Title
A survey of social media data analysis for physical activity surveillance

Authors
Liu, Sam
Young, Sean D

Publication Date
2018-07-01

DOI
10.1016/j.jflm.2016.10.019

Peer reviewed

HHS Public Access

Author manuscript

J Forensic Leg Med. Author manuscript; available in PMC 2019 July 01.

Published in final edited form as:

J Forensic Leg Med. 2018 July ; 57: 33–36. doi:10.1016/j.jflm.2016.10.019.

A survey of social media data analysis for physical activity surveillance

Sam LIU and

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

Sean D. YOUNG

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

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 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.

Keywords

Social Media data analysis; Twitter; Physical activity; public health

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 behaviour 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

Address Correspondence to: Sean D. Young, PhD, University of California, Institute for Prediction Technology, 10880 Wilshire Blvd. Suite 1800, Los Angeles, CA 90024, p: 424-346-4485, sdyoung@mednet.ucla.edu.

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 aim of this paper was to describe 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 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.

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 minutes 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-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. The first approach is to conduct topic modeling, which 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, 13]. 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 can 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, 14]. 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)[15]. It is possible that the "dictionary" does not contain every possible physical activity descriptor, therefore, more advanced methods such as topic modeling (e.g. Latent Dirichlet Allocation) or machine learning methods (e.g. logistic regression, support vector machine) can also be applied to automatically identify physical activity-related tweets[12]. However, these methods still require inputs from domain experts to confirm that the Latent Dirichlet Allocation and machine learning algorithms are extracting relevant information from tweets. Once the proportion of physical activity related tweets are extracted, these data can then be modeled along an existing physical activity dataset to examine relationships and models for predicting and monitoring physical activity levels.

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. The machine learning based approach (e.g. support vector machines, Niave Bayes) typically trains sentiment classifiers using features such as unigrams or bigrams[12]. Researchers have applied both of these sentiment analyses approaches to analyze social media data on a wide range of public health issues such as monitoring public panic during the 2009 H1N1 outbreak[16] and predicting risk for depression[13]. These sentiment analyses methods can also be applied to analyzing social data related to physical activity. A recent study of 15,000 randomly selected tweets found that people often tweeted about health-promoting physical activities (up to 35% of the tweets examined). Most of the physical activity related tweets revealed plans to do physical activity such as aerobic exercise, weight training and stretch exercises[14]. These studies suggest that it is possible to mine social media data and apply sentiment analysis to reveal an individual's attitude and perception of the various aspects of physical activity.

The third method is to conduct social network analysis to determine the influence of one's social network on the person's behavior and health outcomes[12]. A previous study has reported that the prevalence of obesity was influenced more by a person's social network (e.g. friends, family members) rather than geographic distance. Specifically, a person's chances of becoming obese can increase by 57% if he or she had a friend who became obese in a given interval[17]. Since obesity is influenced by physical inactivity, it may be possible to

predict a person's physical activity level by studying their social network as well[1, 7]. In the age of the Internet, and social media platforms such as Twitter, users are enabled to form virtual social networks that can mimic the characteristics of social networks in real-life[12]. Twitter users can declare the people they are interested in following, in which case they get notified when that person has tweeted. The user who is being followed by other users does not have to reciprocate by following them back, which makes the links of the Twitter social network to be directional[12]. These online social network data enable researchers to quantify relationships and their impact on behaviors. A recent study examining the influence of students' Twitter social network on student interaction revealed that students tend to carry out more Twitter interactions with those of similar levels of academic achievement[18]. These results help generate the hypothesis that individuals who are more physically active may belong to a social network that is more physically active. Overall, social network data from Twitter or other social media platforms offer an exciting new area of research that may provide a rich source of data for studying dynamics of individual and group behavior for engaging in physical activity.

Recommendations to advance the field of social media research

Social media-based data have the potential to not only be used as a tool for physical activity surveillance and monitoring, but also be applied to fields such as forensic science. For example, using topic modeling, geo-tagged tweets, and sentiment analysis, social media data may be used to build tools to monitor and predict crime rates, and drug trafficking activates with in a region. Analyzing social media data may help identify individuals who are victims of violence. This may be done by analyzing topics of conversation and sudden changes in sentiment or social media use. Social media data may also provide useful information (e.g. personality, social network, last known location, personal interest) to forensic investigators to help find missing individuals.

In order to advance the field of social media research and apply to other domains such as physical activity research or forensic science, an interdisciplinary team is required. Current methods for extracting social media data require several steps that draw expertise from different fields. Computer scientists and engineers are required to build the software that may be used to extract and store relevant social media data. Domain experts from physical activity or forensic science are needed to help understand the relevant information that should be extracted from social media data (e.g. people's attitude, motivation). Statisticians and epidemiologists are valuable in helping to build sophisticated models that use these social media data to build prediction models. Collaboration across multiple fields of study is a key step to further this field of research.

Additionally, it is important to consider previously established theories and frameworks to guide the development process of surveillance tools using social media data. For instance, ecological model of health can help explain factors influencing physical activity not only at the individual (e.g. beliefs, motivation) level, but also factors related to social (e.g. social support, cultural norms), environmental (e.g. walkability, crime, weather) and policy level (e.g. Education and schools, parks and recreation sector)[11]. Criminology theories may help forensic scientists understand reasons individuals commit a crime[19]. Therefore, these

theories can help guide researchers to identify the types of information that needs to be extracted from social media data and additional data sources (e.g. environmental data, crime rate) in order to build effective prediction and surveillance tools.

It is also important to understand the limitations of social media-based research and to find methods to address these limitations in order to advance the field. A current challenge facing social media research is data accessibility and whether the data can be generalized[20]. Not all social media platforms (e.g. Facebook, Twitter) enable researchers to access user data. The characteristics of individuals that use each platform can also differ, resulting in potential sampling bias. For example, Twitter is fairly well distributed across gender, income and education, but Twitter users tend to be younger and more racially diverse compared with the overall population of Internet users[3]. The method of downloading social media data can also influence the generalizability of the data[20]. For instance, Twitter's application program interface can be used to retrieve up to 1% of all tweets but there is no assurance of a random or representative sample. Alternative methods to download Twitter's full data stream (the firehouse method) exist but this can be extremely expensive and a barrier to access for many academic researchers[20].

Finally, as this emerging field evolves, there is a need to develop a standardized protocol and guidelines for cleaning and analyzing social media data. There are currently multiple methods that can be used to identify search terms, sample data, data cleaning and analysis[12]. In order to develop a standardized method for social media analysis, it is important for researchers to document and publish all methodologies for other investigators to learn and further this field of research.

Conclusion

Surveillance and monitoring of changes in people's behavior and outcomes can help inform government agencies and organizations mobilize appropriate interventions. Social media data has the potential to provide real-time surveillance and prediction of physical activity levels. Current social media analyses approaches that could be used to monitor and predict physical activity levels in real-time include topic modeling, sentiment analysis and network analysis. These approaches can also be applied to the field of forensic science such as building tools to monitor and predict crime rates, and drug trafficking activates with in a region and identifying individuals who are victims of violence. Recommendations to further this field of research include: 1) engaging in inter-disciplinary collaboration and research, 2) taking consideration of previously established scientific theories and frameworks to guide the development process, 3) understanding the limitations of social media data in order to find methods to overcome them, and 4) developing a standardized protocol and guidelines for cleaning and analyzing social media data.

REFERENCE:

1. Young SD, Jaganath D. Feasibility of using social networking technologies for health research among men who have sex with men: a mixed methods study. Am J Mens Health. 2014 1;8(1):6–14. PubMed PMID: Pubmed Central PMCID: PMC3879119. [PubMed: 23407600]

2. Gore RJ, Diallo S, Padilla J. You Are What You Tweet: Connecting the Geographic Variation in America's Obesity Rate to Twitter Content. PLoS One. 2015;10(9):e0133505 PubMed PMID: Pubmed Central PMCID: PMC4557976. [PubMed: 26332588]

3. Social Media Update 2015: Pew Research Center's Internet & American Life Project; 2015. Available from: http://www.pewinternet.org/2015/08/19/the-demographics-of-social-media-users/?utm_expid=53098246-2.Lly4CFSVQG2lphsg-KopIg.0&utm_referrer=https%3A%2F%2Fwww.google.com%2F.

4. Broniatowski DA, Paul MJ, Dredze M. National and local influenza surveillance through Twitter: an analysis of the 2012–2013 influenza epidemic. PLoS One. 2013;8(12):e83672 PubMed PMID: Pubmed Central PMCID: PMC3857320. [PubMed: 24349542]

5. Young SD, Rivers C, Lewis B. Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes. Prev Med. 2014 6;63:112–5. PubMed PMID: Pubmed Central PMCID: PMC4031268. [PubMed: 24513169]

6. Warburton DE, Nicol CW, Bredin SS. Health benefits of physical activity: the evidence. CMAJ. 2006 3 14;174(6):801–9. PubMed PMID: Pubmed Central PMCID: PMC1402378. [PubMed: 16534088]

7. Haskell WL, Lee IM, Pate RR, Powell KE, Blair SN, Franklin BA, et al. Physical activity and public health: updated recommendation for adults from the American College of Sports Medicine and the American Heart Association. Med Sci Sports Exerc. 2007 8;39(8):1423–34. PubMed PMID: . [PubMed: 17762377]

8. Committee. PAGA. Physical Activity Guidelines Advisory Committee Report. Washington, DC: U.S. Department of Health and Human Services, 2008.

9. Prevention CfDCa. Facts about Physical Activity [April 8, 2016]. Available from: http://www.cdc.gov/physicalactivity/data/facts.htm.

10. Caspersen CJ, Pereira MA, Curran KM. Changes in physical activity patterns in the United States, by sex and cross-sectional age. Med Sci Sports Exerc. 2000 9;32(9):1601–9. PubMed PMID: . [PubMed: 10994912]

11. Trost SG, Owen N, Bauman AE, Sallis JF, Brown W. Correlates of adults' participation in physical activity: review and update. Med Sci Sports Exerc. 2002 12;34(12):1996–2001. PubMed PMID: . [PubMed: 12471307]

12. Shamanth K, Morstatter F, Liu H. Analyzing Twitter Data. New York: Springer 2014.

13. De Choudhury MG M; Counts S; Horvitz E Predicting Depression via Social Media. In: Research M, editor. Association for the Advancement of Artificial Intelligence 2013.

14. Kendall L, Hartzler A, Klasnja P, Pratt W. Descriptive analysis of physical activity conversations on Twitter CHI'11 Extended Abstracts on Human Factors in Computing Systems: ACM; 2011 p. 1555–60.

15. Uo Health. Physical activity and health: a report of the Surgeon General. DIANE Publishing, 1996.

16. Chew C, Eysenbach G. Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. PLoS One. 2010;5(11):e14118 PubMed PMID: Pubmed Central PMCID: PMC2993925. [PubMed: 21124761]

17. Christakis NA, Fowler JH. The spread of obesity in a large social network over 32 years. N Engl J Med. 2007 7 26;357(4):370–9. PubMed PMID: . [PubMed: 17652652]

18. Stepanyan K, Borau K, Ullrich C. A social network analysis perspective on student interaction within the twitter microblogging environment In Advanced Learning Technologies (ICALT). IEEE 2010 p. 70–2.

19. Hagan FE. Introduction to criminology: Theories, methods, and criminal behavior: Sage Publications.; 2012.

20. Kim AE, Hansen HM, Murphy J, Richards AK, Duke J, Allen JA. Methodological considerations in analyzing Twitter data. J Natl Cancer Inst Monogr. 2013 12;2013(47):140–6. PubMed PMID: . [PubMed: 24395983]