Investigation on Accident Hotspot Area Using Social Media Data

Shafuan Othman 1, Maisarah Abdul Halim 1, Noraain Mohd Saraf 1, Siti Aekbal Salleh* 1, 2

1 Centre of Studies for Surveying Science and Geomatics, Faculty of Architecture, Planning, and Surveying, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia.
2 Institute for Biodiversity and Sustainable Development, Applied Remote Sensing and Geospatial Research Group, Universiti Teknologi MARA, 40450, Shah Alam, Selangor, Malaysia.

aekbal@uitm.edu.my

Abstract. Fatalities about road accidents become one of the main issues in this country. The precaution step should be taken to create a map of the accident area to create awareness among road users. The problem is data about the accident cases get by a local authority does not provide the specific location of the accident cases. It only shows the number of accident cases in every section in Shah Alam. To solve this problem, data from social media will be used to get the specific location of accident cases around Shah Alam. This research aims to investigate the capability to extract data from twitter for the use of analytic purposes. There are 2 objectives used which is first, to analyse data from social media among the road user about the accident road on the hotspot area for analytic purpose. The second is to produce a map of the accident hotspot area in Shah Alam by using Twitter data. Methods use is divided by two which is one method in terms of coding by using Rstudio and another method is using ArcGIS software. There are 4 analysis was used which is data extraction analysis, data mining analysis, data usage analysis and data achieving analysis were done to produce the 2 types of result which are the location of accident cases that contain specific coordinate and a map of accident hotspot area in Shah Alam. As a conclusion, the objective of the research is successful by producing the 2 types of result and at the same time, the aim of the research was achieved by showed data that extract from Twitter can be used in creating a map of accident hotspot which is one of analytic purpose.

1. Introduction
Fatalities about road accidents become one of the main issues in this country that needs to solve as soon as possible because statistic records that from 2012 to 2013 the accident scene that causes the fatalities is increasing which is 32% [1]. This statistic happens because a lack of awareness about the road accident will cause fatal and lack of knowledge about the hotspot accident area. This issue takes by the installation of AES at the hotspot area to see either it remains same or decrease. The result is the number of violations reduces from 5.59% to 4.01% for a 4-year installation at Jalan Klang Lama
[2]. This result of precaution step really shows the importance to identify the accident hotspot area to perform an effective solution to reducing the number of violations in Malaysia.

Road accidents become a crucial problem in Malaysia and consequently affect the loss of many things including our life [3]. It always happened and the numbers of accident road become higher day by day. The road user never knows the hotspot area in Malaysia to make them more aware of the potential to get an accident in the hotspot area of the accident road. The road traffic accident is caused by several factors associated with the traffic system: road users, road environment and vehicles.

Data about the accident cases are difficult to get by local authority due to privacy policy. Even though data about the accident cases around Shah Alam can be get by the local authority, it still not provides the specific location of the accident cases. It only shows number of accident cases in every section in Shah Alam. To solve this problem, data from social media will be used to create a hotspot map. Data from this twitter was chosen because it has the validity of its data because it goes through the process of sharing data from many people at one time. It is also more economical data as it is obtained from social media sources such as Twitter. The data will extract by using the R programming method to get the amount of road accident and the data will be applied in GIS to produce the map. From this map, it will give other initiatives for public awareness about road accident area. It will make road user care when driving and riding, so it will decrease the amount of road accident. Twitter is relevant to choose for the source of data extraction purpose to get all the data related to the accident scene in Shah Alam. It also has 550 million active users that produce 9100 tweets every second [4]. Hence, this research aims to explore and study the extraction data from Twitter for accident-prone and map the accident hotspot area in Shah Alam.

2. Study Area
The study area is Shah Alam which is located on the west coast of Peninsular Malaysia and it is in the District of Petaling. It was surrounded by the city of Subang on the east, District of Klang on the west. Shah Alam is the capital city of the state of Selangor with the total area 290.3sq km and coordinates for Shah Alam is 3.0733° N and 101.5185° E. Shah Alam was chosen because of several factors in term of the population of people, coverage of internet and others.

![Figure 1. Study area](image)

3. Data Used and Methodology

3.1 Data Used
The basic meaning of Big Data is a huge volume of data that cannot be store and process using traditional approach within the given timeframes. For specific definition of Big data is large amount of data over the traditional management system’s processing capacity to manage and analyse the data in specific time range. There are many examples of big data but in this research, Twitter was chosen as data used.
Twitter is relevant to choose because Twitter which is a micro-blogging service that is used by over millions of people in the world. People can share their message in 140-character-long. It is because of twitter has limitation on length of message [5]. Twitter also becomes the famous platform to share the information about daily event and topic among the user and it only can be use when the users sign in into the social media app.

The data will extract in term of the tweets from Twitter by using the R programming method to get all the tweets related to road accident scene. Rstudio is programming languages and it also for the software environment to statistical computing and graphics. Based on several studies and reviews of utilising GIS for road accident analysis [6], GIS offers an intelligent tool to determine the pattern of accident-prone area. The integration of the social media data into GIS, will enable the capability to analyse data without the constraints of the data need to be physically observed or obtain from the authority [7].

3.2 Methodology
There are 5 phases to complete the objective of research, which is project planning, data collection, data processing, data analysis and results. Figure 2 shows the process of producing the location of accident hotspot area in term of coordinate and a map of hotspot accident area in Shah Alam by using Twitter data. Data processing is process of transform and manipulated the data to produce the useful data that can be used for analysis purpose. Based on this research, it takes a lot of process which is 4 stage of data processing that must be done to get the coordinate as an output.

3.3 Coding for Data Extraction
The coding for extraction data from the Twitter is needed to get the data of tweet about the accident and coordinate. This data gets from the tweets of user in Twitter which is all the tweets have their own coordinate and some of the tweets does not have coordinate because the user does not open the location on the phone when their upload the tweets. Figure 3 show about the coding for extraction data that use in extraction process. Based on the figure, user can change and manipulate the R script to fits with the data that want by the users.
3.4 Coding for Data Mining

Coding for data mining is required to ensure the good result. This process must through the two types of coding process which is data cleaning and word cloud process. Figure 4 show about coding for data mining process. This coding was combined with two process of coding in data mining in R script. Firstly, is writing the packages of tm for the data cleaning and wordcloud for word cloud process. These two packages are necessary to write to make data mining process function. R script on numbers 5 function as to read file that was extracted by using coding for data extraction. R script for number 10 until number 20 used to remove number, remove words and others. User can create an R script that has their own function to remove the unnecessary word depend on user needed. Like shown in figure below number 19 is R script to remove the “english” word. R script number 22 until 25 functions to organizes and publishes data in the form of table related to word tweets by number of tweets. Function of R script number 27 to 35 for word cloud process which is setting the font, size and colour of word show in word cloud result. User also can specify the total number of word tweets like figure below, specified 3 minimum frequency repetitions of word tweets before the word is accepted into the list of tables created by R script number 22 until 25.

Figure 3. Coding for extraction data

Figure 4. Coding for data mining process
3.5 **Twitter Search**

Twitter search is a platform provided by Twitter for users to find tweets related to information on Twitter. This Twitter search will display the icon, screen name of tweet and tweets. It also shows the retweet of tweet by other people which is user copy the tweet by other user and paste on their own account. All tweets are shown decreasing by the number of views, likes, or retweets obtained. Most people tend to use this function to find the trending tweet around the world. Twitter users also can use this Twitter search for various purposes, not only for greatest and latest information only.

![Twitter search interface](image)

Figure 5. Twitter search interface

All tweets searches will be record in the table to list the popular tweet about the accident among people in Shah Alam area. This table will show clearly the most popular road that always recorded accident cases. This data will be compared to the data from extraction process by Rstudio programming software. All of process in Twitter search is not use any Rstudio software, only the Twitter account was used to complete this process.

3.6 **Data achieving process**

Data achieving process is process of list all the popular road that recorded high accident cases around Shah Alam. This table show name of popular road and coordinate of the road. This process starts by using the geocode to find the coordinate of road because not every data that was extracted from Twitter have coordinated. That why the first step is finding the coordinate in term of latitude and longitude. The data from this table will be used to see if the data from the extraction process have same data about the hotspot road with data from twitter search. All coordinates create by using geocode in this process with be used in making the accident hotspot map. Table also show the kilometer of hotspot road and this kilometer get by using geocode.

4. **Analyses**

4.1 **Data Extraction Analysis**

This analysis needs to use the coding from Rstudio Programming. This coding process must go through several coding requirement which is Word String. Word string is used to find the required data of someone by using words that Twitter users might use. For example, in this research, words “Kemalangan” was used to get the data about accident from the Twitter users. All tweets that have made that used words “kemalangan” will be extracted as data. Next process in data extraction analysis is used of number of tweets coding for example is “n=1000”. This coding needs to be used to limit the coding tweets that will be extracted and based on my coding, I used n=1000 to get the 1000 tweets from Twitter.

4.2 **Data Mining Analysis**

Second is data mining analysis is analysis of word cloud packages to identify the trend word that used in tweets need to be done. It shows the popular trend word in term of colour and font size of the...
popular word in tweets. Figure 6 show the result of word cloud process. This process needs to go through data cleaning first, before use of word cloud packages. This process is to clean the unnecessary word that not related to main tweets for making the process of word cloud become easier. Data cleaning will clean the repetition words, similar meaning and hashtags. After all the data cleaning process has done, word cloud packages can be used for data mining analysis. The result of word cloud shows in Figure 6.

![Figure 6. Word Cloud Result](image)

### 4.3 Data Usage Analysis

Data usage analysis was used to make comparison between the data that was extracted with the search tweet in the twitter account. This process shows the data that was extracted up to date with the search tweet. It is important to make sure the data that was used up to date because to make correct map of accident hotspot area in term of the hotspot location of accident in Shah Alam. From the twitter search, there are several locations of accident was finalised become the popular location of accident in Shah Alam.

### 4.4 Data Archiving

Data archiving shows the hotspot location of accident in detail with the name of the road and coordinate of the location. It also categorizes into the number of cases occur on the specific location or road. This analysis was through process of the data extraction analysis, data mining analysis and data usage analysis before data achieving analysis can be done.

![Figure 7. Name of Road with Number of Cases](image)

### 5. Results

#### 5.1 Name of Hotspot Road with Coordinate

Figure 8 show about the name of hotspot roads with exactly coordinate in term of the latitude and
longitude. This table will show the more clearly where the location of the hotspot accident area. Based on the figure 8 show that in New Klang Valley Express (NKVE) have 2 hotspot area which is firstly is 10 km from Shah Alam to Subang with latitude 3.095636 and longitude 101.561994. Second is 13 km with coordinate 3.101481 and 101.582867. Next hotspot road is Federal Highways with 3 locations that recorded high cases of accident on this road. First is 13 km from Klang to Shah Alam with coordinate term of latitude is 3.069544 and longitude is 101.539894. Second is 14 km with coordinate which is 3.073289 for latitude and 101.548614 for longitude. Third is 17 km with latitude is 3.082189 and longitude is 101.575772. After that is North–South Expressway Central Link (ELITE) with 4 popular location of accident cases which is 0.8 km, 1.1 km Shah Alam to Seafield, KM 3.4 Seafield to Shah Alam and KM 5.2 Seafield to Shah Alam. Moreover, Persiaran Kerjaya, Glenmarie, Section U1 which is on Exit 323 from Shah Alam to Subang and Jalan Lapangan Terbang Subang with KM 4.3 from Shah Alam to Subang.

| Name of Hotspot Road                  | Coordinate          |
|---------------------------------------|---------------------|
| New Klang Valley Express (NKVE)       |                     |
| KM 10 from Shah Alam to Subang        | 3.095636 101.561994 |
| KM 13 from Subang to Shah Alam        | 3.101481 101.582867 |
| Federal Highways                      |                     |
| KM 13 Klang to Shah Alam              | 3.069544 101.539894 |
| KM 14 Klang to Shah Alam              | 3.073289 101.548614 |
| KM 17 Shah Alam to Subang             | 3.082189 101.575772 |
| North–South Expressway Central Link (ELITE) |              |
| KM 0.8 Shah Alam to Seafield          | 3.08104 101.55025  |
| KM 1.1 Shah Alam to Seafield          | 3.079316 101.551967 |
| KM 3.4 Seafield to Shah Alam          | 3.066457 101.557108 |
| KM 5.2 Seafield to Shah Alam          | 3.052775 101.563189 |
| Persiaran Kerjaya, Glenmarie, Section U1 |                      |
| Exit 323 from Shah Alam to Subang     | 3.0823 101.564197  |
| Jalan Lapangan Terbang Subang         |                     |
| KM 4.3 from Shah Alam to Subang       | 3.101706 101.580047 |

Figure 8. Name of Hotspot Road with Coordinate

5.2 A Map of Accident Hotspot Area in Shah Alam

The last output for this research is a Map of Accident Hotspot Area that shows in Figure 9. Maps show all road that recorded high accident cases around Shah Alam. The accident area with high accident cases was showed by the red colour symbols while low cases show by blue colour symbols. Data that has gone through the analysis process will be entered into Argis software to generate hotspot map. There are 5 process which is first process is copy features for copy the features from the input features to create a new class. All the geodatabase features class or shapefile will copy. Second process is integrating features for making the features coincident to maintain the integrity of shared features boundaries. This integrates process will detect the features within x, y area. Input vertices when features intersect each other and add repeated coordinate vertices for features within x, y area. Third is collect events for change the event data to weighted point data. Lastly is optimized hot spot analysis for produce the hotspot map. This process produces 11 hotspots area where 2 locations on New Klang Valley Express (NKVE), 3 locations on Federal Highways, 4 hotspot location on North–South Expressway Central Link (ELITE), 1 location on Persiaran Kerjaya, Glenmarie, Section U1 and Jalan Lapangan Terbang Subang. Map show NKVE recorded highest cases accident but along 0.5 km to 4 km, this area shows cold spot because the trends of accidents occur in these km are far from one case to another and do not occur in the same place as often.
6. Conclusion

This study shows the potential used of social media data for mapping incidents pattern. For this case, the accidents cases were chosen as the test subject. A state of art of social media data archiving and road accident analysis is also presented is this study and shows how credible and data obtained from the twitterare beyond boundary. However, a good study of data uncertainty and restriction is required to pin the strategy and the correct guidelines for users/researcher to utilise social media data not just in a more efficient manner, but also producing a highly reliable result. Significantly, the data extracted gave not only information about the occurrence but also the approximation of the event happened in radius. Hence, for this matter, another research opportunity can be seen, as to fill the gaps and how this can be improved in the future. Or, should this is not happening, perhaps researcher can outlined the standard guide for all the big data users, on the do’s and don’ts, or rules of thumb of which big data is suitable for.

Based on this research, there are several recommendations that can be suggested. First is Twitter need to give access for user to extract the data beyond the 7 days because it can make extraction process easier and will increases number of data collection. It also can make user recollect their data if data is missing or corrupted but right now if the data is missing or corrupt, user need to extend their time for data extraction process or will get low amount of data. Next is Rstudio need to enhance their software in term of geocode because the geocode only collects the tweets data around the radius that was mentioned in R script. This radius is circle area where the coordinate was stated on the centre of the circle. It makes not all the study area will be cover because the radius is in the circle. Rstudio need to enhance by provide the radius that can be set by user without create the radius in term of the circle. Last recommendation is using same crowdsource data but in difference social media data like Facebook or Instagram and different cases also. Data related to this accident also has shortcomings related to the distance the accident occurred is less accurate because social media users who share this data do not share relevant data at the scene. It takes a few kilometers to make details of the accident before it is shared on social media such as twitter. That is why this data is not very accurate at the scene.
References

[1] Zainal Abidin, A. N. S., Mohd Faudzi, S. A., Lamin, F., & Abdul Manap, A. R. (2017). *MIROS Crash Investigation and Reconstruction Annual Statistical Report 2007 – 2010* (Issue 239).

[2] H. M. (2018). Awareness Automated Safety System (AwAS) for Red Light Running: After Four Years of Its Implementation (Issue 255).

[3] Jamil, H. M. (2018). *Awareness Automated Safety System (AwAS) for Red Light Running: After Four Years of Its Implementation* (Issue 255).

[4] Shariff, S. S. R., Maad, H. A., Halim, N. N. A., & Derasit, Z. (2018). Determining hotspots of road accidents using spatial analysis. *Indonesian Journal of Electrical Engineering and Computer Science*.

[5] Kambatla, K., Kollias, G., Kumar, V., & Grama, A. (2014). Trends in big data analytics. *Journal of Parallel and Distributed Computing*.

[6] Al-sultany, G. A., & Aleqabie, H. J. (2018). *Events Tagging in Twitter Using Twitter Latent Dirichlet Allocation*.

[7] Shahzad, M. (2020). Review of road accident analysis using GIS technique. International Journal of Injury Control and Safety Promotion, 27(4), 472-481.

[8] Gutierrez-Osorio, C., & Pedraza, C. (2020). Modern data sources and techniques for analysis and forecast of road accidents: a review. *Journal of traffic and transportation engineering (English edition)*.