Challenges of using Twitter as a data source: An overview of current resources

blogs.lse.ac.uk/impactofsocialsciences/2015/09/28/challenges-of-using-twitter-as-a-data-source-resources/

9/28/2015

There are specific challenges to using social media data in academic research, and in particular Twitter data, including ethical, legal and methodological issues. **Wasim Ahmed** *builds on his previous work on acquiring Twitter data and offers a list of some of the challenges researchers may face. He also provides plenty of links to resources for social scientists looking to explore these challenges further.*

In one of my previous blog posts I outlined a number of software applications that could be used to capture and analyse data from Twitter. In this blog post I outline some of the methodological, ethical, privacy, and copyright issues associated with using Twitter as a data source.

Twitter can be used as a source of data for social science research both current and historical in-of-itself, but it can also be used to compliment more traditional data sources such as surveys and interviews. Twitter boasts 316 million monthly active users with 500 million tweets per day. Marc Smith, from the Social Media Research Foundation, at The Next Web conference (2014) notes that although the city squares and plazas of the world are still important, now, more and more people are tweeting and posting about events.

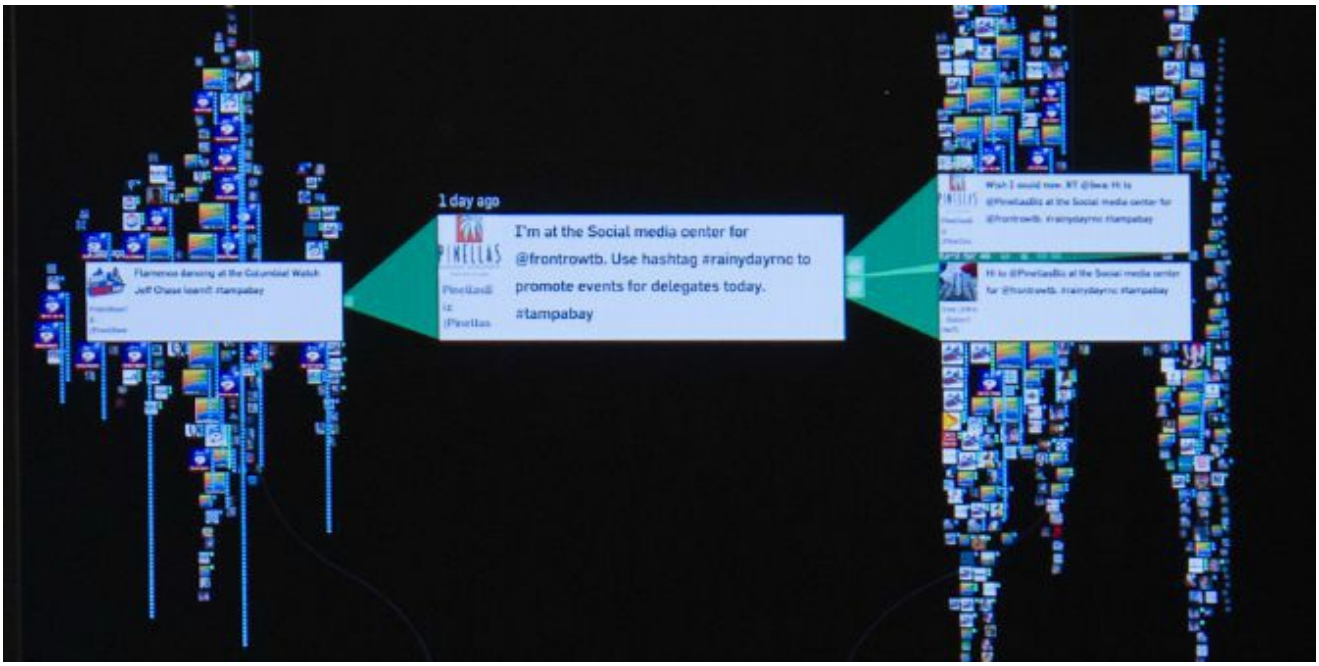

Image credit: Live visualization of Tweets. Social Media Command Center Wikimedia (Creative Commons Attribution 2.0)

Obtaining Twitter data need not require any advanced programming or computer science skills (see my blog post on software applications that can be used for this purpose). However, there are often specific challenges to using social media data in academic research, and in particular Twitter data, which social scientists may face for the very first time. Below is a list of some of the challenges that may be faced when using Twitter as a data source in academic research along with links to resources that provide advice and guidance on these issues:

- **Ethical issues**: In collecting and retrieving data to form large datasets it may not be possible to obtain informed consent from all of the participants, simply due to the volume of tweets retrieved. There are also ethical issues if you decide to reproduce tweets in an academic publication, which have to be handled with care especially concerning tweets related to sensitive topics i.e., obtaining consent before disclosing user IDs or tweets. See NatCen's report on user's views of research using social media.
- **Legal issues:** Sharing of datasets is prohibited under Twitter's API Terms of Service, however, researchers can share the tweet identification numbers, associated with each tweet, which can be used by other researchers to obtain Twitter datasets. If, for any reason, it is not possible to share tweet IDs then sharing the keywords and retrieval time of the data, may allow researchers to obtain a similar dataset. There may also be specific requirements for producing tweets within a publication i.e., following Twitter's guidelines. See Twitter's API Terms of Service.
- **Retrieving datasets:** Use of certain keywords or hashtags may not retrieve all of the data related to a topic. It may help that when brainstorming search queries that as many queries as feasible as possible are selected, and that this dataset is filtered for non-relevant keywords after data-retrieval. This is because missing certain keywords or hashtags could introduce a systematic bias which would lead to a biased sample. See the Demos and Ipos MORI reporton representivity. Datasets are also likely to be limited by the language that is used to retrieve data, for example, using the English keyword *Ebola* to retrieve data related to the Ebola epidemic will not gather data from other countries tweeting about Ebola which may use a different keyword i.e., a different language.
- **Cost:** Twitter data costs a lot of money, and if it has not been possible to retrieve or set up a system to retrieve Twitter data within 7 days of a topic of interest, then it becomes difficult to obtain the data. This is because using the free API ecosystem it is only possible to retrieve Twitter data going back in time 7 days. However, it is be possible to obtain this data using a licensed re-seller of Twitter data. Historical Twitter data can range from not that expensive, to very expensive depending on both the query and time of retrieval. It is possible to generate free estimates for the cost of Twitter data using Sifter.
- **Representivity:** Twitter users are not representative of the national offline population, Twitter users are not even representative of Internet users, and most strikingly Twitter *data* is not representative of Twitter users. This is because not all Twitter users will tweet on a topic of interest, for example, during the Ebola epidemic of last year not all Twitter users would post a tweet related to Ebola. It is also important to remember that it is not always individuals that may be tweeting but also, organizations, and those in a non-personal capacity, for instance journalists. Moreover, as the Demos and Ipos MORI report notes, the data that Twitter produces does not reflect Twitter users, as often a small number of vocal accounts account for a significant proportion of any given dataset. See research by the Pew Internet Research related to the demographic of Twitter users.
- **Spam:** There is a large amount of link-baiting in popular hashtags (i.e., tweets designed for the users to click to be taken to a non-relevant website), and popular topics on Twitter can attract a large amount of spam. It may even be difficult to ascertain whether a user is *real*or a *fictitious*. Often fictitious accounts are set up either to (artificially) increase other users followers (celebrities, or politicians), but are also sold in retweet or favourite packages to fane popularity – where a large amount of users will retweet or favourite a user in large amounts. The extent to which Twitter contains fake accounts, retweets, and favourites is not known exactly, but the fact that these packages are available for cheap and can be found via a Google search suggests that they are popular among users.
- **The unknown:** There are most likely methodological issues around using social media data, in particular Twitter data, within research that at this time are not known. Therefore, caution should be urged when drawing inferences from Twitter data in-and-within itself in this emerging field. Follow updates on NatCen's New Social Media New Social Science (NSMNSS) blog, via their hashtag #NSMNSS, and my research blog.

Resources mentioned in the text above

- Association of Internet Researchers (AoIR)
- COSMOS Online Guide to Social Media Research and Ethics
- New Social Media New Social Science (NSMNSS blog)
- Pew Research Centre
- Research using Social Media; Users' Views
- Sifter (free estimate generation for Twitter data)
- The road to representivity a Demos and Ipsos MORI report on sociological research using Twitter
- Twitters API Terms of Service
- Unlocking the value of social media – a review of research ethics
- Wasim Ahmed, *a blog about my research*

Note: This article gives the views of the author, and not the position of the Impact of Social Science blog, nor of the London School of Economics. Please review our Comments Policy if you have any concerns on posting a comment below.

About the Author

Wasim Ahmed *is a PhD candidate at the Information School, at the University of Sheffield and the Twitter Manager for NatCen's Social Research network New Social Media New Social Science. Wasim has a very successful research blog which includes posts about key trends and issues within social media, but also covers more practical posts on using tools to capture and analyse social media data. Wasim is a keen Twitter user (@was3210), and will be happy to answer any technical (or non-technical!) questions you may have.*