HHS Public Access

Author manuscript

Trends Microbiol. Author manuscript; available in PMC 2015 November 01.

Published in final edited form as:

Trends Microbiol. 2014 November ; 22(11): 601–602. doi:10.1016/j.tim.2014.08.004.

Behavioral insights on big data: using social media for predicting biomedical outcomes

Sean D. Young, PhD, MS[1]

[1]Center for Digital Behavior, Department of Family Medicine, University of California, Los Angeles, Los Angeles, CA, USA

Abstract

Social media "big data" can provide valuable insights about people's behaviors, such as their likelihood of engaging in risk behaviors or contracting a disease. Although in its infancy, advancing this research provides the promise of predicting health-related behaviors to promptly prepare for and respond to public health emergencies and epidemics.

Keywords

Social media; behavioral insights; big data; prediction

Social media technologies have rapidly emerged and become a sustainable necessity in daily life. As quickly as social media arrived, researchers and corporations have realized the value in studying the data from these technologies. This has been particularly true in health and medicine, where social media data provide promise for remote monitoring and surveillance of risk behaviors and disease outbreaks [1–3]. Recent press attention has focused on the potentially unethical treatment and risks associated with social media-based research (http://venturebeat.com/2014/07/06/why-facebooks-user-manipulation-research-study-is-ethically-troubling), making it important to describe how research using social media data can be advanced and provide examples of the potential benefits of this work.

This manuscript provides an overview on how social media data can contribute to the emerging field of 'big data' science, describes current approaches for using social media to monitor and predict health behaviors and disease outbreaks, and provides recommendations on tools and approaches needed to further this field.

The field of big data science is a recently emerging interdisciplinary field bridging researchers in areas as broad as computer science, statistics, genetics, public health/medicine, and the social sciences and humanities (http://www.economist.com/node/15557443). Although there is no clear definition, big data science is typically described as involving datasets characterized by enormity and complexity, as opposed to smaller scale data sets where memory-rich technologies are not needed for processing. Big data science is important because technologies, such as mobile phones, wearable devices, and diagnostic

Address Correspondence to: Sean Young, PhD, MS, 10880 Wilshire Blvd. Suite 1800, Los Angeles, CA 90024, 310-794-8530, youngsean@ucla.edu.

tests, have become increasingly affordable and prevalent, providing large datasets available for merged analyses, including health and medical datasets, genomics data, and social media and technology use data [4]. For example, the Human Genome Project (HGP), completed in 2003, was an international collaboration to sequence the base pairs in the human genome that provided an enormous amount of data (http://report.nih.gov/NIHfactsheets/ViewFactSheet.aspx?csid=45&key=H#H). The HGP was the result of a combined effort, leading to a consolidated dataset that has since been able to be combined with other large datasets for use in areas such as genomics research.

Big data may contain not only relational and structured data (as found in many medical and genetics datasets that use values derived from quantitative measurements), but also unstructured (e.g., free text) data from qualitative assessments and social media conversations [5]. For example, social media sites and search engines can be used to collect unstructured posts, messages, searches, and updates that provide information about users. Social media, such as Facebook and Twitter, allows users to easily and freely communicate with each other by sharing pictures, short messages, website links, and other multi-media communication. Social media sites are becoming an integral component of big data research as a result of the high engagement of social media users (e.g., over 500 million tweets per day on Twitter (http://www.sec.gov/Archives/edgar/data/1418091/000119312513390321/d564001ds1.htm); 2.7 billion likes per day on Facebook (http://gizmodo.com/5937143/what-facebook-deals-with-everyday-27-billion-likes-300-million-photos-uploaded-and-500-terabytes-of-data). These data can be modeled alongside other biomedical datasets and used to predict biomedical outcomes.

Researchers have recently discovered that people's interactions on social media technologies can be analyzed to provide psychological information about attitudes and behaviors, including health-related behaviors [6,7]. Social media users are becoming increasingly comfortable publicly sharing many types of information, including personal stories and health information, providing data that can be extracted, categorized as psychological and behavioral data, and used for analysis in health research. For example, an intervention using social media for HIV prevention found that African American and Latino men who have sex with men (MSM) shared a tremendous amount of personal information through social media, including information on when or whether they had 'come out' and disclosed their status of having sex with other men; experiences of having been homeless; and stigmatization they experienced as a result of being minority MSM [8]. These insights were valuable in showing that people who discussed HIV prevention topics on social media were more likely to request an HIV test in the future, suggesting that policymakers and public health departments can use social media data to predict, prepare for, and respond to health-related events.

Social media data can also be a potential epidemiological tool to monitor risk behaviors and predict disease outbreaks. For example, influenza researchers have used data from Twitter to predict trends in influenza transmission [1]. Broniatowski *et al.* developed an influenza infection detection algorithm to extract and filter tweets associated with influenza during the 2012–13 influenza season. When comparing tweets to actual influenza reports from the

Centers for Disease Control and Prevention, they found their approach detected the weekly change in influenza prevalence with 85% accuracy [1].

Although this work on influenza used lagging indicators of disease (i.e., tweets about symptoms of influenza, suggesting that people had already contracted the virus), researchers working on HIV prevention have extracted psychological and behavioral characteristics from tweets in an attempt to predict behavior and disease transmission. This work showed that (i) tweets can be extracted and identified to suggest that people are currently, or about to engage in sexual- and drug-related risk behaviors, (ii) tweets suggesting the occurrence of these behaviors can be mapped to indicate their origin, and (iii) that these data can be merged and modeled alongside US statistics on actual HIV cases [3]. Results from this study found a significant positive relationship between United States county-level HIV cases and counties with tweets suggesting the occurrence of sexual risk behaviors, controlling for socio-economic status measurements. These results suggest that behavioral health characteristics might be able to be extracted from social media and used for predicting behavior and diseases.

Although social media-based big data methods are feasible and can lead to potentially groundbreaking tools for public health monitoring and interventions, they require (i) the presence of interdisciplinary teams and approaches, and (ii) the availability of large and frequently updated datasets. An interdisciplinary team is needed to conduct research using social media for biomedical big data research because this translational approach integrates the fields of psychology, business, computer science/engineering, and medicine, as well as specific expertise in the field of the main outcomes that will be analyzed (e.g., infectious disease, genomics, cell biology, or cardiovascular disease). For instance, to access social media data, a collaborator with expertise in computer science/engineering must be available to develop the infrastructure for collecting and storing data. In order to characterize and label the free text from people's social media conversations (e.g., labeling tweets by whether they suggest a person will engage in health-related behaviors), specific search terms and models (e.g., natural language processing) need to be extracted and developed with the help of a field/domain expert. For example, a behavioral/social psychologist would be needed if the goal is to extract and label text related to attitudes and behaviors; an anthropologist might be solicited if the goal is to extract text related to cultural representations; an epidemiologist would be valuable if populations and disease statistics were to be extracted from the free text; and a basic scientist or geneticist would be needed to extract and label text related to their fields.

To develop methods of using social media to predict risk behaviors and disease, frequent updates of both social media and biomedical data are needed. Social media data can be queried, or requested, practically in real time; however, for epidemiologic and biomedical data, there is often a delay in time between when people contract a virus or disease and the release of these data. For HIV data, this is particularly an issue, as at the time of this manuscript, the most recent country-level data broadly available for researchers were collected over 5 years ago. To improve these models, this delay must be reduced and biomedical outcomes must be frequently updated. Providing frequent and repeated updates of biomedical data can help to increase both the reliability and validity of the analysis and

provide additional data for testing. For example, although research on Google Trends had shown that Google searches of influenza symptoms can be used to predict outbreaks of the flu, the validity of approaches that use big data to predict disease outcomes has recently been questioned [9]. Because a large amount of data is available on both Google search terms and frequently updated influenza data, these models can be refined and improved.

Social media data are quickly influencing big data research and becoming one of the mainstream tools used in this emerging field. Because these technologies provide rich psychological data and people are freely willing to share personal health information on social media, these technologies will continue to be monitored and explored for their potential in predicting health-related attitudes and behaviors. Understanding the limitations of social media-based big data research (e.g., validity of data, missing data, observational data, representativeness of sample) and methods for how to address these limitations will improve the value of this research in monitoring health behaviors and disease outbreaks. Establishing and documenting methods of using social media in big data research is important so that social media data can more broadly impact fields outside of public health and the social sciences, such as the basic sciences and genomics, where these approaches have not yet been systematically studied.

Acknowledgments

Funding support provided by the National Institute of Mental Health (NIMH).

References

1. Broniatowski DA, et al. National and local influenza surveillance through Twitter: an analysis of the 2012–2013 influenza epidemic. PloS One. 2013; 8:e83672. [PubMed: 24349542]
2. Chew C, Eysenbach G. Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. PloS One. 2010; 5:e14118. [PubMed: 21124761]
3. Young SD, et al. Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes. Prev Med. 201410.1016/j.ypmed.2014.01.024
4. Marx V. Biology: The big challenges of big data. Nature. 2013; 498:255–260. [PubMed: 23765498]
5. Murdoch TB, Detsky AS. THe inevitable application of big data to health care. JAMA. 2013; 309:1351–1352. [PubMed: 23549579]
6. Myslín M, et al. Using twitter to examine smoking behavior and perceptions of emerging tobacco products. J Med Internet Res. 2013; 15:e174. [PubMed: 23989137]
7. Young S, et al. Extrapolating psychological insights from Facebook profiles: a study of religion and relationship status. Cyberpsychology Behav Impact Internet Multimed Virtual Real Behav Soc. 2009; 12:347–350.
8. Young SD, Jaganath D. Online social networking for HIV education and prevention: a mixed-methods analysis. Sex Transm Dis. 2013; 40:162–167. [PubMed: 23324979]
9. Lazer D, et al. Big data. The parable of Google Flu: traps in big data analysis. Science. 2014; 343:1203–1205. [PubMed: 24626916]