Contents lists available at ScienceDirect

Journal of Forensic and Legal Medicine

journal homepage: www.elsevier.com/locate/jflm

A survey of social media data analysis for physical activity surveillance

Sam Liu[*], Sean D. Young

Institute for Prediction Technology, Department of Family Medicine, David Geffen School of Medicine, University of California, Los Angeles, CA, USA

ARTICLE INFO

Article history:
Received 9 May 2016
Received in revised form
25 October 2016
Accepted 28 October 2016
Available online xxx

Keywords:
Social media data analysis
Twitter
Physical activity
Public health

ABSTRACT

Social media data can provide valuable information regarding people's behaviors and health outcomes. Previous studies have shown that social media data can be extracted to monitor and predict infectious disease outbreaks. These same approaches can be applied to other fields including physical activity research and forensic science. Social media data have the potential to provide real-time monitoring and prediction of physical activity level in a given region. This tool can be valuable to public health organizations as it can overcome the time lag in the reporting of physical activity epidemiology data faced by traditional research methods (e.g. surveys, observational studies). As a result, this tool could help public health organizations better mobilize and target physical activity interventions. The first part of this paper aims to describe current approaches (e.g. topic modeling, sentiment analysis and social network analysis) that could be used to analyze social media data to provide real-time monitoring of physical activity level. The second aim of this paper was to discuss ways to apply social media analysis to other fields such as forensic sciences and provide recommendations to further social media research.

1. Introduction

Social media technology, such as Twitter, allows users to communicate with each other by sharing short messages and website links. Users often share their thoughts, feelings and opinions on these social media platforms and as a result, social media data could be used to provide real-time monitoring of psychological and behavior outcomes that inform levels of physical activity.[1,2] A unique aspect about social media data from Twitter is that the posts are public and geo-tagged and thus, all Internet users, including health researchers, can readily access this data. Twitter usage has increased 30% from 2012 to 2014.[3] Currently 1 in 4 adults uses Twitter and usage is expected to increase steadily in the future.[3] Due to the rapid growth in social media use, these sites provide an enormous amount of data (e.g., over 500 million tweets per day on Twitter). The growing body of social media data is becoming a central part of big data research as these data can be modeled alongside other datasets (e.g. biomedical, crime rate) and used to predict outcomes from these datasets. Research has already shown that data from social media technologies can be used for novel approaches to identifying infections disease outbreaks such as influenza transmission[4] and HIV outbreaks.[5] Methods used to analyze social media data for predicting infectious disease outbreaks could be applied to physical activity research and other fields of study such as forensic science. Currently, no studies have described the application for using social media data to predict and monitor physical activity levels. Therefore, the first part of this paper was to describe current social media analysis approaches (e.g. topic modeling, sentiment analysis and social network analysis) that can be used to monitor and predict levels physical activity in real-time. The second part of this paper aim to discuss ways to apply social media analysis to other fields such as forensic sciences and provide recommendations to further social media research.

2. Part 1) methods of analyzing social media data for physical activity surveillance

Regular physical activity is associated with important health benefits and it is critical to chronic disease prevention and management.[6,7] Currently, only about 20% adults in the United States meet the recommended amount of physical activity (at least 150 min of moderate-intensity aerobic activity per week[8]). Based on the latest physical activity survey from Center for Disease Control and Prevention (CDC), the prevalence of physical inactivity varies across the United States.[9] The lack of uniformity in the rate of physical inactivity has made one of the top priorities to conduct studies and surveys to improve our understanding of population-

[*] Corresponding author. University of California, Institute for Prediction Technology, 10880 Wilshire Blvd., Suite 1800, Los Angeles, CA 90024, USA.
E-mail address: samliu@g.ucla.edu (S. Liu).

http://dx.doi.org/10.1016/j.jflm.2016.10.019
1752-928X/© 2016 Elsevier Ltd and Faculty of Forensic and Legal Medicine. All rights reserved.

level physical activity epidemiology. These studies and surveys aim to identify groups and populations who are not engaged in regular physical activity and locations of where physical inactivity occurs.[9–11] These epidemiological studies are extremely valuable in providing data that can be used to target interventions and health promotion efforts to improve physical activity level. However, these current approaches have several limitations. First, there is up to 2 years of lag time in reporting physical activity survey data. Second, it can be challenging to get people to respond to surveys resulting in a sparsity of data. Third, surveys and observational studies can be time-consuming and resource intensive to conduct. Therefore, innovative research approach, such as analyzing social data, may improve the current state of physical activity surveillance in order to help public health organizations to better target interventions aimed at promoting physical activity.

There are currently three main social media analysis approaches that can be used to monitor and predict levels of physical activity in real time.[12] These approaches are not mutually exclusive and thus can be used together to build models for monitoring and predicting physical activity. A summary of these social media analysis approaches and the statistic software analysis tools available in R (© The R Foundation for Statistical Computing) are displayed in Table 1. R is an open-source integrated suite of software that facilities data analysis and graphical display.[13] There are several R packages that have been developed for Twitter and text analysis. A detailed description on data analysis using these R packages is out of the scope of this review. However, a list of resources on how to use these tools for social media data analysis is provided.

2.1. Topic modeling

This approach allows researchers to identify the proportion of tweets related to physical activity within a given region.[12] Previous studies have reported that increased frequency of tweets for a particular topic can suggest behaviors that people are currently about to engage in.[4,5,14] For example, Young et al. (2013) reported that the proportion of tweets related to sexual activity and drug use was significantly positively associated with HIV outbreaks within a county level.[5] Similarly, Broniatowski et al. (2013) reported that tweets related to influenza symptoms were significantly associated with regions of influenza outbreaks in 2012–2013. The authors were able to build a statistical model using Twitter data to detect weekly change in influenza prevalence with 85% accuracy compared with actual CDC influenza reports.[4] Building on these methods, a binary classification model to identify the proportion of physical activity related tweets could be accomplished by identifying whether the tweets contain a word(s) discussing physical activity. Previous studies have partitioned the corpus of tweets into those that discussed physical activities and those that did not using a “dictionary” of physical activity related words.[2,15] This dictionary was created based on the physical activities mentioned in the guidelines for exercise testing published by the American College of Sports Medicine and the Center for Disease Control and Prevention (CDC).[16] The R package “TM” can be used to calculate the frequency of specific keywords in a corpus of tweets.[17] Once the proportion of physical activity related tweets are extracted, these data can then be modeled along an existing physical activity or “gold standard” dataset to examine relationships and models for predicting and monitoring physical activity levels. However, it is possible that the “dictionary” does not contain every possible physical activity descriptor, therefore, more advanced methods such as Latent Dirichlet Allocation (LDA) can also be applied to automatically identify physical activity-related tweets.[12] The R package “topicmodels” enable researchers to conduct LDA analysis.[18] This analysis tool facilitates the automatic discovery of keywords that may relate to physical activity in a corpus of tweets. However, this method still requires inputs from domain experts to confirm that the keywords extracted are relevant.

Table 1
A Summary of Twitter data Analysis Methods and Tools.

Twitter analysis methods	Analysis goals	Analysis tools using R
Topic Modeling	Identify whether the tweet is related to a particular topic (e.g. physical activity, crime).	Topicmodels[18] TM Package[17]
Sentiment analysis	Examine whether the tweet expresses positive, negative or neutral sentiment.	Plyr[20] stringr[21] RTextTools[22]
Social network analysis	Determine the social structure of a particular user.	igraph[26] Statnet[27]

It is critical to examine the construct validity and reliability of the prediction models build using social data against traditional or “gold standard” methods such as survey or observation data. Prediction models trained using social media data will need to be validated and updated continuously overtime in order to adapt to the changes in people's online behaviors, such as the use of specific slangs, emojis and hashtags. Google Flu Trends was a well-known example of using big data generated by Google search queries from 2008 to 2014 to predict influenza activity.[19] Prediction models were trained by fitting the each of the 50 million search queries generated between 2003 and 2007 to the actual influenza cases provided by Center for Disease Control (CDC) between 2003 and 2007. The prediction model was trained to select a list of top queries that provided the most accurate predictions with CDC data. The trained prediction model performed well initially but consistently over estimated the number influenza cases in later years (2011–2013). The overestimation was due to over fitting the data. Consequently, some of the search terms that were not related to the flu, such as “high school basketball”, were also strongly corrected with CDC's data by pure chance. Furthermore, the Google flu trend prediction model did not take into account changes in people's search behaviors (e.g. changes in search terms).[19] As a result, the development of Google Flu Trend provided valuable lessons for future research in this area. Predictions models build using social media data needs to be continuously validated and updated over time.

2.2. Sentiment analysis

The second method is to conduct sentiment analysis to determine whether an individual's attitude or perception towards a topic is positive, negative or neutral. The two commonly used approaches to determine whether a tweet expresses a positive or negative sentiment are the lexicon-based approach and the machine learning-based approach.[12] The lexicon-based approach determines a tweet's sentiment using a dictionary of positive and negative words. There are several lexicon dictionaries that have been created for Twitter sentiment analysis such as SentiWordNet and Hu and Liu's lexicon. A sentiment score can be calculated for each individual tweet by comparing the number of positive and negative words against the lexicon dictionary (sentiment score = positive sentiment score − negative sentiment score). A more positive sentiment score indicates a tweet has more positive words. In contrast, a more negative sentiment score means a tweet has more negative words. A score of zero means that the tweet has a neutral sentiment. The R packages “plyr”[20] and “stringr”[21] can be