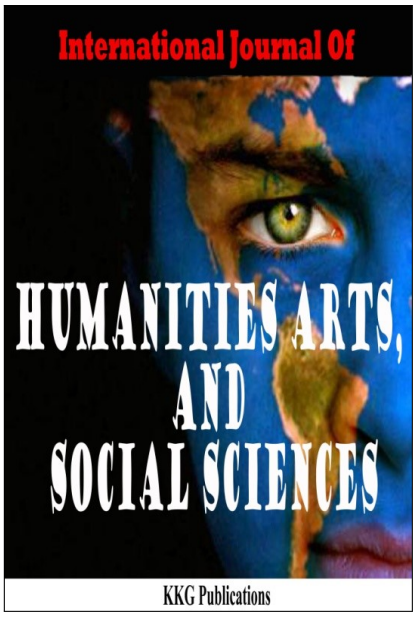


Key Knowledge Generation
Publication details, including instructions for author and subscription information:
http://kkgpublications.com/social-sciences/

Analysing Public Perceptions of International Events by using Geo-located Twitter Data

A. G. VAN DER VYVER [1], DUNCAN GILLIES [2]

[1] Monash, Johannesburg, South Africa
[2] Entelect, Johannesburg, South Africa

Published online: 15 April 2017

To cite this article: Vyver, A. G. D. V., & Gillies, D. (2017). Analysing public perceptions of international events by using geo-located twitter data. *International Journal of Humanities, Arts and Social Sciences, 3*(2), 64-70.
DOI: https://dx.doi.org/10.20469/ijhss.3.20004-2

To link to this article: http://kkgpublications.com/wp-content/uploads/2017/3/IJHSS-20004-2.pdf

KKG PUBLICATIONS

International Journal of Humanities, Arts and Social Sciences
2017, 3(2), 64-70

IJTES

ANALYSING PUBLIC PERCEPTIONS OF INTERNATIONAL EVENTS BY USING GEO-LOCATED TWITTER DATA

A. G. VAN DER VYVER [1]*, DUNCAN GILLIES [2]

[1] Monash, Johannesburg, South Africa

[2] Entelect, Johannesburg, South Africa

Keywords:
Public Percption
Social Media
Twitter
Big Data

Received: 18 December 2016
Accepted: 04 February 2017
Published: 15 April 2017

Abstract. The growth in data generated by social media platforms like Twitter provides a wealth of potential information waiting to be extracted (or mined) - traditionally with a price tag. With the recent advancements in Open Source technologies, specifically Big Data, within the Information Technology world, businesses have started to gather as much information as possible about their customers and market space. The Big Data platform, Hadoop, has become extremely proficient at managing social media data ingestion, storage and processing, due to its ability to use both structured and unstructured data. The aim of this study is to demonstrate a Big Data environment running on Open Source technologies, in order to explore the possibilities of performing geo-located sentiment analytics on Twitter data. Subsequent to this, the link between events and changes in population sentiment was investigated. In this study, an average of 47% of the total tweets ingested were geo-locatable to a country. The Open Source Big Data software was able to demonstrate the reliability of the environment, as well as identify possible limitations to having an environment setup like the one used in this study. A number of research sub-questions were answered, one of which provided information suggesting causality between an event and the change in a populations sentiment when focusing on the events specific topic on Twitter. By performing sentiment analytics on the Twitter data, potential influential users were identifiable for each use case, while allowing additional analytics to be performed and so highlight themes and trends within the data. Three use cases will be concisely addressed in this paper. The first is the Oscar Pistorius Trial (legal), the second is the FIFA World Cup of 2014 (sport), and the last one, a movie titled Maze Runner.

INTRODUCTION

Twitter is a social media platform that allows users to log on and express a thought or idea, whilst other users are allowed to view the comment and comment on it in turn. Twitter is becoming extremely popular with both the young and the old as a form of social networking, information gathering and a means to express one's opinions.

As social media become more accessible and more widely used by people, the amount of data generated is increasing dramatically. Weil (2010) indicated that Twitter users send over 350,000 tweets (messages with up to 140 characters) per minute, which, in turn, generates even more data about the message and the user this is known as metadata.

Everything from the user's location, to the number of Followers, where the person is from, the time and date of the tweet, the tweeter's profile name, the number of re-tweets and more information, is generated with each tweet sent. All these data can be analysed to gain insight about people and their trends, habits, thoughts, sentiments or opinions on a huge variety of topics. Three use cases will be addressed in this paper. Tweets relating to the Oscar Pistorius Trial (legal), the FIFA World Cup of 2014 (sport), and the third, a movie titled Maze Runner were sampled and analysed.

LITERATURE REVIEW

Twitter

Twitter is an online social networking and micro-blogging service, which allows users to write short messages or tweets as text and to send these to the system. There is a limit of 140 characters per tweet. Users are also able to follow other users of Twitter and view their tweets. Unregistered users are only able to read tweets posted, but are not able to respond to a tweet or to quote (re-tweet) the message (Boyd, Golder & Lotan, 2010). Milstein and O'Reilly (2009) argue that Twitter shares common characteristics with other communication tools, such as email, blogs, text messages or SMSs. As tweets are so short - which makes them easy to read and write - they can be "seen" as headlines. These messages are all in the public domain and anyone is able to view any person's tweets. Users have the option to subscribe to or become a follower of other users who interest them, and they are subsequently notified when that

*Corresponding author: A. G. Van Der Vyver
†Email: braam.vandervyver@monash.edu

user releases a tweet. The more appealing or interesting the tweeter (the person) is, the higher the number of people who are likely to follow them. The ability to tweet via a computer, tablet or mobile phone all add to the usability and real-time communication ability that this type of social media application tries to create. All these characteristics make Twitter appealing to a diverse range of users for a number of reasons, and it is proving a most useful communication tool for both business and personal needs Milstein and O'Reilly (2009).

Big Data

The term Big Data first appeared towards the end of the 1990's and has become a buzz word in the last few years in the Information Systems field (O'Leary, 2013; Zainuddin, Norhuda, Adeib, Mustapa & Sarijo, 2015). The term Big Data can be defined as follows: "Big Data symbolizes the aspiration to build platforms and tools to ingest, store, and analyse data that can be voluminous, diverse, and possibly fast changing" (Chaudhuri, 2012).

Sagiroglu and Sinanc (2013) describe Big Data as massive data sets that are large and varied, and that have a complex structure in terms of the difficulty of storing, processing and analysing the data in order to gain insight. The keywords 'volume', 'variety' and 'velocity' are often associated with Big Data, and can be seen as appropriate descriptors of the key characteristics of this term (Laney, 2001). However, Torres (2013) argues that there are in fact five Vs that characterise Big Data: he adds 'veracity' and 'value' to the original three. However, for purposes of this study, only the three Vs are explored, i.e. volume, velocity and variety.

Traditional methods of storing and analysing data are unable to cope with the volume, velocity or variety of data that are generated from a social media platform like Twitter (Kleiner, Stam & Pekari, 2015). Therefore, Big Data technologies have been designed and developed for use in the type of data processing activities associated with social media platforms.

This type of technology excels at processing massive volumes of data at near real-time speed and it has the ability to store both structured and unstructured data seamlessly - while being able to run on commodity hardware (Agrawal et al., 2012).

METHODOLOGY

Research Questions

The primary research question of this study is as follows:

What are the possibilities of analysing geo-located Twitter data using Open Source Big Data technologies?

In order to address this question, the following sub-questions will be asked:

Is it possible to filter Twitter data by keywords to extract relevant data from the database of Twitter created all the time.

What capabilities or limitations exist when running the entire analytics environment on Open Source technologies?

Can the data collected be geo-located and what can be deduced from a select population's sentiment?

What trends exist, if any, in Twitter data's geo-located sentiment?

What possibilities would exist for a company, in relation to a product or brand, if a company were able to perform sentiment analysis on Twitter data?

Can key social influencers be identified and what are their sentiments?

Can marketing campaigns or significant events impact a population's sentiment?

Case studies

Case Study Research Design was used for this study's methodology, as the study was an outcome-based Information Science project, whose aim was to evaluate sentiment analytics on geo-located Twitter data using Open Source Big Data environment. This study offers levels of evaluation and iteration by detailing specifics throughout the study's lifecycle and adding to the problem domain by investigating the Open Source, scalable nature of the studied task.

Quantitative research is used due to the data being collected at a numerical level and various types of post-aggregation processing being performed on the data. Population sentiments were processed, quantified and generalized, in order to formulate facts and uncover trends in the main research project.

Data Collection

Flume was used to collect data from Twitter's Streaming API. "Apache Flume is a tool/service/data ingestion mechanism for collecting, aggregating and transporting large amounts of streaming data such as log files, events (etc...) from various sources to a centralized data store" (Apache Flume an Introduction).

Flume is an Apache project that is part of the Hadoop environment that is used for data ingestion. Flume is a reliable system that is able to efficiently collect, aggregate and move data from many different sources to HDFS.

Flume was used in this study, as the data collector. Flume is able to ingest social media streams and store the data within Hadoop, where it can be analysed. Apache Flume is rapidly becoming the standard for social media problems (McClary, 2013). Therefore it was selected for the major work in this study.

It collected the data by using designated keywords that were each associated with the different use cases, as described below.

Each use case required a separate Flume agent to be set up and run, in order to collect only the relevant Twitter data. A variety of different use cases were chosen to represent a wide variety of data - from different subject matter, to different total size of data. The six selected use cases were:

- FIFA World Cup This use case ingested FIFA World Cup data, which: provided the largest data set, with close to 1 billion tweets; and showed a sample of the massive volume of Twitter data that can be processed.
- Maze Runner This use case ingested The Maze Runner movie release weekend data, which focused on a short data collection period and pulled in a small data set.
- Oscar Pistorius Trial This use case ingested all data regarding the Oscar Pistorius Trial, which provided a variety of sentiments and, specifically, a negative sentiment for a number of countries, as well as trending topics and themes.

The keywords used to collect the tweets for each use case were:

- #soccerworldcup, #footballworldcup, #fifaworldcup, #worldcup2014
- #Themazerunner, #mazerunner
- #Oscartrial, #oscarpistoriustrial, #oscarpistorius

Data collection in this way allowed each use case to have separate data segments - and therefore separate result sets when querying the data. Each use case required a separate Flume agent to ingest data, thus allowing each data workflow to run individually and minimising faults.

Sentiment Analysis

Sentiment analysis was conducted by using a data dictionary. The Data Dictionary used - also known as a lexicon Kumar, Morstatter, Zafarani and Liu (2013) and Izhar, Baharuddin, Mohamad and Wan Hasnol (2016) - was an English sentiment lexicon consisting of 6800 words and containing: known words from the English language, their part of speech (e.g. noun, verb, adjective, etc.), and a sentiment (i.e. positive, negative or neutral). A polarity was given to each sentiment (i.e. positive = +1, negative = -1 or neutral = 0) which allowed the aggregation of words, sentences and tweets to get an overall tweet polarity that was then converted back to a sentiment for readability (i.e. positive >0, negative <0 or neutral = 0). Each tweet was segmented into sentences using a full stop as the delimiter; thereafter each sentence was segmented into words, using a space character as the delimiter. Each word was mapped to the Data Dictionary, in order to apply a polarity code; if the word in the tweet could not be mapped to the Data Dictionary, a neutral polarity was given to the word. Each tweet was aggregated using the polarities; this resulted in the tweet being rated greater than, less than or equal to 0. If the tweet was rated greater than 0, the tweet was considered a positive tweet. If the tweet was rated less than 0, the tweet was considered a negative tweet. If the tweet was rated equal to 0, the tweet was considered a neutral tweet.

RESULTS AND DISCUSSION

General

The duration of data collection, the number of tweets and the portion of geo-locatable tweets for each of the use cases were as follows:

TABLE 1
Data Collection Duration

Use Case	Duration	Number of Tweets	Percentage of Geo-Locatable Tweets
FIFA World Cup	4 weeks	± 960,000,000	69.55%
The Maze Runner	3 days	± 70,000	75.25%
Oscar Pistorius Trial	3 weeks	± 180,000	47.22%

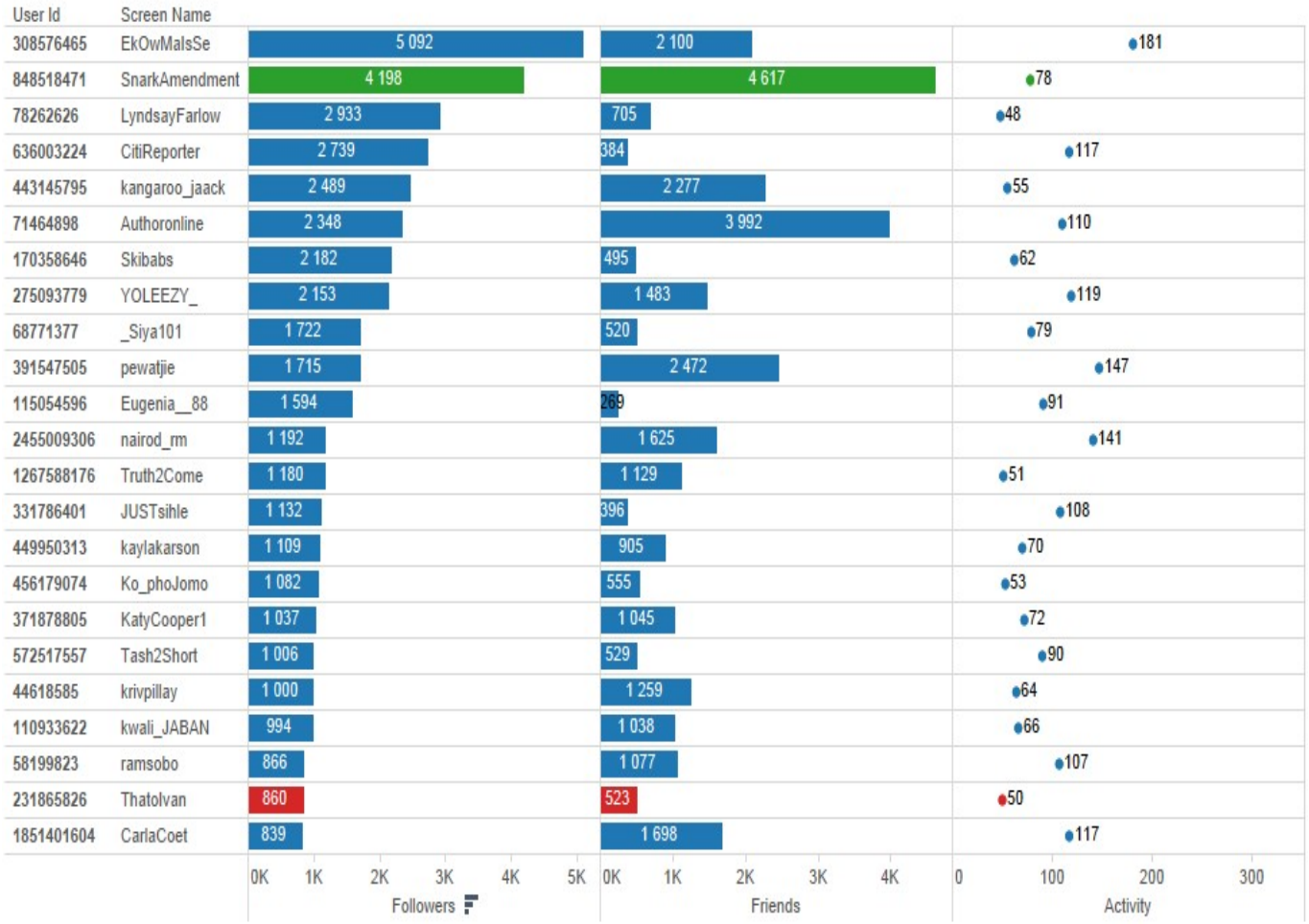
The following color-coded sentiment key was used:

FIGURE 1
Sentiment Key

Key Influential Users

Key influential users are identifiable based on the number of Followers and Friend counts that are extractable from tweet metadata. An influential user is seen as an information controller, as the user controls a certain amount of information flowing throughout Twitter and has a large reach to other users on Twitter (Kumar et al., 2013). Figure 5 shows the top influencers that were active during the data collection period for each use case, as well as the Friends count, their Followers, and their activity levels. The number of followers an influential user had indicated people's potential interest in the user as well as the user's potential influence. This translates in figure 1 top influencer's sentiment reflecting a large portion of the overall sentiment for this use case discussed below. Influential users do, however, not always serve as an accurate indicator of the general mood.

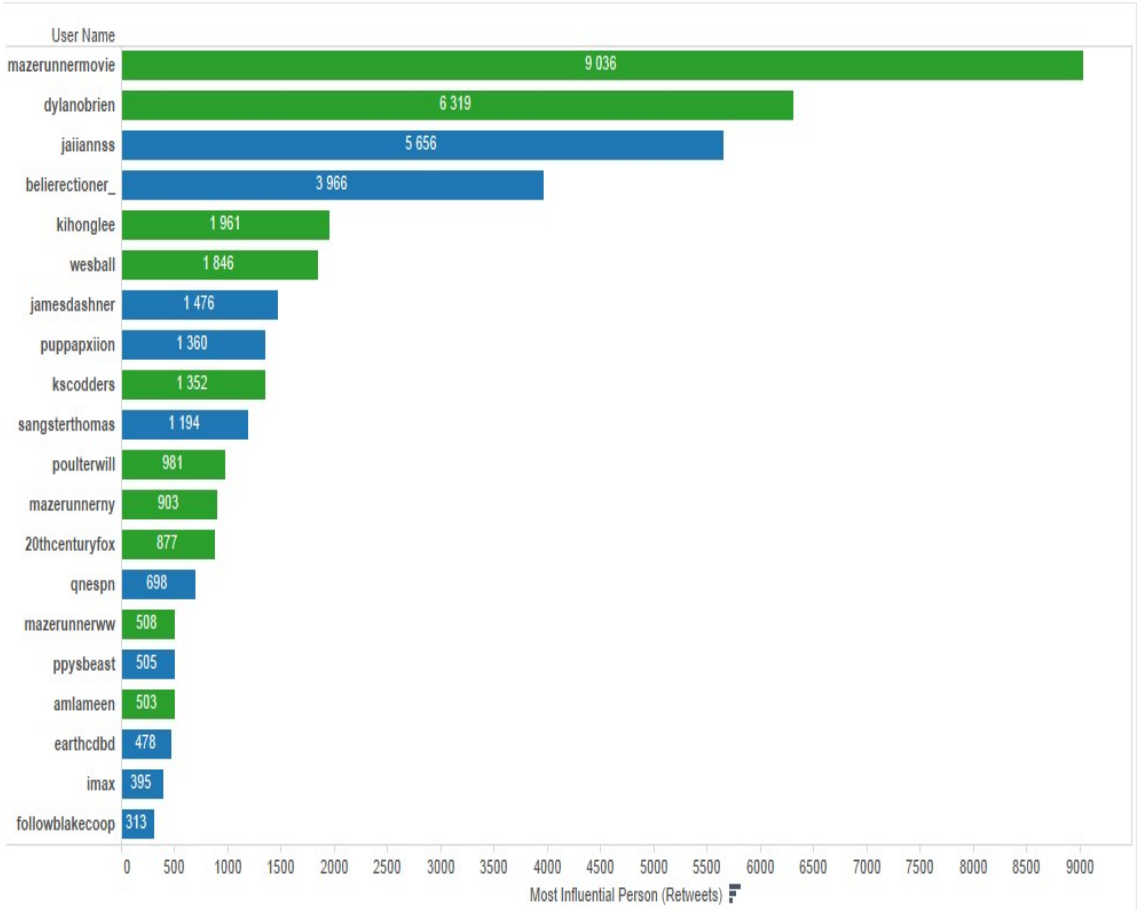
FIGURE 2
Oscar Pistorius Trial - Influential Users



Re-Tweeted Users

The researchers analysed the most re-tweeted users and their sentiment, suggesting the user that had the most influential tweets and the conversations started by these users within different use cases. Looking at the number of re-tweets a user reflects shows that people found one or more tweets from this user interesting and wanted to share these, which means there was influence or potentially a mockery. Although this item is similar to the most influential user, in this case, the results show which tweeter had the most opinions mentioned and conversations started - and not necessarily the user's ability to control the flow of information. Fig. 3 depicts the re-tweeting patterns for the most popular tweets relating to the movie, Maze Runner. Figure 3 shows the key influential tweets by the number of re-tweets per user throughout The Maze Runner use case; and the conversations generated by these users.

FIGURE 3
The Maze Runner - Most Re-Tweets



Geo-Location Maps

In this paper geo-location maps are used to depict sentiment by country. The researchers set themselves the task to determine whether or not enough tweets are geo-located to provide significant evidence of a population's mood for a country, in order to investigate any possible trends that exist within geo-location data for specific countries and/or sentiments. The data were geo-located using fields from the metadata of a tweet, and a time-zone map, in order to transform city names and places into country level detail.

FIGURE 4
FIFA World Cup - Geo-Location Map (in Millions)

Figure 4 displays the text vertically, in order to show all the values on a single map. It shows the massive volume of tweets (in the millions) regarding the FIFA World Cup, with some countries (e.g. the United States and the United Kingdom) generating close to 100 million tweets each. The top 20 countries have been selected to generate a sample indication of the geo-located tweets. All countries except one reflect a positive sentiment about the FIFA World Cup. The Oscar Pistorius case offers a glaring example of mixed fortunes in terms of the sentiments displayed by the various countries.

FIGURE 5
Oscar Pistorius Trial - Geo-Location Map

CONCLUSION

In order to answer the main research question that relates to identifying the possibilities of analysing geo-located Twitter data using Open Source Big Data technologies, a number of research sub-questions were developed and explored. Research sub-questions 1 and 2 were successfully answered, in the sense that the research study showed the ability to collect Twitter data using an Open Source environment using Topics, hash-tags, and it being able to geo-locate the tweets.

Research sub-questions 4, 5 and 7 were explored and were answered in the sense that Twitter users' opinions were identified, and the ability to geo-locate the tweets and identify trends was proved. However, this raised an important question regarding whether or not Twitter data are a true representation of a population's sentiment. This was because although the research sub-questions were answered by identifying the possibility for companies to analyse Twitter sentiment, there was no clear indication in the data that Twitter sentiment is an unbiased representation of a population's mood. Hence this result can only be seen as directional at best. The data relating to the Oscar Trial indicate that many of the countries had a negative sentiment, with the majority of use case themes also reflecting a negative sentiment. It could then be argued that, given that the Oscar Trial was negatively received by a large majority of the country, as indicated by other recognised media sources, the Twitter data may represent, or be suggestive of a country's sentiment for the topic in question because of the corroboration provided by the other recognised media sources.

Research sub-question 6 was answered by showing the ability to identify influential users who control the flow of information for the different use cases. However, the data did not show that an influential user changed another Twitter user's sentiment, due to the data not showing whether or not other specific Twitter users themselves have access to other Twitter users' tweets. What the results do suggest, however, is that future studies could focus on influential users and influential tweets, in order to show a clearer picture of the profile influencers, based on influential tweets, i.e. re-tweets, as this means that a user must view and deliberately re-tweet a message. By focusing on both these aspects and identifying the sentiment of the influential user, and that of the tweets being re-tweeted, would allow a communicator to gain a better understanding of social media presence of its product/event.

REFERENCES

Agrawal, D., Bernstein, P., Bertino, E., Davidson, S., Dayal, U., Franklin, M., ... & Jagadish, H. V. (2012). *Challenges and opportunities with big data: A community white paper developed by leading researchers across the United States*. Computing Research Association, Washington, DC, WA.

Boyd, D., Golder, S., & Lotan, G. (2010). Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. Paper presented at *43rd Hawaii International Conference on System Sciences (HICSS),* Honolulu, HI.

Chaudhuri, S. (2012). How different is big data? Paper presented at *28th International Conference on Data Engineering (ICDE),* Washington, DC, WA.

Izhar, T. A. A., Baharuddin, M. F., Mohamad, A. N. & Wan Hasnol, W. M. H. (2016). Using ontology for goal-based query to evaluate social media data. *Journal of Advances in Humanities and Social Sciences, 2*(2), 108-118.

Kleiner, B., Stam, A., & Pekari, N. (2015). *Big data for the social sciences* (FORS Working Paper No. 2015-2). FORS, Lausanne, Switzerland.

Kumar, S., Morstatter, F., Zafarani, R., & Liu, H. (2013). Whom should I follow?: Identifying relevant users during crises. Paper presented at *Proceedings of the 24th ACM Conference on Hypertext and Social Media* (pp. 139-147). Paris, France.

Laney, D. (2001). *3D management: Controlling data volume, velocity, and variety*. Retrieved from https://goo.gl/C5iHn

McClary, D. (2013). *Acquiring big data using Apache flume*. Retrieved from https://goo.gl/2qguqN

Milstein, S., & O'Reilly, T. (2009). *The twitter book*. Sebastopol, CA: O'Reilly Media.

O'Leary, D. E. (2013). Artificial intelligence and big data. *IEEE Intelligent Systems, 28*(2), 96-99.

Sagiroglu, S., & Sinanc, D. (2013). Big data: A review. Paper presented at *International Conference on Collaboration Technologies and Systems (CTS),* (pp. 42-47), San Diego, CA.

Torres, O. H. (2013). *Big data evolves*. Retrieved from https://goo.gl/CNGDI1

Weil, K. (2010). *Measuring tweets* Retrieved from https://goo.gl/A8RJsM

Zainuddin, N. A., Norhuda, I., Adeib, I. S., Mustapa, A. N., & Sarijo, S. H. (2015). Artificial neural network modeling ginger rhizome extracted using rapid expansion Super-Critical Solution (RESS) Method. *Journal of Advances in Technology and Engineering Research, 1*(1), 1-14.

— This article does not have any appendix. —