Using Twitter for Demographic and Social Science Research: Tools for Data Collection

Authors: Tyler McCormick[1,2,4,5], Hedwig Lee[1,5], Nina Cesare[1], Ali Shojaie[3,2,4]

[1] Department of Sociology, University of Washington
[2] Department of Statistics, University of Washington
[3] Department of Biostatistics, School of Public Health, University of Washington
[4] Center For Statistics And The Social Sciences, University of Washington
[5] Center for Studies in Demography and Ecology, University of Washington

ABSTRACT
Despite widespread success by information/computer scientists in using social media data to *predict* specific aspects of human behavior (e.g., suggest Facebook friends), little attention has been paid to using social media data to extract demographic information from users in order to *understand* behaviors and attitudes from the perspective of social and behavioral scientists. This paper develops a scalable, sustainable toolkit for social science researchers interested in using Twitter data to examine behaviors and attitudes. We offer new approaches for the extraction, processing, and analysis of data from social media. We begin by describing how to collect Twitter data using a novel targeted sampling scheme that, drawing insights from statistics and epidemiology, extracts information most likely to be of interest to social and behavioral scientists. We then describe and evaluate a method for processing data to retrieve information reported by users that is not encoded as text (e.g., details of images). We end by offering suggestions for statistical analyses.

INTRODUCTION
Social media networks, such as Twitter and Facebook, provide exciting opportunities that, according to a recent issue of the American Sociological Association (ASA) magazine, can “open up a new era" of social science research.(1) Social media and the data extracted from social media has gained a growing interest among many researchers attempting to better understand the nature and power of social media in influencing social relationships and behavior.(2-7) Although social science researchers have begun to use Twitter to document changing moods and other sentiments and opinions on the aggregate level,(5, 8-11) the potential of such data for demographic research has yet to be realized.

For the vast majority of social scientists currently collecting data using a combination of surveys or case-study methods, social media data present a completely new perspective on data collection. Whereas surveys ask respondents to recall behaviors or sentiments retrospectively, social media data afford the opportunity to observe behaviors and human interaction in real-time and on a large scale. With appropriate infrastructure, scientists can analyze and begin presenting results within a matter of months (or sooner), rather than the years typically required for a survey. Social media data are also distinct from data derived from surveys often used by social and behavioral scientists because they allow researchers to collect reports of behaviors that are unsolicited and unprompted by a researcher. One could even argue that these data provide a better reflection of day-to-day social experiences. Indeed, Twitter interactions have been described as persons "want[ing] to know what the people around them are thinking and doing and feeling, even when co-presence isn't viable" and "shar[ing] their state of mind and status so that others who care about them feel connected.”(12) Despite these possibilities, social scientists often see such data as inaccessible for social science research and solely relevant to computer and physical scientists. The same ASA article (1) laments, “most of the social and behavioral science using online data is coming from computer and information scientists who do not always have the training required to ask the right questions, or to recognize unfounded assumptions and socially unjust

ramifications."(7)

A further hindrance arises as, currently, each investigator must devise her/his own sampling strategy for data collection. Indeed, very little social science research has been able to systematically collect data from Twitter.(3, 9, 13) This prospect is especially challenging given the numerous differences between Twitter data and the surveys at the heart of most formal training in sampling. Using traditional surveys, for example, researchers see comparatively few respondents but have a great deal of control over what information respondents provide. Respondents then provide information of interest to the researchers, but the limited sample size may not produce enough variability to study less commonly observed phenomena in their entirety. Twitter in contrast, is completely unelected but offers unprecedented exposure to variability. On the other hand, the uncontrolled nature of information sharing on Twitter necessitates that such data be verified. Removing the actual and perceived barriers that prevent social scientists from using social media data offers new research opportunities for social scientists and increases the potential for interdisciplinary research between engineers, computer scientists or statisticians with social and behavioral scientists.

This paper will describe the process of developing a scalable, sustainable data infrastructure that facilitates access to social media data by social scientists. In particular, we describe how researchers interested in using Twitter can extract useable demographic and other behavioral/social information from it to answer relevant social science questions. To help illustrate this process, we examine a specific behavior, reporting the intention to not vote in the 2012 presidential election, to help elucidate the data collection process. We hope to show that it is possible to use Twitter data for social science research by following systematic protocols for data extraction.

In the following sections, we outline our three-pronged approach of data extraction, processing, and analysis. We begin with an introduction to Twitter with a discussion of the resources, and challenges, associated with using such data. We then describe our data extraction strategy that uses a case-control sampling framework to produce samples of users from the Twitter application programming interface (API). Next, since Twitter users typically do not report demographic information directly, we describe a processing strategy that allows us to gather this information from users' photos. At the heart of the strategy is a framework for using Amazon Mechanical Turk's[1] to efficiently code large volumes of images. Finally, we discuss a statistical analysis strategy that accounts for the impact of the case-control framework on coefficients of interest. To outline the Twitter data analysis toolkit we use the example of intending not to vote for president.

METHODS

Before beginning data collection using Twitter, we first address two key questions. First, we describe the nature of Twitter data as a vast and emerging resource with known limitations. Next, we contextualize Twitter data in the framework of the typical social science research paradigm to provide a more complete sense of when Twitter data may be useful.

Description of Twitter

Twitter is a microblogging platform that allows users to record their thoughts in 140 characters or less. The text-based content of these messages may include personal updates, humor, or thoughts on media and politics. This concise format allows users to update their blogs multiple times per day, rather than ever few days as is the case with traditional blogging platforms.(14) Besides projecting their thoughts

[1] https://www.mturk.com/mturk/welcome

independently, users can communicate with one another either through private messages or by using the *@reply* command, and contribute to broader conversations by including a *hashtag* identifier in their tweet. Tweets from those who the user follows are displayed sequential feed that is updated in real time.

Self-presentation (15) on Twitter is developed through active conversation rather than fixed profiles. To generate this conversation, Twitter users project their thoughts toward an imagined audience of networked individuals, some of whom bear reciprocal ties to the users themselves and some of whom do not. This interesting mix of public and private attention requires users to maintain a balance between transparency and authenticity when their tweets.(16)

Twitter and Social Science Research

The sequence of events for a typical social science/demographic research project, in brief, usually proceeds as follows: Step 1: Inception of an idea; Step 2: Conducting background research; Step 3: Formulating a hypothesis/problem statement; Step 4: Testing a hypothesis via secondary analysis or data collection and then performing analyses; Step 5: Interpreting the data to draw conclusions.

Steps 1 to 3 will likely determine if Twitter data will serve as the ideal data source to test a researcher's specific hypothesis. No matter what data source used, researchers need to consider the benefits and limitations of data sources to make their decision about which best suits the research goals.

There are many situations where Twitter data may be ideal for a research project. For example, using the most current data to examine attitudes and behaviors (e.g., voting intentions or happiness); using a large amount of data to examine a rare events or small groups (e.g., members of small political or religious groups, persons presenting extreme attitudes or ideas, or the LGBT community); pretesting to determining if a behavior or attitude not currently in a survey is evident in Twitter; examining behaviors and attitudes where social desirability bias in an official survey may occur (e.g., racist attitudes or anti-immigrant sentiments); examining collective experiences based on a timely event (e.g., teacher strikes, terrorist attacks or natural disaster) and collecting large amounts of data on a limited research budget. Note that these examples are not meant to represent all possible scenarios for the use of Twitter data. Indeed, future social science researchers will surely find new and innovative ways to use Twitter and other social media data.

Extracting Data from Twitter: Using the Twitter API and Case-Control Design

Description of Twitter Application Programming Interface (API)

An application programming interface (API) is a standardized system of programming instructions that allows web platforms to access and share information from one another.[2] In the same way that the web page's interface provides the user directives for interaction, the API helps guide communication between web programs. Like many other web tools, Twitter has released its API for researchers and other web developers to use. Using an external data hub, this project will utilize instructions provided within the Twitter API to crawl, collect and store information about users and tweets.

Extracting Data from Twitter

In most demographic surveys individuals are selected randomly without regard for whether or not they have the attribute of interest (e.g., planning not to vote). A researcher can then compare the people

[2] http://dev.twitter.com

who have the attribute to those who don't. This works well unless the attribute is rare, in which case you get lots of non-attribute people and very few people with the attribute. From a statistics perspective that means the variance of your estimators is big even though the total sample size is very large because the number of people with the attribute is still small. We can face this challenge by using Twitter data, though this strategy is compounded by the lack of a convenient way to access randomly selected individuals through the Twitter API.

We propose using a case-control sampling framework for extracting data from Twitter. Under the case-control design you select an exhaustive set of people with the attribute (all in the most extreme case), known as *cases* and you compare them to a randomly selected subset of the people without the attribute (the *controls*). For this to work, the *cases* have to be comparable (from the same population) as the *controls*. This framework leverages the power of the Twitter API, with which, as we describe below, it is easy to select users based on a specific attribute or behavior. This method ensures that we see enough presumably rare cases to reasonably estimate variability within the population.

We begin by using the Twitter API to search for the specific attribute or behavior or interest (e.g., a hashtag, reported behavior or location). We will use multiple search queries, including, but not limited to, "I am not voting;" "I'm not voting;" "I am not going to vote;" and "I will not vote"[3] to identify individuals who do not plan to vote in the 2012 presidential election. Using the search interface of the API we can also distinguish individuals who report that they will not vote from those who disagree with a particular candidate (who might say "I'm not voting for Romney" for example) by excluding tweets containing certain phrases ("for Romney" in this example). We can also exclude many of those who are discussing voting in other contexts (for a contestant on a television show, for example) in a similar fashion. We give several examples in Exhibit 1.

Through this process we will collect information on Twitter users who are voluntarily reporting that they are not participating in the 2012 presidential election. Such information is of interest to social scientists because it builds upon existing literature featuring the use of social media as a tool for examining political climates and predicting electoral outcomes.(8, 10, 17) However, while this research provides insight on how social media discourse reflects constituents' preferences, there is still work to be done in terms of recognizing who is willing to participate politically and express these preferences. This work helps to fill this gap by discussing how Twitter data can be used to better understand this hard-to-reach population.

The unit of analysis for this study is the individual, or subject, as illustrated in Exhibit 1. These data consist of texts from tweets, time, and (often) location. We will also describe below how to generate demographic information (e.g., age, gender, race/ethnicity). The size of Twitter makes it possible to gather samples on an unprecedented scale (tens of thousands of users or more), even for very rare groups.

Using this information, we could describe the characteristics of individuals who plan not to vote in the upcoming presidential election. A more informative analysis, however, would *compare* individuals who plan not to vote with either (i) the population overall or (ii) those who do plan to vote. The second step in our case-control design is to select a population of controls for comparison. In the case of those planning not to vote, a natural comparison group is individuals who do plan to vote, which can be

[3] Note that these search queries are not inclusive (e.g., we will also search for "I'm not voting for president" and "I'm not voting for in the presidential election") and will be adapted based on additional sensitivity analysis.

collected using the methods described above. In other cases, however, the relevant comparison group will be the population of Twitter users overall. In these situations, we will use recent developments in "graph crawling" algorithms to produce random samples.(18, 19) The approach begins with a set of individuals, known as "seeds." From the seeds, the algorithm recruits individuals by randomly selecting one of more users that are connected to the seeds. The process continues until the dependence on the seeds has diminished and the resulting sample has the properties of simple random sample.

Coding data from Twitter: Amazon Turks

In the previous section we described how Twitter data can be obtained using the Twitter API. However, demographic information about actors is not typically reported on Twitter. Thus, we propose to code basic demographic information from users' profile images using the online, on-demand workforce of the Amazon Mechanical Turks (https://www.mturk.com/mturk/welcome).(20) Amazon Mechanical Turk is a marketplace for work that requires human rather than artificial intelligence. This service offers an efficient way to obtain answer to specific questions by defining new tasks (known as a HITs or Human Intelligence Tasks), which can be performed by online workers. Our HITs will involve showing an online worker (known as a Turk) an image downloaded from Twitter and asking a series of approximately 5 questions (e.g., ``If there is a person in the photo, is the person male or female?"). We will ask multiple Turks to view each image, allowing each Turk to view each image only once. Images to which two Turks give conflicting responses will be shown to a third Turk.

A key goal of our project is examining how to most efficiently, and reliably use this resource. This process will involve piloting various phrasings of questions, selecting appropriate prices, and developing infrastructure to download and code images on a large scale. In our preliminary work, we found that approximately 90% of active users have some demographic information that could be used for coding.

Analysis framework

In the previous sections we have discussed a case-control strategy for extracting data from Twitter and using the Mechanical Turks to process this data. We finish by presenting a statistical analysis strategy that accounts for the case-control design. From a survey sampling perspective, the case-control design amounts to over-sampling the cases. This process results in a sample with a much greater fraction of cases than would be present in a simple random sample. Further issues arise as there are various ways of constructing the group of controls. The process of "matching" the cases with controls can introduce confounding results if the features of the controls are not appropriates modeled. We lean heavily on work from epidemiology and public health in presenting a conditional logistic regression approach that adjusts for these two issues. Additional details of this approach can be found in other literature.(18, 19)

DISCUSSION

In this paper, we present a toolkit for extracting, processing, and analyzing data from Twitter. We believe that social media data, such as Twitter, present an opportunity for a fundamentally different approach to social science research. As with all new data collection, Twitter has certain limitations. There are, of course, limitations and issues with Twitter data collection that we are not able to address in this work. Our goal, however, is to address enough of these challenges to make Twitter an accessible resource for a larger fraction of social scientists and, in doing so, explore the contexts in which Twitter data has the greatest potential contribution in social science research.

Exhibit 1: Example Twitter users and comments. Note that the majority of users have pictures that could be, via our data processing techniques, coded for basic demographic information. In preliminary experiments, we find that nearly 90% of active users have this information.

REFERENCES

1. Golder S, Macy M. Social Science with Social Media. ASA Footnotes, 2012;40(1).
2. Brickman Bhutta C. Not by the Book. Sociological Methods & Research 2012;41(1):57-88.
3. Heaivilin N, Gerbert B, Page JE, Gibbs JL. Public Health Surveillance of Dental Pain via Twitter. Journal of Dental Research 2011;90(9):1047-1051.
4. Moreno MA, Grant A, Kacvinsky L, Egan KG, Fleming MF. College Students' Alcohol Displays on Facebook: Intervention Considerations. Journal of American College Health 2012;60(5):388-394.
5. Golder SA, Macy MW. Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures. Science 2011;333(6051):1878-1881.
6. Lowe JB, Barnes M, Teo C, Sutherns S. Investigating the use of social media to help women from going back to smoking post-partum. Australian and New Zealand Journal of Public Health 2012;36(1):30-32.
7. Valkenburg PM, Peter J, Schouten AP. Friend Networking Sites and Their Relationship to Adolescents' Well-Being and Social Self-Esteem CyberPsychology & Behavior 2006;9(5):584-590.
8. Yardi S, Boyd D. Dynamic Debates: An Analysis of Group Polarization Over Time on Twitter. Bulletin of Science, Technology & Society 2010;30(5):316-327.
9. Naaman M, Becker H, Gravano L. Hip and trendy: Characterizing emerging trends on Twitter. Journal of the American Society for Information Science and Technology 2011;62(5):902-918.
10. Diakopoulos NA, Shamma DA. Characterizing debate performance via aggregated twitter sentiment. In: Proceedings of the 28th international conference on Human factors in computing systems. Atlanta, Georgia, USA: ACM; 2010. p. 1195-1198.
11. Reips U-D, Garaizar P. Mining twitter: A source for psychological wisdom of the crowds. Behavior Research Methods 2011;43(3):635-642.
12. boyd d. Twitter: "pointless babble" or peripheral awareness + social grooming? . In. http://www.zephoria.org/thoughts/archives/2009/08/16/twitter_pointle.html; 2009.
13. Krishnamurthy B, Gill P, Arlitt M. A few chirps about twitter. In: Proceedings of the first workshop on Online social networks. Seattle, WA, USA: ACM; 2008. p. 19-24.
14. Java A, Song X, Finin T, Tseng B. Why we twitter: understanding microblogging usage and communities. In: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis. San Jose, California: ACM; 2007. p. 56-65.
15. Goffman E. The Presentation of Self in Everyday Life New York, NY: Doubleday; 1959.
16. Marwick AE, boyd d. I Tweet Honestly, I Tweet Passionately: Twitter Users, Context Collapse, and the Imagined Audience. New Media & Society 2010.
17. Conover M, Ratkiewicz J, Francisco M, Gonçalves B, Flammini A, Menczer F. Political polarization on twitter. In: 5th International Conference on Weblogs and Social Media (ICWSM), 2011; 2011; Barcelona, Spain; 2011.
18. Breslow N. Design and Analysis of Case-Control Studies. Annual Review of Public Health 1982;3(1):29-54.
19. Breslow NE. Statistics in Epidemiology: The Case-Control Study. Journal of the American Statistical Association 1996;91(433):14-28.
20. Ipeirotis PG, Provost F, Wang J. Quality management on Amazon Mechanical Turk. In: Proceedings of the ACM SIGKDD Workshop on Human Computation. Washington DC: ACM; 2010. p. 64-67.