Bam @bamsprdy_9 · 8 Feb 2015 View translation

@petajkt @detikcom **Pulogebang jakarta timur banjir setinggi lutut orang dewasa**

*@petajkt @detikcom **Pulogebang Jakarta timur** flood level up to knee-high*

Figure 6.1: The first sample of the fail identification of location name (there is no keyword precede Pulogebang Jakarta timur)

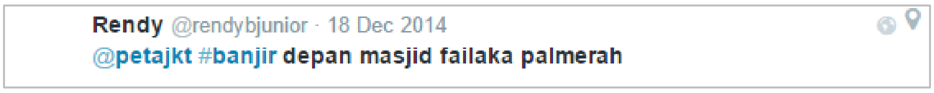


ari wahyudi @adjie_yudi27 · 8 Feb 2015

@petajkt **hujan deras didaerah Islamic Centre Jakarta Utara plus Banjir** 10-20cm didepannya masih bisa dilewati kendaraan bermotor.

*@petajkt pouring rain **inthearea** of Islamic centre Jakarta utara plus is flooding around 10-20 cm but still passable by the vehicle*

Figure 6.2: The second sample of the fail identification of location name ("di" or "in" is a keyword to identify the location name but it written without space)

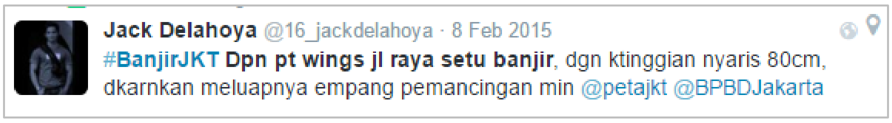


Rendy @rendybjunior · 18 Dec 2014

@petajkt #banjir depan masjid failaka palmerah

@petajkt #banjir in front of ***failaka*** *mosque palmerah*

Figure 6.3: The third sample of the fail identification of location name (the identified location name is 'mosque palmerah' instead of 'failaka mosque')



Jack Delahoya @16_jackdelahoya · 8 Feb 2015

#BanjirJKT Dpn pt wings jl raya setu banjir, dgn ktinggian nyaris 80cm, dkarnkan meluapnya empang pemancingan min @petajkt @BPBDJakarta

#BanjirJKT flood ***Frnt*** *pt wings jl raya setu, with flood level almost 80cm because of the overflow of the fishing pond @petajkt @BPBDJakarta*

Figure 6.4: The forth sample of the fail identification of location name ("dpn" is abbreviation of "depan" or *in front of* in English)

The second evaluation using F-measure is the identification of flood level. As mentioned in Chapter 5.3, all Tweets that mentioned a flood level are categorized as relevant Tweets. To evaluate the result, then all the Tweets were checked manually to calculate the F-measure test of the flood level Tweets as shown in Table 6.3.

Table 6.3: The contingency table of the flood level Tweets

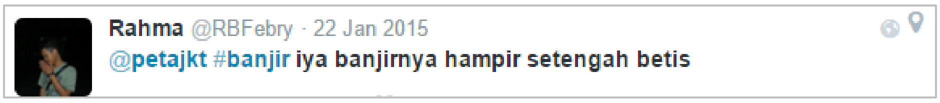
	identified has a valid flood level	not identified has a valid flood level
The Tweet has a valid flood level	212	30
The Tweet has invalid flood level	4	913

Precision, recall, and F-measure computation:

Precision	= 212/(212+30) = 87%
Recall	= 212/(212+4) = 98%
F-measure	= 92%

The incorrect cells (False Positive and False Negative) shows a small number of Tweets and the correct cells (True Positive and True Negative) shows a high number of Tweets. Therefore, the F-Measure of level flood identification is high. It is proven that the system is feasible in identifying a flood level Tweet.

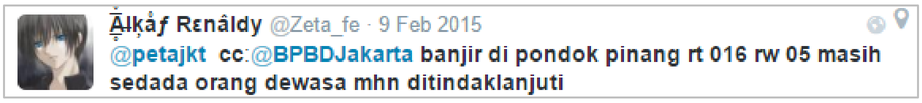
There are several words related to part of the human body that usually used to depict a flood level were not available in the system. Therefore, the system failed to identify the Tweet even though it has a valid flood level (FP). Those words are the calf, chest, and neck as seen in the example of Figure 6.5, 6.6, 6.7 respectively. Another case is the use of the unstructured keyword as seen in Figure 6.8.



Rahma @RBFebry · 22 Jan 2015
@petajkt #banjir iya banjirnya hampir setengah betis

@petajkt #banjir Yes the flood almost to half of ***calf***

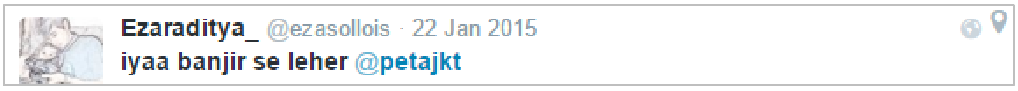
Figure 6.5: The first sample of the fail identification of flood level



Ǟlķåƒ Rεnâldy @Zeta_fe · 9 Feb 2015
@petajkt cc:@BPBDJakarta banjir di pondok pinang rt 016 rw 05 masih sedada orang dewasa mhn ditindaklanjuti

@petajkt cc:@BPBDJakarta flood at pondok pinang rt 016 rw 05 still as high as ***chest*** *of adult. Please respond*

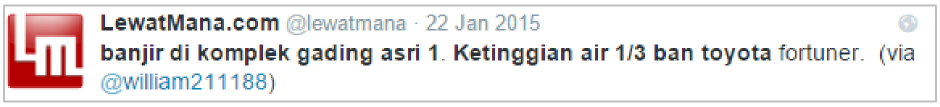
Figure 6.6: The second sample of the fail identification of flood level



Ezaraditya_ @ezasollois · 22 Jan 2015
iyaa banjir se leher @petajkt

Yes flood up to the ***neck*** *@petajkt*

Figure 6.7: The third sample of the fail identification of flood level

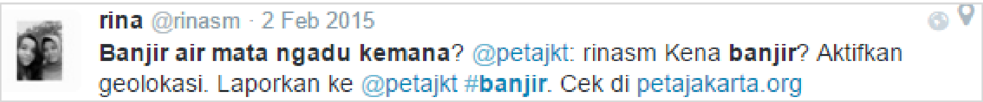


LewatMana.com @lewatmana · 22 Jan 2015
banjir di komplek gading asri 1. Ketinggian air 1/3 ban toyota fortuner. (via @william211188)

Flood at komplek gading asri 1. The water level up to ***1/3 tire of toyota fortuner.***

Figure 6.8: The forth sample of the fail identification of flood level

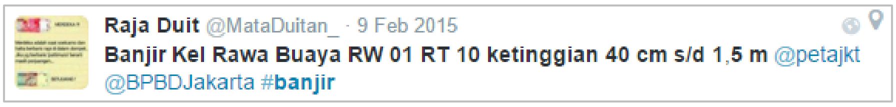
On the other hand, the Tweet has invalid flood level but identified has a valid flood level (FN) was caused by two factors. First, the use of word "banjir" (flood) is not related to the flood situation. In the Indonesian language, the word "banjir" can be used to exaggerate a situation. For example "banjir air mata" (flood of tears) as seen in Figure 6.9. Second, the system failed to recognize the unstructured format number of flood level as seen in Figure 6.10. It should be noted that the system identified a number before metric unit as flood level and the system converted the meter unit into centimeter unit. However, the system could not recognize the number in front of decimal division in a float type number. For example, when the user wrote 1,5 meter (a half meter), the system only read the "5 m" part. Therefore, it was converted into 500 centimeters, not 1500 centimeter.



rina @rinasm · 2 Feb 2015
Banjir air mata ngadu kemana? @petajkt: rinasm Kena banjir? Aktifkan geolokasi. Laporkan ke @petajkt #banjir. Cek di petajakarta.org

Flood of tears where should I complain? @petajktK rinasm do you suffer by flood? Activate geolocation. Send the report to @petajkt #banjir. Check at petajakarta.org

Figure 6.9 : The fifth sample of the fail identification of flood level



Raja Duit @MataDuitan_ · 9 Feb 2015
Banjir Kel Rawa Buaya RW 01 RT 10 ketinggian 40 cm s/d 1,5 m @petajkt @BPBDJakarta #banjir

Flood at Kel Rawa Buaya RW 01 RT 10 up to 40 cm s/d 1,5 m @petajkt @BPBDJakarta #flood pic.twitter.com/teQpia7fu3

Figure 6.10: The sixth sample of the fail identification of flood level

The third evaluation using F-measure is the identification of the relevant and off-topic Tweets. The similar output is shown in the contingency Table 6.4. Since the True positive is high, as a result, the precision, recall, and F-measure is high as well.

Table 6.4: The contingency table of the relevant and off-topic Tweets

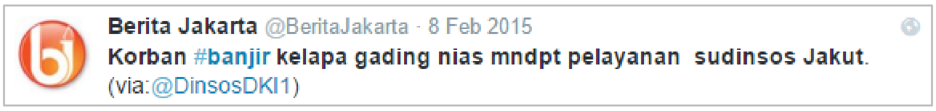
	identified as relevant Tweet	not identified as relevant Tweet
The Tweet has a valid relevant information	495	63
The Tweet has invalid relevant information	39	562

Precision, recall, and F-measure computation:

Precision	= 495/(495+63) = 89%
Recall	= 495/(495+39) = 92%
F-measure	= 90%

As seen in the contingency table, the error identification is grouped into two conditions; the False Negative (FN) condition is the Tweet that has invalid relevant information but identified as relevant Tweet, and The False Positive (FP) condition is the Tweet that has valid relevant information but not identified as relevant Tweets.

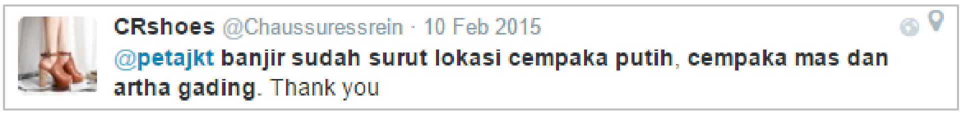
After checking manually, FN was caused by two factors; first, the system identified a flood keyword and location name. However, the real Tweet was about evacuation refugees, updating situation and asking the condition as seen in Figure 6.11, 6.12, 6.13 respectively. Second, the system identified a flood keyword and flood level within the Tweet as seen in Figure 6.14 but in fact, it was just a noise Tweet.



Berita Jakarta @BeritaJakarta · 8 Feb 2015
Korban #banjir kelapa gading nias mndpt pelayanan sudinsos Jakut. (via:@DinsosDKI1)

#flood refugees of kelapa gading nias getting the services from sudinsos Jakut

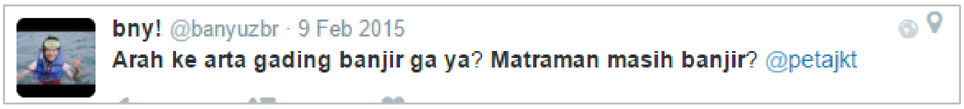
Figure 6.11: The first example of error identification of relevant Tweets ("banjir" is flood keyword and "kelapa gading nias" is location name)



CRshoes @Chaussuressrein · 10 Feb 2015
@petajkt banjir sudah surut lokasi cempaka putih, cempaka mas dan artha gading. Thank you

@petajkt the floods has receded at cempaka putih, cempaka mas and artha gading. Thank you

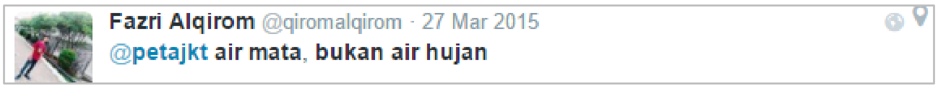
Figure 6.12: The second example of error identification of relevant Tweets ("banjir" is flood keyword and "cempaka putih, cempaka mas, and arta gading" are location name)



bny! @banyuzbr · 9 Feb 2015
Arah ke arta gading banjir ga ya? Matraman masih banjir? @petajkt

Does the way to arta gading is flooded or not? Is Matraman still flooded? @petajkt

Figure 6.13: The third example of error identification of relevant Tweets ("banjir" is flood keyword and "arta gading, matraman" are location name)

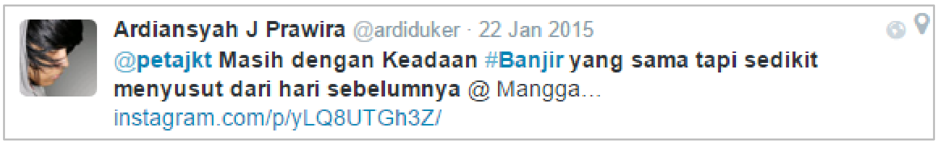


Fazri Alqirom @qiromalqirom · 27 Mar 2015
@petajkt air mata, bukan air hujan

@petajkt water of tears, not water of rain

Figure 6.14: The forth example of error identification of relevant Tweets ("air" is flood keyword and "mata" is flood level keyword)

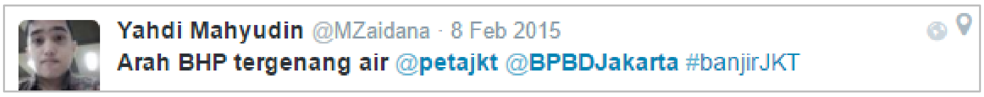
On the other hand, FP error was caused by two factors; first, in Figure 6.15 the system failed to recognize as the relevant Tweet because it was not using the correct format (not mentioned a location or a flood level). Second, in Figure 6.16 the system failed to recognize as the relevant Tweet because the system failed to recognize "BHP" as the location name.



Ardiansyah J Prawira @ardiduker · 22 Jan 2015
@petajkt Masih dengan Keadaan #Banjir yang sama tapi sedikit menyusut dari hari sebelumnya @ Mangga…
instagram.com/p/yLQ8UTGh3Z/

@petajkt still in the same #flood condition but a little bit receded than yesterday

Figure 6.15: The fifth example of error identification of relevant Tweets



Yahdi Mahyudin @MZaidana · 8 Feb 2015
Arah BHP tergenang air @petajkt @BPBDJakarta #banjirJKT

Toward to BHP is still inundated @petajkt @BPBDJakarta #banjirJKT

Figure 6.16: The sixth example of error identification of relevant Tweets (BHP is the location name but the system fails identify it)

6.2. Geocoding of Location Content

The quality of the Geocoding output are related to the location content resulted from the filtering. As mentioned in Chapter 5.4, from 1159 Tweets the system identified 670 Tweets have the location content or location name. The system converted those location names into geographic coordinates using OSM nominatim and Google Geocoding as the geocoding services. The system marked the approximate result obtained from Google Geocoding as the ambiguous location name.

Figure 6.17 shows Google Geocoding contributes 57% from the total location name that successfully geocoded whereas OSM only 7%, about 19% location names are ambiguous, and 16% location names are failure identified.

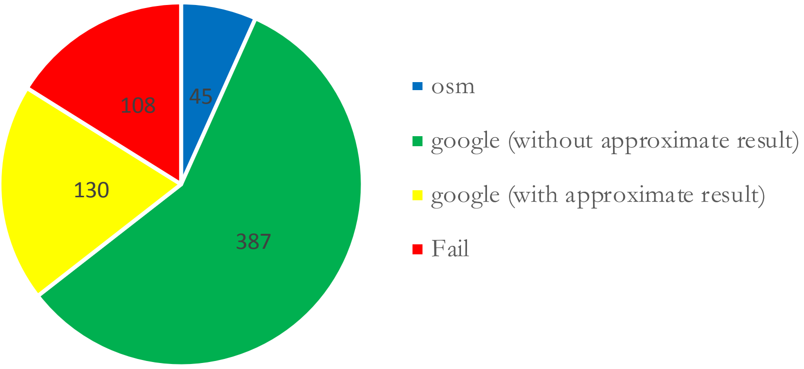


Figure 6.17: The proportion of the Geocoding output

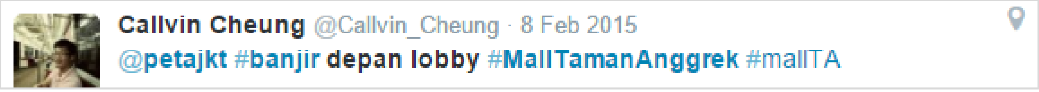
Furthermore, the output was checked manually to analyze the quality of each service. Table 6.5 shows OSM has the better results than Google Geocoding with 64% coordinates of the Tweet (primary geocoding) is close to the coordinates of its location content (secondary geocoding), no output is outside of Jakarta boundary, the average distance between Tweet location and geocoding is 2 Km, and no incorrect location name was geocoded by the system.

On the other hand, Google Geocoding has 52 % coordinates of the Tweet is close to the coordinates of its location content, 5 secondary geocoding are outside of Jakarta, the average distance between Tweet location and geocoding is 13 Km, and 45 incorrect location names were geocoded by this service.

Table 6.5: The comparison output between OSM and Google Geocoding

Geocoding Services	Tweets successfully geocoded	Tweets close to the geocoding output	The output outside of Jakarta	The average distance between input and output	False location name but successfully-geocoded
OSM	45	29	0	2072 meter	0
Google (without approximate result)	373	195	5	13661 meter 6434 meter (without the outlier)	45
Google (with approximate result)	135	-	-	-	-

Evidently, there is an outlier of the output from Google Geocoding. The distance of this outlier is too far, which is 2,803,158 meter. The location name of this outlier obtained from the Tweet in Figure 6.18. The system identified "lobby" as the location name since "depan" is the keyword precede the location name. Even though it is not a valid location name, the Google Geocoding found the similar location name that very far from Jakarta. Therefore, if the outlier is omitted from the average distance calculation, then Google Geocoding produce the average distance from the Tweet location is about 6 Km. Moreover, if the Tweets with incorrect location name are omitted from the average distance calculation, then it produce the average distance is about 3 Km.


Callvin Cheung @Callvin_Cheung · 8 Feb 2015
@petajkt #banjir depan lobby #MallTamanAnggrek #mallTA

@petajkt #flood lobby front #MallTamanAnggrek #mallTA

Figure 6.18: The Tweet that produce incorrect location name

Figure 6.19 and Figure 6.20 are the distance between primary geocoding and secondary geocoding obtained from OSM and Google geocoding respectively. Both figures show that most of the Tweets have the distance from the content-location are below 5 km.

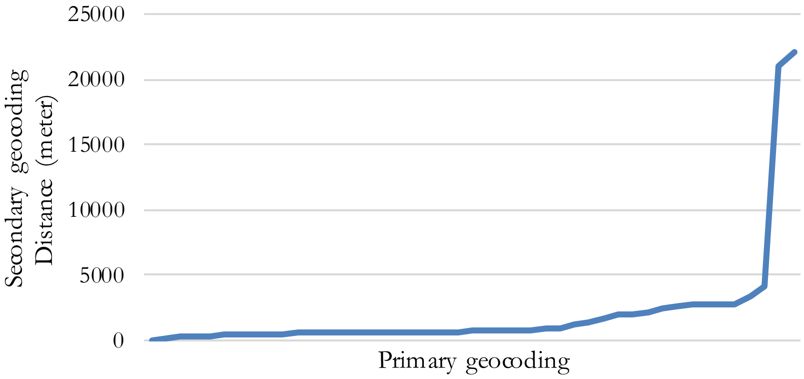


Figure 6.20: The distance between primary geocoding and secondary geocoding obtained from OSM

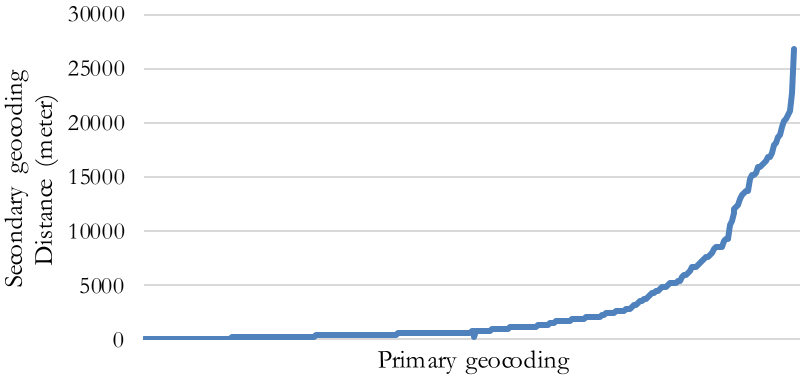


Figure 6.19: The distance between primary geocoding and secondary geocoding obtained from Google geocoding

In order to check the output of geocoding, checking manually of secondary geocoding output was conducted. The checking only for Tweet that has a true location content obtained from filtering process. The result grouped in Table 6.6. First group is secondary geocoding which located in the true location as given by the location content. Second group is secondary geocoding which located in the same administrative area with the location content. Third group is false secondary geocoding but the primary geocoding is located in the same area with the location content. Last group is geocoding which both primary and secondary are located in a different location from the content. For each group, the distance between primary geocoding and secondary geocoding divided by three; less than one kilometer, one until five kilometers, and more than five kilometers.

Table 6.6: The quality checking of secondary geocoding

No	Category	< 1 Km	1-5 Km	> 5 Km
1	**True secondary geocoding**	209	45	8
2	**True secondary geocoding (in administrative level)**	6	22	0
3	**False secondary geocoding, True primary geocoding**	0	27	31
4	**False secondary geocoding, False primary geocoding**	0	11	28

The table shows a large number of true secondary geocoding has a distance of less than one kilometer from the primary geocoding. In contrast, a small number of true secondary geocoding has a distance of more than five kilometers from the primary geocoding. This means, most of the location contents are located in the same area where the Tweet was sent.

Some true secondary geocoding that located more than one kilometer from the primary geocoding indicates disambiguate of the actual location. This mean, a location content could be different from where the Tweet is sent. Therefore, determining the actual location from a Tweet is an essential part of an automated extraction tool (Sultanik & Fink, 2012).

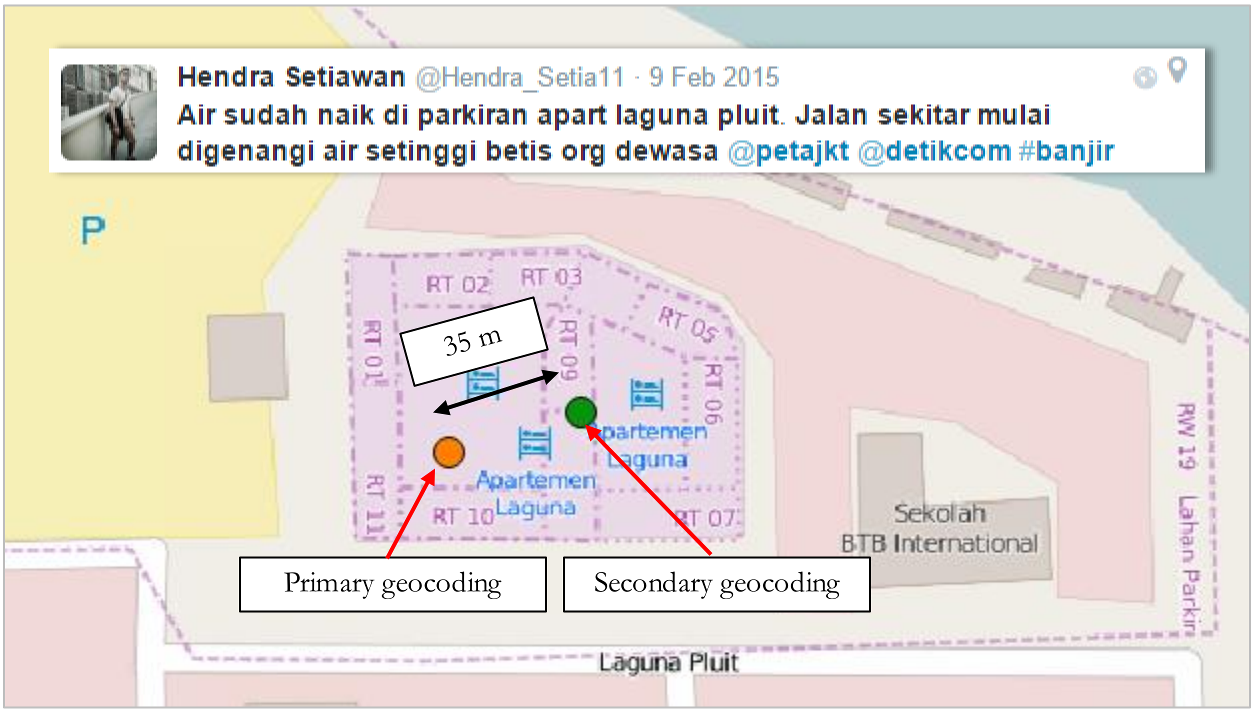
Regarding the quality of the geocoding services, the F-measure approach is calculated based on the contingency Table 6.7. The result of F-measure shows the services is reliable to geocode a location name.

Table 6.7: The contingency table of geocoding service

	True geocoding	False geocoding
valid location name	290	97
location name not valid	0	45

Precision = 290/(290+97) = 74%
Recall = 290/(290+0) = 100%
F-measure = 85%

Figure 6.21 is the example of a secondary geocoding that very close with its primary geocoding. The system was identified "apart laguna" as location name, and geocoding service was returned the same location ("apartemen laguna") as secondary geocoding.



Hendra Setiawan @Hendra_Setia11 · 9 Feb 2015

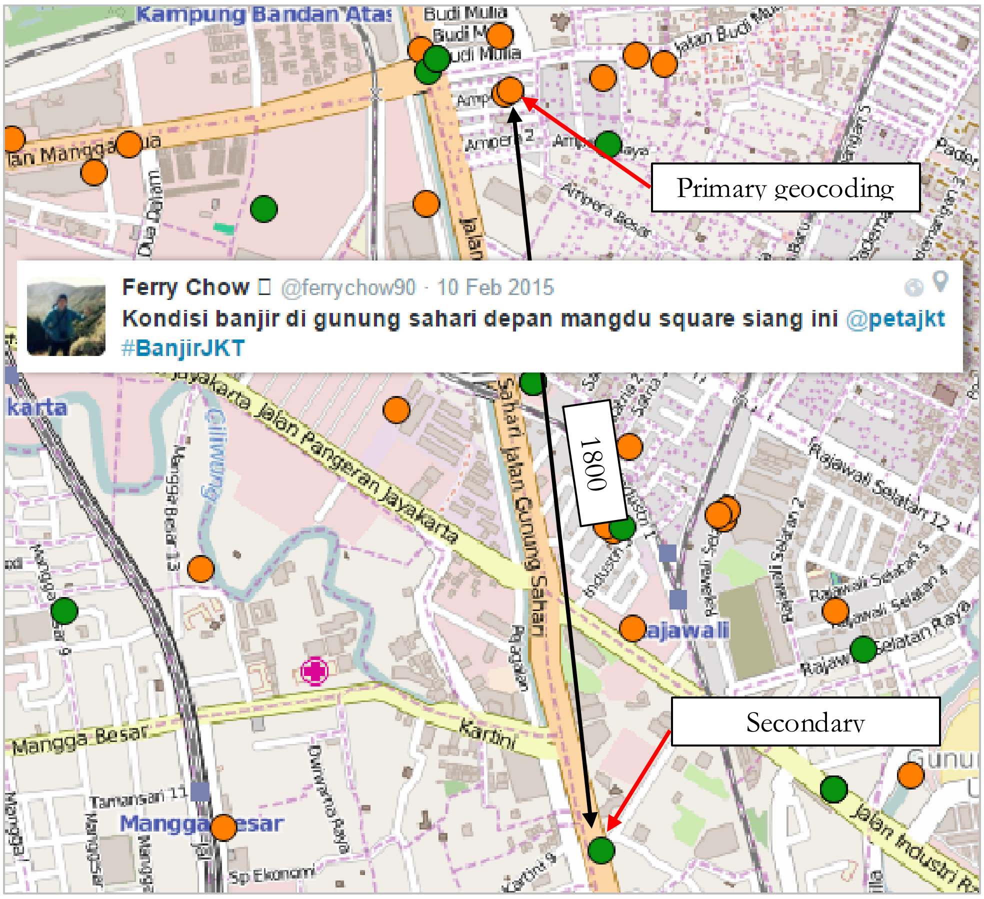
Air sudah naik di parkiran apart laguna pluit. Jalan sekitar mulai digenangi air setinggi betis org dewasa @petajkt @detikcom #banjir

Water has up at the parking lot of ***apart laguna*** *pluit. The nearby road started to inundate up to calf @petajkt @detikcom #banjir*

Figure 6.21: The sample of geocoding output (secondary geocoding) that very close to the geocoding input (primary geocoding)

Figure 6.22 is the example of both primary geocoding and secondary geocoding apparently are in the same location but the distance is relatively far. The system was identified "gunung sahari" as a location name and the geocoding service was recognized it as the name of a road. Since it is a long road, then the distance between both geocoding is high as well.

On the contrary, Figure 6.23 shows the example of the failure of geocoding service in identifying a location name. The system was identified "Kramat Jati" as a location name but Google geocoding was returned "Kramat Asem" as the secondary location. As referred by Roongpiboonsopit & Karimi (2010) Google will expand the searching strategy by adding the city/area if it fails in finding the address. Even though the location name was changed, Google geocoding identified it as a "good match". This is the answer what makes the 45 false location name was geocoded by Google. After checking manually, if the distance between primary geocoding and secondary geocoding was larger than five kilometers, it has a possibility returning a false address.

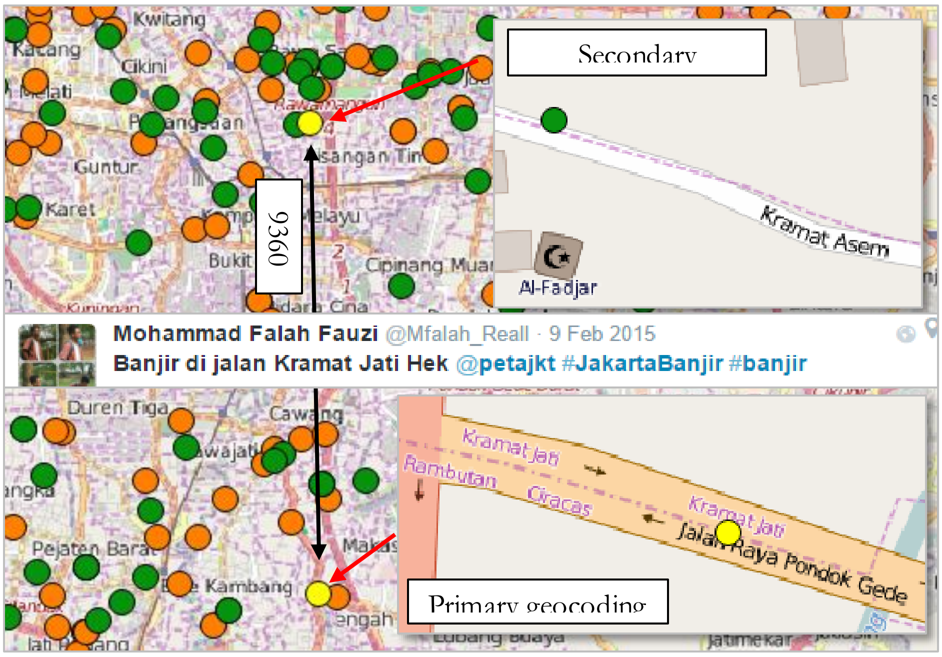


Ferry Chow □ @ferrychow90 · 10 Feb 2015

Kondisi banjir di gunung sahari depan mangdu square siang ini @petajkt #BanjirJKT

Flood at gunung sahari in front of mangdu square this afternoon @petajkt #BanjirJKT

Figure 6.22: The example of primary geocoding and secondary geocoding in the same location but the distance is far



Mohammad Falah Fauzi @Mfalah_Reall · 9 Feb 2015

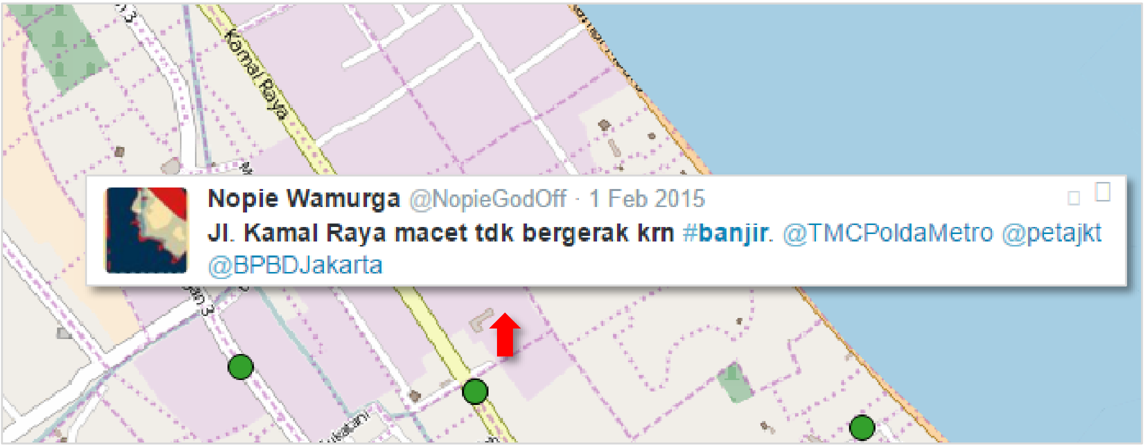
Banjir di jalan Kramat Jati Hek @petajkt #JakartaBanjir #banjir

Flood at jalan Kramat Jati Hek @petajkt #JakartaBanjir #banjir

Figure 6.23: The example of failure in identification the location name by Google geocoding

Regarding the uncertainty of flood area boundary obtained from the authoritative data, the output of secondary geocoding can be used to check flood location that not covered by the dataset. After random checking, many of flood location located in the outside boundaries.

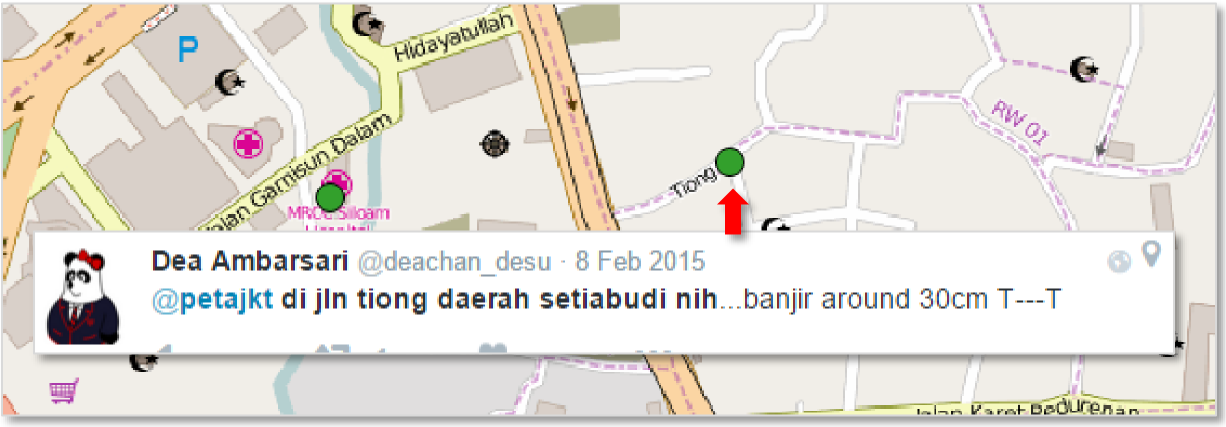
For example, Figure 6.24 shows a Tweet mentions that "Jl.Kamal Raya" is flooded. The secondary geocoding is located at "Kamal Raya" as well. However, it is not covered by the flood area dataset. The distance from the secondary geocoding to the flood boundary is 800 m. Another example is shown in Figure 6.25. a Tweet mentions that there is a flood at "jln tiong" but it is not covered by the flood area dataset.



Nopie Wamurga @NopieGodOff · 1 Feb 2015
Jl. Kamal Raya macet tdk bergerak krn #banjir. @TMCPoldaMetro @petajkt @BPBDJakarta

*Traffic jam at **Jl. Kamal Raya.** is stuck because of #flood @TMCPoldaMetro @petajkt @BPBDJakarta*

Figure 6.24: The first example of flood location is not covered in the authoritative data



Dea Ambarsari @deachan_desu · 8 Feb 2015
@petajkt di jln tiong daerah setiabudi nih...banjir around 30cm T---T

*@petajkt at **jln tiong setiabudi** area.. flood around 30cm*

Figure 6.25: The second example of flood location is not covered in the authoritative data

6.3. Evacuation Shelter Selection

As mentioned in Chapter 5.5, there is one Tweet failed to find the closest shelter. Thus, it failed to get direction to the shelter as well. When it is looked in Figure 6.26, the nearest path from the Tweet is unconnected. The system calculated the optimum path from the Tweet to the shelters. However, the nearest path from the Tweet is unconnected. Therefore, it failed to get a shelter

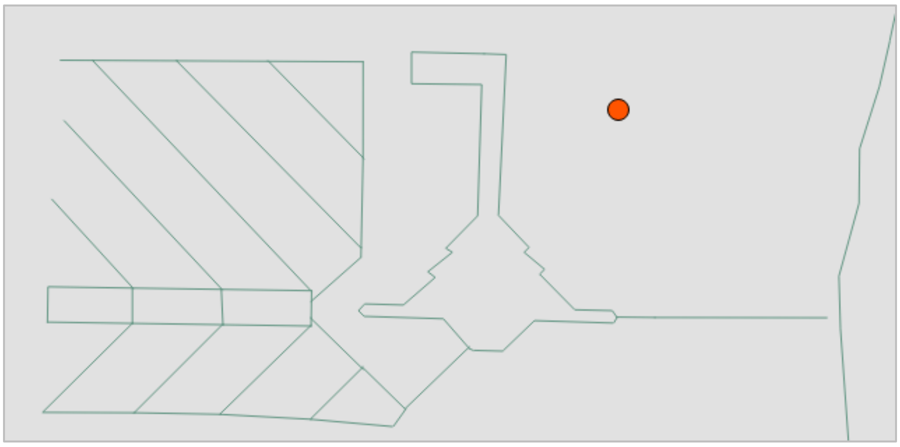


Figure 6.26: The Tweet that failed to identify a shelter because of the unconnected path

6.4. System Performance

In order to check the performance of the system, the timer in every process (Filtering, geocoding, and shelter selection) were inserted into the code. Therefore, the time-consuming of each Tweet was recorded in the database. Table 6.8 shows the average of time-consuming in each Tweet processing. The system was tested on the same personal computer but it was used two different speed internet connections. The slower connection was influenced the Geocoding process since it was needed a connection to the online Geocoding services. On the other hand, the average time-consuming of the other two processes are quite similar.

Table 6.8: The average time consuming of each process

Process of	Average Time Consuming (second)	
	Fast Internet Connection 37.60 Mbps	Slow Internet Connection 8.73 Mbps
NLP	1.56	1.62
Geocoding	0.3	4.09
Shelter selection	2.44	2.52

7. CONCLUSIONS AND RECOMMENDATION

7.1. Conclusions

This study shows that correctly filtered content from Twitter during flood disaster provide a promising source of reliable flood information. Even though the available authoritative data is obtained from the aggregation of neighborhood, it can be used to see the spatial pattern of the output. The comparison shows the spatial distribution of filtered tweets was more concentrated in the flood areas. In general, the results can help answer the following of research questions:

1. What is the characteristic of Twitter content in flood disaster?

   According to the requirement analysis and the result, the relevant information about flood report is mentioned as a flood location or a water level. Aside from that, another information that possibly referred in a Tweet is related to evacuation refugees or shelter, updating flood condition, and noise content. The system classified this information as off-topic Tweets since it concerned only about a Tweet that inform a flood report as a relevant Tweet.

2. What is the use case to be implemented?

   The conceptual design is explained in Chapter 3. The design explained a use case scenario of the concept. The workflow of the scenario consists three main process of utilizing Twitter content in flood disaster. First, filtering the content to obtain detailed information about location content and flood depth level. Second, geocoding of the location content to assess the closeness between geocoding of Twitter and geocoding of its location content. Third, identifying the closest shelter as a feedback information.

3. Which natural language processing method to use or modify in filtering the Twitter content?

   To filter the information of a Tweet, the NLP method that was used and modified is Named Entity Recognition (NER). A combination of predefined keywords, toponym, Part Of Speech Tagging (POS-T) and rule-based method were used to identify a location entity. Predefined keywords and rule based method also used in the identification of a flood level and a relevant Tweet. Using the considered methods allow us to obtain the necessary details from the Tweets.

4. What is the evaluation result of filtering the Twitter content?

   After evaluation using F-measure approach, the result shows that the system has a good quality in Identifying both the detail information of the content and relevant Tweets. The content are location name and water level that afterward are used to classify relevant Tweet and off-topic Tweet. Several factors caused the incorrect output; the incompleteness of toponym, the incompleteness of the predefined keywords, the incompleteness of rule assignment, and the use of the unstructured text such as abbreviation.

5. How to assess the geocoding of Twitter and the geocoding of its location content?

   A location content obtained from filtering is used to evaluate the closeness between coordinates of the Tweet as primary geocoding and coordinates of its location content as secondary geocoding. The location content is sent to geocoding services to obtain the secondary coordinates.

   After Studying the literature and experimenting with several samples of location name, it was decided to use OSM nominatim and Google geocoding as the services. According to the literature review and the experiment, OSM provides a better accuracy. Thus, it used as primary geocoding service in the system.

   The output of metadata from Google geocoding can return an approximation of the quality of geocoding. It used to mark a Tweet that has an ambiguous location content.

6. What is the evaluation result of assessing the geocoding of Twitter and the geocoding of its location content?

   Regarding the quality of the output, apparently OSM has a better result than Google. After checking manually, no false location content was geocoded by OSM, and the distance between primary geocoding and secondary geocoding is relatively close to each other.

   On the other hand, Google geocoding has a result that can be misleading. It showed after evaluation; Google geocoding can return different location name from the input. It happened because if Google cannot find a given address, then it will find a similar location name even only partially. The output shows if the distance between primary geocoding and secondary geocoding is too far (usually more than 5 km), it has a possibility that the geocoding service may return a false output.

   After evaluation using F-measure, the result shows the geocoding services provide an opportunity to return a good output as long as the location content is a valid location name.

7. What is the method to use or modify for finding a route to the nearest shelter under flood disaster conditions?

   The nearest shelter identified based on the optimum path from where the Tweet was sent. The optimum path was calculated by Dijkstra algorithm on pg-Routing. The selected shelter was determined if it fulfills the requirements which are; the shelter has no possibility being inundated and not too far from its location. A shelter has no possibility inundated if the shelter is far from a good Tweet report.

   The good Tweet mean it has a relevant information and close to its location content. The two conditions indicate the Tweet contains a true information. In other words, flooding likely occurs in the location from where the Tweet was sent. A shelter that is located next to the Tweet (in this case using a boundary 1 km) has a score of possibility inundated. The score is increased if there is another Tweet in that area. The process is iterated every time a new Tweet comes in to the system. The inundated shelter should be eliminated from the selection process. In the end, the nearest shelter can be identified

8. How can these directions be communicated back to the information seeker?

   After identifying the nearest shelter, the route to the shelter can be identified. The information of the route used as a feedback to the seeker. In this case, the seeker is the Twitter user who sending the information about flood. Considering Twitter is a social media for broadcasting a text message, the route should be converted into a text format. The route is determined based on the name of each path and the azimuth to identify a point of the compass for the un-named path.

   As explained in point 8, the process identification of inundated shelter was iterated every time a new Tweet comes to the system. Consequently, the selected shelter might turn out to be an inundated shelter. Therefore, the user should be informed about the changed shelter.

7.2. Limitation and Recommendations

There are several limitations of this research. Based on that, recommendations are proposed for the future works. The limitation and recommendation are in the following points:

1. Currently, the system is not running in a real-time manner since it used the data from the last year monsoon season. The data was used due to the need of obtaining a sufficient number of Tweets. The real-time system can be implemented using the Twitter Streaming API. The API can replace the current data collection process.

   In a real-time condition, the system can help the relief agency to classify a relevant Tweet quickly since the current system "Peta Jakarta" do the validation of information by cross check manually. Moreover, the in-situ information about water level can be used to develop an inundation modelling in a real-time condition. As referred by Zhu (2010), water level is one of parameters for flood inundation modelling. Using the flood model, the inundated boundary can be predicted shortly.

2. As mentioned in the evaluation, there are several part of human body that usually used by user to describe a flood level. The words are the calf, chest, and neck. However, these words are not included in a predefined keyword. It caused the system failed in identification the flood level if the words mentioned in a Tweet. Therefore, these words can be included in the predefined keywords to enhance the quality of the filtering.

3. As explained in the evaluation of Geocoding content, the Google geocoding provides a risk to return a wrong location. Using a ratio similarity between a given address name and the output address can be considered to overcome the problem. Adding more online geocoding service into the system is also possible. However, it will influence the time consumption.

4. The shelter dataset provides a large number of shelters plan. The system helps to identify a shelter that has a possibility being inundated based on a Twitter location. Currently, the system only considers the distance to a relevant Tweet as a parameter for identifying possibly inundated shelters.

   As a recommendation for the future work, adding different datasets such as Digital Elevation Model and flood prone area map are necessary to improve the possibility. For example, if a shelter is lower than a relevant Tweet, then it has a possibility inundated, or if a shelter is located in a flood prone area, then it has a possibility inundated as well. The method for identifying an inundated shelter can be used by a relief agency to provide preliminary data in real-time. However, it is recommended to do the verification to check the validity. Moreover, Twitter users can be involved for contributing information about the availability of the shelters. Because, based on the requirement analysis such information is possible obtained from the user.

5. Another limitation is the system not considering an inundated path in the selection of the shelter. Because, if all path around the Tweet is identified as an inundated path, then it will be difficult to identify the nearest path from a Tweet. For the recommendation, the inundated path can be identified using the similar method to the identification of inundated shelter. However, the inundated path can be ignored in identifying the nearest path from the Tweet.

LIST OF REFERENCES

Adam, 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. http://doi.org/10.1504/IJWBC.2011.041206

Alçada-Almeida, L., Tralhão, L., Santos, L., & Coutinho-Rodrigues, J. (2009). A multiobjective approach to locate emergency shelters and identify evacuation routes in urban areas. *Geographical Analysis*, *41*(1), 9–29. http://doi.org/10.1111/j.1538-4632.2009.00745.x

BPBD Jakarta. (2015). Recapitulation of Floods event February 2015 (in Indonesian). Retrieved January 21, 2016, from http://bpbd.jakarta.go.id/assets/attachment/document/Rekapitulasi_Kejadian_Banjir_Bulan_Februari_2015.pdf

Bruns, A., & Liang, Y. E. (2012). Tools and methods for capturing Twitter data during natural disasters. *First Monday*, *17*(4). http://doi.org/10.5210/fm.v17i4.3937

Budi, I., Bressan, S., Wahyudi, G., Hasibuan, Z. A., & Nazief, B. A. A. (2005). Named Entity Recognition for the Indonesian language: Combining contextual, morphological and part-of-speech features into a knowledge engineering approach. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 3735 LNAI, pp. 57–69). http://doi.org/10.1007/11563983_7

Cameron, M. A., Power, R., Robinson, B., & Yin, J. (2012). Emergency situation awareness from twitter for crisis management. *Proceedings of the 21st International Conference Companion on World Wide Web - WWW '12 Companion*, (July 2015), 695. http://doi.org/10.1145/2187980.2188183

CCCMCluster. (2014). The Mend Guide Comprehensive Guide for Planning Mass Evacuations in Natural Disasters. Retrieved September 14, 2015, from http://www.globalcccmcluster.org/system/files/publications/MEND_download.pdf

Crampton, J. W., Graham, M., Poorthuis, A., Shelton, T., Stephens, M., Wilson, M. W., & Zook, M. (2013). Beyond the Geotag? Deconstructing "Big Data" and Leveraging the Potential of the Geoweb. *SSRN Electronic Journal*. http://doi.org/10.2139/ssrn.2253918

Dartmouth Flood Observatory. (2008). 2007 Global Register of Major Flood Events. Retrieved November 11, 2015, from http://www.dartmouth.edu/~floods/Archives/2007sum.htm

Davis, C. A., & de Alencar, R. O. (2011). Evaluation of the quality of an online geocoding resource in the context of a large Brazilian city. *Transactions in GIS*, *15*(6), 851–868. http://doi.org/10.1111/j.1467-9671.2011.01288.x

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik*, *1*(1), 269–271. http://doi.org/10.1007/BF01386390

Dinakaramani, A., Rashel, F., Luthfi, A., & Manurung, R. (2014). Designing an Indonesian Part of speech Tagset and Manually Tagged Indonesian Corpus. *International Conference on Asian Language Processing*, 2–5.

DKI Jakarta Capital Government. (2014). PetaJakarta. Retrieved June 9, 2015, from http://petajakarta.org/banjir/en/

Elwood, S., Goodchild, M. F., & Sui, D. Z. (2012). Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice. *Annals of the Association of American Geographers*, *102*(3), 571–590. http://doi.org/10.1080/00045608.2011.595657

Flanagin, A. J., & Metzger, M. J. (2008). The credibility of volunteered geographic information. *GeoJournal*, *72*(3-4), 137–148. http://doi.org/10.1007/s10708-008-9188-y

geopy. (2016). geopy. Retrieved January 14, 2016, from https://github.com/geopy/geopy

Ghahremanlou, L., Sherchan, W., & Thom, J. A. (2015). Geotagging Twitter Messages in Crisis Management. *The Computer Journal*, *58*(9), 1937–1954. http://doi.org/10.1093/comjnl/bxu034

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., … Smith, N. a. (2011). Part-of-speech tagging for Twitter: annotation, features, and experiments. *Human Language Technologies*, *2*(2), 42–47. http://doi.org/10.1.1.206.3224

GIS BPBD DKI Jakarta. (2015). Hystorical Floods Map 2013-2015 (in Indonesian). Retrieved January 21, 2016, from http://gis.bpbd.jakarta.go.id/maps/81

Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal*, *69*(4), 211–221. http://doi.org/10.1007/s10708-007-9111-y

Google Developers. (n.d.). The Google Maps Geocoding API. Retrieved January 14, 2016, from https://developers.google.com/maps/documentation/geocoding/intro

Hahmann, S., Purves, R., & Burghardt, D. (2014). Twitter location (sometimes) matters: Exploring the relationship between georeferenced tweet content and nearby feature classes. *Journal of Spatial Information Science*, *9*(9), 1–36. http://doi.org/10.5311/JOSIS.2014.9.185

Holderness, T., & Turpin, E. (2014). *Assessing the Role of Social Media for Civic Co-Management During Monsoon Flooding in Jakarta , Indonesia Tomas Holderness & Etienne Turpin SMART Infrastructure Facility University of Wollongong*. SMART Infrastructure Facility, University of Wollongong. Retrieved from https://dl.dropboxusercontent.com/u/12960388/WhitePaper_vWeb.pdf

Idris, N. H., Jackson, M. J., & Ishak, M. H. I. (2014). A conceptual model of the automated credibility assessment of the volunteered geographic information. *IOP Conference Series: Earth and Environmental Science*, *18*, 012070. http://doi.org/10.1088/1755-1315/18/1/012070

Jiang, B. (2012). Volunteered Geographic Information and Computational Geography: New Perspectives*. In *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information in Theory and Practice* (pp. 1–13). http://doi.org/10.1007/978-94-007-4587-2

Jung, J. J. (2012). Online named entity recognition method for microtexts in social networking services: A case study of twitter. *Expert Systems with Applications*, *39*(9), 8066–8070. http://doi.org/10.1016/j.eswa.2012.01.136

Klein, B., & Castanedo, F. (2013). Emergency Event Detection in Twitter Streams Based on Natural Language Processing. *UCAmI, Springer LNCS 8276*, 239–246. Retrieved from http://link.springer.com/chapter/10.1007/978-3-319-03176-7_31

Landwehr, P. M., & Carley, K. M. (2014). Social Media in Disaster Relief. In *Data Mining and Knowledge Discovery for Big Data, Vol. 1* (Vol. 1, pp. 225–257). http://doi.org/10.1007/978-3-642-40837-3

Lickfett, J., Ashish, N., Mehrotra, S., Venkatasubramanian, N., Irvine, U. C., & Green, J. (2008). The RESCUE Disaster Portal for Disasters and Emergency Response. *Spring*, (May), 787–796.

Liu, X., Zhang, S., Wei, F., & Zhou, M. (2011). Recognizing Named Entities in Tweets. *In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, *1*(2008), 359–367. Retrieved from http://acl.eldoc.ub.rug.nl/mirror/P/P11/P11-1037.pdf

Madry, S. (2015). The Emerging World of Crowd Sourcing, Social Media, Citizen Science, and Remote Support Operations in Disasters (pp. 117–121). http://doi.org/10.1007/978-1-4939-1513-2_9

Middleton, S. E., Middleton, L., & Modafferi, S. (2014). Real-time crisis mapping of natural disasters using social media. *IEEE Intelligent Systems*, *29*(2), 9–17. http://doi.org/10.1109/MIS.2013.126

NLTK Project. (2015). Natural Language Toolkit. Retrieved January 14, 2016, from http://www.nltk.org/

O'Reilly. (2005). What Is Web 2.0. Retrieved from http://oreilly.com/pub/a/web2/archive/what-is-web-20.html

Oliver, D., Tiwari, R., Evans, M. R., & Shekhar, S. (2014). Disaster Response and Relief, VGI Volunteer Motivation in. In *Encyclopedia of Social Network Analysis and Mining* (pp. 370–380). http://doi.org/10.1007/978-1-4614-6170-8_57

OpenStreetMap Wiki. (2016). Nominatim. Retrieved January 14, 2016, from http://wiki.openstreetmap.org/wiki/Nominatim

Ostermann, F. O., & Spinsanti, L. (2011). A Conceptual Workflow For Automatically Assessing The Quality Of Volunteered Geographic Information For Crisis Management. *Agile 2011*, 1–6.

pgRouting Community. (n.d.). pgRouting. Retrieved January 14, 2016, from http://pgrouting.org/

Poser, K., & Dransch, D. (2010). Volunteered Geographic Information for Disaster Management With Application To Rapid Flood Damage Estimation. *Geomatica*, *64*(1), 2010.

PostGIS. (n.d.). PostGIS. Retrieved January 14, 2016, from http://postgis.net/

PostGis. (n.d.). St_Azimuth. Retrieved March 6, 2016, from http://postgis.net/docs/ST_Azimuth.html

Powers, D. (2011). Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation. *Journal of Machine Learning Technologies*, *2*(1), 37 – 63.

Python Software Foundation. (2015). psycopg2. Retrieved January 14, 2016, from https://pypi.python.org/pypi/psycopg2

Rashid, H., Haider, W., & McneilL, D. (2007). Urban riverbank residents' evaluation of flood evacuation policies in Winnipeg, Manitoba, Canada. *Environmental Hazards*, *7*(4), 372–382. http://doi.org/10.1016/j.envhaz.2007.09.006

Reza Fazeli, H., Nor Said, M., Amerudin, S., & Zulkarnain Abd Rahman, M. (2015). A Study of Volunteered Geographic Information (VGI) Assessment Methods For Flood Hazard Mapping: A Review. *Jurnal Teknologi*, *4*, 185–190.

Ritter, A., Clark, S., Mausam, & Etzioni, O. (2011). Named entity recognition in tweets: an experimental study. *EMNLP '11 Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1524–1534. Retrieved from http://dl.acm.org/citation.cfm?id=2145595

Robin. (2009a). Natural Language Processing. Retrieved January 2, 2016, from http://language.worldofcomputing.net/nlp-overview/natural-language-processing-overview.html

Robin. (2009b). Parts of Speech Tagging. Retrieved January 3, 2016, from http://language.worldofcomputing.net/pos-tagging/parts-of-speech-tagging.html

Robin. (2009c). Tokenization. Retrieved January 2, 2016, from http://language.worldofcomputing.net/tokenization/tokenization-overview.html

Roche, S., Propeck-Zimmermann, E., & Mericskay, B. (2011). GeoWeb and crisis management: issues and perspectives of volunteered geographic information. *GeoJournal*, *78*(1), 21–40. http://doi.org/10.1007/s10708-011-9423-9

Roesslein, J. (2009). Tweepy. Retrieved January 14, 2016, from http://www.tweepy.org/

Rojas, G., & Muñoz, V. (2014). Twitter-Based Geocollaboration: Geovisualization and Geotagging of Microblogging Messages (pp. 181–198). Springer International Publishing. http://doi.org/10.1007/978-3-319-04028-8_13

Roongpiboonsopit, D., & Karimi, H. a. (2010). Comparative evaluation and analysis of online geocoding services. *International Journal of Geographical Information Science*, *24*(April 2015), 1081–1100. http://doi.org/10.1080/13658810903289478

Rouse, M. (2014). REST (representational state transfer) definition. Retrieved January 10, 2016, from http://searchsoa.techtarget.com/definition/REST

Sanyal, J., & Lu, X. X. (2009). Ideal location for flood shelter: A geographic information system approach. *Journal of Flood Risk Management*, *2*(4), 262–271. http://doi.org/10.1111/j.1753-318X.2009.01043.x

Schade, S., Díaz, L., Ostermann, F., Spinsanti, L., Luraschi, G., Cox, S., … De Longueville, B. (2011). Citizen-based sensing of crisis events: sensor web enablement for volunteered geographic information. *Applied Geomatics*, *5*(1), 3–18. http://doi.org/10.1007/s12518-011-0056-y

Shelton, T., Poorthuis, A., Graham, M., & Zook, M. (2014). Mapping the data shadows of Hurricane Sandy: Uncovering the sociospatial dimensions of "big data." *Geoforum*, *52*, 167–179. http://doi.org/10.1016/j.geoforum.2014.01.006

Sheth, A. (2009). Citizen sensing, social signals, and enriching human experience. *IEEE Internet Computing*, *13*(4), 87–92. http://doi.org/10.1109/MIC.2009.77

Singh, P. S., Lyngdoh, R. B., Chutia, D., Saikhom, V., Kashyap, B., & Sudhakar, S. (2015). Dynamic shortest route finder using pgRouting for emergency management. *Applied Geomatics*, *7*(4), 255–262. http://doi.org/10.1007/s12518-015-0161-4

Starbird, K., & Stamberger, J. (2010). Tweak the tweet: Leveraging microblogging proliferation with a prescriptive syntax to support citizen reporting. *Proceedings of the 7th International ISCRAM Conference – Seattle, USA, May 2010*, (May), 1–5. http://doi.org/10.1111/j.1556-4029.2009.01243.x

Steinberg, F. (2007). Jakarta: Environmental problems and sustainability. *Habitat International*, *31*(3-4), 354–365. http://doi.org/10.1016/j.habitatint.2007.06.002

Sui, D., & Goodchild, M. (2011). The convergence of GIS and social media: challenges for GIScience. *International Journal of Geographical Information Science*. http://doi.org/10.1080/13658816.2011.604636

Sultanik, E. a., & Fink, C. (2012). Rapid Geotagging and Disambiguation of Social Media Text via an Indexed Gazetteer. *Proceedings of the 9th International ISCRAM Conference*, (April), 1–10. Retrieved from http://www.iscramlive.org/ISCRAM2012/proceedings/190.pdf\nhttp://www.iscramlive.org/portal/iscram2012proceedings

Team Mirah Sakerti. (2010). Why flood in Jakarta. Retrieved December 10, 2015, from http://bpbd.jakarta.go.id/assets/attachment/study/buku_mjb.pdf

Teske, D. (2014). Geocoder Accuracy Ranking. In *Communications in Computer and Information Science* (Vol. 500, pp. 30–44). http://doi.org/10.1007/978-3-662-45006-2

Twitter. (2016). Twitter Public API. Retrieved January 14, 2016, from https://dev.twitter.com/rest/public

Twitter.inc. (2015). Twitter Search API. Retrieved January 14, 2016, from https://dev.twitter.com/rest/public/search

Twitter.inc. (2016). GET search/tweets. Retrieved January 14, 2016, from https://dev.twitter.com/rest/reference/get/search/tweets

Vieweg, S., Hughes, A. L., Starbird, K., & Palen, L. (2010). Microblogging during two natural hazards events. *Proceedings of the 28th International Conference on Human Factors in Computing Systems - CHI '10*, 1079. http://doi.org/10.1145/1753326.1753486

Wang, T., Huang, R., Li, L., Xu, W., & Nie, J. (2011). The Application of the Shortest Path Algorithm in the Evacuation System. *2011 International Conference of Information Technology, Computer Engineering and Management Sciences*, 250–253. http://doi.org/10.1109/ICM.2011.119

Watanabe, K., Ochi, M., Okabe, M., & Onai, R. (2011). Jasmine: A Real-time Local-event Detection System Based on Geolocation Information Propagated to Microblogs. *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, 2541–2544. http://doi.org/10.1145/2063576.2064014

Ye, M., Wang, J., Huang, J., Xu, S., & Chen, Z. (2012). Methodology and its application for community-scale evacuation planning against earthquake disaster. *Natural Hazards*, *61*, 881–892. http://doi.org/10.1007/s11069-011-9803-y

Yin, J., Lampert, A., Cameron, M., Robinson, B., & Power, R. (2012). Using social media to enhance emergency situation awareness. *IEEE Intelligent Systems*, *27*(6), 52–59. http://doi.org/10.1109/MIS.2012.6

Zhang, L., & He, X. (2012). Route Search Base on pgRouting, *115*, 1003–1007. Retrieved from http://www.springerlink.com/content/k89k311r025g063l/

Zhu, J. (2010). GIS based urban flood inundation modeling. *Proceedings - 2010 2nd WRI Global Congress on Intelligent Systems, GCIS 2010*, *2*, 140–143. http://doi.org/10.1109/GCIS.2010.264

APPENDIX

Table 1: The comparison of relevant Tweets and off-topic Tweets based on filtering category

Category	All Tweets	Relevant Tweets	Off-Topic Tweets
Total Tweets	1159	534	625
Tweets contain location name	670	501	169
Tweets contain flood level	216	216	0

Table 2: The number of relevant and off-topic Tweets per month

Month	Total Tweets	Relevant Tweets	Off-topic Tweets
2014-12	160	49	110
2015-01	275	100	170
2015-02	699	374	325
2015-03	25	11	14
Total	1159	534	625

Table 3: The Geocoding output based on the category

Category	All Tweets	Relevant Tweets	Off-Topic Tweets
Fail	117	80	37
Approximation output	135	101	34
Success	418	320	98
Tweets that close to its content location	224	169	55