Exploration of Road Traffic Tweets for Congestion Monitoring

Lim Cheng Yang[1], Bhawani Selvaretnam[1], Poo Kuan Hoong[1], Ian K. T. Tan[1], Ewe Kok Howg[2], Lau Heng Kar[2]
[1]*Faculty of Computing and Informatics, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Selangor, Malaysia.*
[2]*Intel Microelectronics (M) Sdn Bhd, Bayan Lepas Free Industrial Zone, Phase 3, Halaman Kampung Jawa,*
11900 Bayan Lepas, Penang, Malaysia.
cylim.pg@gmail.com

***Abstract*— Online social network services such as Twitter and Facebook have gained popularity in recent years with continuous increase of users. This is especially true for Twitter, a popular micro-blogging service that enables users to send tweets which contain valuable data in real-time. Real-time tweets information can be used in many areas and one of the least explored areas is crowdsourcing of road traffic conditions. We have found that not many people tweet about traffic conditions; however, there are formal sources that keep their accounts updated with the latest traffic info. In this paper, we present an analysis of tweets that are related to the traffic conditions in Malaysia. Detailed analysis was conducted to understand the structures and the nature of the traffic tweets. Based on our analysis, we found that the real-time nature of the tweets is useful in reporting road traffic conditions and such information will be useful to the road-users.**

***Index Terms*— Road traffic; Twitter; Natural language processing; Data mining.**

I. Introduction

Microblogging is one of the most popular ways for interacting with friends or people in their circles in recent times [1]. Twitter is one of the popular microblogging services, founded in 2006, and currently records over 320 million active users that generate over 500 million tweets per day. Micro-blogging is a form of communication for the users to socialize with other users by posting status describing an event through the Web. Microblogging services were used to share or express their opinions, status, and disseminate real-time information. Those posts contain noisy yet valuable information for the community if it is correctly extracted. Twitter produces large volumes of data daily, contributed by their users and the data covers a large variety of fields and opinions towards a product or an event. These data can be collected and processed to be valuable information that can be used for different purposes depending on one's interest. There are multiple attempts on harvesting data from Twitter like prediction of the results of an election, prediction of earthquake and getting market feedback of a product. However, there many more variety of data that can be harvested from Twitter.

Microblogging acts as a broadcasting medium that exists in the form of blogging. It has been growing rapidly over the years and numerous microblogging sites have emerged over the years. Microblogging is somewhat similar with the traditional blog despite its content size. Like traditional blogs, it allows the users to exchange content with their audiences, however the content is significantly shorter compared to traditional blogs. Microblogging has a great impact to society and one of the significant effects is online branding [2]. However, in this paper, we will be focusing on Twitter on road traffic as the tweets' content are compact, more towards text based compared to other microblog, and the data are easier to be obtained.

Twitter is one of the popular microblogging services that is unique compared to other microblogging services. Twitter is basically a massive information network that is made up of short, 140-character messages [3]. The 140-character messages or posts are called Tweets. Twitter currently records over 500 million accounts and which 302 million accounts are active. As such, the information generated from the users are enormous. Twitter's service handled 1.6 billion search queries per day. Due to its unique Short Message Service (SMS) like feature, it has been described as "the SMS of the internet" and was one of the ten most visited websites in 2013. Unlike most of the other social networking sites like Facebook or Friendster, Twitter uses "follow" system. Twitter users can follow another user's account to view their tweets on their homepage stream or being followed by another user. The Twitter users can change the privacy settings and can make their tweets to be shown to the public or only to those that they have approved as followers. When a user followed another user, the tweets from the other user will be shown in the user's homepage. In Twitter, they have their own Twitter lingo that is derived from a common tweeting practice. Among them they are: 'RT' for retweets, '@' followed by the identified user name means mentioning the user, '#' followed by a word stands for hashtags. RT is one of the common terms one used when retweeting ones' tweets and would like to share the specific tweets to their audiences. Hashtags are special words or phrase that usually are used to classify and group the tweet into the relative tweet groups. When a user clicks on a hashtag, Twitter will automatically query for the hashtag phrase and show all the related tweets which have the same hashtags.

Crowdsourcing is a common approach to gather traffic information from the users. Map applications like Waze gathers and predicts the traffic data from the user's phone location and speed of travel. However, there are only minor attempts of extracting traffic data from microblogging sites.

Twitter users share different variety of statuses which include the traffic details, such as road closures, traffic congestions, accidents, and even speed traps. Occasionally, Twitter users will tweet about road traffic conditions when they are stuck in it, with the intention to let the users that follow them know the road traffic events that just happened. In addition, there are other sources like radio stations that broadcast road traffic information for the benefit of road users. The aim of this research is to obtain the respective road traffic states of the locations and make use of the data to convey messages to interested users. The real time traffic reports will help many people to avoid getting trapped in the traffic and able to plan their routes.

The purpose of this paper is to analyze thoroughly the structures of the traffic tweets which will be beneficial to the researchers who are interested in the traffic management. The analysis of the components of the tweets will be reported. We have collected 65,413 tweets over three-month period and these are analyzed.

This paper is organized as follows. Section II reports on the related work done using Twitter's data. Section III analyzes the nature of the traffic tweets in terms of its languages, frequency and etc. While Section IV outlines the sources of the traffic tweets and discusses its structure and the components

II. Related Works

In recent years, with the improved economy, the number of road users had been increasing. As such, traffic congestion became an unavoidable issue and it has caused the road users angst. This issue had caught the eye of the society and there emerge a varieties of method to counter this problem. One of the ways to deal with this problem is by using crowdsourcing. Waze used crowdsourcing to obtain traffic data from its users. By simply driving around with the application installed in the smartphone, Waze collects the driving data from the users and translate it into real-time traffic updates. There were other methods like planting sensors like cameras in the roads to collect traffic data. However, we realized that Twitter could also act as a sensor and the traffic tweets posted could act as a signal if it could process accurately.

Due to its enormously large potential data that can be extracted from Twitter, it draws the attention of researchers from various disciplines to extract valuable data in their respective domains they are working in. Among them, there are attempts of research on the user's behavior on Twitter, determining the current trend based on the tweets, prediction of results of a certain event and so on. As mentioned before, Twitter is a massive data network which consists of different varieties of data and could be proved useful in different field. [4] used the tweets of the targeted users to predict the election results. In their studies, they have used 104,003 political tweets that were published weeks prior to the federal election of the national parliament in Germany. They collected all the tweets that contains the names of the parties represented in German parliament or selected protruding politicians of these parties. They used an automated text analysis software to extract the sentiment of the collected tweets. The tool they used was LIWC [5] a text analysis tool developed to assess the sentiment polarity of text samples. The results they obtained were close to the actual election results.

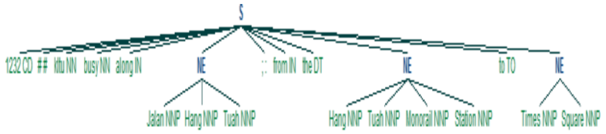

Figure 1: NER results with POS Tags from NLTK Python

III. Handling Tweets

We have identified and analyzed 24 official sources in Twitter that reports traffic conditions in Malaysia, however, only 19 of them are usable as some of them already stop reporting. We used Tweepy, an open source Twitter API to collect the tweets in JSON format from both the official sources and informal sources. We have collected 65,413 tweets over 3 months. The tweets are noisy in nature as there are inconsistent capitalization of the location names, inconsistent spacing and different abbreviation used for a same location name. We further classified the tweets into 3 categories, which is multiple location in single tweet; single location in single tweet; and multiple tweets reporting for a single location.

A. Tweet Language

A mix of different cultures often creates a mixture of languages in a sentence particularly from informal sources. According to [11], dealing with multilingual text requires additional resources like parallel corpora for every language, which makes it hard to deal with. Unlike the informal sources, the tweet of the formal sources is either written in pure English language or Malay language. Looking only at the formal sources, there are 69% of the tweets are in Malay and 31% of the tweets are in English. However, due to the rich culture in Malaysia, there might be unique terms that appeared in the tweet's text. If translation is necessary, we would suggest to use some custom machine translation program rather than conventional translation program. This will produce a better result than conventional translating software as user are able to train their own translation pairs. This would also be able to solve the problem of the variation of unique terms of the tweet. Figure 1 is a result produced when we try to examine the tweets Part of Speech (POS) tags to find out what that are the location names are commonly tagged as. From Figure 1, we can see that most of the locations are Noun phrases. As for the terms used to describe traffic conditions, they couldn't be classified into a common tag. This is because there are more than one ways to describe the traffic condition.

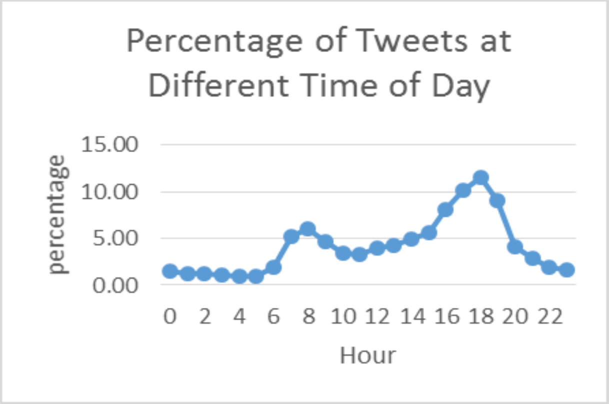

Figure 2: Percentage of Tweets averagely on different time

B. *Tweet Frequency*

We have performed time analysis of the tweets and split the tweets into their respective hours. From Figure 2, we could deduce that the peak time of the tweets are around 7 am to 9 am and hiked ridiculously around 4 pm to 7 pm. 4pm to 7pm is the time where usually the workers finish their work and started going home and 7am to 9am is the time where workers starts to travel to their workplace, these 2 time frames are the most crucial time for a worker. However, the tweets at 4 – 7 pm is significantly higher than the morning session. We deduce that this phenomenon happens as workers are more concern about how much time they need to spend to reach home after a long day's work than time needed to go to work. In terms of days, there are more tweets on the weekdays compared to the weekends. This phenomenon is mostly caused by the people who care less about the traffic when they are not working. In addition, some of the official sources doesn't tweet on certain public holidays, like Christmas.

We have sampled 813 of the traffic tweets manually and we have found that the percentage of the positive tweets are significantly lesser than the negative tweets, which is 17% and 71.7% respectively. Positive traffic states indicate that the location is clear or smooth while negative traffic states indicates the traffic is congested. It could be explained that people are more interested in the information of locations that are congested. Normally, users would assume that the locations are not congested if it is not tweeted.

C. *Tweets Filtering*

Even official sources have tweets that are unrelated to the traffic condition. Advertisement tweets and news are usually mixed in with the traffic tweets and collected together. The filtering can be done by keyword filtering, which is filter out some common words that aren't supposed to be seen in a traffic tweets e.g. "Touch & Go", "cergas", "layari" and etc. According to [8], they use conditioning to filter out the unrelated tweets. They made sure that the tweets must contain at least one location word and one verb and discarded the ones that doesn't fulfil this condition. However, we proposed that researchers should use machine learning method to filter out the tweets instead of filtering tweets by certain keywords and conditions, as the tweets from informal sources sometimes might not state the traffic condition with a clear verb. For example, the sentence "The traffic here is making me sleepy" implies that the traffic is congested but it doesn't use any clear traffic state words.

In addition, we have manually inspected the tweets and we found that the tweets can be categorized into 3 categories, which is: one tweet contains one location, one tweet contains multiple location, and multiple tweet talking about one location. It is quite common for the multiple tweet one location to happen as there are a lot of retweets that occurs in Twitter. They are either users that want to share that information or retweeted by retweet bots.

D. *Special Characters*

We have noticed that some of the Twitter's tweets contains non-ASCII characters when we try to process it. When the tweets are extracted, characters are in the character code format. These characters may or may not hold any meaning in the text. We have sampled 7k tweets and extracted a list of common non-ASCII characters that they have used in the traffic tweets. The details are tabulated in Table 1. The special characters may cause some of the process to fail if it is not handled properly. To prevent such things to happen, we suggest to change the encoding format of the tweet text before processing it.

IV. TRAFFIC TWEET ANALYSIS

We have classified the sources of the tweets into formal sources and informal sources for analyzing purposes. Formal sources are tweets from the organization that manages the roads, government, or radio stations. The tweets from the official sources are comparably more formal and have some sort of pattern to it compared to those tweets from normal users. As mentioned before, the reason the formal sources tweets about traffic besides preventing traffic congestion is to get user's attention. Thus, there will be some unrelated tweets like advertisements or news tweeted by them. We have identified 24 formal sources that reports the traffic states. Malaysia is a multi-cultural society that consist mainly of Malays, Chinese and Indians. They heavily influenced each other's daily life and mainly the language spoken.

Table 1
Special Characters and its meaning

Character Code	Actual Character	Represents
\u2013	-	To
\u2022	·	Bullets
\u2026	…	Etc.
\u2019	‘	‘
\u00b7	<tab>	Tab
\u00a0	<spacing>	Blank Space
\u00b2	Superscript 2	Power of 2
&amp	&	And

Table 2
Sample of Tweet Components

Element Type	Sample Terms
Time	10.30; 1030; 1030hrs; 1030 am
Location	Mid Valley; Templer Intersection; Bkt.Jalil, Bukit Jalil, Paradigm Mall, TPM
Directions	From…to; along; heading; dekat
State	Slow; delay; penambahan masa
Hash Tags	#kltu; #jbtu; #penangtu
URLs	https://t.co/pDSfP19Nyl
User Mentions	@kltrafficupdate

In generally, the mixture of different languages were used and this becomes a unique trait for the Malaysians. As of Malaysia's rich culture, the tweets were written in either English, Malay or mixed language. In contradictory, some of the language used by the informal sources are mixed of both languages. Despite their difference in their tweets' structure, the components in the tweet text have a certain pattern to it. In general, we have identified the common components that consists in a tweets. They are: Time, Location, Directions, State, Hashtags, URL and User Mentions. The elements in the tweets and their significance are as described as below

A. Time

The most common component that are present in the tweets is the time tagging of the traffic tweet. It is also a very useful in the research as some of the traffic tweets are retweets or a delayed tweet by the users and the time will help to identify the time of the traffic state. The time in the tweet poses as a great role as traffic condition is a spatio-temporal event, which means it have a time frame for the event to start and end. As there are retweets, by the time the tweets being retweeted there may be a possible chance that the circumstances had changed, thus we suggest that the time that is written on the tweet should be compared with the time where the tweet is written.

B. Locations

Locations can be a point, an area, a road, or a link between two areas. For example, a point location could be a roundabout, a building or an intersection point. Links are like from location A to location B, "From Sungai Arak to Paradigm Mall". Location, in fact is the most important component in the tweet as if the location was not properly mention, the tweet has no value and couldn't be used. In addition, we realized that there is a common phenomenon where the actual location names are not mentioned in a tweet, instead, they mentioned the local names of the location. This would cause confusion to whatever process that wants to perform on the tweets.

C. Directions

The keywords that are used to state the directions are like "from…to", "along", "heading" and etc. They are quite essential as there is a possibility that not both of the directions of the routes are congested.

D. State

Traffic states are the most important component in the tweet. It states whether a certain location is congested or smooth. The terms used to describe the traffic states in formal sources uses a certain clear words e.g. jam, slow moving, congested and etc. However, for informal sources, they may use words that doesn't clearly states the state of the traffic, for example: "This traffic is making me sleepy". The sentence itself implies that the traffic is congested but it doesn't state it clearly.

E. Others

Hashtags are one of the common components that are being used in microblogging nowadays. The original usage of the hashtags is to simplify the searches and grouping of a certain topic, however the hashtags now are being abused and simply used. The formal sources have a certain sets of hashtags e.g. #kltu, #jbtu, #penangtu and etc. Although they don't hold any significant meaning for the tweets for formal sources, some of the hashtags are useful in the tweets of informal sources.

URLs aren't a common component of the tweets but there are 24% of the tweets consists of URL. Among them, only a small portion of them are Facebook link, which contains extra traffic information. This is because of the limitation of the Twitter of 140 characters per tweet, it couldn't accommodate enough space for some official sources that specifies multiple locations and their traffic state. The other URLs are unrelated tweets like links to a webpage that contains either news or advertisement.

User mention is a feature of Twitter which users use to link a tweet to other users. Among the other components, user mentions are the least common and holds the least value. It could be removed and doesn't change the meaning of the tweet.

Table 3
Keywords searched and summary of results returned s

Terms Searched	Language	Description of returned results
Jam	English	Returns tweets about time or tweets about butter jam
delays	English	Returns tweets about event postponed or event delay
Slow	English	Returns tweets about slow process
Accident	English	Returns tweets about accidents all over the world
Traffic	English	Returns traffic tweets but the frequency is low
Terhalang	Malay	Returns tweets about vacation being blocked and etc.
Lancar	Malay	Returns tweets about things being/ going smooth
Kemalangan	Malay	Returns tweets about various accidents, not just related to road accident
trafik	Malay	Returns traffic tweets

User mentions are usually used in a retweeted tweets and mention the original author. In addition, there exists some "retweet bots" which they query tweets from Twitter by searching for a specific terms and retweet them in a certain period.

Table 2 show actual sample of the terms extracted from the tweets collected. From Table 2, we can see that for some cases, there are different terms that were used to mention a same location. This is due to the rich culture and dialects in Malaysia, affecting different forms of abbreviation to emerge. This issue will be further discussed in the next section.

For informal sources, we have used numerous attempts to collect the tweets. One of the way we used is to deduce a few common traffic state keywords from the official tweets we have collected and search for the tweets. However, due to the sparse nature of the tweets, there are some problem emerged when we query for the informal tweets.

There are a lot of unrelated tweets that have nothing to do with traffic conditions. As of there are many meanings when it comes to the terms, for example: jam may be defining as “traffic jam” which means the traffic congested or it could mean something like “butter jam”. In short, there exist this problem when it comes to crawling for informal tweets. The details of the problem are shown in Table 3. All the data in Table 3 were searched by running a query in Twitter website. In addition, Twitter Streaming API doesn’t allow running location tracking and keywords filtering at the same time so it is best to take note of it. From what we have observed, it seems that the informal sources are lacking in terms of numbers compared to the formal sources in the same time frame

V. Conclusion

To be able to correctly use the tweets to extract the traffic conditions in Malaysia, there are a lot of factors that needed to be taken into considerations. The rich culture in Malaysia had made the tweets harder to be process as they consist mixed languages and slang of different dialect. Thus a proper translation has to be done beforehand to ensure the data extracted are accurate. The abbreviation must be taken care of properly as it may change the meaning of the tweets. In short, our analysis shows that it is possible to extract data from the tweets to predict the traffic in Malaysia. Future work of our analysis is to draw out and implement a detailed framework to process these raw data into valuable data in order to ease the traffic congestion and help the targeted users save time and costs.

Acknowledgment

This article and the findings are based on work supported by the Collaborative Research in Engineering, Science & Technology (CREST) Funding under grant number P04C3-14. Any opinions given, findings, and conclusions or recommendations that is stated in this article are those of authors and do not necessarily reflect those of CREST.

References

[1] Zhao, D., & Rosson, M. B., “How and why people Twitter: the role that micro-blogging plays in informal communication at work,” *In Proc. international conference on Supporting group work,* ACM, 2009, pp. 243-252.

[2] Jansen, Bernard J., Mimi Zhang, Kate Sobel, and Abdur Chowdury. "Micro-blogging as online word of mouth branding." *In CHI'09 Extended Abstracts on Human Factors in Computing Systems*, ACM, 2009, pp. 3859-3864.

[3] Java, A., Song, X., Finin, T., & Tseng, B. "Why we twitter: understanding microblogging usage and communities," In *Proc. 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, ACM, 2007, pp. 56-65.

[4] A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. "Predicting elections with twitter: What 140 characters reveal about political sentiment," *ICWSM* 10 (2010), pp. 178-185.

[5] Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. “The development and psychometric properties of LIWC2007,” 2007.

[6] Sakaki, T., Okazaki, M., & Matsuo, Y., “Earthquake shakes Twitter users: real-time event detection by social sensors,” *In Proc. 19th international conference on World wide web*, ACM, 2010, pp. 851-860.

[7] Endarnoto, S. K., Pradipta, S., Nugroho, A. S., & Purnama, J., “Traffic Condition Information Extraction & Visualization from Social Media Twitter for Android Mobile Application,” *in 2011 IEEE Int. Conf. Electrical Engineering and Informatics (ICEEI)*, pp. 1-4.

[8] Wanichayapong, N., Pruthipunyaskul, W., Pattara-Atikom, W., & Chaovalit, P., “Social-based traffic information extraction and classification,” *In 2011 IEEE 11th Int. Conf. ITS Telecommunications (ITST),* pp. 107-112.

[9] Liu, M., Fu, K., Lu, C. T., Chen, G., & Wang, H., “A search and summary application for traffic events detection based on Twitter data,” *In Proc. 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, ACM, 2011, pp. 549-552.

[10] Tejaswin, P., Kumar, R., & Gupta, S., “Tweeting Traffic: Analyzing Twitter for generating real-time city traffic insights and predictions*,” In Proc. 2nd IKDD Conference on Data Sciences*, ACM, 2015, p. 9.

[11] Mihalcea, R., Banea, C., & Wiebe, J. M., “Learning multilingual subjective language via cross-lingual projections,” 2007.