International Journal of Scientometrics, Informetrics and Bibliometrics
ISSN 1137-5019

VOLUME 18-19 (2015): ISSUE 1. PAPER 1

Evaluating the Comprehensiveness of Twitter Search API Results: A Four Step Method

Mike Thelwall

Statistical Cybermetrics Research Group, School of Technology
University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1SB
United Kingdom
E-mail: **m.thelwall@wlv.ac.uk**

Abstract

The Twitter Search API (Applications Programming Interface) is a free service that allows software to automatically submit searches to Twitter and retrieve matching tweets. It is widely used to gather tweets for social science and other research, although this is not its main purpose. It does not guarantee to be comprehensive, however, so this article introduces a simple method to check the coverage of its results for narrowly focused topics. An application of the method shows that the results are incomplete, put possibly only due to the filtering out of duplicate, potentially offensive or conversational content

Keywords

Twitter, social media, Twitter API

Introduction

The social web has provided a large source of data for social science research in the form of the comments posted by users to various sites, such as Twitter, on multiple topics (Beer, 2012; Zimmer, & Proferes, 2014). For example, researchers have investigated social web sites for evidence of specific topics or activities, such as earthquake warnings (Sakaki, Okazaki, & Matsuo, 2010), news stories (Thelwall, Buckley, & Paltoglou, 2011), crisis communication (Hughes, & Palen, 2009), politics (Vilares Calvo, Thelwall, & Alonso, 2015) although predicting elections is controversial (Ceron, Curini, & Iacus, 2015; Jungherr, Jürgens, & Schoen, 2012; Tumasjan, Sprenger, Sandner, & Welpe, 2010), link analysis (Orduña-Malea, Torres-Salinas, & Delgado López-Cózar, 2015) and television viewing (Pittman, & Tefertiller, 2015). Twitter is a particularly attractive source because of its wide social userbase, its international coverage (Wilkinson & Thelwall, 2012), although it is very uneven (Orduña-Malea, Torres-Salinas, & Delgado López-Cózar, in press), its high volume of posting, and the provision of free simple access to it by Twitter through its Applications Programming Interface (API). This API does not guarantee to give comprehensive search results but is volume limited so that it cannot be used to download all tweets. Moreover, it does not reveal the process used to select the tweets returned. This unknown sampling method is problematic for research uses of the data because it may have biases that affect the research findings.

The Twitter Search API returns only tweets from the most recent seven days and is rate-limited (see next section). Each researcher investigating Twitter must gather their own tweets, however, because the organisation prohibits data sharing by academics (Wisdom, 2013). Users needing older or more comprehensive data can purchase it from an official data reseller, such as DataSift (until August 2015: DataSift, 2015) or Gnip, or from a commercial analytics provider, such as Pulsar, but there are no academic discounts on Twitter's basic charge rates.

A simple method is needed to assess the comprehensiveness of a set of tweets returned by a query because Twitter's algorithms do not guarantee comprehensive results, are likely to change over time, because the results may vary by query, and because it is good research practice to assess the comprehensiveness of data sets used, when possible. In response, this article introduces a new simple method to evaluate a result set. The method is available in the free software Webometric Analyst. The technique exploits the comprehensiveness of the Twitter Data API timeline queries. These allow the most recent approximately 3000 tweets from a user to be downloaded, even if they are older than seven days. The assumption behind the new method is that people are likely to make multiple posts on narrowly focused topics. Hence downloading the timelines of users that have posted at least once on the topic may reveal additional posts from them that may not have been returned by the original queries. This method is described in detail below and applied to three different data sets and evaluated qualitatively.

The Twitter APIs

Twitter is a massive source of public opinion and comments about a wide variety of topics and is a potentially very valuable source of data for research. As a result of this, it is important to have high quality effective tools for gathering tweets.

Twitter maintains several public APIs that can be used to access Tweets, including one based on a random sample, one based on searching an unspecified section of Twitter, and one comprehensive source of

current tweets.

The *Firehose API* (Twitter, 2015a) allows all tweets to be automatically downloaded as they are posted to Twitter but is only available to a limited set of users. These are presumably commercial users with a special arrangement that pay substantial price for their access and usage rights.

The *Streaming API* allows random tweets to be automatically downloaded in the form of "a small random sample of all public statuses" (Twitter, 2015b). This seems to return up to a maximum of 1% of current statuses, but less at busy times (Purohit, Hampton, Bhatt, Shalin, Sheth, & Flach, 2014). Twitter previously provided a 10% streaming API to businesses, known as its gardenhose service (GNIP, 2011; Twitter, 2015e). The Streaming API can also be filtered with search terms and other parameters and its coverage may be comprehensive if the filter is restrictive enough. Commercial accounts (i.e., with "elevated status") can also retrieve older data (Twitter, 2015e).

The *Search API* (Twitter, 2015c) does not explicitly limit the volume of tweets downloaded but limits the number of queries submitted (Twitter, 2015d) to a maximum per time interval (e.g., 15 or 180 per 15 minutes). It downloads tweets matching the query from the previous up to two weeks, in contrast to the Firehose and Streaming APIs, both of which only return current statuses (i.e., the most recent tweet) and not previous tweets.

The Search API is the focus of the current article. If an application can use the Search API to download all available tweets matching its queries without triggering the rate limit then, in theory, its set of tweets will be comprehensive for the queries. The Twitter documentation suggests otherwise, however, because "the Search API is focused on relevance and not completeness. This means that some Tweets and users may be missing from search results" (Twitter, 2015c, para 2; see also GNIP, 2011). The lack of guaranteed comprehensiveness is a problem for research uses of the Twitter API (Khanaferov, Luc, & Wang, 2014; Levine, Mann, & Mannor, 2015; Tsuya, Sugawara, Tanaka, & Narimatsu, 2014) and particularly for social science or medical applications (Black, Mascaro, Gallagher, & Goggins, 2012; Small, Kasianovitz, Blanford, & Celaya, 2012; Sugumaran & Voss, 2012) because of the unknown sampling method.

One previous article has systematically analysed Twitter API search results, but this used two very high volume queries (sex and sports) and so its finding that increasing the time period between data collection from 1 minute per request to 2 or 3 minutes per request results in a reduced number of total results is unsurprising. Nevertheless, the decrease in results was very small and they did not check for results missed by all of the queries (Black, Mascaro, Gallagher, & Goggins, 2012).

Research Questions

The following research questions are part of the overall goal of evaluating the effectiveness of the new four step method, which is described below in the Methods section.

1. Is the method capable of identifying matching tweets not returned by Twitter API searches for low volume topic-focused queries?
2. Are the missing results always of types that are not relevant for most academic analyses, such as spam, retweets or duplicate content?

Methods

The method first collects the tweets from the Twitter Data API using the standard queries running over a specific period of time, then filters out tweets that fall outside of the time frame covered. To ensure that no tweets are excluded that might have been returned by the API, the search logs must be checked to ensure

that the searches are started before the beginning of the time period analysed and at some stage before the start was not curtailed by the rate limit. This is necessary because if the rate limit is met then some tweets may have been lost for this reason. In addition, the rate limit must not have been met during the data collection period to ensure no lost data. If the rate limit was met during the period then no data will have been lost to rate limiting if a) after the end of the data collection the tweets are dated after the end of the analysis time frame and b) at no stage during the data collection the time lag between data collection and the date of the tweets returned is less than seven days. The latter conditions ensure that no data is lost to rate limiting because each query returns tweets in chronological order using the Twitter ID of each tweet and starting with the ID following the largest previously found. This produces the standard set of topic tweets.

The second stage of the method is to download all tweets from the timelines of the users that posted any relevant tweets in the data set. Timelines are downloadable from a different part of the API than searches and are retrieved in reverse chronological order and so timeline downloading can be curtailed for any user once tweets have been returned from before the start of the survey period. This is important as a practical step because timeline downloading is rate limited and so downloading the timelines of many users can be prohibitively time consuming.

The third stage is to filter the second stage tweets to identify those that match the original queries and time period. Assuming that no data has been lost at any stage, the results of stages 1 and 3 should be identical.

The fourth stage is to compare the stage 1 and 3 results and qualitatively evaluate the differences.

Table 1. The four stage method to check for missing tweets that match a query.

Stage	Description
1	Extract tweets matching a query from the Twitter API, ensuring no tweets have been lost due to rate limiting or gaps during the data collection
2	Download all Tweets from the timelines of all users with at least one tweet in stage 1.
3	Extract stage 2 tweets that match the query and time period of data collection.
4	Compare the stage 2 and stage 3 tweets and investigate the differences qualitatively.

The above four step method (Table 1) is limited in that it will not find any tweets from users that had no tweets in the original data set. It is impossible to find all these without buying a complete set of matching tweets from Twitter or a data reseller, however.

The following queries were chosen to represent different types of search but to be relatively narrowly focused in order to increase the chance that the same person sends multiple matching tweets. Ten queries were originally chosen but one produced no missing tweets and so an eleventh query was chosen to replace it. This is an ad-hoc collection because there does not seem to be a reasonable way to create a representative set

of queries.

#POC15: Wellcome Press Officers Conference

#EUMacro: Macroecology meeting 14-16 June 2015

#DistractinglySexy: Anti-chauvinism in science humour hashtag

Philae: Spacecraft lander on asteroid

Chomsky: Leading academic

@imperialcollege: Major UK university account

@Pulsar_Social: Social media company

@UNDP: United Nations Development Programme

@EverydaySexism: Account monitoring sexism

"Gajendra Chauhan": Students in India protest against the appointment of a Hindu nationalist politician to an important post

“United Nations”: Major international organisation.

Results

Most of the eleven queries returned close to the maximum number of tweets although there were differing numbers of users per tweet (Table 2). Not all of the tweets returned by the queries were also found in the timelines. This discrepancy may cover tweets that were replies to other tweets and would not make sense on their own.

Table 2. An analysis of the overlap between tweets matching a query and tweets returned by the timelines of users with at least one tweet matching the query (June 2015).

Query	Date and time of first tweet returned by query	Date and time of last tweet returned by query	Number of tweets returned by query	Number of timelines (users) of tweets returned by query	Tweets in timelines of tweets returned by query	Number of tweets returned by query and also in timelines
#POC15	Jun 12 09:12:36	Jun 12 15:16:18	95	34	10417	95
#EUMacro	Jun 08 12:55:30	Jun 17 06:20:00	1572	237	74418	1568
#DistractinglySexy	Jun 13 04:17:05	Jun 13 06:24:48	1987	1012	273653	1983
Philae	Jun 14 22:11:48	Jun 14 21:50:17	1987	1761	861637	1971
Chomsky	Jun 12 08:27:29	Jun 14 10:43:59	1989	1716	215620	1776
@imperialcollege	Jun 04 16:11:33	Jun 13 18:12:34	639	473	147430	602

@Pulsar_Social	Jun 08 13:10:16	Jun 16 21:36:51	154	86	28392	153
@UNDP	Jun 13 08:49:02	Jun 14 18:39:49	1983	1328	507285	1820
@EverydaySexism	Jun 09 08:08:33	Jun 14 06:11:21	1984	1481	496454	1781
"Gajendra Chauhan"	Jun 12 10:47:09	Jun 13 10:55:14	1989	1269	751798	1942
“United Nations”	Jun 14 14:58:55	Jun 15 06:22:03	1985	1566	414337	1902

In all cases except one, at least one tweet was found in a timeline that matched the original query and was posted during the time window of the query but was not retuned by the query (Table 3). There does not seem to be a pattern in terms of the type of query affecting the percentage of missing results. In most cases, one of the following was a logical reason for the tweet being omitted from the query results.

- RTs: The query was a retweet so the original matching tweet would be the logical tweet to include.

- No words in tweet: The tweet consisted purely of hashtags, hyperlinks and/or usernames and so could be withheld by twitter as potential spam.

- The tweet was not in English and so would not match the original query, which was set to English only.

- The tweet was a reply to a previous tweet and so might be withheld by Twitter because it might make sense on its own (e.g., one tweet was “No” and a hashtag and username).

- The tweet was not included but another identical tweet or a RT of was included in the query sample.

- The tweet contained a likely spam word or strong language and might be withheld as potentially offensive.

Four tweets were missing from the query results but did not match any of the above conditions. Nevertheless, these all had a possible reason for omission.

- #DistractinglySexy: Tweet contains several hashtags and a hyperlink to site with many adverts, so might have been classed as Spam by Twitter.

- Philae, Chomsky: Tweet hyperlinks to another tweet – may be part of a conversation although this is not clear from viewing the tweet.

- @UNDP: Tweet contains a broken link.

Table 3. An analysis of missing tweets found in users’ timelines matching the query parameters but not returned by the query. Percentages reported are for the number of tweets found in timelines matching the query as a fraction of all tweets from both sources matching the query. More than one reason may apply to a tweet, except in the case of the Other category.

Query	Timeline	RTs	No*	Not	Reply	RT or	Strong	Other

	tweets missing from queries		words in tweet	English		copy in set	language	
#POC15	0 (0%)	0	0	0	0	0	0	0
#EUMacro	25 (2%)	12	9	10	3	0	0	0
#DistractinglySexy	81 (4%)	51	0	56	1	0	0	1
Philae	25 (1%)	7	0	20	0	1	0	1
Chomsky	18 (1%)	2	0	14	0	2	0	1
@imperialcollege	5 (1%)	1	0	4	0	0	0	0
@Pulsar_Social	6 (4%)	2	0	4	0	0	0	0
@UNDP	20 (1%)	14	0	5	0	0	0	1
@EverydaySexism	26 (1%)	1	1	16		3	4**	0
"Gajendra Chauhan"	75 (1%)	29	0	38	0	9	0	0
“United Nations”	8 (0%)	2	0	2	0	6	0	0

*other than hashtags, hyperlinks, usernames
**pussy, dickhead, masturbating, rape

Discussion

The results clearly demonstrate that keyword search results from the Twitter API are an incomplete sample of tweets, even for queries that do not trigger the Twitter API rate limit. This seems to be true for all types of query (hashtags, phrases, keywords, usernames), without major differences between them. The excluded results appear to be either duplicates or poor quality in some way and so it seems that the most important tweets may still be returned by keyword searches. The results do not prove this, however, but only demonstrate that such a conclusion is consistent with the data, at least for the eleven queries analysed here. The findings are not conclusive for three reasons. First, although the proposed reasons for omission seem to be plausible, they are speculative in the absence of confirmation from Twitter. Second, it is possible that tweets are excluded from the API by user, and perhaps on an arbitrary basis, and the method here would not reveal this. This seems to be unlikely, however, except for spam users and users not tweeting in English, because Twitter seems to process tweets predominantly as separate entities rather than as collections from users. Although Twitter states that some users may be missing from the results (Twitter, 2015c, para 2), it has excluded some tweets from users from which other tweets were excluded, suggesting that the user is not the fundamental unit of analysis for Twitter. An exception may occur for users that are blacklisted as spammers, however, who would presumably have all of their (spam) tweets removed. Third, there may be other queries or other types of query for which many more matching tweets are not returned by Twitter API searches.

Another way to check the comprehensiveness of the results returned by the API is to compare them to the tweets reported by the web interface of Twitter for the same query. This is possible by repeatedly scrolling down the results of the web page, which triggers updating the list of tweets and this can be continued until all the tweets are shown or the tweets are older than the time period. These tweets can then be compared to the API tweets. This comparison is not straightforward to the formatting of the results by Twitter, but is possible. Comparisons for the #EUMacro sample using this method did not find any of the tweets from the timeline that were not in the search API results. This is consistent with the conclusion that the tweets missing from the search results may have been deliberately excluded as duplicate or spam. If true, then the proportion of

tweets removed for a query is likely to vary, for example with commercial or popular terms that attract a lot of spam being more affected by the spam removal than other types of queries.

Conclusions

The results give some confidence in the use of the Twitter API to generate reasonably comprehensive sets of tweets for a query, albeit with some low quality and duplicate tweets removed. The removal of such tweets is unlikely to be problematic for most research uses of Twitter, except for Spam detection, and so it seems reasonable for researchers to continue to use it. The main research drawback of the filtering is that it is a hidden processing step that affects the sample to be analysed and this type of unknown is undesirable in research. Nevertheless, in the absence of practical alternatives and in the light of the numerous sampling problems that almost always affect social science research, it does not seem to be a major problem. This assumes that the ongoing initiative to archive Twitter at the Library of Congress is unsuccessful, does not grant remote access, or provides tweets that are too old for research uses (see also: Zimmer, 2015).

The four step method introduced here to check Twitter's coverage of a particular topic has shown that it is capable of identifying missing tweets, although it is not capable of estimating the proportion of missing tweets. Its weaknesses are that is not capable of identifying individual tweeters that are completely missing from the results and it is not suitable for topics on which individuals are unlikely to post multiple tweets. Nevertheless, it may be worth applying in future research as a simple step to check for more aggressive filtering by Twitter, such as the exclusion of tweets from the results without an obvious reason. It would be interesting to apply similar methods to assess indirect sources of tweets used by researchers, such as Topsy (Zimmer, & Proferes, 2014), to see how they compare.

References

Aboukhalil, R. (2013). Using the Twitter API to mine the Twitterverse. **XRDS: Crossroads, The ACM Magazine for Students,** 19(4), 52-55.

Beer, D. (2012). Using social media data aggregators to do social research. **Sociological Research Online**, 17(3), 10.

Black, A., Mascaro, C., Gallagher, M., & Goggins, S. P. (2012). Twitter zombie: Architecture for capturing, socially transforming and analyzing the Twittersphere. In **Proceedings of the 17th ACM international conference on supporting group work** (pp. 229-238). New York: ACM Press.

Ceron, A., Curini, L., & Iacus, S. M. (2015). Using sentiment analysis to monitor electoral campaigns method matters— Evidence from the United States and Italy. **Social Science Computer Review**, 33(1), 3-20.

DataSift (2015). Twitter Ends its Partnership with DataSift – Firehose Access Expires on August 13, 2015. **http://blog.datasift.com/2015/04/11/twitter-ends-its-partnership-with-datasift-firehose-access-expires-on-august-13-2015/**

GNIP (2011). Guide to the Twitter API – Part 3 of 3: An overview of Twitter's Streaming API. **https://blog.gnip.com/tag/gardenhose/**

Hughes, A. L., & Palen, L. (2009). Twitter adoption and use in mass convergence and emergency events. **International Journal of Emergency Management**, 6(3-4), 248-260.

Khanaferov, D., Luc, C., & Wang, T. (2014). Social network data mining using natural language processing and density based clustering. In 2**014 IEEE International Conference on Semantic Computing (ICSC2014)**, Los Alamitos, CA: IEEE Press (pp. 250-251).

Jungherr, A., Jürgens, P., & Schoen, H. (2012). Why the pirate party won the German election of 2009 or the trouble with predictions: A response to Tumasjan, A., Sprenger, T.O., Sander, P.G., & Welpe, I.M. "predicting elections with twitter: What 140 characters reveal about political sentiment". **Social Science Computer Review**, 30(2), 229-234.

Levine, N., Mann, T. A., & Mannor, S. (2015). Actively learning to attract followers on Twitter. arXiv preprint arXiv:1504.04114.

Orduña-Malea, E., Torres-Salinas, D., & Delgado López-Cózar, E. (2015). Hyperlinks embedded in Twitter as a proxy for total external in-links to international university websites. **Journal of the Association for Information Science and Technology**, 66 (7), 1447-1462.

Pittman, M. & Tefertiller, A.C. (2015). With or without you: Connected viewing and co-viewing Twitter activity for traditional appointment and asynchronous broadcast television models. **FirstMonday** 20(7) **http://firstmonday.org/ojs/index.php/fm/article/view/5935** DOI: 10.5210/fm.v20i7.5935

Purohit, H., Hampton, A., Bhatt, S., Shalin, V. L., Sheth, A. P., & Flach, J. M. (2014). Identifying seekers and suppliers in social media communities to support crisis coordination. **Computer Supported Cooperative Work (CSCW)**, 23(4-6), 513-545.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes Twitter users: real-time event detection by social sensors. In **Proceedings of the 19th international conference on the world wide web** (pp. 851-860). New York: ACM Press.

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.

Sugumaran, R., & Voss, J. (2012). Real-time spatio-temporal analysis of West Nile Virus using Twitter data. In P**roceedings of the 3rd International Conference on Computing for Geospatial Research and Applications** (p. 39). New York: ACM Press.

Thelwall, M., Buckley, K., & Paltoglou, G. (2011). Sentiment in Twitter events. **Journal of the American Society for Information Science and Technology**, 62(2), 406-418.

Tsuya, A., Sugawara, Y., Tanaka, A., & Narimatsu, H. (2014). Do cancer patients tweet? Examining the Twitter use of cancer patients in Japan. **Journal of medical Internet research**, 16(5), e137. doi: 10.2196/jmir.3298

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment. **International Conference on Weblogs and Social Media (ICWSM210)** Los Alamitos: IEEE Press (pp. 178-185).

Twitter (2015a). Firehose. **https://dev.twitter.com/streaming/firehose**

Twitter (2015b). GET statuses/sample. **https://dev.twitter.com/streaming/reference/get/statuses/sample**

Twitter (2015c). The Search API. **https://dev.twitter.com/rest/public/search**

Twitter (2015d). API Rate Limits. **https://dev.twitter.com/rest/public/rate-limiting**

Twitter (2015e). Count. **https://dev.twitter.com/streaming/overview/request-parameters#count**

Vilares Calvo, D., Thelwall, M., & Alonso, M.A. (2015). The megaphone of the people? Spanish SentiStrength for real-time analysis of political tweets. **Journal of Information Science**, 41 (6), 799-813. doi: 10.1177/0165551515598926

Wilkinson, D., & Thelwall, M. (2012). Trending Twitter topics in English: An international comparison. **Journal of the American Society for Information Science and Technology**, 63(8), 1631-1646.

Wisdom, D. (2013). How Twitter gets in the way of knowledge. BuzzFeed News. **http://www.buzzfeed.com/nostrich/how-twitter-gets-in-the-way-of-research**

Zimmer, M., & Proferes, N. J. (2014). A topology of Twitter research: Disciplines, methods, and ethics. **Aslib Journal of Information Management**, 66(3), 250-261.

Zimmer, M. (2015). The Twitter Archive at the Library of Congress: Challenges for information practice and information policy. **First Monday**, 20(7). **http://journals.uic.edu/ojs/index.php/fm/article/view/5619/4653**

Received 8/July/2015
Accepted 22/July/2015