The Ebola epidemic on Twitter: challenges for health informatics

Wasim Ahmed[1], Peter Bath[2]

[1] Health Informatics Research Group, Information School, University of Sheffield, Sheffield, UK, wahmed1@sheffield.ac.uk

[2] Health Informatics Research Group, Information School, University of Sheffield, Sheffield, UK, p.a.bath@sheffield.ac.uk

Keywords

Twitter, infectious diseases, social media, ebola, epidemics

1. Introduction

There are a number of fundamental challenges faced when using Twitter to gather public perceptions, views, opinions, and thoughts on infectious disease outbreaks. Twitter has over 284 million active users with 500 million Tweets sent every day [1]. Studies that have analysed data from Twitter for sentiments related to infectious disease outbreaks have reported that Twitter offers an excellent way to sample large populations to help understand the public concerns of people [2,3,4]. By investigating user sentiments on Twitter, health authorities can potentially track and monitor disease patterns and thus be in a better position to disseminate health information [2, 3, 4] at a local level. Earlier studies on infectious disease outbreaks have reported the benefits of using Twitter for sentiment detection, e.g., for the Swine Flu outbreak of 2009 [2, 3, 4]. However, previous research has not made explicit some of challenges that researchers are likely to experience when conducting research on Twitter. The objective of this poster, therefore, is to outline some of the issues that are likely to be encountered.

2. Methods

The outbreak of EVD began in Guinea in December 2013 which then spread to Liberia and Sierra Leone in West Africa. The first diagnosed infection of EVD outside of Africa was reported in the U.S on September 30th 2014 [5]. This research has collected data from the free API ecosystem that Twitter provides, but has also collected firehose data, that is to say all the publically available tweets. This research [6] is using both free off the shelf tools such as Chorus [7], TAGS [8] and Mozdeh [9], as well as industry software such as DiscoverText [10].

3. Results

This research has identified the following key challenges when working with a large sample of Twitter data:

1. Information quality: popular hashtags or keywords can attract large volumes of spam, link baiting and activities from automated accounts, including bots.
2. Ethical and privacy issues: When a large volume of Twitter data is gathered it may not be possible to ask for informed consent from all users. Tweets cannot be quoted verbatim in most research papers without consent, and obtaining consent to do this can be difficult. Additionally, on the 9th of February 2015 it was announced that it would be possible to once again search the whole of Twitter through Google. When analysing Twitter data on health-related issues, there are additional sensitivities relating to people potentially sharing private information, vulnerable people and individuals in extreme circumstances.
3. Validity and reliability: Datasets gathered using Twitter cannot be shared with other researchers, as this violates Twitter's Terms of Service, therefore much of the research on Twitter cannot be reproduced by other researchers.

4. Representative sample: Twitter data is not representative of the general offline population. Therefore, certain groups and sets of individuals may be over-represented in a dataset, and others will be under-represented.
5. Feasibility: Obtaining a complete set of Twitter data is not feasible for most research studies. Moreover, maintaining firehose data through costly subscription services is also not possible or feasible.

4. Conclusions

This abstract has argued that, although Twitter is a viable platform for studying health sentiment, there are fundamental challenges involved in Twitter research. Awareness of these challenges will help understand the limitations of research on Twitter data and other social media. Future research will analyse gathered Tweets on EVD, and report the methodology employed in filtering the data.

5. Acknowledgments

The authors wish to gratefully acknowledge the contribution of Dr Farida Vis.

References

[1] Twitter, (N.D). About Twitter, Inc. | About. [Online] Available at: https://about.twitter.com/company [Accessed 1 Feb. 2015].
[2] Signorini A, Segre AM, Polgreen PM. (2011) The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic. PLoS ONE 6(5): e19467. doi:10.1371/journal.pone.0019467
[3] Chew, C., & Eysenbach, G. (2010). Pandemics in the age of Twitter: Content analysis of tweets during the 2009 H1N1 outbreak. PLOS ONE, 5(11).
[4] Szomszor, M., Kostkova, P., & St Louis, C. (2011). Twitter informatics: Tracking and understanding public reaction during the 2009 Swine Flu pandemic. In Proceedings - 2011 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2011 (Vol. 1, pp. 320–323). doi:10.1109/WI-IAT.2011.311
[5] WHO. (2015). WHO | Ebola virus disease. [ONLINE] Available at: http://www.who.int/mediacentre/factsheets/fs103/en/ [Last accessed 20/01/2015].
[6] Ahmed, W. (2015). Using Twitter to gain an insight into public views and opinions for the Ebola epidemic. [Blog] *A blog about my PhD research*. Available at: https://wasimahmed1.wordpress.com/ [Accessed 24 Mar. 2015].
[7] Mozdeh (2014). Mozdeh Twitter Time Series Analysis. [ONLINE] Available at: http://mozdeh.wlv.ac.uk/ [Accessed 23 Dec. 2014].
[9] Chorus. (2014). Project site for the Chorus Twitter analytics tool suite. [ONLINE] Chorusanalytics.co.uk. Retrieved from: http://chorusanalytics.co.uk/ [Last accessed 23/12/14].
[9] Hawksey, M. (2014). TAGS. [ONLINE] TAGS. Available at: https://tags.hawksey.info/ [Last accessed 23/12/14].
[10]Discovertext.com, (2015). Home | discovertext. [ONLINE] Available at: https://www.discovertext.com/ [Last accessed 12/02/2015].