Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Psychology

Research opportunities at the intersection of social media and survey data

Emma S Spiro

This article reviews literature related to the use of social media, specifically Twitter, to study health behavior. It will discuss the potential for studies that link social media data with survey data. After detailing study design considerations and outlining guidelines for work in this area, it will propose opportunities for novel contributions to health research.

Address
Information School, Mary Gates Hall Ste. 370, University of Washington, Seattle, WA 98195, USA

Corresponding author: Spiro, Emma S (espiro@uw.edu)

Current Opinion in Psychology 2016, **9**:67–71

This review comes from a themed issue on **Social media and applications to health behavior**

Edited by **Melissa A Lewis** and **Clayton Neighbors**

For a complete overview see the Issue and the Editorial

Available online 17th December 2015

http://dx.doi.org/10.1016/j.copsyc.2015.10.023

2352-250X/© 2015 Elsevier Ltd. All rights reserved.

Introduction

As social media continue to become integrated into daily life, the extensive records these systems archive as part of normal operation promise to change the available avenues of inquiry in the social and behavioral sciences [1]. Researchers from a variety of disciplines are using social media data to explore human behavior [2,3•,4–6,7••,8]. A growing number of studies explore questions within the health domain [9••], where social media have been used to explore the effects of peer influence on health behaviors [10•,11], to aid in the detection of mental health concerns [12,13••], and to quantify deviant behaviors [14–16]. Other work has looked at the use of social media as a tool for health communication [17], particularly within the domains of public health [18–20,21•,22], health information seeking [23] and general health status or well-being [24,25••,26]. While the majority of these studies are observational and descriptive, recent work has also employed large-scale, randomized experimental trials to infer causal relationships [27].

The focus of this review is Twitter, a platform designed to allow users to find out what other people are doing [28,29]. Within the site, 140-character messages, known as *tweets*, are posted by individual users and then delivered to that person's content subscribers, known as *followers* [30,31]. The larger stream of public tweets is searchable through the website interface. As of 2015, almost 25% of online adults use Twitter [32•], and the platform has particularly high penetration among Internet users 18–29 years old [33]. Research utilizing Twitter data specifically has dramatically increased in recent years as scholars recognize the wealth of observational data available [34,35,29]. However, while social media provide large-scale data in terms of the number of observations or cases, data are often quite poor in terms of the user features or characteristics contained. Very limited data is available about user socio-demographic characteristics, for example, driving researchers to infer such attributes from the data that is available. These techniques have become more prevalent within the literature [36,37].

For social scientists, the lack of individual-level attributes limits the scope of questions that can be addressed using social media data. Cavazos-Rehg *et al.* [38••], for example, offer a detailed look at tweets about drinking behaviors for prominent Twitter users, but they are unable to determine whether pro-drinking or anti-drinking content is more prevalent in certain socio-demographic groups, compared to others. While prior work exploring 'big social data' makes important contributions to understanding how these data provide insight into health behaviors, many studies analyze tweets in isolation (see e.g. [39•]) limiting the generalizability of the work and its applications to specific domains (e.g. youth and young adult populations) [40].

The limitations of Twitter and other social media data, however, are precisely the benefits of more traditional forms of data collection—surveys and questionnaires—where detailed individual-level attributes are self-reported by participants. Likewise, the limitations of survey data (e.g. prohibitive resource cost for large-scale efforts) are the strengths of big social data. In combining both methodological approaches, researchers can curate rich, complex observations of health behavior at scale. As such, this paper offers an alternative framework for utilizing social media data in studies of health behavior—a user-centric approach that builds from survey data. It considers the advantages and opportunities of research at this intersection, offering new directions for scientific inquiry.

Linking Twitter data to survey data

Survey research is widespread in the social and behavioral sciences [41]; both the advantages and disadvantages of

survey research are well documented. Yet limited work lies at the intersection of surveys and social media. Social media have been used to aid participant recruitment [42,43], but few studies [44••,13••] have collected data about survey participants' activity on social media platforms. Survey and big social data methods complement each other by alleviating weaknesses in the other approach. Moreover, linking social media to survey data affords the ability to associate self-reported health behaviors with those expressed (directly or indirectly) online [45].

If health behaviors be detected, measured and inferred from observations on social media platforms, large-scale, early intervention and behavior change strategies may be viable [46,9••]. This is the core aim of many studies of social media and health behavior. However, without validation these techniques—which require 'ground truth' data to which to compare inferred health behaviors—these claims fall short. Survey data can provide comparison data; and while self-reported health behaviors have notable constraints (e.g. social desirability and participant recall concerns) the potential value of this work is substantial.

In practice, combining Twitter data with survey is not without challenges. Study participants can be asked to volunteer social media identities as part of contact information on a survey or questionnaire. Participants may or may not choose to volunteer this information. However, in the case of Twitter, this information (account usernames) is all that is currently needed to collect large amounts of data on the online social behaviors of study participants.

User-centric data collection on Twitter

Opportunities for novel research contributions at the intersection of social media and survey data abound. In order for researchers to capitalize on these opportunities it is necessary to design data collection systems that augment survey and questionnaire data with social media data. This section briefly discusses some of the challenges that arise in this endeavor, offering design guidelines and other considerations.

Defining the study population. Studies that utilize big social data, such as tweets, face notable issues related to the dual problems of defining the study population and sampling that population. There are multiple approaches to sampling data from Twitter which are in part determined by the restrictions of the Twitter application programming interface (API)—the standard access point for data collection most tools utilize (e.g. see [47,48•]). The Twitter API offers programmatic access at the level of the tweet or the user. While prior work tends to use data collected from the public stream of tweets, we in contrast propose methods that utilize user-centric entry points.

One of the advantages this framework is that the population of interest is well-defined—defined by the survey component of the study—avoiding issues that arise when attempting statistical sampling strategies in social media environments (which are in many cases near impossible due to unknown factors, such as the set of possible cases). Bringing to bear all the standard techniques for sampling populations already employed in the social and behavioral sciences, one obtains a sample of participants, each of whom is asked for their social media identifiers (e.g. account username of Twitter). Not only does this alleviate sampling bias towards highly active users (as is present in tweet-based samples), but it also grants the possibility of obtaining a representative sample of the population of interest.

Linking survey and social media data is entirely contingent on the willingness of survey participants to share their social media identify. Whether non-response bias and/or response rates are a severe limitation remains an open question. During a pilot study of this approach, our project found that approximately 20% of study participants were willing to share social media handles as part of their contact information before taking part in a study of risky alcohol and sex-related behaviors (PI:Lewis, R01AA021379). As work in this area continues, these limitations will be further explored.

Longitudinal monitoring of behavior on Twitter. While detailed description of data collection tools is outside the scope of this paper, this section provides sufficient detail for researchers to understand the overall aims of this approach. Data collection begins with a well-defined study population, with accompanying social media account information collected during survey administration. Twitter account usernames (`user_name`) allow for an initial access point to collect participants' social media data. Twitter usernames are not static and can be changed by the user while still maintaining the same account. Therefore, it is recommended that researchers obtain a unique account identifier for each participant soon after survey data has been collected; this unique identified will be referred to as a `user_id`. IDs can be used to systematically sample Twitter data.

Three types of observational data are available via the Twitter API: Firstly, information about the user account; secondly, tweets; and finally, social tie information. Each component must be queried in turn. One can obtain a user's most recent tweets at the time of query.[1] Data collection can subsequently occur at regular intervals to continue longitudinal observation. If data collection proceeds over time the researcher can ensure (almost) complete data on each participant during that period; every tweet posted during the observation period can be

[1] At the time of writing the limit was 3200 posts.

Download English Version:

https://daneshyari.com/en/article/879330

Download Persian Version:

https://daneshyari.com/article/879330

Daneshyari.com