D Í G I T O S
Revista de Comunicación Digital

Twitter research for social scientists: a brief introduction to the benefits, limitations and tools for analysing Twitter data

Investigación de Twitter para científicos sociales: una breve introducción de los beneficios, limitaciones y herramientas para el análisis de datos de Twitter

Javier Ruiz-Soler
javier.ruiz.soler@eui.eu
European University Institute
http://orcid.org/0000-0003-2203-1134

Recibido: 07/11/2016
Aceptado: 23/01/2017

ABSTRACT

The analysis of social media is currently very important due to the unprecedented quantity of information. Twitter is becoming an indispensable source of information for researchers aiming to implement big data in their projects. However, despite the potential field of research opened by that Twitter data, it contains some risks a researcher must be aware. In this paper I present on the one hand the benefits and caveats of research conducted on Twitter, and on the other hand the constraints of Twitter data collected from the Application Programming Interfaces (APIs). There are, therefore, three major methodological problems identified: (i) representation bias: it is very difficult to make general assumptions using research based on Twitter. (ii) language challenge: users can write in many different languages. It implies that when collecting data, some cautions need to be taken in order to accurately gather the data we need, (iii) data bias: Depending of the data needed, one API might be a better fit than other. The main aim in this paper is to discuss these methodological constraints from a theoretical point of view. I propose, as a starting point, possible solutions to overcome them, or at least reduce their impact in the research.

KEY WORDS

Twitter, big data, API, social sciences

RESUMEN

El análisis de las redes sociales es actualmente muy importante debido a la cantidad sin precedentes de información disponible. Twitter se está convirtiendo en una fuente indispensable para los investigadores que tienen la intención de implementar big data en sus proyectos. Sin embargo, a pesar del potencial de investigación abierto por los datos provenientes de Twitter, estos contienen una serie de riesgos que un investigador debe de reconocer. En este artículo presento por un lado los beneficios y las limitaciones de la investigación realizada en Twitter, y por otro lado las restricciones de los datos de Twitter recogidos de las interfaces de programación de aplicaciones públicas (API). Existen, por tanto, tres problemas metodológicos importantes identificados: (i) sesgo de representación: es muy difícil hacer suposiciones generales usando Twitter como base de investigación. (ii) cuestión idiomática: los usuarios pueden escribir en idiomas diferentes. Esto implica que al recolectar los datos, algunas precauciones deben ser tomadas con el fin de recoger con precisión los datos que necesitamos. (iii) sesgo en los datos: dependiendo de los datos necesarios, una API podría ser un mejor que otra. El objetivo de éste artículo es el de discutir estas limitaciones metodológicas desde un punto de vista teórico. Propongo además, como punto de partida, una serie de posibles soluciones para atajar las limitaciones, o en algunos casos limitar su impacto en la investigación.

PALABRAS CLAVE

Twitter, big data, API, Ciencias Sociales

1. INTRODUCTION

Twitter was launched in October 2006.[1] Falling into the category of social media, it has been conceptualized in very different ways. Twitter has been considered as a "microblogging tool", (Gayo-Avello, 2015) as a Short Message Service (Kwak et al., 2010), and as well as a "social network" (Eleta and Golbeck, 2014). Today Twitter, with 315 active millions of users per month and publishing more than 500 million tweets per day, is a platform focused on information and entertainment in real time (Smith, 2016).

Research conducted on Twitter is a growing field of study. Different academic disciplines have explored Twitter in order to illuminate the potential of the elements and networks one can find inside the platform. The research conducted on Twitter is vast, and has been applied to a variety of domains: the spread of diseases (Sadilek et al., 2012), political events (Burgess and Bruns, 2012), the ideology of users (Barbera, 2015), interaction in natural disasters (Sakaki et al., 2010), and even the prediction of elections (DiGrazia et al., 2013; Shi et al., 2012; Tumasjan et al., 2010), or stock markets (Bollen and Mao, 2011).

However, these studies have been facing severe limitations of data collection and analysis. On the one hand, Twitter is a perfect digital socioscope, providing unlimited amount of data. On the other hand, collecting and processing data involves difficulties which are not easy to deal with. This paper answers what Twitter API one must use in order to collect data, depending on the type data requested. In addition I describe what challenges as a social scientists we face collecting this type of data.

The purpose of this paper, therefore, is to expose and discuss the advantages, but more specifically problems of the Twitter platform as source for research, but also the gathering of Twitter data using the public APIs.[2] I describe these limitations and possible ways to overcome them. This paper is a theoretical approach of these questions, which I consider basic for a social scientists who dives in the world of big data, becoming a computational social scientist.

2. DEFINING TWITTER

Twitter as a social platform provides a unique and characteristic platform for social interaction. Its unique degree of transnational communication and the open interactivity among users make the platform an ideal public arena with, in principle, no restrictions (Ahmed, 2015a; Almuhimedi et al., 2013; Cantijoch, 2014; González-Bailón et al., 2014). Its asymmetric and open principle of "following" users without mandatory reciprocity, coupled to a very open APIs, make it an ideal medium for its study. These "dynamic" analyses, which typically map networks of tweets, content, or

[1] According to Alexa Web metrics, Twitter was the tenth web site most visited as of September 2016, after Google, Facebook, YouTube, Baidu, Yahoo, Amazon, Wikipedia and Qq.

[2] An API is a serial of instructions for functionalities and procedures in order to be used for an external program software. Commonly mentioned as *the back door of a web, platform or service*, it is where programmers can have access to a platform through the use of specific scripts, and to request information from the platform for diverse usage (Mitchell, 2015).

interaction, owe their popularity to the availability of the material and the possibility for researchers to analyse its contents.

In addition, Twitter counts with specific technical characteristics, forming its own technical slang. The first and more known characteristic is its brevity and simplicity of 140 characters allowed for posting a tweet. Furthermore, Twitter counts with the employment by the users of other elements such as direct replies to tweets (@replies), the address to other accounts (@mentions), and the diffusion of information (RT retweets). In addition, the interactions made by the users shape networks around issue publics under a hashtag (#hashtag[3]), forming conversations and communities. These hashtagged topics form themselves networks of topics, named issue publics (Bruns and Highfield, 2016).

Facebook and Twitter are the most popular places where these interactions take place. One of the key differences between Facebook and Twitter is that most of the content on Twitter is publicly accessible via the Twitter API or through resellers such as GNIP and Datasift, whereas most Facebook content is private. Thus, Twitter has emerged as the single most powerful big data source available to social scientists for collecting fine-grained time-stamped data of interaction for local, regional and global events, in addition to individual level.

Twitter is in summary a huge database available with numerous possibilities for research. The footprints left by its users while interacting with each other can be collected and analysed. As an example of its richness for data mining, the metadata in each tweet contains not only the text, but also forty five different variables such as number of followers, favourites, language, geographic location, etc.

3. GATHERING DATA FROM TWITTER: SCOPE AND LIMITATIONS

The analysis of social media is currently very important due to the unprecedented quantity of information. It is becoming an indispensable source of information for researchers aiming to understand events and movements of all kind (economic, political, cultural) and the behaviour of individuals (Cantijoch, 2014, p. 146). However, despite the potential field of research opened by Twitter as a data source, there are some methodological caveats important to take into consideration (Gayo-Avello, 2015). I describe these limitations and discuss how to handle them. There are three major methodological problems when doing research on Twitter: (i) representation bias, (ii) language challenge, and (iii) data bias.

3. 1. REPRESENTATION BIAS

Researchers have underlined the weak representativity of Twitter for inferences of general publics (Ahmed, 2015a; González-Bailón et al., 2014). In other words, it is very difficult to make general assumptions using research based on Twitter for different reasons.

[3] Twitter users make use of the hashtag symbol # before a relevant keyword or phrase (no spaces) in their Tweet to categorize those Tweets and help them show more easily in Twitter Search. Clicking on a hashtagged word in any message shows you all other Tweets marked with that keyword. Hashtags can occur anywhere in the Tweet – at the beginning, middle, or end. Hashtagged words that become very popular are often Trending Topics (Twitter 2017).

First, Twitter has small usage compared to other platforms (as Facebook) or news media (as TV, newspaper or radio). Twitter has around 315 millions of active users worldwide (Chaffey, 2016). Nevertheless, the number of accounts is very limited compared to other platforms, and even less if we look at real active accounts[4]. For instance, above 79% of the European population had Internet access in 2015 (Eurostat, 2016). Of those, around 80% had a social media account. Facebook takes the first place with 70% with a Facebook account. Twitter comes second with 31%. (Chaffey, 2016; Pew Research Center and Washington, 2014).

In addition, Twitter usage differs significantly across the world. In some countries it is widely used, while in others its use is marginal (Street et al., 2015). These differences of usage and within different countries, makes it very hard to take Twitter as a representative sample of general population. It is simply not big enough, and not used enough. Some research have tried to export inferences from offline samples to online samples (see for instance Dunbar et al., 2015). The problem with this type of research is that did not take into consideration that Twitter datasets are samples of a sample: a sample of population collected inside another sample (specific Twitter data).

Second, the "digital divide" raises additional concerns about generalizing any knowledge from online to offline populations. The Twitter population tends to be younger, better educated, and more affluent (Duggan, 2015). This situation raises important questions about the potential for reproducing and even amplifying social stratification from Twitter to general population (Golder and Macy, 2015, p. 12). Twitter has its own sociodemographic characteristics. It is not only different from the offline worlds, but also from other platforms online, such as Facebook or Instagram.

However, although it seems that Twitter might not be suitable for general inferences, there are still grounded arguments in favour of Twitter as a tool for analysis. The Twitter world is not identical to the offline world, but it is entirely real. Users who desire status, admiration, social approval, and attention in their offline relationships will bring those desires with them to Twitter. Individuals must navigate many of the same social obstacles online as they do offline when they seek information, political support, friendship, romance, or consumer goods (Golder and Macy, 2014; Mejova et al., 2015). Consequently, the argument that Twitter is not widespread enough becomes irrelevant if we assume that that is not possible to make general inferences to the entire population from Twitter.

In addition, despite its low level of users compared to other social media platforms, Twitter has attracted so much research attention due to its openness, interaction system and innate transnationality (Ahmed, 2015a; Almuhimedi et al., 2013; Cantijoch, 2014; González-Bailón et al., 2014). Indeed, its openness and structure compared to other social media and network platforms is making Twitter a growing field of study for different disciplines.

[4] Active account means to have published a tweet in the last month at least once, and that the account is not from a spamming algorithm or bot.

Therefore Twitter should be considered as one more online platform available to people, independently of whether it is used as a first option or not. In that sense, it is correct that an inference concluded from research on Twitter data cannot be generalized to the entire population. However, research on Twitter can lead to show inferences on how Twitter is used for sampled populations or topics. For example, rather than to explore how Europeans use Twitter to interact about certain European topics, we should look into how Twitter users make use of it for the interaction of European topics. As an example of the disagreement of using Twitter for general inferences, the research that tried to replicate or to predict events from Twitter data such as the result of national elections, has encountered fierce academic resistance (Gayo-Avello, 2012).

3. 2. LANGUAGE CHALLENGE

The second warning that needs to be discussed is the 'language challenge' of Twitter. Researchers have indicated that studies on online social networks, especially Twitter hashtags, have the need to focus exclusively in some specific languages (English, French or Spanish, to mention some), trying to frame those language communities and applying different analysis on each one of the languages (Ackland, 2013; Hermes, 2006). The reason is that some hashtags are used in different languages. As an example of this problem, we have issue publics on Twitter (hashtags) referring to the European Union. #UE hashtag is the same hashtag for French as an abbreviation for the European Union (Union européene) than for Spanish (Unión Europea). That is, interaction among users might take place in different languages, forming language bubbles. Therefore it is possible that gathering Twitter data implies collecting tweets in different languages, and it could skew the results.

On the other hand, however, filtering languages while gathering Twitter data from a specific event might misrepresent data samples as the full data of an event cannot be captured. In order to make sure that analysis is coherent and robust the full data should be extracted, if possible. Indeed, multi-language users are key nodes facilitating the transmission of information between different language communities (Cheng and Wicks, 2014). Moreover, some topics may simultaneously attract different hashtags. For instance, #EU is the hashtag used in English as an abbreviation of European Union, while in other languages the hashtag is different. For example in Spanish it is #UE (Unión Europea).

In other words, language issues are important, especially if the study is based on collecting tweets. To overcome this possible problem, researchers need to adapt their approaches, and to design a strategy to capture the data they are interested in. If the researcher needs to collect tweets from a specific event or topic, he or she must be aware that users might use different hashtags for the same event, and therefore the need to design a strategy to capture all these tweets in different hashtags. On the other hand, a hashtag used for a specific event might content different user typology and different languages. Therefore a strategy to separate the tweets is needed.

3.3. DATA BIAS: GATHERING VALID DATA

The third problem –data gathering– might be considered the most important limitation because it is a structural constraint. This problematic issue is divided in to two: (i) the

methods, skills and training needed to gather and process the data, and (ii) the process of retrieving data from Twitter.

3. 3.1. DATA RETRIEVAL: LEARNING DATA MINING

First, data gathering from online social networks like Twitter requires a great deal of computational knowledge and computer programming. This caveat, however, is not exclusive for Twitter. Social scientists are now becoming computer scientists, gaining knowledge about the different processes to gather and process data with programming languages or computing environments. The first wave of online studies was dominated by physical, computer, and information scientists. However, now social scientists have also started to access online data and even big data of different kind. Nevertheless, there is a basic distinction between those with a more technical and engineering background, and those with a social scientist and humanistic background. While the former know how to collect, process and manage data, they usually lack the theoretical background to know where to look, which questions to ask, what to hypothesize or what the results may imply (Golder and Macy, 2015, p. 5). In contrast, the latter might lack skills and training in programming languages and data structures for the process and manipulation of data. However they are able to formulate the right questions that can be answered with the analysis of data.

Therefore, the first to take into consideration for a social scientist is to be able to gather, process and manage data. Social scientists have already a strong theoretical background and they know where to look and what to ask. Although these technical skills of data process and handling can be learned and polished through training, the learning curve for those with no experience could be very steep. In the case of Twitter, a researcher with very little knowledge of data formats, data gathering, storage and retrieval might encounter difficulties working with projects of such nature. It is true that depending on the research project, data can be extracted manually, for example from specific accounts and specific dates. However, when considering whether to gather millions of tweets, the manual process is discarded and researchers must rely on powerful software to manage such a large quantity of data. An example of the impossibility of manual processing, some researchers have, for instance, processed 500 millions tweets for a specific project (Golder and Macy, 2011).

Analysing big data from Twitter can be challenging, too, since making sense of big data requires a lot of effort. Usually, data collected is raw and needs pre-processing to clean out the variables that are not useful for research purposes. The pre-process and first exploration of data generally define if data is suitable for the project or not. In the case of Twitter, spam tweets, bots for the following/followee networks, languages and special characters need to be filtered before any in-depth analysis.

Although programming languages such as Python and command line software are very commonly used in the environments of big data, there are commercial products that have been developed to pre-process and analyse Twitter data. They have a more friendly approach than command line software since they count with a graphic user interface[5]. However, the researcher must still know the structural technicalities behind

[5] NodeXL and Tableu are some examples of data collection and analysis.

the data. Otherwise, it could be the case that he or she overlooks some relevant aspects of the data nature, thereby harming the whole research process.

Although there are commercial solutions for data collection and analysis, their monetary cost is very high. However, there are plenty of free tools to gather and analyse data in different programming languages: for example StreamR in R language, and Tweepy[6] in Python, to mention some of them. An alternative might be the Digital Methods Initiative Twitter Capture and Analysis Toolset (DMI-TCAT) (Borra and Rieder, 2014) from Amsterdam University, which uses a set of tools to retrieve, collect and analyse tweets in various ways. DMI-TCAT provides robust and reproducible data capture and analysis, and it interlinks with existing analytical software.

3. 3.2. DATA RETRIEVAL: LEARNING WHAT ARE WE MINING

A second problematic question refers to the limitations associated with how Twitter provides the data when collecting it through the public APIs. This limitation is independent of the approach adopted by the researcher, or even the programming skills for data gathering. When Twitter provides the data requested, it contains several structural limitations that are impossible to overpass because of the restrictions Twitter impose when gathering data. However it is possible to restrict or minimize these limitations. The researcher, therefore, must be aware of these limitations or restrictions. The limitations associated with data retrieval refer to two issues (i) the number of tweets that can be extracted, and (ii) the impossibility of replicating the dataset. To explain these limitations we need to describe the methods to extract data from Twitter (Ahmed, 2015a; Almuhimedi et al., 2013; González-Bailón et al., 2014).

There are three main methods to gather data through the Twitter API(s): Firehose, REST and Stream[7]. Each of these has different procedures to extract specific data (Hansen et al., 2011). The first method, Firehose, allows full access to Twitter data without any limitation. The Firehose API provides 100% of Twitter data in real time. Despite the suitability of this method for research, Firehose is not generally used due to its high monetary costs. Only large companies or institutions with high monetary resources might make use of it. In addition, Firehose is not available directly. That is, it is not public per se. Only through third party companies, such as GNIP and Datasift, can researchers have access to the Firehose API (Layton, 2015; McKinney, 2013; Mitchell, 2015; Twitter, 2016).

The other two methods (REST and Stream) are public and easy to access. REST provides access to read and write Twitter data. The possibilities of the REST API variables are immense: reading author profiles and followers' data, extracting settings, languages, etc. It also allows, with the Search API, to extract tweets containing specific keywords (words, phrases or hashtags), geographical boundaries and user IDs.[8]

In comparison to Firehose, it contains, however, some limitations. The REST API has

---

[6] For a list of command line scripts approved by Twitter visit https://dev.twitter.com/overview/api/twitter-libraries [Consulted: 8th February 2017]

[7] Exists also the Decahose which provides 10% random sample of the real time Firehose through streaming connection.

[8] The Twitter Search API is part of Twitter's REST API.

rate limits: a researcher cannot take a full following list of users, unless it waits for the Twitter API to provide access every 15 minutes[9]. In addition, the search request can only go back in time one week, and it only provides a sample of up to 1% of the capacity of the Firehose. That is, the API will return at most 1 percent of all the tweets produced on Twitter at a given time. Once the number of tweets matching the given parameters eclipses 1 percent of all tweets on Twitter, Twitter begins to sample the data returned to the user.

Figure 1. The REST API and its process of Twitter data collection

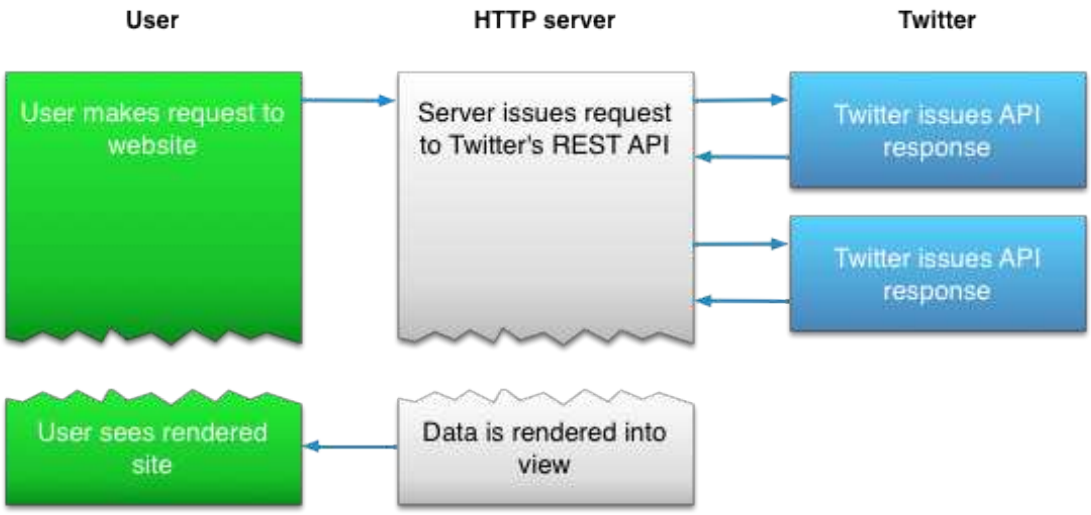

Image from Twitter Website

The third method to gather Twitter data (Stream) consists in leaving the API call open for a certain period of time, collecting data "on live". Stream API can be set up to stream tweets with specific keywords (words, phrases or hashtags), geographical boundaries and user IDs. Like the Search API, Stream provides data up to 1% of the capacity of the Firehose.

There is no limitation on the time that the call can be kept open. However, it requires more resources in programming and infrastructure than Search API. Since Stream API is a constant open call to the Twitter API, there is a need to prepare additional coding in the programming script in case of connection problems appear during the time the call is open. Usually, Stream API is open for several hours, days, weeks, or even months. During the time the call is open, there might be difficulties, and this is why additional coding is needed, in order to continue data gathering in the case of an Internet connection failure. In addition to auxiliary coding, Stream API requires extensive hard disk space as the data gathered might be large. On average, a million tweets require around one gigabyte of hard disk space (Lutz, 2013; McKinney, 2013; Mitchell, 2015; Twitter, 2016). A common solution is to have an external server, or a server provided by a university to store the collected data.

In sum, only one method provides the total amount of data (the 100% with the Firehose

[9] All requests to REST API are rate limited and all of them have different rate limits, depending of the data requested. A table of the rate limits is available online: https://dev.twitter.com/rest/public/rate-limits [Consulted: 8th February 2017]

API), while the other two methods collect up to1% of tweets, depending on the filters imposed by Twitter. This represents two problems. First, a limitation of the collection of tweets, which can misrepresent the data sample. It is, therefore, extremely difficult to extract valid conclusions of very large events (elections), or global events (Olympic games) of large datasets, when the data gathered is up to 1% of the total amount of data for that event.

Figure 2. The Stream API and its process of Twitter data collection

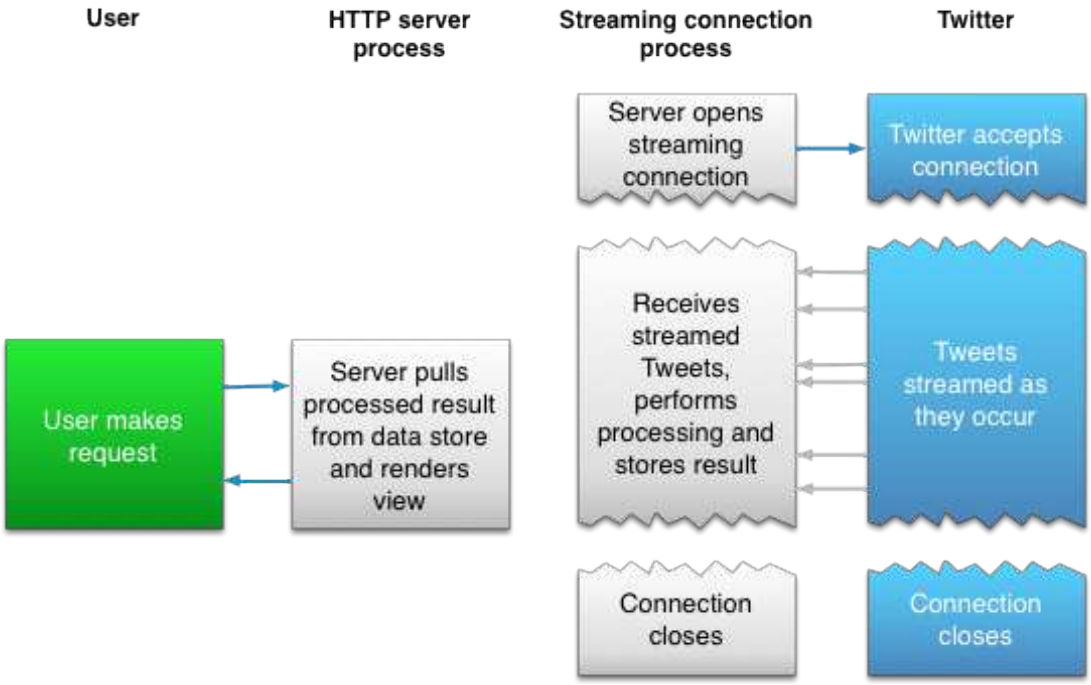

Image from Twitter website

Second, the functioning of both public APIs makes it impossible to replicate the same data gathering. Indeed, a second data gathering for the same time and keywords on Twitter might provide different data, if during that period the percentage of total tweets is higher than the 1% allowed. This is due that one time the data gathering hits the 1% barrier, Twitter starts limiting the data. Additionally, there is no information how Twitter delimitates or randomizes the data collection after this 1% point. Therefore, is technically impossible, when collecting data and hitting the 1%, we could extract the same tweets. This is a major limitation since researchers would be unaware of both the nature of the population they are aiming to analyse and the specific sampling methods used by the public APIs to satisfy researchers' need of samples.

However a clarification is important. It does not mean that analysis from the same data sample cannot be replicated. The problem is the gathering of the same data sample, which cannot be replicated if the data collection is hitting the 1%. Two data samples collecting tweets from the same hashtag, and hitting the 1% will contain different data.

3.3.3. FINAL RECOMMENDATIONS FOR GATHERING TWITTER DATA

It seems clear that two of the three methods for gathering Twitter data (Search and

Stream) might provide biased or different data samples, due to the imposed limitations of 1% of the capacity provided by Firehose. If researchers cannot afford using Firehose due to its high cost, only two methods are left: Search and Stream API. Which of the two APIs might be interest for data collection depends in what kind of data is needed.

In a recent blogpost, (Ahmed, 2015b) compared two out of the three methods through an experiment. Ahmed monitored during a period of three days in January 2015 the Search and Firehose APIs with the keyword "Ebola"[10]. The result was 195,713 tweets with the Firehose API (the 100% of tweets), while a total of 155,086 tweets were gathered with Search API (79%). This experiment shows us two important remarks. First, the keyword "Ebola" reached more than 1% of the total amount of tweets from the Firehose API in certain periods of time. This is due that the Search API collected less than the 100% of Tweets provided by the Firehose. Second, and most importantly, there is no mechanism to know if a keyword or hashtag will overpass the 1%, and when it can happen.

In addition, in Ahmed's experiment, the sampled data acquired 79% of possible tweets. We only know it because we can compare it with the number of tweets provided by Firehose API. However, this comparison is generally not possible. Most research using Stream or Search APIs to gather data will not be able compare the results with those provided by the Firehose API, so it is impossible to know how big the population of tweets really is. In other words, not having any previous information about the tweets' population size when using public APIs implies a structural uncertainty about the validity of the sample,

Other studies (González-Bailón et al., 2014; Morstatter et al., 2013, 2014) have addressed the issue of data sampling between Search, Stream and Firehose APIs. Their research has shown that while Search API is limited and very restricted due to the impossibility of having access to tweets more than a week back in time, the Stream API does not place such a limitation on data gathering as long as the API call is open and does not exceed the 1% barrier. This means that the Stream API is able to fully collect 100% of data if the request for data does not overtake 1% Firehose's capacity.

In addition, research has shown that the more time the API call is open for data gathering, the less biased the sample is, acquiring a higher percentage of accuracy in comparison with the one provided by Firehose (Ahmed, 2015a; Gayo-Avello, 2015; Huberman et al., 2008). This is because up to a certain point, even if the Stream API is not gathering 100% of tweets, the possibility of biasing the sample is very low: once the Stream API has been running for a certain period of time, it is very unlikely that the data will change enormously. Once the collection of data has been run for a while it is less probable that its basic nature changes. Nevertheless, it is not clear how long it needs to be the Stream API call open to reduce the possibility of bias to a level which is acceptable. As a result, it is decisive for researchers to have enough information in advance regarding the structure of the event, keyword, or hashtag to collect in order to decide when and how long the Stream API call must be open.

[10] We refer to keywords for words without containing the symbol "#". We could ask Twitter to provide all tweets containing the keyword "Ebola", or the hashtag #ebola. With "Ebola" we will get tweets containing "Ebola" and #Ebola. With #Ebola we will get tweets containing #Ebola, but not "Ebola".

Table 1. Comparison of main Twitter APIs systems

API	Cost	Access	Limitations	Rate Limits	Type of Data	Infrastructure
Firehose	High (depends of the quantity of data)	Private	None	None	All	Low
REST	Free	Public	1% and 7 days for Search API	Yes	Tweets and specific variables	Medium
Stream	Free	Public	1%	-	Tweets	High

In summary, the two public methods for gathering data, research has shown that Search API is a good option for first exploring small datasets and short events, and to retrieve specific variables (such as profiles descriptions, followers, favourites, etc.). In making use of the REST API, research points out that rate limits must be taken into consideration. For example, following/followee networks with hundreds or thousands of accounts can take up to several weeks of data collection.

For a higher accuracy of samples for medium or large datasets, Stream API is the best option. If the research objective is to monitor tweets on specific real-time events, even just for more than 7 days, using Stream API might be the best option. Still, two major issues shall be noted. First, researchers need to know in advance what data they need to collect, and when. Second, the keyword(s) or hashtags to stream must be appropriately set up. That is, in order for the Stream API to provide accurate and valid data, it is worthless to collect data from very general keywords or hashtags such as #elections, #usa, etc. The amount of data will be enormous, not even related to the interest of the research project. In addition, it could reach the 1% barrier at any moment, and therefore starting to delimitate our data samples.

4. CONCLUSIONS

In this paper, I have described and critically reflect on two related issues: the benefits and constraints of Twitter as a platform for research, and a discussion in how to gather and process data with a comparison of the public Twitter APIs.

Twitter as a platform is an excellent resource used every day for research purposes for social scientists of different disciplines such as political sciences, communications, phycology, etc. However, as we have seen, it also includes some issues that need to be acknowledged by the approaching this platform for research. On the one hand it has the challenges of representativity of Twitter. On the other hand we have the idiomatic questions and the challenges it creates when collecting tweets from events.

In addition, I discussed the questions of the Twitter APIs. Although Twitter provides an excellent variety of APIs to gather data, researchers need to be aware of their specific limitations and how to address them. In that sense, the application of a methodology such as network analysis, content analysis, or sentiment analysis to Twitter data can only be of interest if applied to proper data. It is and intrinsic circle where data and methods need to match.

Responding to the question I arise for this paper, I propose to use each of the two public APIs –REST and Stream- for distinct purpose. REST API for the query of specific variables such as profile descriptions, profile pictures, followers, lists, etc. In addition, the Search API –which is part of the REST API- is useful for small datasets of tweets containing not more than 7 days in the past. On the other hand, the Stream API is useful for collecting big datasets of events during periods of time. Of course if one would have access to the Firehose, the question of which one, if REST or Stream API, is superfluous. However, access to the Firehose rarely occurs, unless with one counts with strong economic funding.

Nevertheless, what data and how much appropriate? It all depends of the kind of project and research question. What it is clear is that one must be aware of various elements of the Twitter data that can affect the research before starting to gather data. Not only the technicalities of the APIs but also how to manage and process this data. It is almost impossible to not to face any data limitation when gathering tweets or any other Twitter variable because of the different problematics here presented. These limitations are intrinsic: they come with the job of data mining. Nevertheless, the important is to try to limit these constraints to the minimum expression. And for that, a good knowledge and understanding of Twitter data mining is necessary.

More research is needed comparing empirical datasets, especially discerning how much two data samples differ when hitting the 1% limit. In addition, complementary research can be conducted if Twitter changes the rules of the APIs, as it happened in the past (Twitter, 2012). A discussion about what has been changed, and what implications these changes might have for computational social sciences researchers might be fundamental for the future of the research based on Twitter.

Finally, in this paper I did not enter into the ethical discussion of sharing Twitter datasets. Such question is out of context in this paper. However, future research should address different ways to anonymize Twitter datasets, as it is a requirement by the Twitter terms: datasets are forbidden to be shared due to privacy issues. This issue has implications for the publishing of research using Twitter data, as it is forbidden to share datasets without anonymizing it for replication purposes.

REFERENCES

Ackland, R. (2013): *Web social science: concepts, data and tools for social scientists in the digital age*. London: SAGE.

Ahmed, W. (2015a): "Challenges of using Twitter as a data source: an overview of current resources". Available at https://wasimahmed.org/2015/09/20/challenges-of-using-twitter-as-a-data-source-an-overview-of-current-resources/ [Accessed 15 February 2017]

Ahmed, W. (2015b): "A comparison of Twitter APIs across tools". Available at https://wasimahmed.org/2015/06/04/a-comparison-of-twitter-apis-across-tools/ [Accessed: 15 February 2017]

Almuhimedi, H., Wilson, S., Liu, B., Sadeh, N., Acquisti, A. (2013): "Tweets are forever: a large-scale quantitative analysis of deleted tweets". In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*. ACM, pp. 897–908.

Barbera, P. (2015): "Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data". In *Polit. Anal*, vol. 23, pp. 76–91. doi:10.1093/pan/mpu011

Bollen, J., Mao, H. (2011): "Twitter Mood as a Stock Market Predictor". In *IEEE Computer Society,* vol. 44, pp. 90–93.

Borra, E., Rieder, B. (2014): "Programmed method: Developing a toolset for capturing and analyzing tweets". In *Aslib Journal of Information Management*, vol. 66, pp. 262–278. doi: 10.1108/ajim-09-2013-0094

Bruns, A., Highfield, T. (2016): "Is Habermas on Twitter?". In Bruns, A., Enli, G., Skogerbø, E., Larsson, A.O., Christensen, C. (Eds.): *The Routledge Companion to Social Media and Politics*. London: Routledge, pp. 56–73.

Burgess, J., Bruns, A. (2012):"(Not) The Twitter Election: The dynamics of the #ausvotes conversation in relation to the Australian media ecology". In *Journalism Practice*, vol. 6, pp. 384–402. doi:10.1080/17512786.2012.663610

Cantijoch, M., (2014): Analysing social media data and web networks. New York: Palgrave Macmillan.

Chaffey, D., (2016): "Global Social Media Statistics Summary 2016". Available at http://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/ [Accessed: 17 February 2017]

Cheng, T., Wicks, T., (2014): "Event Detection using Twitter: A Spatio-Temporal Approach". In *PLoS ONE,* vol. 9(6), e97807. doi:10.1371/journal.pone.0097807

DiGrazia, J., McKelvey, K., Bollen, J., Rojas, F., (2013): "More Tweets, More Votes: Social Media as a Quantitative Indicator of Political Behaviour". In *PLoS ONE,* vol. 8, e79449. doi:10.1371/journal.pone.0079449

Duggan, M., (2015): "The Demographics of Social Media Users." In *Pew Research*. Available at: http://www.pewinternet.org/2015/08/19/the-demographics-of-social-media-users/ [Accessed: 17 February 2017]

Dunbar, R.I.M., Arnaboldi, V., Conti, M., Passarella, A., (2015): The structure of online social networks mirrors those in the offline world. In *Social Networks*, vol. 43, pp. 39–47. doi:10.1016/j.socnet.2015.04.005

Eleta, I., Golbeck, J., (2014): "Multilingual use of Twitter: Social networks at the language frontier." In *Computers in Human Behaviour*, vol. 41, pp. 424–432. doi:10.1016/j.chb.2014.05.005

Eurostat, E.C., (2016): "Digital economy and society". Available at http://ec.europa.eu/eurostat/statistics-explained/index.php/Digital_economy_and_society_statistics_-_households_and_individuals [Accessed: 17 February 2017]

Gayo-Avello, D., (2015): "What do we mean when we talk about Twitter political opinion?" In *The Plot*. Available at: http://www.the-plot.org/2015/06/26/what-do-we-mean-when-we-talk-about-twitter-political-opinion/ [Accessed: 17 February

Gayo-Avello, D., (2012): "No, you cannot predict elections with Twitter". In *IEEE Internet Computer Society*, vol 16, pp. 91–94.

Golder, S.A., Macy, M.W., (2015): "Introduction". In Mejova, Y., Weber, I., Macy, M.W. (Eds.), *Twitter: A Digital Socioscope*. Cambridge: Cambridge University Press, pp. 1–20.

Golder, S.A., Macy, M.W., (2014): "Digital Footprints: Opportunities and Challenges for Online Social Research" Available at http://www.annualreviews.org/doi/10.1146/annurev-soc-071913-043145 [Accessed: 2 February 2017).

Golder, S.A., Macy, M.W., (2011): "Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures". In *Science Magazine*, vol. 333, 1878–1881. doi:10.1126/science.1202775

González-Bailón, S., Wang, N., Rivero, A., Borge-Holthoefer, J., Moreno, Y., (2014): "Assessing the bias in samples of large online networks." In Social Networks, vol. 38, pp. 16–27. doi:10.1016/j.socnet.2014.01.004

Hansen, D.L., Schneiderman, B., Smith, M.A., (2011): *Analysing social media networks with NodeXL: insights from a connected world*. London: Elsevier.

Hermes, J., (2006): "Citizenship in the Age of the Internet". In *European Journal of Communication*, vol. 21, pp. 295–309. Doi: 10.1177/0267323106066634

Huberman, B., Romero, D.M., Wu, F., (2008): "Social networks that matter: Twitter under the microscope". In *First Monday,* vol 14(1). Available at: http://firstmonday.org/article/view/2317/2063 [Accessed: 17 February 2017]

Kwak, H., Lee, C., Park, H., Moon, S., (2010): "What is Twitter, a social network or a news media?" In *Proceedings of the 19th International Conference on World Wide Web*. ACM, pp. 591–600.

Layton, R., (2015*): Learning data mining with Python: harness the power of Python to analyze data and create insightful predictive models*. Birmingham: Packt Publishing Ltd.

Lutz, M., (2013): *Learning Python*. Sebastopol, CA: O'Reilly.

McKinney, W., (2013): *Python for data analysis*. Sebastopol, CA: O'Reilly.

Mejova, Y., Macy, M.W., Weber, I., (2015): *Twitter: a digital socioscope*. New York, NY: Cambridge University Press.

Mitchell, R., (2015): *Web scraping with Python: collecting data from the modern web*. Sebastopol, CA: O'Reilly.

Morstatter, F., Pfeffer, J., Liu, H., (2014): "When is it biased? Assessing the representativeness of twitter's streaming API". In *Proceedings of the 23rd International Conference on World Wide Web*. ACM, pp. 555–556.

Morstatter, F., Pfeffer, J., Liu, H., Carley, K.M., (2013): "Is the sample good enough? Comparing data from twitter's streaming api with twitter's firehose". In ArXiv Soial and Information Networks. Available at: https://arxiv.org/abs/1306.5204 [Accessed: 17 February 2017]

Pew Research Center, (2016): "Social Networking Fact Sheet". Available at: http://www.pewinternet.org/fact-sheet/social-media/ [Accessed 17 February 2017]

Sadilek, A., Kautz, H.A., Silenzio, V., (2012): "Predicting Disease Transmission from Geo-Tagged Micro-Blog Data", in Conference on Artificial Intelligence. Available at: http://www.aaai.org/ocs/index.php/AAAI/AAAI12/paper/view/4844 [Accessed 17 February 2017]

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 World Wide Web*. ACM, pp. 851–860.

Shi, L., Agarwal, N., Agrawal, A., Garg, R., Spoelstra, J., (2012): "Predicting US primary elections with Twitter." Available at http://snap.stanford.edu/social2012/papers/shi.pdf [Accessed 17 February 2017

Smith, K., (2016): "44 Astonishing Twitter Stats and Facts for 2016, In *Brandwatch*. Available at: https://www.brandwatch.com/2016/05/44-twitter-stats-2016/ [Accessed 17 February 2017

Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M., (2010): "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment." In *Fourth International Conference on Weblogs and Social Media*, pp. 178–185. Available at: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1441 [Accessed 17 February 2017]

Twitter, (2016): "Documentation Twitter". In *Twitter Developers*. Available at: https://dev.twitter.com/overview/documentation [Accessed 17 February 2017]

Twitter, (2012): "Changes coming in Version 1.1 of the Twitter API", In *Twitter Blogs*. Avaliable at: https://blog.twitter.com/2012/changes-coming-in-version-11-of-the-twitter-api [Accessed 8 February 2017].