How to cite this paper:

Anis Naseerah Shaik Osman, Sharifah Sakinah Syed Ahmad, & Halizah Basiron. (2017). Impact of twitter on human interaction in Zulikha, J. & N. H. Zakaria (Eds.), Proceedings of the 6th International Conference on Computing & Informatics (pp 16-22). Sintok: School of Computing.

IMPACT OF TWITTER ON HUMAN INTERACTION

Anis Naseerah Shaik Osman[1], Sharifah Sakinah Syed Ahmad[2], and Halizah Basiron[3]

[1]*Universiti Teknikal Malaysia Melaka*, annis9way@gmail.com
[2]*Universiti Teknikal Malaysia Melaka, sakinah@utem.edu.my*
[3]*Universiti Teknikal Malaysia Melaka, halizah@utem.edu.my*

ABSTRACT. The aim of this research is to examine whether Twitter impacts its users by using data science approach. A comparison between the data collection using Twitter Search API via RStudio and NodeXL has been made. After comparing the two methods, data collection using NodeXL is proven to be the better method for obtaining the sufficient information by comparing the durations of the Tweets retrieved within a fixed amount of Tweets, and limitation of the data that is allowed to be retrieved using the respective methods. The data collected from NodeXL is then used to generate word clouds by using the 'wordcloud' package in RStudio to get the insights on what the user mostly Tweet about in the period. Universiti Teknikal Malaysia Melaka (UTeM) Twitter data is taken for comparing the two methods. The ones with the most words appear in the word cloud is preferred as it gives more information. Lastly, a statistical analysis is used to visualize the activeness of the user's timeline via Tableau so that more analyses can be done on how much the user interacts in Twitter. Rapid KL is one of the main public transport operators in Malaysia where they use Twitter as their official feed for news and info. Therefore, myrapidkl's Twitter timeline is being used as an experiment to examine the data from the Tweets statistically. The usage and interaction pattern is being analysed to know the activeness of user. This research could also be used to track children, missing person and criminals activities via social network. Apart from that, companies or organisations could use this research to analyse the responses to their products and services.

Keywords: Twitter, data science, word cloud, RStudio, NodeXL, Tableau

INTRODUCTION

A lot of data can be obtained from social network as it continues to grow space around the world. These data can be exploited and be beneficial to many parties with the use of the right tools, analysis, and aggressive mindset to get the information needed (Huberman, 2008). Social networks are websites and applications that enable the registered users to construct and share their content to participate in social networking (Agarwal, 2011). There are many impacts of social network to our society (Jones, 2013) (Junco et.al, 2010). A study conducted by National Institute of Health discovers that adolescents with solid, positive direct relationships may be those most frequently using social media and social network as an additional venue to associate with their companions.

This project aims to investigate the impact of Twitter on human interaction online while using data from social network sites. Learning about the impact of social network in human interaction will not be legitimate without using data from the site itself. By getting this project done, much more information can be obtained to study their interactions online in a better way. This project could be used to study their pattern on how they use the social network and who they interact with online which enables them to be monitored and be extra cautious with the surrounding especially when it is related to cyber crime.

A few sets of Tweets from two organizations were taken and will then be simulated using word cloud and statistical analysis to verify the hypothesis of the project. A data science model is used as a guide to get information from the Twitter data. Charts are produced by the end of this project to get the visualization of the experiments executed on the data.

Three primary objectives are being discussed in this project. The first objective is to explore the suitable method for collecting Twitter data online. The right method needs to be used to obtain the actual data so that the data gathered will be sufficient to answer the objectives. Next, is to explore on data science methodology - data analytics lifecycle, using social network data. Data science is becoming more demanding each day, as it is a platform to get varieties of information that could be a good use for many associations.

The last objective is to investigate whether social network application like Twitter, which could give impact on human interaction online using real-time data, obtained from the Internet. An interaction between users and organisations could happen if they have query or different views on things, and they may discuss it on Twitter. Therefore, two organisations' Twitter data were being used to correlate on social network and its impact on human interaction online.

METHODOLOGY

Data analytics lifecycle methodology is used to analyse data using data science approach. According to Schmarzo, data analytics lifecycle is designed for data science and big data problems that act as a strategy to guide its user in solving complication (Schmarzo, 2012). Data analytics lifecycle consists of six repetitive steps as shown in Figure 1 below.

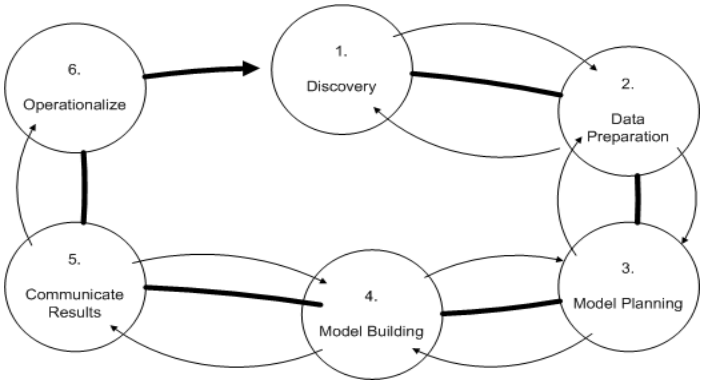

Figure 1. Data Analytics Lifecycle

This first phase is discovering the problem faced, which leads to picking the right title for the project. Research and Development Working Group did some studies and found out that the

topic of Big Data for Accessibility is one of the most wanted research topics slated for July 2015. Apart from that, Malaysian Communications and Multimedia Commission were looking for proposals in a particular study area which includes impact of new media on the population. After considering these studies along with a numerals studies done by researchers regarding on data science and social network, a title combining both of these aspects are chosen as it would give a lot of useful output to a lot of parties.

This phase focuses on obtaining, understanding and do pre-process on the data. This project collects Twitter data using two methods which are by Twitter Search API via RStudio and NodeXL. A study by Fred Morstatter shows that API performs worse than randomly sampled data, especially at low coverage when compare the statistical properties (Morstatter et.al 2013). It depends strongly on the coverage, which means that the more Retweet or likes the Tweet got, the more chances it will get in getting picked in the sampling. It is estimated that streaming API users may receive from 1% of the Tweets to over 40% of the Tweets at a time whereas an article written by Wasim stated that there are easier ways to extract Twitter data compared to Twitter Search API (Wasim, 2015). He suggested NodeXL as one of the tools available to social scientist for social network data. NodeXL tool enables user to crawl data from Twitter apart from analysing the data in table and graph form to make the visualization of the data clearer.

In NodeXL, only five columns were taken in this project that are Vertex 1, Vertex 2, Relationship, Relationship Date, and Tweets. Pre-processing in RStudio is done after the dataset is imported into it, which includes removing punctuation, eliminating the common English words (stop words), transform all letters to lowercase letter, removing the URL of a website and strip whitespace from the Tweets.

Model planning involves planning on which algorithm, model and deciding on which tools should these models be applied. Techniques and approach decided need to meet the objectives to get the right outcome. This step includes data exploration where all the data that have been pre-processed are being used as mentioned by Shah in his research (Shah, 2016). The right model needs to be chosen based on the end goal. The first part of the model planning is generating two word clouds for comparing which method in obtaining the Twitter data is better. After done some evaluation on word clouds, the better method is used to make a more detailed analysis by using word cloud with a longer duration and statistical analysis by using Tableau. The statistical analysis is being done to correlate the impact of Twitter on human interaction.

Model building is the phase where the execution of the previous plan is taken place. The data or inputs need to be sufficient, the model used is accurate to meet the goal, and the environment used the best that is to get the problem or question solved efficiently (EMC Corporation, 2014). Two algorithms are utilized in this project using three tools. The first algorithm is word cloud algorithm (Kodali, 2015). Word cloud is an image composed of words or term on a particular subject where the size of each word indicates its frequency in the subject. The Tweets that are being utilised in the making of this word cloud are from both Twitter Search API and Tweets extracted using NodeXL. Tweets from these two tools are used to compare which one is the better tool that could collect Tweets sufficiently by analyzing the results in RStudio. Statistical analysis is the next algorithm used in this project. Statistical is the research of the collection, analysis, presentation, interpretation, and organization of data. The information taken from NodeXL will be employed in the making of statistical analysis in Tableau.

RESULT AND DISCUSSION

In order to compare two of these methods, the same Twitter user's page with the same amount of Tweets is collected. Tweets are gathered from NodeXL first. The Twitter user page used is UTeM_my which is the official UTeM Twitter page. There are 114 data collected in NodeXL. The same amount of data will then be used to collect data from Twitter API. Word cloud is then being created to compare which method produced the more information. The results are shown as below.

Figure 2: Word cloud generated using Twitter Search API (left) and NodeXL (right)

Figure 2 shows the results obtained in getting the word cloud by using Twitter Search API and NodeXL. Data collected by Twitter API are from 30-06-2016 to 08-08-2016 (40 days), whereas data collected from NodeXL are from 30-07-2016 to 08-08-2016 (10 days). All settings for word cloud creation for both methods are same as one another. However, the word cloud using NodeXL is more informational as there are more words in the word cloud. A table has been created in comparing the two methods in crawling Twitter data. The result of the comparisons is shown in Table 1.

Table 1. The Overall Differences Between Nodexl And Twitter Search API

Twitter Search API	Characteristics	NodeXL
Longer duration	Durations	Shorter duration
The more the data requested, the longer the duration is which means, the fewer data obtained	Limitations	Limited to either around 1000 data or 10 days depends on whichever comes first
Only one Retweeted Tweets are included in the dataset	Advantage	All Tweets within the limitation amount are taken
-Only samples are available based on favourites and amount of Retweets - Few packages installation needed in R	Disadvantage	-Retweeted Tweets need to be removed manually

The next result that can be discussed is how the impact of Twitter on human interaction can be seen and analysed in this project. These parts are divided into two sections where the first part focuses on data cloud to know what topics or words that are being frequently used when human

interact with an organisation. The second part is by using Tableau to get the statistical form of information to know when is the most active time for people to interact via an organisation Twitter.

Tweets throughout the month of May were collected, and Retweets were removed to create a word cloud. The word cloud will then be analysed on the interaction between an organisation with users. The word cloud in Figure 2 shows the result of a month's worth of Tweets from RapidKL. There are about 222 words in the word cloud above from more than 4000 Tweets collected. The words like "lrt", "bus" and "bas" are visible as RapidKL provide the services of public transportation. Apart from that, most words in the word cloud above show the place's name such as "kuala" "puchong", and " klang". It shows that RapidKL Twitter is a platform to communicate with users regarding on the locations, updates on the routes or vehicles' conditions. Words that indicate the time-related words can be seen as they are communicating on the arrival time of the public transportation based on its route. The significance of this result is that many people are tweeting about the places, time and their opinions on the services provided that could be further inspected. By creating the word cloud, RapidKL could know in general the quality of their services and how effective are their interactions with users.

Figure 2. Word cloud generated from an accumulation of data from nodexl for a month

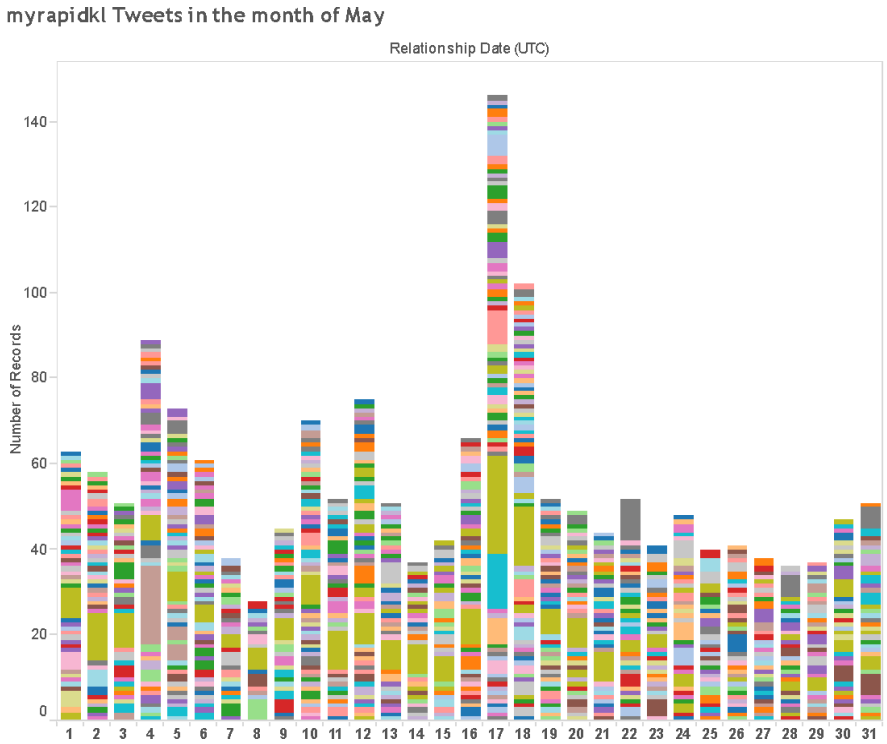

Figure 3. Number of tweets that are on myrapidkl timeline daily throughout the month of may.

The statistical analysis in this project focuses on the time-related data obtained from NodeXL for the Twitter user 'myrapidkl'. The reason why this data is being used is that the data obtained is cumulative of a month's worth of data and it can tell a lot about the pattern on when are the time where there is more interaction between the organisation and the users on the timeline. Tweets that were accumulated are from the month of May. The overall bar chart is shown in Figure 3. Vertex 1 which has different colours in Figure 3 refers to users that are on myrapidkl Twitter timeline. There are 941 users in total. The most Tweets on myrapidkl timeline are on 17th followed by 18th May. A big news was introduced on the timeline on 17th which means that there are more people talk about the matter on that day and the day after. There are overall more than 40 Tweets for each day of the month and can be said that this myrapidkl Twitter timeline is a popular page where users and the admins interact with one another. The significance of this result is that it enables to track the activeness of users on Twitter in the sense of time used to interact with a particular user and the peak time of the month or day when users interact the most. This output can be used as an interpretation of user's activeness on social media.

CONCLUSION

There are plenty of information that can be obtained by collecting data from social network. People have been using social network on a daily basis to communicate with each other, and this information could reveal a lot of answers to any queries. There is an app with a similar foundation

called Bark where it monitors a child's activity on connected accounts and alerts the parents if the app detects a potential issue like signs of depression or cyber bullying.

Both NodeXL and Twitter Search API are great free tools that could be used in collecting Tweets. Based on the experiment conducted, it is safe to say that NodeXL is better at crawling online Tweets as it takes all the data needed within a period and easier at collecting the data. It could be easily said that the more an individual spends time on a social network, the more it gives impact to them. There are most probably that if a user is being mentioned in a Tweet or others reply their Tweet, the user will continue responding and therefore they are engaged in the conversation.

REFERENCES

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R. (2011) Sentiment Analysis of Twitter Data, Proceedings of the Workshop on Languages in Social Media. Portland, Oregon. 30-38.

Ahmed, W (2015), Using Twitter as a data source: An Overview of Current Social Media Research Tools. Retrieved from http://blogs.lse.ac.uk/ impactofsocialsciences/2015/07/10/social-media-research-tools/

EMCCorporation. (2014). Data Analytics Lifecycle.

Huberman, Bernardo A., and Romero, Daniel M. and Wu, Fang. (2008). Social Networks that Matter: Twitter Under the Microscope. Peer Reviewed Journal. 14. 4-6.

Jones, H. (2013). Social Media's Affect on Human Interaction, accessed 7 January 2016. Retrieved from https://www.hastac.org/blogs/haley117/2013/06/07/social-medias-affect-human-interaction

Jones, S. (2014). Twitter and Society. Digital Formation. 89. Broadway, New York: Peter Lang Publishing. 29-33.

Junco, R., Heiberger, G., & Loken, E. (2010). The Effect of Twitter on College Student Engagement and Grades. Journal of Computer Assisted Learning. 2. 119-132.

Kodali, T. (2015). Building Wordclouds in R | R-bloggers. Retrieved from http://www.r-bloggers.com/building-wordclouds-in-r/

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is The Sample Good Enough? Comparing Data From Twitter's Streaming API with Twitter's Firehose. Arizona State University.

Peer, E., Samat, S., Brandimarte, L., & Acquisti, A. (2015). Beyond the Turk: An Empirical Comparison of Alternative Platforms for Crowdsourcing Online Behavioral Research. Social Science Research Network.

Schmarzo, B (2012), Bill's Most Excellent Data Scientist Adventure, accessed 13 May 2016, Retrieved from https://infocus.emc.com/william_schmarzo/bills-most-excellent-data-scientist-adventure/

Shah, R. (2016). Phases of Data Analytics Lifecycle, accessed 14 May 2016. Retrieved from http://quicktechie.com/cs/data-science-q-a/157-phases-of-data-analytics-lifecycle