
URL:
This version was downloaded from Northumbria Research Link: http://nrl.northumbria.ac.uk/34322/

Northumbria University has developed Northumbria Research Link (NRL) to enable users to access the University’s research output. Copyright © and moral rights for items on NRL are retained by the individual author(s) and/or other copyright owners. Single copies of full items can be reproduced, displayed or performed, and given to third parties in any format or medium for personal research or study, educational, or not-for-profit purposes without prior permission or charge, provided the authors, title and full bibliographic details are given, as well as a hyperlink and/or URL to the original metadata page. The content must not be changed in any way. Full items must not be sold commercially in any format or medium without formal permission of the copyright holder. The full policy is available online: http://nrl.northumbria.ac.uk/policies.html

This document may differ from the final, published version of the research and has been made available online in accordance with publisher policies. To read and/or cite from the published version of the research, please visit the publisher’s website (a subscription may be required.)
The Ebola epidemic on Twitter: challenges for health informatics

Wasim Ahmed¹, Peter Bath²

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 publicly 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