Journal of Medical Internet Research
ISSN 1438-8871
The leading peer-reviewed journal for health and healthcare in the Internet age
About Search Review Archive Current Issue Submit Membership Editorial Board

J Med Internet Res. Nov 2014; 16(11): e250.
PMCID: PMC4260066
Published online Nov 14, 2014. doi: 10.2196/jmir.3532

The Reliability of Tweets as a Supplementary Method of Seasonal Influenza Surveillance

Anoshé A Aslam, MPH,[1] Ming-Hsiang Tsou, PhD,[2] Brian H Spitzberg, PhD,[3] Li An, PhD,[2] J Mark Gawron, PhD,[4] Dipak K Gupta, PhD,[5] K Michael Peddecord, PhD,[1] Anna C Nagel, MPH,[1] Christopher Allen, MA,[2] Jiue-An Yang, MA,[2] and Suzanne Lindsay, PhD[1]

[1]Graduate School of Public Health, San Diego State University, San Diego, CA, United States
[2]Department of Geography, San Diego State University, San Diego, CA, United States
[3]School of Communication, San Diego State University, San Diego, CA, United States
[4]Department of Linguistics, San Diego State University, San Diego, CA, United States
[5]Department of Political Science, San Diego State University, San Diego, CA, United States

Ming-Hsiang Tsou, Department of Geography, San Diego State University, Storm Hall 313C, 5500 Campanile Drive, San Diego, CA, 92115, United States, Phone: 1 619 594 0205, Fax: 1 619 594 4938, Email: mtsou@mail.sdsu.edu.
Corresponding author.
Corresponding Author: Ming-Hsiang Tsou ; Email: mtsou@mail.sdsu.edu

Sentinel means "a thing that acts as an indicator of the presence of disease."

Abstract

Background

Existing influenza surveillance in the United States is focused on the collection of data from sentinel physicians and hospitals; however, the compilation and distribution of reports are usually delayed by up to 2 weeks. With the popularity of social media growing, the Internet is a source for syndromic surveillance due to the availability of large amounts of data. In this study, tweets, or posts of 140 characters or less, from the website Twitter were collected and analyzed for their potential as surveillance for seasonal influenza.

Objective

There were three aims: (1) to improve the correlation of tweets to sentinel-provided influenza-like illness (ILI) rates by city through filtering and a machine-learning classifier, (2) to observe correlations of tweets for emergency department ILI rates by city, and (3) to explore correlations for tweets to laboratory-confirmed influenza cases in San Diego.

Methods

Tweets containing the keyword “flu” were collected within a 17-mile radius from 11 US cities selected for population and availability of ILI data. At the end of the collection period, 159,802 tweets were used for correlation analyses with sentinel-provided ILI and emergency department ILI rates as reported by the corresponding city or county health department. Two separate methods were used to observe correlations between tweets and ILI rates: filtering the tweets by type (non-retweets, retweets, tweets with a URL, tweets without a URL), and the use of a machine-learning classifier that determined whether a tweet was “valid”, or from a user who was likely ill with the flu.

Results

Correlations varied by city but general trends were observed. Non-retweets and tweets without a URL had higher and more significant ($P<.05$) correlations than retweets and tweets with a URL. Correlations of tweets to emergency department ILI rates were higher than the correlations observed for sentinel-provided ILI for most of the cities. The machine-learning classifier yielded the highest correlations for many of the cities when using the sentinel-provided or emergency department ILI as well as the number of laboratory-confirmed influenza cases in San Diego. High correlation values (r=.93) with significance at $P<.001$ were observed for laboratory-confirmed influenza cases for most categories and tweets determined to be valid by the classifier.

Conclusions

Compared to tweet analyses in the previous influenza season, this study demonstrated increased accuracy in using Twitter as a supplementary surveillance tool for influenza as better filtering and classification methods yielded higher correlations for the 2013-2014 influenza season than those found for tweets in the previous influenza season, where emergency department ILI rates were better correlated to tweets than sentinel-provided ILI rates. Further investigations in the field would require expansion with regard to the location that the tweets are collected from, as well as the availability of more ILI data.