American Journal of Infection Control 43 (2015) 563-71

Contents lists available at ScienceDirect

American Journal of Infection Control

journal homepage: www.ajicjournal.org

Major article

What can we learn about the Ebola outbreak from tweets?

Michelle Odlum EdD, MPH, Sunmoo Yoon RN, PhD *

School of Nursing, Columbia University Medical Center, New York, NY

Key Words:
Ebola outbreak
Social media
Data mining

Background: Twitter can address the challenges of the current Ebola outbreak surveillance. The aims of this study are to demonstrate the use of Twitter as a real-time method of Ebola outbreak surveillance to monitor information spread, capture early epidemic detection, and examine content of public knowledge and attitudes.

Methods: We collected tweets mentioning Ebola in English during the early stage of the current Ebola outbreak from July 24-August 1, 2014. Our analysis for this observational study includes time series analysis with geologic visualization to observe information dissemination and content analysis using natural language processing to examine public knowledge and attitudes.

Results: A total of 42,236 tweets (16,499 unique and 25,737 retweets) mentioning Ebola were posted and disseminated to 9,362,267,048 people, 63 times higher than the initial number. Tweets started to rise in Nigeria 3-7 days prior to the official announcement of the first probable Ebola case. The topics discussed in tweets include risk factors, prevention education, disease trends, and compassion.

Conclusion: Because of the analysis of a unique Twitter dataset captured in the early stage of the current Ebola outbreak, our results provide insight into the intersection of social media and public health outbreak surveillance. Findings demonstrate the usefulness of Twitter mining to inform public health education.

The current Ebola virus disease (EVD) outbreak is of major global concern and is classified by the World Health Organization (WHO) as an international health emergency. Beginning in December 2013 in the eastern section of the Republic of Guinea,[1-3] new cases were reported regularly by the Ministries of Health in Guinea, Liberia, Sierra Leone, and recently Senegal (daily direct flights to John F. Kennedy International Airport [New York, NY], Washington Dulles International Airport [Dulles, VA], and Charles de Gaulle Airport [Roissy-en-France, France]).[4-6] With a cumulative case total of >13,567 (case) and 4,951 deaths,[7,8] the first major West African outbreak of the most virulent Zaire strain of EVD is now the largest EVD outbreak to date.[9] Local, regional, and international agencies are challenged to contain the epidemic, reduce fatalities, and allay the climate of fear.[9,10] However, ongoing disease containment and surveillance is difficult because of the current outbreak. Furthermore, in resource-limited settings, barriers to optimal public health outbreak surveillance exist.[4] With Ebola in the United States and the recent New York City diagnosed case,[11] there is valid cause for concern of spread in developed countries. In populated cities such as New York City, contamination is a sobering reality, and with its rodent population out numbering the humans, endemic Ebola is not outside the realm of possibility. Rodents are a main reservoir of viral hemorrhagic fevers. Similar to Lassa fever, another hemorrhagic disease, the mode of transmission is direct exposure to excreta of infected rats.[12] The certainty of EVD containment in the immediate future is not known. To improve compliance with measures of prevention and control, several priority actions are recommended for strengthened surveillance systems. These include the use of emerging technologies to support early warning systems for communication between agencies and the general public.[1]

* Address correspondence to Sunmoo Yoon, RN, PhD, School of Nursing, 630 W 168th St, Mail Code 6, New York, NY 10032.

E-mail address: sy2102@columbia.edu (S. Yoon).

Author contributions: Yoon and Odlum designed the study. All of the authors drafted the manuscript, critically revised the manuscript for important intellectual content, had full access to all of the data in the study, and take responsibility for the integrity of the data and the accuracy of the data analysis.

Disclaimer: This study used publically available data, and analyses meet the criterion for exemption §46.101(b)4 research, involving the collection or study of existing data, documents, records, pathologic specimens, or diagnostic specimens, if these sources are publicly available, or if the information is recorded by the investigator in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects.

Conflicts of interest: None to report.

0196-6553/$36.00 - Copyright © 2015 by the Association for Professionals in Infection Control and Epidemiology, Inc. Published by Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ajic.2015.02.023

The Department of Communicable Diseases Surveillance and Response at the WHO endorsed the conceptual framework of public health surveillance and action.[13] This framework demonstrates how public health surveillance and action relate through data information messages.[14] Public health action includes acute (epidemic-type) responses, defining a framework for both surveillance and action during emergent outbreak situations.[15] In the event of re-emerging rare infections, such as EVD, active systems for surveillance and acute response are used to halt transmission.[14] Although emergent diseases are to be viewed in relation to control strategies and national surveillance, resources are often limited. If surveillance systems are not timely, complete, efficient, or adaptable, gaps in knowledge may occur.[16] Optimal outbreak surveillance for public health action should comprehensively uses multiple modalities for data collection, analysis, and dissemination.[17] Knowledge transfer tools to support strategies for outbreak control are necessary.[15]

Social network sites (SNSs) allow users to play active roles in the reporting and dissemination of news events. Users share insights, opinions, and apprehensions, while disseminating interpretations of health events outside a public health context.[17,18] There is an increasing need to both develop and share health information which is essential in outbreak surveillance efforts. Surveillance through electronic mediums, such as the Internet, provides tremendous opportunities for public health practice.[17] Twitter, one of the most popular SNSs, is a microblogging application allowing for communication through 140 characters called tweets. Streams of tweets can contain useful information, with news events documented, shared, and discussed.[19] Twitter users interact through direct messages or solicited replies that can be largely disseminated through forwarding (retweeting), allowing for rapid and broad propagation.[18] Over 645 million registered Twitter users exist globally with a distribution of >58 million tweets daily.[20] The community of Twitter users reflects a diverse and rapidly growing global population.[21,22] Twitter is viewed as an emerging broadcast medium for information and news regarding public health events, evidenced by its usefulness during the influenza A virus subtype H1N1 pandemic planning activities.[18] Twitter's capacity for broad reach, timeliness, and low overhead has the potential to capture epidemic trends, gather information, and disseminate knowledge.[17,23] The utility of Twitter supports its potential to impact public health outbreak surveillance efforts in new and innovative ways.

Taking advantage of this unique opportunity to examine the use of a powerful SNS tool during a public health crisis, the aim of our study was to provide a snapshot of EVD-related tweets in the midst of the current outbreak to monitor trends of information spread, examine early epidemic detection, and determine public knowledge and attitudes.

METHODS

Tweet corpus

Tweets mentioning EVD were collected daily from Twitter (https://twitter.com/) via a Google Chrome–based version of NCapture (QSR International; Melbourne, Australia) from July 24-August 1, 2014. Key words used for the identification of EVD-related data included #Ebola, #EbolaOutbreak, #EbolaVirus, and #EbolaFacts. Key word selection for our data corpus was informed by Twitter search trends and suggested search functions. Data elements collected for each tweet included content, time stamp, geographic location with latitude and longitude codes from the sender's Internet Protocol (IP) address and self-identified address, user name, message type (unique or retweet), and the number of followers (number of disseminated). For example, if a celebrity sent out a tweet message, which was cited by 100 people, it would be calculated as follows: the number of posted tweets marked as 101 (unique = 1, retweet = 100), and the number of disseminated tweets is counted as 54 million in this study (dissemination = 54 million).

Trends of information spread

We investigated the trends of geographic spread of EVD information within Twitter. To evaluate EVD temporal patterns of information dissemination, the number of posted (unique and retweet) and disseminated tweets was characterized by date in an early stage of Ebola outbreak. Descriptive statistics, including the volume of posted (unique and retweet) and disseminated tweets, were traced according to geographic location. The poster's locations were indicated using an interactive, data visualization, and business intelligent software (Tableau 8.1; Tableau, Seattle, WA). Time series analysis using an exponential smoothing algorithm[24] was used to identify trends of how fast the tweets mentioning Ebola were disseminated. The trend model of the dissemination speed of tweets mentioning EVD was visualized.

Content analysis

A content analysis was conducted to capture public perceptions of EVD and to reduce noise. We used an natural language processing approach in the analysis of EVD tweet content. To identify topics of collected tweets, we cleaned symbols and Web addresses and transformed text to a vector form and N-gram and reduced the dimensionality of the volume using Notepad++ (Notepadd++: Don Ho, Paris, France Weka: University of Waikato, Hamilton, Waikato, New Zealand) and Weka software. The detailed steps of tweet cleaning, preparation, and refinement are described in the author's (S.Y.) other article.[25] The N-gram forms (unigram, bigram, and trigram) of tweet messages were clustered based on content similarities for topic detection. K-means algorithm was then applied using Weka. Clusters were visualized to summarize the detected topics.

RESULTS

Trends of information spread

A total of 42,236 tweets mentioning the recent EVD outbreak were posted (16,499 unique and 25,737 retweets) and disseminated to 9,362,267,048 people from July 24-August 1, 2014 (Fig 1). On July 24, the baseline of 382 posted tweets (128 unique and 254 retweets) was disseminated to 1,502,743 Twitter users. On July 26, with the announcement of the EVD infection of an American physician, the baseline number increased 8-fold to 3,222 posted tweets (1,574 unique and 1,648 retweets), with a 644-fold increase in disseminated tweets (967,404,925). The number of tweets decreased for 3 days and started drastically increasing on July 30 with the announcement of the Sierra Leone emergency declaration, the U.K. foreign secretary's official message, and the Peace Corp pulling volunteers.[26] After the first press announcement from the Centers for Disease Control and Prevention (CDC) on July 28,[27] the number of tweets increased 2-fold. Within 3 days of this announcement, the number of tweets continued to spread quickly with increasing news events, including the WHO's $100 million plan; the CDC's level 3 notice for Liberia, Sierra Leone, and Guinea; and the infected American physician's return with EVD. On July 31, 2014 (3 days after the first CDC announcement), EVD news items were disseminated to 4,864,972,879 Twitter users, 63 times higher than the initial number.