methods

Data Collection on Twitter

5
CHAPTER

Devin Gaffney and Cornelius Puschmann

where does the data come from?
describing the #streaming, #REST, and
#search APIs to social scientists

In addition to being a versatile communications platform to users around the globe, Twitter is also an excellent source of current information. Data extracted from Twitter is used by researchers with different backgrounds (pollsters, marketers, academics from different disciplines) to answer a variety of questions, ranging from simple information about particular users or events (How many followers does a given user have? Who is the most active user tweeting under a certain hashtag?) to complex queries (Which users are central in a large network? How does information propagate among groups of users?). Some studies examine select individuals or small communities, while others require large volumes of information collected over long periods. Depending on the aims, different tools can be used to collect data—from Web-based analytics services that combine collection, analysis, and visualisation, to directly mining the Twitter API and interpreting the data using a dedicated statistics package.[1] Collecting data as part of a project, whether directly through the API or by using a dedicated software package, remains one of the most challenging aspects of Twitter-based research. While the technical and methodological requirements may seem daunting at first glance, an in-depth knowledge of the tools and the kind of data available through them can address many common concerns. In this chapter, we provide an overview of different techniques and their respective advantages and limitations. First, we discuss collecting data via Twitter's API, both directly and using a set of software packages, and then we turn to the question of how to integrate Twitter data into common social scientific study designs.

THE TWITTER APIS

Rather than offering a single API, three different Twitter data interfaces are available to researchers wanting to query the service: the Streaming API, the REST API, and the Search API. With few exceptions, the corpus of research generated to this point has relied on data collection through one of these three sources.

THE STREAMING API

The Streaming API is likely the most widely used data source for Twitter research. Typically, large-scale quantitative analyses of Twitter data are based on raw data collected through this source (Hong, Convertino, & Chi, 2011; Wu, Hofman, Mason, & Watts, 2011). It is worth pointing out that the Streaming API is a highly unorthodox kind of resource compared to how most other APIs function. In more traditional configurations, an API is "pull" based—the researcher requests a page of data from the server by requesting a URL, at which point the server returns the requested page. The Streaming API, however, is "push" based—that is, data is constantly flowing from the requested URL (the endpoint), and it is up to the researcher to develop or employ tools that maintain a persistent connection to this stream of data while simultaneously processing it. This stream of data is provided exclusively as a live poll, meaning that the moment a tweet is posted on Twitter, it becomes available. Because streaming data is supplied in the fashion of a live polling system not designed for historical analysis, research that takes a diachronic perspective is much more difficult than it would be via a traditional pull system, as the researcher must essentially operate on Twitter's schedule. When studying forms of relatively spontaneous organisation, such as the 'Arab Spring', 'Occupy', and 'Indignados' movements, data collection is especially difficult, as it may only be in hindsight that the event is recognised as such and its beginnings become significant (cf. Juris, 2012; Lotan et al., 2011; Vallina-Rodriguez et al., 2012 for such research). Studies of scheduled events, such as elections, require the researcher to be conscientious in establishing a stream for collecting data, ideally long before the event, in order to compile an analytically useful corpus. A central future challenge to the academic community will be to conduct complex and multifaceted analyses despite such restrictions, rather than tailoring research questions to data availability.

The Streaming API: Representative sampling

Fortunately the need to capture live data does not apply in the same way to all research contexts, and much of the research on Twitter to date asks more general questions, for example, by focussing on the platform's macroscopic structural properties (Kwak, Lee, Park, & Moon, 2010), or by describing how users conceptualise their communicative practices (Baym, Chapter 17 in this volume; Marwick & boyd, 2011). While Twitter is used in many countries and languages, user communities differ significantly in relation to their size, composition, and usage habits. Researchers should be keenly aware of seemingly small details that may be reflected in the data, for example, usage spikes over the course of a day, or fluctuations in activity during the weekend in comparison to workdays. Collecting data for prolonged periods of time is always preferable when possible, even if not all the data collected is used in the analysis. Many active users do not tweet daily, or perhaps even weekly, while others are highly active and skew the representativeness of a sample accordingly. Finally, not all quantitative research of Twitter is based on the contents of tweets: works such as Cha & Haddadi (2010) employ the social graph data available and focus entirely on follower-followee relations, rather than message content.

The Streaming API: Bandwidth limitations

The Streaming API is delivered in three bandwidths: "spritzer", "gardenhose", and "firehose", which deliver up to 1%, 10%, and 100% of all tweets posted on the system, respectively. By default, any regular user account on Twitter is granted spritzer access to the system, which is frequently sufficient for research purposes. The gardenhose is granted occasionally to users with defensible and compelling reasons for increased access, and the firehose is only available as a component of "a business relationship" with Twitter directly or through authorised re-sellers (Singletary, 2012). For the spritzer and the gardenhose, the percentage cap comes into effect only when more than the respective percentage of all tweets match the conditions placed on the stream. If, for example, a researcher is collecting data for a small conference, spritzer access will be sufficient to capture every tweet posted under the conference hashtag, since in only the most extreme cases will tweets about such an event exceed 1% of all traffic on the platform. It should be pointed out that according to the Twitter documentation, the respective sample sizes are based on the entirety of all information posted to Twitter as it is streamed, rather than on a subset of tweets to which a certain filtering criterion (e.g., a hashtag or keyword) applies. When in doubt about whether all desired content has been captured, researchers should check if their query returns a number of results close to 1% or 10% of the APIs' current total throughput. The streaming service also returns status messages indicating how many tweets have been missed if a cap has come into effect, notifying the researcher of the total number of tweets missed since the poll began.

The Streaming API: Endpoints and parameters

Beyond the three bandwidth options, the Streaming API offers two different methods, *sample* and *filter*, as points of access to data. *Sample* simply provides up to 1% or 10% of all tweets, selected at random. While the data has never been independently verified as random, it is generally assumed that it is of an acceptable degree of randomness. Crucially, two samples taken at the same point in time are identical, making reproducibility of results possible (Bruns & Liang, 2012; Hecht, Hong, Suh, & Chi, 2011). Inside the *filter* method, the *track*, *follow*, and *locations* parameters can be used to select specific results from the stream.

Track allows for researchers to search for multiple comma-delimited terms to be sent into Twitter's streaming request as an option. When Twitter receives the request, it only returns tweets that include those words, separated by non-word characters, to the researcher. In all but very few cases, the result will not exceed 1% of the total traffic on the platform, making this method combined with the *track* parameter a convenient way of compiling a keyword- or hashtag-based corpus.

Follow returns only tweets from a set of users represented by their collective comma-delimited user IDs. Currently, this parameter allows for collection from up to 5,000 accounts. Researchers intent on studying specific communities of users may find this method particularly useful for researching known groups of people for extended periods of time.

Locations provides an ideal access point for researchers interested in geographically bounded research. As of this writing, approximately 1% of all traffic on Twitter is "geotagged"—that is, an additional metadata object is appended to a tweet indicating its geographic origin. These geotagged tweets are represented as points, or latitude/longitude pairs that indicate a precise location, and polygons, or rectangles drawn by four pairs of points that can be as small as a city park or as large as a province. While the data is only 1% of all traffic, it is likely that this proportion will increase in the future, making it a highly attractive instrument for various kinds of geographically bounded research.

THE REST API

The REST (REpresentational State Transfer) API provides a set of methods for data interaction that is fundamentally different from the Streaming API, using the more traditional pull model. In total, over a hundred active methods are available in the REST API, few of which have been explored for research purposes. Using a combination of methods, the social graph data of a group of users can be assembled through this system, i.e. information on who is following whom and other data beyond the immediate content of individual tweets. Specifically, given a user of interest, two methods (*followers/ids* and *friends/ids*) can return listings of other users that follow or are followed by the user of interest at up to 5,000 user IDs per request. Further useful methods in the REST API provide access to trending topics, allow batch user lookups with groups of user IDs, and generally perform functions which are interesting in concert with the information that can be collected through the Streaming API.

REST API: Rate limiting

The REST API carries one heavy restriction: it is a rate-limited resource. Just as the Streaming API's spritzer and gardenhose levels of access are artificially limited to only a portion of traffic, the rate limit is in place largely to ensure reasonable traffic expectations for Twitter's infrastructure. This makes it very difficult for researchers to collect the data they desire in a timely manner, particularly in the case of REST requests. In the past, an un-authenticated computer could make 150 requests per hour. When any account logged into Twitter via Open Authentication (OAuth), this rate limit increased to 350 requests per hour. For using the followers/ids or friends/ids methods, this meant that at best, only 150 (or 350, when logged in) users could be processed for each method per hour. As of March, 2013, however, these limitations will be further reduced to approximately 60 requests per hour, and only OAuth requests to the API will be honoured (Sippey, 2012). If, for example, the researcher collects information about a user who has 563 friends and 178 followers, it would be possible to collect all friends and followers with one request to each of these methods, for a total cost of two requests in a one-hour window.

If that user has 5,630 friends and 1,780 followers, the amount would increase to three requests (as friends/ids can, at most, return 5,000 accounts per page of data). Previously, researchers could apply for IP-based and account-based white-listings, which, if granted, increased this limitation to 20,000 requests per hour. Twitter has since ended this practice, and does not hand out new white-listings. boyd and Crawford (2012) have put forward a compelling argument around the artificial class division that is created with this distinction of high-throughput accounts and everyone else (see also Puschmann & Burgess, Chapter 4 in this volume). While there are ways to circumvent limitations—for example, by setting up large networks of computers that collect data in tandem—such practices are actively monitored by Twitter, and violators are punished by blacklisting their accounts.

THE SEARCH API

Early analyses of Twitter, such as Gaffney's (2010) work on the Iran election of 2009, were largely based on the Search API. Originally, this was the only point of access for searching for tweets that mentioned hashtags, and was therefore widely used for event-based research. Similar to the REST API, it is a pull-based resource, and essentially replicates the functionality of Twitter's search function. While, in theory, some historical collection of data is still possible through the Search API, in practice its utility is severely limited. Data loosely falls off of the search system within a week of being posted, and no reliable information is available on its completeness. Twitter actively discourages use of the Search API and plans to discontinue it in the near future, as it is costly to maintain and was never intended for high-throughput real-time data dissemination.

TOOLS

While virtually all access to Twitter data takes place through one of the APIs (and often via a combination of several API methods), a number of tools exist to simplify this process. Rather than having to make API calls directly, researchers can use them to specify what data they want to collect. Client-based programs such as The Archivist or TAGS come with the constraint that they must be run on a regular basis from the user's computer to collect data. By contrast, server-based collection methods such as yourTwapperKeeper and Twitter Database Server run around the clock, collecting data whenever it is made available. This process is further simplified by Web-based services such as 140kit, that provide more in-depth analytical capabilities than services not designed for research, but at the same time restrict download access to tweets to conform with Twitter's Terms of Service, which bar the republication of full tweets without the company's consent (see Beurskens, Chapter 10 in this volume). Finally, data resellers such as Gnip and DataSift provide extensive historical data without the challenges of collection, but at a premium.

Table 5.1: Software Packages for Twitter Data Collection

Tool	Requires Hosting?	Requires Programming?	Provides Raw Data?	Provides Analytics?	Paid Service?
140kit	No	No	No	Yes	No
140kit Source Code	Yes	Yes	Yes	Yes	No
yourTwapperKeeper	Yes	Yes	Yes	No	No
The Archivist	No	No	No	Yes	No
TAGS	No	No	Yes	Yes	No
Twitter Database Server	Yes	Yes	Yes	No	No
Gnip	No	No	Yes	Yes	Yes
DataSift	No	No	Yes	Yes	Yes

THE ARCHIVIST

The Archivist (TA) is a free and open-source desktop application that runs on Windows based on Microsoft's .NET framework, and is among the simplest tools for saving and analysing tweets. In contrast to most other available tools, TA does not require a Web server. Each instance can collect tweets that include a certain keyword or hashtag, retrieved through the Search API. Use of the Search API makes TA subject to its limitations, a problem likely to become more severe in the future.

Retrieving large volumes of information or historical data is not generally possible, and for continuous retrieval, the collecting machine must run constantly. TA is recommendable only for small collections of tweets that can be manually verified for consistency, such as small hashtag archives and individual user streams. For any research requiring a reliable sample that cannot be manually verified, using TA or any other desktop software cannot be advised, as issues of latency, bandwidth, and stability are likely to impact the quality of the sample.

TAGS

The Twitter Archiving Google Spreadsheet (TAGS) is a Web-based script that can be used for the cloud-based collection of tweets. Running under Google Spreadsheets, TAGS is able to retrieve Twitter data through the REST API and consequently is subject to rate limiting, especially when used without authenticating with Twitter. It sidesteps some of the limitations of The Archivist by being hosted, while at the same time not requiring users to run their own server. Like The Archivist, TAGS performs a number of statistical operations on the extracted data, facilitating analysis. While it is not necessary to install any software, TAGS has an interface that is slightly less intuitive than that of The Archivist and requires a Google account.

YOURTWAPPERKEEPER

YourTwapperKeeper (YTK) is one of the most popular tools available to researchers wanting to simplify the process of extracting data. Written in PHP, YTK is among the most accessible projects currently available. It leverages both Twitter's Streaming API and the Search API to collect tweets that match a given term. In order to run the code, a researcher must employ an active PHP connection and be able to run a pair of scripts—the stream (Streaming API requests) and crawl (Search API requests) scripts. Additionally, it requires a MySQL database in order to store the information collected.

YTK (like its precursor TwapperKeeper) has been used in a number of research projects (Papacharissi & de Fatima Oliveira, 2012; Wilson & Dunn, 2011), and is frequently cited as an appropriate route for data collection (Bruns & Liang, 2012). There are some drawbacks to employing YTK, however. Most importantly, it only captures a small portion of the range of metadata currently available through the Streaming API with each tweet. Of the data that Twitter provides, YTK only collects the tweet's text and a few basic metadata attributes, which are then stored in a single table, rather than saving all available information and performing preprocessing to facilitate analysis. As a result, questions related to links, geographic places, and less orthodox questions focussed on particular metadata attributes cannot be easily investigated unless the researcher post-processes the data or alters their installation's source code.

TWITTER DATABASE SERVER

Twitter Data Server (TDS) is another server-based solution for collecting hashtag data. Like YTK, it is based on PHP and the MySQL database server. It provides only an absolutely minimal interface for browsing the captured data, instead relying on the user's ability to interact with the MySQL tables that contain the captured tweets through the command line or via a third-party utility such as the popular phpMyAdmin. While running and interacting with TDS is somewhat more complicated than YTK, TDS appears to consume less computational resources and capture additional fields, such as resolved URLs, not provided by YTK.

140KIT

140kit[2] is a Web-based tool for the analysis of Twitter data. Unlike the other services mentioned, 140kit ensures that all metadata fields are collected. In its online version, no raw data is allowed to be downloaded due to Terms of Service limitations imposed by Twitter on public datasets (Twitter, 2012). The software is available as a hosted platform and as a stand-alone package similar to yourTwapperKeeper (140kit Source Code), though the language that the program is written in, Ruby and Ruby on Rails, is less prevalent than PHP and may be more difficult to implement, depending on the kind of Web server available. Like YTK and TDS, it also requires a computer to run the software constantly in order to stream data. Additionally, it only supports the Streaming API, though researchers can extend functionality if required. In contrast to most other analytics tools, 140kit is built specifically with researchers in mind as its target audience, and may thus be more suitable for answering questions outside of commercial contexts.

GNIP AND DATASIFT

Gnip is one of the best tools available to researchers in terms of data quality, but it comes at a premium. The company collects data through firehose access to Twitter, and re-sells this data to both the research community and private businesses. Access for 10% and 50% of the Streaming API, alongside several other datapoints of note (such as integration with Klout scores), costs US$5,000 and US$30,000 per month, respectively (Small, Kasianovitz, Blanford, & Celaya, 2012). While the availability of historical data makes the service potentially relevant, the considerable costs of the service and a lack of complete metadata (similar to YTK) may deter researchers from using it.

Similar to Gnip, DataSift provides much of the same functionality, though its pricing scheme differs, with the lowest level of access costing US$3,000 per month at the time of writing. Also similar to Gnip, some metadata fields that may be of interest to the researcher are omitted, though it also adds other datapoints of note (again, integration with Klout scores). A downside of both services is that by targeting businesses, not all of the information that may be relevant to researchers is made available through them.

CONCLUSION: LIMITATIONS

The limitations of social scientific research based on Twitter data stem from constraints which impact research projects on different levels. As with any other methodology, not all types of data and forms of analysis align themselves equally well with all kinds of research questions. Because of its tendency to be data- rather than question-driven, much of the current quantitative research on Twitter focusses on measuring and comparing specific structural parameters in very large data samples, sometimes with little regard for the theoretical salience of these parameters. This is understandable before the backdrop of Big Data research as a fundamentally new approach to finding patterns, relationships, and links between elements, rather than paradigmatically theorising the meaning of said elements beforehand (Lazer et al., 2009). An ideal study should be well grounded in a specific set of research questions and query the data in accordance with them. In contrast to traditional instruments such as surveys and conventional content analysis, it is important to note that even the exploratory phase of research is markedly quantitative when exploring social media. Since searching, filtering, and ranking are the only feasible way to make masses of content readable to the human researcher, they form a logical first step in any analysis, even in qualitative studies. At the same time, quantitative research should present data as it pertains to the questions asked, rather than simply because it is possible and large volumes of data have been collected. Furthermore, there is the question of how representative Twitter users are of the overall population—both on Twitter and beyond it. Adoption rates and usage strategies differ greatly, casting doubt on claims of representativeness. When making judgments about populations of Twitter users based on tweets, those users who mainly read but hardly post may be overlooked, while the significance of highly vocal users may be given too much weight. Inferences about the population at large based on Twitter are difficult as a result of this inherent skew, yet without generalisation the potential for sociological research is limited, in spite of much enthusiasm for Twitter as a data source (e.g., Golder & Macy, 2012).

A third and distinct set of limitations is technology-based. As has been pointed out, there is no way of checking how completely a given data set captures what flowed through Twitter at the time that it was compiled. Without firehose access, researchers rely entirely on Twitter to provide a representative sample of what is there. At the same time, it is important to note that this is not solely because of Twitter's need to monetise its data, but a result of the unique challenge of building an infrastructure powerful enough to store such vast quantities of information in real time. Incomplete data sets can hamper an analysis, yet asking in hindsight for a complete archive of tweets related to a past event is impossible. Not only does this make new research difficult, it also makes absolute reproducibility extremely hard to achieve, with obvious implications for the validity of research results.

Why quantify to begin with? Qualitative sociological research on Twitter comes with its own unique potentials, but also with its own set of constraints, for example, with regard to privacy. Arguably, Twitter's strength lies in the ability to gain interesting insights from short and often highly context-bound messages, yet these are also difficult to interpret and carry a range of meanings for different stakeholders. While a "deep", qualitative approach is more nuanced than computational procedures, it is also severely limited in its scale. By choosing an object that can be studied in detail, the researcher makes specific choices about her object of study. On the other hand, the "shallow" aggregation of data always risks arriving at judgments that are ill-supported because they are based on incorrect or overreaching implicit assumptions. Relying on one's informed experience from other contexts when considering Twitter as a source of data, and pragmatically deciding how research objectives can be aligned with the technical and methodological challenges at hand will usually produce the best result.

NOTES

[1] The acronym API stands for application programming interface. APIs are data interfaces offered by many Web platforms. Their main purpose is to provide software developers an unambiguous, data-only version of a site's content for use in their own software.

[2] One of the authors of this chapter is the principal developer of 140kit.

REFERENCES

boyd, d., & Crawford, K. (2012). Critical questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662–679.

Bruns, A., & Liang, Y. E. (2012). Tools and methods for capturing Twitter data during natural disasters. *First Monday*, *17*(4), 1–8.

Cha, M., & Haddadi, H. (2010). Measuring user influence in Twitter: The million follower fallacy. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (ICWSM '10)* (pp. 10–17). Menlo Park, CA: AAAI Press.

Gaffney, D. (2010). #iranElection: Quantifying online activism. *Proceedings of the WebSci10: Extending the Frontiers of Society On-Line* (pp. 1–8). 26–27 Apr. 2010. Raleigh, NC.

Golder, S., & Macy, M. (2012). Social science with social media. *ASA Footnotes*, *40*(1), 7.

Hecht, B., Hong, L., Suh, B., & Chi, E. H. (2011). Tweets from Justin Bieber's heart: The dynamics of the 'location' field in user profiles. *Proceedings of the International Conference on Human Factors in Computing Systems (CHI '11)* (pp. 1–10). Vancouver, British Columbia, Canada: ACM Press.

Hong, L., Convertino, G., & Chi, E. H. (2011). Language matters in Twitter: A large scale study characterizing the top languages in Twitter characterizing differences across languages including URLs and hashtags. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM '11)* (pp. 518–521). Menlo Park, CA: AAAI Press.

Juris, J. S. (2012). Reflections on #Occupy Everywhere: Social media, public space, and emerging logics of aggregation. *American Ethnologist*, *39*(2), 259–279.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media? *Proceedings of the 19th International Conference on the World Wide Web (WWW'10)* (pp. 1–10). Raleigh, NC.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A-L., Brewer, D., . . . Van Alstyne, M. (2009). Computational social science. *Science*, *323*(5915), 721–723.

Lotan, G., Graeff, E., Ananny, M., Gaffney, D., Pearce, I., & boyd, d. (2011). The revolutions were tweeted: Information flows during the 2011 Tunisian and Egyptian revolutions. *International Journal of Communication*, *5*, 1375–1405.

Marwick, A., & boyd, d. (2011). To see and be seen: Celebrity practice on Twitter. *Convergence: The International Journal of Research Into New Media Technologies*, *17*(2), 139–158.

Papacharissi, Z., & de Fatima Oliveira, M. (2012). Affective news and networked publics: The rhythms of news storytelling on #Egypt. *Journal of Communication*, *62*(2), 266–282.

Singletary, T. (2012). How do I get firehose access? *Twitter Developer Forums*. Retrieved from https://dev.twitter.com/discussions/2752

Sippey, M. (2012). Changes coming in Version 1.1 of the Twitter API. *Twitter Developer Blog*. Retrieved from https://dev.twitter.com/blog/changes-coming-to-twitter-api

Small, H., Kasianovitz, K., Blanford, R., & Celaya, I. (2012). What your tweets tell us about you: Identity, ownership and privacy of Twitter data. *International Journal of Digital Curation*, *7*(1), 174–197.

Twitter. (2012, 25 June). Terms of service. Retrieved from http://twitter.com/tos

Vallina-Rodriguez, N., Scellato, S., Haddadi, H., Forsell, C., Crowcroft, J., & Mascolo, C. (2012). Los Twindignados: The rise of the Indignados movement on Twitter. *Proceedings of the ASE/IEEE International Conference on Social Computing (SocialCom'12)* (pp. 1–6). Amsterdam, The Netherlands.

Wilson, C., & Dunn, A. (2011). Digital media in the Egyptian revolution: Descriptive analysis from the Tahrir datasets. *International Journal of Communication*, *5*, 1248–1272.

Wu, S., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Who says what to whom on Twitter. *Proceedings of the 20th International Conference on the World Wide Web (WWW '11)* (pp. 705–714). New York, NY: ACM Press.