Comparison of Twitter APIs and tools for analysing Tweets related to the Ebola Virus Disease

Wasim Ahmed

Information School, University of Sheffield, Regent Court, 211 Portobello, Sheffield, South Yorkshire S1 4DP UK**.** Email: wahmed1@sheffield.ac.uk**.**

Peter A. Bath

Information School, University of Sheffield, Regent Court, 211 Portobello, Sheffield, South Yorkshire S1 4DP UK**.** Email: p.a.bath@sheffield.ac.uk**.**

Abstract

In order to obtain data via Twitter, researchers may use a variety of software or they may ask for a custom tool to be created by software developers. However, different software may use different Application Programming Interfaces that may provide varying levels of Twitter data. In this study we compare data on Tweets about the Ebola virus obtained via the Firehose, and Search APIs respectively over a 3-day period in January 2015. We found that searchers with the keyword 'Ebola' were gathering up to 79% of all tweets via the Search API using Mozdeh, and 74% of all tweets via Chorus. The complete set of tweets was 195,713 Tweets (100%) obtained via Texifter and subsequently stored in DiscoverText.

Keywords: *Twitter, API, Tools.*

Introduction

The majority of research on Twitter is centred on using two particular APIs, the Search API, and the Streaming API (Gaffney and Puschmann, 2014). Twitter's Search API provides access to tweets that have occurred, i.e., users can request tweets that match specific 'search' criteria, similar to how an individual user would conduct a search directly on Twitter. The Streaming API, however, is a push of data as tweets occur in near real-time. The Streaming API is provided through three bandwidths: the 'spritzer', 'garden-hose', and the 'Firehose' which can deliver 1%, 10% and 100% of all tweets over a given time period, respectively (Gaffney and Puschmann, 2014).

There is a lack of evidence-based research examining the amount of tweets that are provided via the Search API and the Firehose. As it may not always be possible for researchers to obtain costly Firehose data for the purposes of reliability and validity, it is important to determine how much of the Firehose sample is returned via the free API ecosystem.

Methods

Twitter data relating to the Ebola virus outbreak were retrieved via Texifter (n.d), Chorus (n.d), and Mozdeh (n.d) which use the Firehose, Search API, respectively. Data were collected from 03/01/15 to 05/01/15, the period when the British nurse, Pauline Cafferkey, was in hospital suffering from Ebola virus. Word clouds of each dataset were produced using NVivo. Word clouds are a visualization of word frequency as a weighted list, with frequently-occurring words appearing larger.

Results

We found that from 03/01/15 to 05/01/15 searchers with the keyword 'Ebola' were gathering up to 79% (155,086) of all tweets via the Search API using Mozdeh (n.d), and 74% (145,348) of all tweets via the using Chorus (n.d). The complete set of tweets was 195,713 (100%), which was obtained via Texifter and stored in DiscoverText (n.d). Figures 1 to 3, below, are the NVivo word cloud outputs for the data using these three methods.

Figure 1 – Firehose API via Texifter

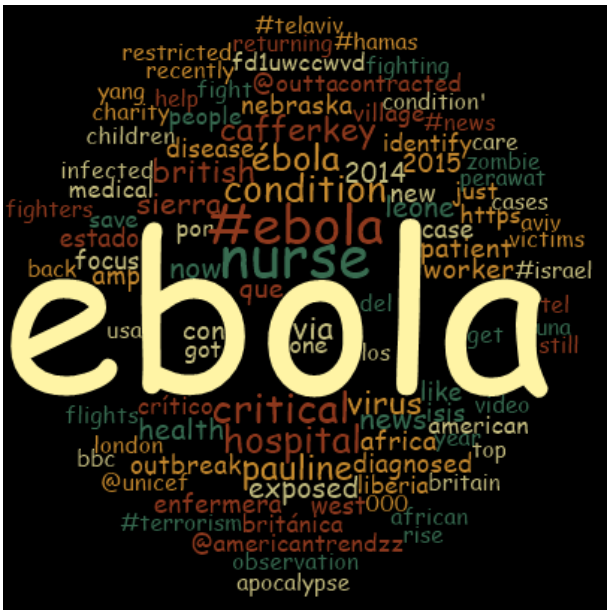


Figure 2- Search API (Mozdeh)

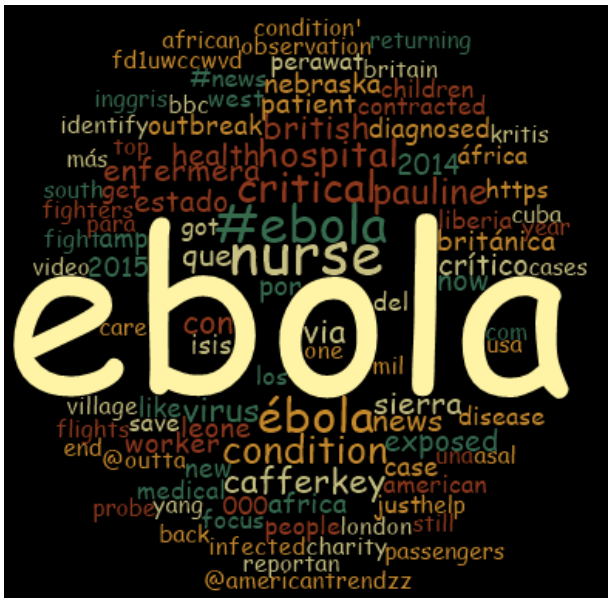
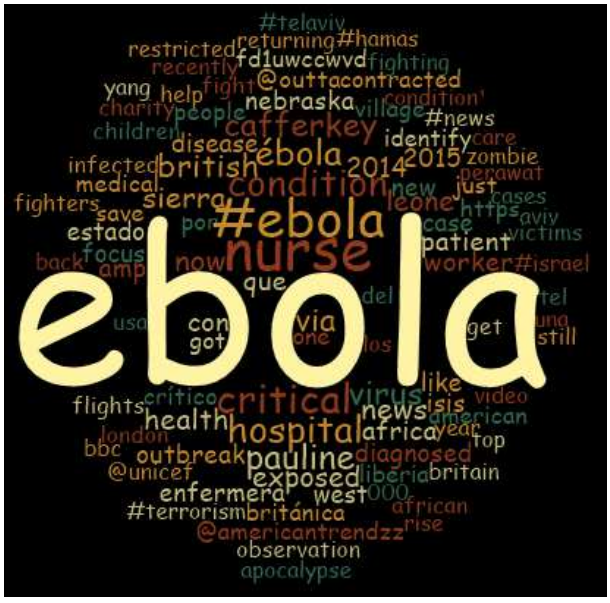


Figure 3 – Search API (Chorus)



Conclusions

These results suggest that if a limited amount of search queries are used and data is retrieved over a relatively short period of time that Twitter may well provide a sufficient amount of tweets for most research questions. However, González-Bailón et al (2014) have found that the structure of samples may be affected by both the type of API and the number of hashtags that are used to retrieve the data. Therefore, depending on the number of keywords and hashtags used, the amount of tweets retrieved is likely to vary. Further research will seek to examine datasets using more quantitative methods. Moreover, this research monitored Twitter over a 3-day period, a more comprehensive study may monitor Twitter for a significantly longer period of time; however, this may not be feasible due to the high cost of Firehose data.

References

Chorus. (n.d.). Project site for the Chorus Twitter analytics tool suite. [online] Chorusanalytics.co.uk. Retrieved from: http://chorusanalytics.co.uk/ [Last accessed 03/06/2015].

Discovertext.com, (n.d.). Home | discovertext. [online] Available at: https://www.discovertext.com/ [Last accessed 03/06/2015].

Gaffney and Puschmann. (2014). Data Collection on Twitter. In Jones, S (Eds.) *Twitter and Society* (pp.55-67). New York, NY: Peter Lang.

Mozdeh (n.d.). Mozdeh Twitter Time Series Analysis. [online] Available at: http://mozdeh.wlv.ac.uk/ [Accessed 23 Dec. 2014].

Author biography

Wasim Ahmed is a PhD student in the Information School at the University of Sheffield.

Peter Bath is a Professor of Health Informatics in the Information School at the University of Sheffield.