Big data opportunities for social behavioral and mental health research

This item was submitted to Loughborough University’s Institutional Repository by the/an author.

Citation: GRUEBNER, O. ... et al. 2017. Big data opportunities for social behavioral and mental health research. Social Science and Medicine, 189, pp. 167-169.

Additional Information:
• This paper was published in the journal Social Science and Medicine and the definitive published version is available at https://doi.org/10.1016/j.socscimed.2017.07.018.

Metadata Record: https://dspace.lboro.ac.uk/2134/26030

Version: Accepted for publication

Publisher: © Elsevier

Rights: This work is made available according to the conditions of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) licence. Full details of this licence are available at: https://creativecommons.org/licenses/by-nc-nd/4.0/

Please cite the published version.
Big data opportunities for social behavioral and mental health research

Oliver Gruebner¹,², Martin Sykora³, Sarah R. Lowe⁴, Ketan Shankardass⁵, Sandro Galea⁶, SV Subramanian⁷

1 Humboldt-Universität zu Berlin, Geography Department, Berlin, Germany
2 University of Zurich, Epidemiology, Biostatistics, and Prevention Institut, Zurich, Switzerland
3 Loughborough University, School of Business and Economics (SBE), Centre for Information Management (CIM), Loughborough, UK
4 Montclair State University, Department of Psychology, Montclair, NJ, USA
5 Wilfrid Laurier University, Department of Health Sciences, Waterloo, Ontario, Canada
6 Boston University, School of Public Health, Boston, MA, USA
7 Harvard T.H. Chan School of Public Health, Department of Social and Behavioral Sciences, Boston, MA, USA
Social media and mental health

Social media can be defined as online applications with which users share and consume content and engage with each other in various ways. Sykora (2012) reviews various general taxonomies of, and data that can be extracted from, social media platforms, such as Twitter, Facebook and Snapchat, and related user-generated content (e.g. text/image/video posts, comments, likes, ratings, etc.). For example, Twitter can yield at least 38 potential data features, including explicit information (i.e., data that can be directly analyzed) and extractable information (i.e., additional data that involve some measure of coding on the part of the researcher). Explicit and extractable information can be generated both for Twitter users (e.g., number of total tweets, followers, and followees; and age, gender, ethnicity) and posts (e.g., hashtags, URL links, number of retweets; and content, sentiment, language). However, most of this information has minimally been utilized in research to date (Sinnenberg et al., 2017). In a recent paper, Murphy (2017) explains what data are available from Twitter and provides code and step-by-step instructions to retrieve them.

A growing number of research studies focus on the methodological aspects of the use of social media data as indicators of mental health and its causes. Conway and O’Conner (2016) provide a good overview on the current state of social media usage for mental health research. As they illustrate, methodologies such as Natural Language Processing (NLP) and Machine Learning (ML) have been applied in several studies. NLP deals with computational processing of human natural language in form of texts such as from social media and ML are data driven computational algorithms that can “learn”
from data to recognize patterns and make predictions. Evidence has shown that
sentiment, affect-related states, basic emotions (e.g., anger, fear, or happiness), and
mental health disorders, such as post-traumatic stress disorder (PTSD) (Coppersmith et
al., 2014) and depression (De Choudhury et al., 2013) can be identified effectively and at
scale based on text classification using such methods as NLP or ML (Bollen et al., 2011;
Dodds PS, Harris KD, Kloumann IM, Bliss CA, Danforth CM, 2011; Gerrard et al.,
2017; Glass and Colbaugh, 2011; Peng et al., 2017; Ravi and Ravi, 2015; Sykora et al.,
2013; Thelwall et al., 2010). Other research has combined advanced sentiment analysis
(EMOTIVE) and space-time scan statistics (SaTScan) to identify elevated concentrations
of negative emotions (e.g. fear, sadness, confusion) as a way to assess mental health risk
after mass traumatic events, including Hurricane Sandy in 2012 (Gruebner et al 2017
accepted) and the Paris terrorist attacks in 2015 (Gruebner et al., 2016).

The study by XXX in this volume of Social Science & Medicine provides a
compelling example of how social media can be used to examine mental health. In this
case, data were employed to provide new insights into the underlying processes by which
celebrity status, traditional media coverage, and respective Twitter activities led to
increases in suicides in the Japanese population. The authors integrated different forms of
data, using not only Twitter, but also traditional media and vital statistics in their
analysis. This facilitates comparison with the more standardized approach of using survey
data. One notable finding was that prominent suicide deaths were followed by an increase
in population suicides only when there was considerable Twitter activity of users
discussing these suicides. In contrast, when there was no substantial Twitter activity of
users regarding these deaths, suicide counts did not exhibit any notable increase, even
when traditional media reported on them. The authors concluded that social media users are likely to be exposed to a large amount of information and the phenomenon of “social learning,” that is, imitating social behavior perceived acceptable within a user community, could account for increased suicidal behavior in the general population.

**Mental health research opportunities afforded by social media**

As the study of XXX on suicide demonstrates, the analyses of social media can provide insight into areas where data were previously not available (i.e., user responses in Twitter versus traditional media responses). Social media thereby offers the opportunity to open new frontiers in the social behavioral and health sciences as it also provides data where more purpose-designed studies either cannot be launched in time, or would be prohibitively difficult to conduct. For example, in trauma-related mental health research, investigators typically collect event specific information only after the specific event under study (e.g., a mass traumatic event, such as a natural disaster or terrorist attack). Missing, then, is potentially important information on the time before and during the event. More generally, mental health researchers often lack data on longer periods of time to systematically assess the relationship between core determinants (e.g., poverty, urban neighborhood conditions) and mental health. Because social media data is inherently longitudinal, it could also facilitate investigation of causal relationships. Furthermore, social media data is available in real-time, facilitating surveillance and prediction of mental health risk (Kuehn, 2015; Sadilek et al., 2016).
All the above further demonstrates that social media research on mental health could be done using more advanced sentiment analysis across spatial-temporal domains, social networks dimensions, and multiple levels of analysis. For example, future studies of social media could extend the usage of advanced sentiment analysis to conduct ecological momentary assessments (EMA) (Shaughnessy et al., 2017), that is, evaluation of dynamic, near real-time experiences such as emotions (e.g., anger, fear, happiness) or stress (Shaughnessy et al., 2017; Sykora et al., 2015; Thelwall, 2017). EMA approaches have several strengths, including minimizing recall bias and enabling longitudinal analysis of data (Shaughnessy et al., 2017).

Studies involving geo-social data, that is social media data with geographic location and time information available, have the potential to tap into the methodological and theoretical developments of the geographic information sciences, including understanding how one’s actual socio-ecological context (e.g., neighborhood poverty, available green spaces) affects mental health. For example, Robertson and colleagues demonstrate how these data can be used to measure personal activity spaces and predict home and work locations for individuals (Robertson et al., 2017). Integrating external information about these personalized activity spaces and following users over time would allow for the investigation of possible causal relationships between socio-ecological factors and mental health indicators. Further, analyzing the movement patterns of geo-social data may also facilitate the identification of mobile (e.g. traveling, relocating, or migrating) populations. This essentially includes tracking mobile populations over time and assessing psychological outcomes in relation to changing socio-ecological contexts. In addition, spatially explicit modeling and mapping could be used to identify geographic
areas of mental health risk, guiding us to where we may most effectively intervene to support health and mental wellbeing.

Social networks, that is, the quality and quantity of connectedness amongst users, such as users’ comments to other users’ tweets could also be investigated during or in the aftermath of a certain event (e.g., a natural disaster). We might ask questions, such as, how do people support others? And more generally, what are the effects of social-media specific dynamics of content sharing (e.g., replies, likes, etc.) on mental health? Moreover, combining social networks with activity spaces could provide further insights into the connectedness of individuals and groups, how information is shared among and across groups, and how influential specific individuals and groups are. Further, this kind of analysis has potential to improve our understanding on the differential contributions of the multiple dimensions of social support (e.g., emotional or practical support, or social integration) and related mental health outcomes.

Social media data could be further studied at multiple levels of analysis facilitating investigations of combined individual level and community level exposure together. While single posts can be assessed atomically (e.g., in ecological momentary assessment approaches), they can also be assessed at the individual level, such that we could derive information on a user’s social media usage over time as well as longitudinal trajectories of mental health indicators, for example, symptoms of depression and schizophrenia (Sinnenberg et al., 2017). Social media data have also been used to assess socio-demographic data such as gender, age, or ethnicity (Sloan et al., 2015), which could be used to explain mental health risk at the individual level. Locational information could be used to link posts and users to communities, and integrate user information with
community-level information, such as neighborhood poverty, green space, or traffic noise. This would allow for the joint analysis of multiple levels of influence of mental health risk.

**Challenges of social media for behavioral research**

Although there is a myriad of opportunities available to researchers interested in incorporating social media data for behavioral science research, there are nonetheless several challenges inherent to this work. A key limitation is that the data are not representative of the general population and may be biased in several ways. There is a likely spatial bias to be expected when working with geo-social data. Recent evidence suggests substantial variance in how often and at which times users tweet with their geo-location on and that this variance differs across age, gender, ethnic, and socio-demographic groups (Sloan and Morgan, 2015). Further, several municipalities or public and private spaces within them (e.g., parks, restaurants, airports) provide free wireless networks, which suggests that the representativeness of social media will also vary across settings.

Further, we may expect socio-demographic bias in the quantity and quality of tweeting behavior, such that younger users might be posting more often and on different topics as compared to older users. Such variability may also be seen across different gender, ethnic, or socio-economic groups (Sloan et al., 2015). For example, a “digital divide,” in which access to computers and the Internet is reduced in lower income groups within a population has been documented in several jurisdictions, suggesting that people
of lower income may be under-represented in studies using social media. Yet, studies have also demonstrated that people with lower income, or who are marginalized in other ways, and who do use social media, may use it more intensively (Haight et al., 2014; Junco et al., 2010), which complicates this pattern and any assessment of bias.

In addition, we may expect contextual bias, perhaps due to local social norms (as may be defined by immediate friends, local trends, or national regulations) that might define which social media platforms are the more prominent means of communicating with peers and others. For example, when comparing social media data across international borders, we need to be aware that some social media platforms are more prominently used in one country versus another, such as Twitter in the United States versus Sina Weibo in China.

Along with developing means to addressing the potential biases inherent to social media data, future studies should continue to work on measures that increase the validity and reliability of explicit and extracted information from social media by, for example, randomly inviting social media users to participate in a structured questionnaire that could be evaluated against explicit and extracted information on users and their posts.

Given the above discussion of current and potential future uses of social media data, there are of course numerous significant and important considerations around privacy, informed consent, and related ethical issues. A good starting point for a discussion around the various ethical concerns when working with social media can be found for example in Shilton and Sayles (2016), Sax (2016), Zimmer and Proferes (2014), or Moreno et al. (2013).
Conclusions

The current study of XXX is of particular interest as it opens up several pathways for the social and medical sciences to investigate health-related behavior in the ever-increasing populations engaging with social media platforms. This includes continuing to work on validation of extracted measures from social media and to fully exploit available social media information to systematically assess mental health risk at multiple levels of analysis, through social networks, and across spatial-temporal dimensions.

References


