MOX-Report No. 16/2018

An elephant in the room: Twitter samplingmethodology.

Calissano, A.; Vantini, S.; Arnaboldi, M.

MOX, Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox-dmat@polimi.it http://mox.polimi.it

An elephant in the room: Twitter sampling methodology.

Anna Calissano[a], Simone Vantini[b], Michela Arnaboldi[c]

February 22, 2018

[a] MOX, Dipartimento di Matematica, Politecnico di Milano
Piazza Leonardo da Vinci 32, I-20133 Milano, Italy
E-mail: anna.calissano@polimi.it
[b] MOX, Dipartimento di Matematica, Politecnico di Milano
Piazza Leonardo da Vinci 32, I-20133 Milano, Italy
E-mail: simone.vantini@polimi.it
[c] Dipartimento di Ingegneria Gestionale, Politecnico di Milano
Via Lambruschini, 4/B, I-20156 Milano, Italy
E-mail: michela.arnaboldi@polimi.it

Abstract

The usage of social media data is spreading among the broad scientific community: 30000 papers dealing with this type of data are indexed in Scopus in the last decade. On the one hand, this data are very appealing, creating a rich bucket of information. On the other one, gathering them through a repeatable sampling strategy is increasing in complexity (or maybe becoming impossible?). The aim of this paper is to map the scientific community awareness about the sampling strategies used to download on-line data, focusing on the most studied social media: Twitter. This review unveils two unexpected results: the downloaded data are typically far from being randomly sampled, and around 99% of papers does not explicitly declare the sampling strategy used to download the data. These two facts pose some worrisome doubts about the trustworthiness of all the results presented in this stream of literature.

1 Introduction

Social media use has exploded in the last decade with more then 3000 million active user per month considering just Facebook, Twitter, and Youtube (Socialbakers). The explosion has generated a plethora of data, seized by business companies, consultancy industry, and academic researchers. Within this framework, an exhaustive and improved statistical analysis of digital data requires combination of different skills, such as a well designed sampling strategy, an accurate and conscious model design, and a critic interpretation of the results. Due to obvious time and cost constraints, the research team very rarely presents deep knowledge in all the three competences listed above, that are needed to face the multi-disciplinary complexity of digital data analysis. Indeed, on one extreme, there are researchers (mostly with background in computer science) fully aware about the procedures to extract data from on-line or any other digital sources, but less aware about methods for statistical analysis. At the other edge, there are researchers (mostly with background in statistic) whose deep knowledge in statistic analysis does not come with a sufficient knowledge about digital data sampling tools. This knowledge gap might produce results that can be neither repeatable nor reliable, because derived from datasets whose sampling strategy is unknown, or obtained via arguable usage of statistics tools. The amount of studies possibly exposed to this risk is not negligible. In fact, more than 30.000 papers in one decade (2008-2018) have focused on social media data, crossing different disciplines (Scopus). Through this paper, we address this priority problem of social media sampling, focusing on Twitter, that is currently the most studied. More specifically, the literature review has two aims. The first is reviewing the state of the art of awareness and consideration of sampling problems for Twitter data, while the second is identifying and summarizing solutions proposed by the few studies which face this problem.

To unfold our argument the paper is organized as follows. Section 2 provides an overview on data downloading methods for Twitter platform. In Section 3, we detailed the procedure to perform the literature review and we illustrate some general results. Section 4, we briefly group the papers not investigating this problem, while in Section 5 we detailed the solutions proposed by researchers to face this problem. Finally, we draw some conclusions.

2 Twitter

In the digital era, an high number of statistical studies substitute or integrate the classical survey with the usage of web data, from web survey to general web traces. Among all the possibilities, social media data are often selected as a useful and interesting source in different field. In fact, social networks offer both social and content information, creating a rich bucket of information from different perspectives. Among the different social media, Twitter is one of the most studied platform for two main reasons. Firt of all, it is both a social media and a blog platform at the same time, creating a new paradigm of communication (Janses et. al (2009); Java et al. (2007)). Twitter users can express their idea within 140 characters, sharing their posts with their Twitter friends. Note that from September 2017 posts can exceed the 140 characters. Each user can chose friends to follow and select users to be followed by. These users' connections create on-line relationships and give chance to start and spread discussions. This peculiarity makes Twitter platform interesting for both content analysis (Barnaghi et al. (2016); Dass et al. (2016); Milioris et al. (2015); Gazar et al. (2016), etc.) and social network analysis (Lu et al. (2013); Rahimi et al. (2015), etc.). The other main reason why researchers select Twitter among the other platform is its open access philosophy. By open access, we mean two things. Firstly, all the profile on Twitter are public by default (Twitter Inc.), unlike Facebook. This philosophy is perfectly coherent with the definition of micro-blog, namely a platform used to maximize posts visibility. The second and more interesting aspect is that via Twitter API everyone can download Twitter contents under some restrictions (Twitter Developers). The open-access data is an additive value that makes Twitter one of the most used platform among scholars.

2.1 Twitter Data Access

"The Twitter Platform connects your website or application with the worldwide conversation happening on Twitter." (Twitter Developers). As anyone can read on the Twitter Developer website, Twitter allows developers to access the twitter platform via APIs. The API is a pre-defined method to interface among different softwares. For what Twitter concerns, programmers can access Twitter in an automatic way, connecting the personal computer to the Twitter database. The action the programmer can do are basically two: writing and reading. By writing, the programmer can publish posts using the API connection. By reading, the programmer can "read" the posts who is interested in. The reading function is the one used for collecting the data.

The accessibility to Twitter database depends on how Twitter store its massive data. Twitter data storage is organized in two main steps: the *temporary* and the *permanent* storage. In a temporary repository, every new post is held for one week. After one week, all seven-days-old posts are moved from the temporary repository and move to another permanent database. Saying that, data can be collected in two ways: via *Rest API*, or via *Streaming API* . For an explanation of the Rest API we refer to Valkanas, et al. (2014) and to few indications found in Twitter Developers. " The REST APIs provide programmatic access to read and write Twitter data. Create a new Tweet, read user profile and follower data, and more". It allows access to the data warehouses, both the temporary and the permanent repository. REST APIs use HTTP requests (i.e. GET, POST) to perform the communication between the end user and the Twitter service. These APIs support multiple query types. There are different restrictions imposed to the queries that can be clustered into four typologies, according to Valkanas, et al. (2014):

- Rate restrictions, i.e. the number of queries of a specific type that the developer can issue within the 15 minute window;
- Maximum Result Size, i.e. the upper bound on the results of a particular query. For instance, even if a user has posted 5000 tweets, we are only able to access the most recent 3200.
- Probing Result Size, i.e. the number of results that we can retrieve each time we probe the service with that particular query. For instance, a query for a users time-line will return at most 200 tweets.
- Maximum Query Size, i.e. the number of objects that we can query simultaneously with a single probe to the service. Typically this is 1, (e.g., 1 tweet each time, using its id), but there are some exceptions (e.g., look up at most 100 users).

On the other hand, the Streaming APIs collect the real time Twitter flow, from the temporary repository. Through this API, one can receive data as a flow of tweets. The API returns a percentage of all public posts, though not uniformly. Consequently, data received through this API may reflect fluctuations of the actual stream, e.g., increase /decrease of posts, temporal patterns of user interactions, etc.

In addition to the different data warehouse, i.e the temporary and the permanent repository, and different type of APIs, i.e. the Rest and the Streaming API, there is a third aspect based on the business model of Twitter. Some APIs are for *free* and others are for *payment*. The payment services are offered by Gnip (Gnip). In Table 1, all the possible APIs connections are listed according to these categories.

	Rest API	**Streaming API**
Free Solutions	Search API	Spritzer API
	Sample API	Gardenhose API
Payment Solutions	Search 30 Days API	Firehose API
	Full Archive Search API	Decahose API
	Historical Power Track	

Table 1: Different type of APIs offered by Twitter, divided by costs and type of sample.

In this paragraph, we go through all the option listed in Table 1, as reported in Twitter Developers. The Search API “allows queries against the indices of recent or popular Tweets and behaves similarly to, but not exactly like the Search feature available in Twitter mobile or web clients, such as Twitter.com search. The Twitter Search API searches against a sampling of recent Tweets published in the past 7 days.” Search allows to filter and download part of the Tweets in the temporary on-week data warehouse. As you can read on the website “before getting involved, it is important to know that the Search API is focused on relevance and not completeness”. The Search 30 days API is a payment version of Rest API, with a longer time frame. It ”provides a rolling 30 days of access to historical Twitter data”. The type of queries are about the volume of the data requested or attributes of the data itself. The Historical Power Track and the Full Archive API give complete and comprehensive access to every publicly available Tweet from the first Twitter ever in March 2006. The difference between these two APIs are several including the type of queries supported and the default time limits of the Full Archive. There is the possibility of gathering a random sample thanks to the Sample API. It “returns a small random sample of all public statuses”. All the others APIs are Streaming, free and payment ones. The Gardenhose returns a free randomly sampled 10% of the whole unfiltered Twitter stream while the Spritzer the 1 % of the whole. The Decahose delivers a 10% random sample of the real-time Twitter Stream. This is accomplished via a real-time sampling algorithm which randomly selects the data, while still allowing for the expected low-latency delivery of data as it is sent through the Firehose by Twitter. It is the payment version of Gardenhose. Finally, Twitter Firehose “delivers 100% of Tweets in real-time through a streaming connection. Full Firehose streams provide 100% of the publishers real-time Firehose to your app, with no additional limitations” (Gnip). Among all these different options, Firehose appears to be the best solution. However, the prize of this API mismatches with most researchers' butget.

Moreover, Twitter data can be gathered stratifying for content, users or all the other meta-data of interest. Thus, queries can focus on users Sofean et al. (2012) Lu et al. (2013), keywords (e.g. Pichl et al. (2014) Nagar et al. (2014)) or hashtags , geo-localisation (e.g. Hwang et al. (2013) Oussalah et al. (2013)) or entities related to the Tweet content, such as urls (Cao et. al (2014)) etc.

3 Literature Review

In the recent years, the number of papers using Twitter data has been soaring. There are around 14000 papers on Scopus (Scopus) discussing about Twitter in different field, from Mathematics to Computer Science, form Agriculture to Medicine, from Social Science to Natural Science. Parallel to this high number of works, there is a niche literature that is trying to understand the Twitter sampling strategy and measure the Twitter samples representativeness, moving forward from the seminal paper by Morstatter et al. (2013). The problem of on-line data sampling and more in general Big Data quality is well known in the statistical community (Dovrandi et al. (2017), De Veaux et al. (2016), Elliott et al. (2017)). The bias of the Twitter sample is an open and murky issue which the scientific community is silently aware of. It is actually an elephant in the room. The aim of this paper is to disclose the problem of sample quality and offer a clearer overview on what do we really know about Twitter data.

This overview is based on a scoping literature review created via keywords and snowball approach on Scopus, (Scopus). The Scopus search engine looks for strings into Title, Keywords and Abstract. 14581 paper were found by filtering with Twitter keyword, and 6895 with Twitter Data. To pick the papers interesting for our review, i.e. the one downloading twitter data and performing analysis, we tailor four searching strings: "Rest Twitter API", "Streaming Twitter API", "Sample Twitter API", and "Sample Twitter Bias". The first two help selecting papers explicating the downloading techniques. The third and the fourth focus on sampling strategy in general and sample bias. With these strings we gathered 65 papers. Moving from the first papers collected, we selected the relevant papers cited by or citing the papers, following the so called snowball approach. The results is a collection of 109 papers. The first and shocking results is that around the 1.5% among the papers using Twitter data declare their sampling methodology into Title, Keywords and Abstract.

3.1 The Purpose of Twitter Data Analysis

Before introducing the results of this review, a first and essential distinction needs to be done. A first result is the evidence of two categories of papers: *Twitter4Real* and *Twitter4Twitter*. In the Twitter4Real group, we collect all the papers that use Twitter to describe any off-line world problems. Largely diffused examples are election predictions using Twitter data (McGregor et al. (2017), Rezapour et al. (2017)), studies concerning virus or illness diffusion (Majak et al. (2017), Hwang et al. (2013), Towers et al. (2015)), sentiment and opinion analysis on specific events or topics (Barnaghi et al. (2016), Dass et al. (2016)), studies of socio-economic phenomena such as traffic incident detection (Gu et al., 2016) or money exchange rate prediction (Janetzko, 2014), and many other issues concerning off-line world. On the other hand, papers in the Twitter4Twitter group focus on studying usages and patterns of Twitter platform. The majority of authors of these papers come from a Computer Science background. These works focus on hashtags usage (Doong, 2016), spams detection (Chen et al., 2015), sharing function among users (Ahn et. al, 2015), and many other problems related with the Twitter world. Note that papers describing on-line phenomenon integrating Twitter data with other sources are listed in Twitter4Real category (Pichl et al. (2014), Arakawa et. al (2010), Saveski et al. (2016)). The aim of this overview is to study the awareness of the scientific community concerning Twitter sample quality. To reach this goal, we focus on Twitter4Twitter papers. In fact, understanding how Twitter data are gathered is the key aspect in studying the bias of the analysis. Measure the bias between Twitter on-line world and off-line one is a step further (Lamy et al. (2016), Almeida et. al (2015) Carley et. al (2016)), involving social and anthropological issues. in fact, the sub-sample distortion is influencing the quality of the results also in the Twitter4Real papers.

3.2 Review Results

As shown in Figure 1, the 109 selected papers are all dated in the last 6 years, showing an extremely recent attention on the problem. We collect the papers until December 2016, that might cause an underestimation of papers from 2016. The same figure shows the number of papers in the two groups, demostrating a continous interesting in both purpose.

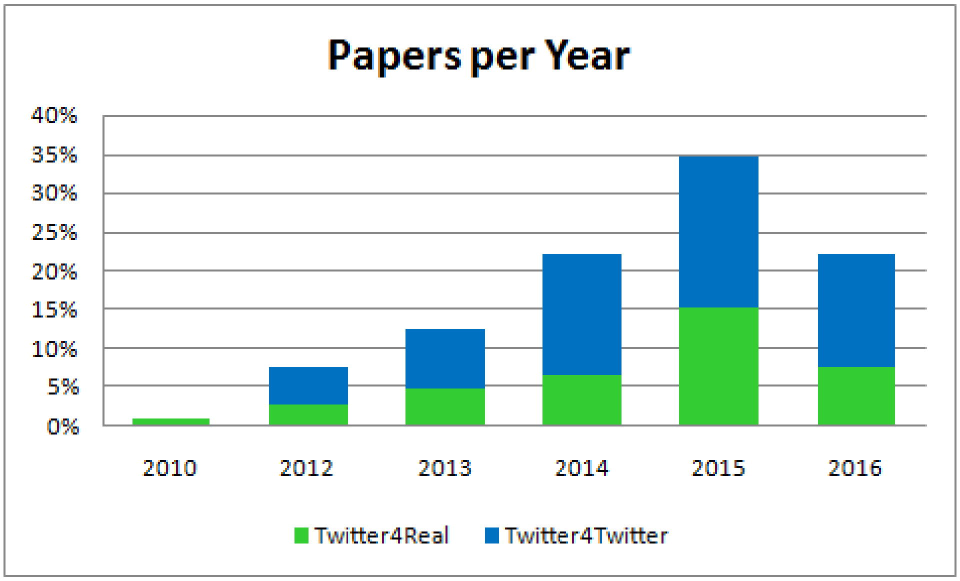

Figure 1: The barplot shows the number of papers collected per year of publication. Bars are colored according to the Twitter usage label. Twitter4Twitter category contains all the papers studying different aspects of Twitter media, while Twitter4Real contains papers using Twitter to study off-line world issues.

In Figure 2, we can see the distribution of the first scientific area associated by Scopus to every paper. Twitter is used as a data source in many different fields for many different reasons. In the category "Others", all the papers belonging to Sociological or Biological Studies are listed, such as Eisenstein et al. (2014) or Eom et al. (2015), or Human Science such as Al-garadi et. al (2016). As expected, the majority of papers describing Twitter lays in the field of Computer Science.

In Figure 3, papers are divided into three groups. Papers dealing with the problem of data quality and investigating the issue are 17% and tagged in the *Awareness and Action* category. Works conducting the analysis on

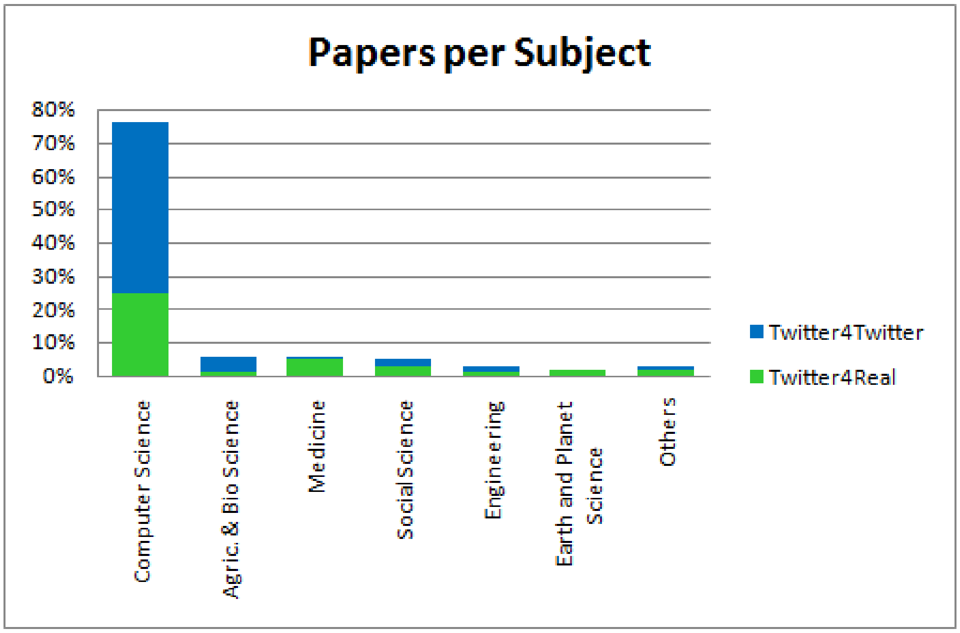

Figure 2: The barplot shows the number of papers in every field category associated by Scopus. The Others category contains all the subject with less then 1% papers such as Art and Humanities, Psychology and Business and Accounting. Bars are colored according to the Twitter usage label. Twitter4Twitter category contains all the papers studying different aspects of Twitter media, while Twitter4Real contains papers using Twitter to study off-line world issues.

the data sampled via API, without showing any concern on sample bias are in the *No Awareness No Action* category which counts the 77% of papers. The remain 6% of papers are tagged as *Awareness No Action* and enlighten the problem of Twitter APIs bias, but they do not propose any solution. The second pie graph in Figure 3 displays the number of papers clustered in Twitter4Real and in Twitter4Twitter.

The aim of the analysis is to generally understand how literature face the problem of Twitter data sampling. We focus the attention on the Twitter4Twitter level of analysis. Under these perspective, only 13 papers of 109 seems to care about the sampling strategy and consequent sample quality: the ones in the Awareness and Action bin. These papers will be detailed described in the next Section 6.

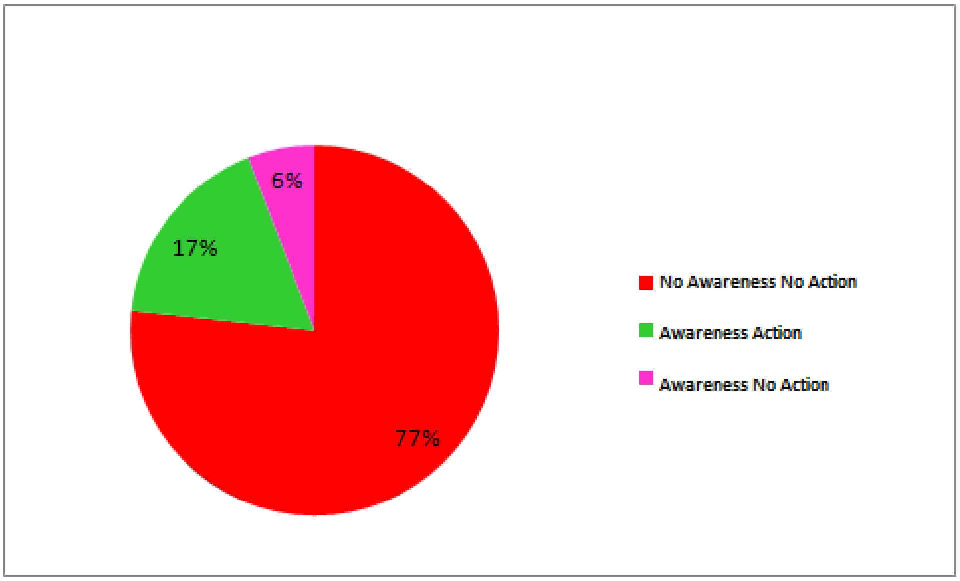

Figure 3: Pie graphs represent the percentage of papers considering the sample bias (Awareness and Action) or not (No Awareness No Action). (Awareness No Action) are the papers citing a work dealing with Twitter Sample Bias but not proposing any solution.

4 "No Awareness" and "Awareness but No Action"

The papers in *No Awareness No Action* and *Awareness but No Action*, are declaring the type of APIs queries they use but they do not tackle the problem of data quality downloaded via API. These papers face different problems in different disciplines, so the cluster of the papers is quite hard. We list below the papers accordingly to the purpose of the analysis and eventually the method. Note that the papers in the two groups are listed together, because they are not introducing solution to the problem tackled in this review.

Works in Twitter4Twitter group focus on studying three main aspects: hashtags, network analysis or content analysis. Doong (2016) and Alvanaki et. al (2013) study hashtags or keywords distributions, Ordua Malea et al. (2015) analyses hyperlinks usage on Twitter, while Harvey et al. (2015) is even proposing a personalized hashtags recommended system. Ahn et. al (2015), Avrachenkov et. al (2014), Lu et al. (2013) and Rahimi et al. (2015)

focus more on the network analysis of Twitter. Content analysis is largely diffuse and spreads from topic detection (Milioris et al. (2015) Bunrside et al. (2014) Chepurna, et al. (2015) Gazar et al. (2016)) to Sentiment Analysis (e.g. Barnaghi et al. (2016), Dass et al. (2016), Ntzel et al. (2012), Kang et al. (2012), Cavalin et al. (2015)). These algorithms are applied to track product reputation Das et al. (2014), to evaluate elections Fink et al. (2013) or to conduce socio-political analysis Magdy et al. (2015) or to implement a question-answering database Cavalin et al. (2016). As one can see in Figure 2, a high number of Twitter4Twitter papers belong to the Computer Science environment. Many of these papers propose easy access architectures for gathering data from APIs (e.g. Oussalah et al. (2013); Marcus et al. (2012); Roegiest et al. (2016); Bechini et al. (2016)), well-posed queries syntax (such as Palma et al. (2015); Togias et al. (2012)) or new platforms performing social network analysis Davis et al. (2016). Other computer science papers focus on spams and advertise analysis (e.g. Chen et al. (2015); Hazra et al. (2015); Cao et. al (2014); Zhang et al. (2014)) and brand reputation or trend detection (such as Benhardus et al. (2013); Arvanitidis et. al (2014)).

Twitter4Real papers instead focus on urban landscape analysis, health field problems and social analysis. Papers concerning urban analysis goes from on-line city diversity measurement (Förster et al. (2015)), to on-line location naming (Chan et al. (2014)). Gu et al. (2016) and Wanichayapong, et al. (2014) focus on traffic problem. A high number of works use posts on Twitter as useful database to predict and understand health problem such as disease spreading or prediction and drug diffusion (Yin, et al. (2015); Nagar et al. (2014); Hwang et al. (2013); Sofean et al. (2012); Lamy et al. (2016); Towers et al. (2015); Dos Reis et al. (2015); Missier et al. (2016)). Parallel to the health applications, there is a large number of papers studying social behaviours in general (e.g. Webb et al. (2016); An et. al (2015); Barber et al. (2015); González-Bailón et al. (2014)). Some of these Social Science papers focus on emergency and mass event reaction or prediction (Purohit et al. (2013); Olteanu et al. (2015); Boecking et al. (2015); Compton, et al. (2013); Buntain et al. (2016)) or social event identification in general (Kenett at al. (2014) Paltoglou (2015)), other focus on linguistic variation and lexical changes (Eisenstein et al. (2014); Eisenstein et al. (2015)). Hogan (2013) argues about the technical and practical issues making Twitter both mirroring of social behaves and hiding many inequalities. Under the social science cap, Al-garadi et. al (2016) studies the papers studying cyber-bullism on Twitter. Ei Elmongui et al. (2015) tries to infer users' home address from their geo-localized Twitter usage. As underlined in Section 2, Twitter is a micro-blog platform, thus is an important platform for sharing news. Malik et al. (2016) studies activities of news organization and the general Twitter usage as a news sharing platform. Morgan et al. (2013) proposes an analysis of ideological bias of news on Twitter, while Uddin et al. (2012) tries to distinguish facts from rumours. As a news platform, politic is one of the hottest topic, with many papers studying election polls, parties' supports, political polarization and other aspect of political expression on Twitter (Ceron et al. (2016); Coletto et al. (2015); Fink et al. (2013); Hanna et al. (2013); Eom et al. (2015); Borge-Holthoefer et al. (2015); Hagar et al. (2016); Wells et al. (2016)). In the Twitter4Real category, there are also some papers combining different on-line data sources, such as Hargittai et al. (2012), that studies how Twitter usage influence other on-line activities. Some of these papers propose a comparison among Twitter and other social media Saveski et al. (2016), or conduce analysis of on-line data including Twitter platform dataset Arakawa et. al (2010) or Osborne et al. (2014). While Pichl et al. (2014) is combining Spotify and Twitter data implementing a Music Recommended system. There are two papers difficult to be clustered: Janetzko (2014), studying the EUR/DOL exchange or Giordano et al. (2015), implementing a World of Things vision, where internet, physic and human resources interacts. There are also papers proposing data download technique (Oussalah et al. (2013)) or Big Data sampling approach Lu et al. (2013), but not concerning the level bias we are focusing on.

5 Awareness and Action: Solution Proposed

Among the papers belonging to both *Awareness and Action* and *Twitter4Twitter* category, i.e. the ones investing the our target problem, we identify two different approaches: *Comparing* and *Sampling.* Papers in the *Comparing* group tries to describe the sampling strategy via comparing results obtained with the same analytic tools but different datasets, downloaded with the same queries but via different APIs. In the *Sampling* group, the approach focuses on proposing different sampling strategies either to fool the APIs limitation or to understand the APIs sampling strategy. In the following two section we will go through all these papers. Note that all the information reported follows the authors' declaration.

5.1 Comparing

The comparison among different results depends on several aspects, such as APIs, queries, and analysis. In fact, the type of the analysis conduced on data change the features the authors focus on. In addition different types of APIs and queries lead to completely different sub-samples.

Morstatter et al. (2013) is the seminal paper of this line of research and it is one of the few paper matching a Streaming API with the expensive complete Firehose dataset, conducing a significative comparing study. The idea is to download Firehose dataset belonging to one specific week and meanwhile downloading via Streaming API three dataset filtered by keywords, by user, and by geo-localization. To conduce a fair comparison, the Firehose is filtered with the same queries and randomly sub-sampled to get the same dimension of the Streaming datasets. The obtained sub-sample is called Random Set. Firstly, they focus on the temporal distribution of data, inferring that on average but not homogeneously, the 43% of the whole data stream is sampled per day, but some peaks are structurally over or under represented. Secondly, they compare the hashtags rank via Kendal's tau showing that only focusing on the most cited, the rankings are coherent in the two datasets. They also conduced content analysis. Topic matching between Firehose and its Random subsample is better than the match between Firehose and Streaming API sample. For what concern the network property of the dataset, the UserXUser network was built. In the streaming, only the 50-60% of users are well represented also. If you focus on the geo-located tweets, Morstatter et al. (2013) affirm that circa the 90% of tweets are downloaded via Streaming API.

Mostratter performs a second focusing on both comparing and sampling (Morstatter et al. (2014)). The first objective of this second paper is to obtain the same results of the previous paper using not the expensive Firehose dataset but the free Sample API (Note that the Sample API should give the users back a random 1% of the whole Twitter stream with no possibility of adding filters). In fact, if the Sample API is a random selection of the Firehose, it can be freely used as a benchmark. The comparison is made on hashtags through Kendall's tau and they prove that the Sample API is coherently representing the hashtags rank of the Firehose. Using the Sample API as benchmark, they have a look at the hastags time distribution of the Streaming API, underlining some over and under estimations. Accordingly Morstatter et al. (2014), time evolution analysis of hashtags conduced on Streaming API is bias. The second part of the paper focus on launching different queries either from different location or in different time. They demonstrate that Streaming APIs queries are independent from the location they are launched, giving back the same datasets. The second example tests if same queries of Streaming API gives back the same dataset, in a certain time windows. They launched same query starting one after the other and analyse the dataset in the time overlap via Jaccard index: dataset from different APIs result to be the same.

Similarly to the techinique just described, the paper by Kenneth is focusing on the design of sampling methodology, proposing a distributional approach with a statistical flavour Kenneth et al. (2014). They compare several Streaming APIs launched using the same queries (i.e. keywords filter) at the same time: 14 samples are obtained. The idea is to suppose a theoretical sampling model and to compare it with the empirical measurements. If the samples are supposed to be randomly gathered from the stream, the presence or absence of tweet in one of the samples should be modelled as a Bernoulli variable, with the parameter given by the ratio of the sample and the whole data dimensions. They discovered the overlap between samples is about 96%. Thus, the samples are surely not randomly sampled, but barely deterministic. Consequently, no extra data can be gathered using multiple identical queries launched at the same time.

Another interesting study about Streaming API is Wang et al. (2015), that proposes a comparison between a ”complete” dataset and two free access Streaming APIs. To gather the ”complete” dataset, all the post by users geo-localized in Singapore are downloaded using Rest API. The same users are listed in the queries launched via Spritzer and Gardenhose APIs. These two free access options are supposed to sample the 1% and the 10% from the Stream of data. They prove that the effective sampling ratio between the Rest API and the two free stream access options is smaller than declared (circa 0.9% and 9% respectively). However, if we focus on frequency, the sampled datasets maintain the same scaling behaviour of the users tweeting frequency distribution, but overestimate the low-frequency proportions. The authors carry out analysis on the Tweets contents, declaring that text and URLs are satisfactorily captured by the sampled data. Finally, they focus on users' network. The streaming samples covers the most active users while there is a less coverage of the low-frequency users.

In Valkanas, et al. (2014), a comparison between Streaming API (1%) and Gardenhose API (10%) is conduced. Note that the Spritzer is the old version of the Streaming API: the policy are the same. The first comparison is performed on geo-located Tweets. Every Tweet can be geo-located by the users, thus the data can be downloaded using geo-location. The percentage of geo-tagged tweets are the same in the two dataset. The second level of the analysis is based on the Tweets' content. Via Lexicon Based Algorithm, they find out that the sentiment trend along time is the same for the two datasets. Retweets, a.k.a. RT, are another interesting parameter to study both the network and the popularity. The top 100 re-posted tweets are the same in the two sample. Finally they have a look at the UserXUser network based on the retweet net. The Largest Connected Component size does not share the same pattern in the two samples increasing the time interval.

For what Twitter networks concerns, González-Bailón et al. (2014) propose a detailed analysis. In this contest, there are three datasets: Search API dataset downloaded from the UK with a 6 keywords query, Streaming API datasets downloaded from Spain with 70 keywords query and a subsample of the Streaming API dataset, containing the keywords used in the Search API. There are many different ways to built networks from Twitter data. In this work, the network considered are built on three different features: RT , @ mentions (including RT), @ mentions (excluding RT). The users' coverage in all the possible combination of networks and datasets is computed using Jaccard index . The 6 keywords dataset are more similar even if the APIs are different. Note that the bigger the time windows considered to built the network, the better the users' coverage. In conclusion, whatever the network is, the streaming APIs are better for capturing the peaks, without ensuring the peripheral activity coverage.

Xu et al. (2015) offer a comparison between the Sample API and the Streaming API. Sample API is a random 1% of the whole dataset. Both dataset are filtered with two keywords: Flu and Ebola. As expected, the impact of sampling on tweeting volumes affects the most on smaller dataset linked with less popular topics. In fact, in the case of ̈flu; the observed data dimension deviates from the projected random volume. In addition, the deviation is larger for small time granularity. Considering a small time interval, the distortion might be bigger than considering a longer interval. An interesting comparison is drawn between the twitter distribution across users. Results show under estimation of the users who tweet multiple posts. This interesting discovery shows that spammers detection may be more difficult considering this under estimation. In this paper, a comparison between the RT graph is also conduced. The main result is that Sample API dataset does not include all the edges of re-tweeting relationships, due to random sample design. Finally, they propose a sampling community technique to obtain a full graph representation. Comparing with random and stratified sampling, community sampling achieve the closest tweet distribution across users.

5.2 Sampling

This paragraph illustrates papers belonging to the *Sampling* group, focusing on proposing different sampling strategies either to overcome the APIs limitation or to understand the APIs sampling strategy. The majority of the papers in this group belongs to Computer Science field. These works propose easy access sampling methodologies or try to infer the Twitter's selection techniques.
An interesting results on selection technique is illustrated in Kergl et al. (2014). Considering several downloaded datasets, the main idea is to study how the twitter IDs are generated, in order to understand in which order the data are sampled by Twitter. Note that every tweet is associated to a unique ID. After a detailed analysis of the IDs composition, the authors discover that part of the ID string is the time stamp of the Tweet's creation. Studying the time stamp, an analysis of the time distribution of posts can be done. The Spritzer, declared to be a random 1% of the whole stream, is actually a sample of Tweets always belonging to a specific interval of 10 millisecond. The same happens for the Gardenhose. The declared random 10% is a set of Tweets posted in a 100 millisecond. They conclude that the sample is not uniformly sampled in time, as Twitter declare.

Zafar, et al. (2015) propose an innovative sampling strategy for content analysis: a sampling approach based on following Experts. Twitter provide a list of Experts in different context, from Music to Neurology and many others. To measure the quality of the samples a comparison with Spritzer and the Expert dataset in the same time window is proposed. As expected, expert posts Tweets with higher quality content and more polarized Tweets. Due to the popularity of the experts on the network, their posts are more retweeted. Expert dataset contain a lower number of spams (i.e. there is the 12 % less spam URLs in the Expert dataset and there are no spam users). The comparison of this two dataset offers an interesting point of view on ideas and news diffusion on Twitter network. Most of the time, posts concerning important events and news firstly appear in the Experts' dataset as a RT from a post of a common users (rarely sampled in the 1% Spritzer). After an Expert's relevant post, an high number of RT posts appear in the Spritzer sample.

In White et al. (2012), they propose a sampling architecture, with a different interpretation of the Tweet selection. They cite the following: “The status id module 100 is taken on each public status, that is from the Firehose. Modulus 0 is delivered to Spritzer, and values 0-10 are delivered to Gardenhose”. The count starts when connection is activated. A part from different interpretation of sampling strategy by White et al. (2012) and Kergl et al. (2014), the selection seems not to be random. The authors propose a user-friendly sampling architecture. The idea is to develop a network architecture using 30 different and individual IPs in a round robin proxy fashion. The aim is to collect a number of Tweets close to the 100 % of Tweets. The comparison is done thanks to a counting of the number of duplicated posts and the Twitter declaration of number of Tweets.

Sampson, et al. (2015) propose another sampling strategy to “Surpassing the Limit”. Firstly they try to understand if the declared number of missing data for a query with maximum number of keywords (i.e. 500 keywords) is credible. To demonstrate the fact that the number of missing data is not a fixed threshold, they launch a query from a single crawler and the same query split in different crawlers. Excluding the duplicate Tweets, the number of Tweets collected is far more that the expected limit. Saying that, the authors try to built a system to collected the maximum number of Tweets, selecting the most diffuse keywords. The architecture is divided in two step. Firstly they create a cluster of the words appearing most of the time together. Secondly, they distribute the clusters to all the crawler in Round-Robin fashion, Spectral Cluster or Round Robin K means. All these techniques increase the number of data collected via APIs.

Gisselbrecht et al. (2016) proposes a contextual bandwidth solution for sampling the posts mostly related to a topic of interest by following the most active users on this topic. The sampling approach proposed faces a problem of Streaming API: at each query, only 5000 users can be followed. The idea is to select a topic of interest and use a contextual bandit algorithm to select the right users at each query (i.e. the ones who talk the most about the chosen topic). At each step, the algorithm assign to the users a score, computed via text classification of its posts (e.g. LDA and SVD algorithm). The list of users discussing the most about the topic is update step after step and fed by a parallel Sample API download. This paper is an extension of Gisselbrecht et al. (2015).

6 Conclusions

This review aimed at resuming the scientific community awareness on social-media sampling strategies, focusing on Twitter platform. On Scopus (Scopus) were found 17150 papers performing an analysis of Twitter Data, until December 2016. Focusing on papers declaring the sampling strategy, we get a list of 109 papers of which only 13 are aware about issues related to data downloading strategies and either try to investigate the problem or propose possible solutions. We conclude that Twitter sampling strategies appear to be variegated and hard to navigate for a statistically educated audience. At a first sight, it seems that there are different types of free APIs that allow for queries. However, a deeper investigation shows that none of them seems to perform a random sampling or following certain fixed design strategies. There indeed are only two APIs giving back random samples, with some severe draw-backs. The Sample API is free and it offers a random 1% but no filter is allowed, so it turns out to be useless for any focused analysis. The Firehose instead allows to download the whole Twitter stream related with specific queries. However, its costs are not affordable for the majority of the scientific community. To overcome this problem, some articulate sampling strategy techniques are proposed, but they are far from intuitive for an audience without a strong computer science background (Sampson, et al. (2015); Gisselbrecht et al. (2016)), and for this reason are not definitely mainstream.

Nevertheless, the big finding of this review lies in the remaining thousands of papers. Indeed, circa the 99% of papers dealing with social media data does not even declare the sampling strategy used to download data analyzed in the paper. This evidence causes this entire stream of literature to be not repeatable and it also unveils doubts about its trustworthiness. These facts undermine the very foundations of scientific method, in this fast growing, complex, and cross disciplinary research field. This is definitely an *elephant in the room*, that the statistic community has the duty to expose.

References

Ahn, Hyerim, and Ji-Hong Park (2015) *The structural effects of sharing function on Twitter networks: Focusing on the retweet function* Journal of Information Science

Al-garadi,Mohammed Ali, Kasturi Dewi Varathan, and Sri Devi Ravana (2016) *Cybercrime detection in online communications: The ex-*

perimental case of cyberbullying detection in the Twitter network. Computers in Human Behavior 63: 433-443.

Almeida, Jussara M., and Gisele L. Pappa. (2015) *Twitter Population Sample Bias and its impact on predictive outcomes: a case study on elections.* Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

Alvanaki, Foteini, and Sebastian Michel. (2013) *Scalable, continuous tracking of tag co-occurrences between short sets using (almost) disjoint tag partitions.* Proceedings of the ACM SIGMOD Workshop on Databases and Social Networks. ACM, 2013.

An, Jisun, and Ingmar Weber (2015) *Whom should we sense in social sensing-analyzing which users work best for social media now-casting.* EPJ Data Science 4.1: 1.

Apoorva, G., et al. (2016) *An approach to sentiment analysis in Twitter using expert tweets and retweeting hierarchy* Microelectronics, Computing and Communications (MicroCom), International Conference on. IEEE

Arakawa, Yutaka, Shigeaki Tagashira, and Akira Fukuda (2010) *Relationship analysis between user's contexts and real input words through Twitter.* 2010 IEEE Globecom Workshops. IEEE

Arvanitidis, Alexandros, et al. (2014) *Branty: A Social Media Ranking Tool for Brands.* Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer Berlin Heidelberg, 2014

Avrachenkov, Konstantin, et al. (2014) *Quick detection of high-degree entities in large directed networks.* IEEE International Conference on Data Mining. IEEE

Barnaghi, Peiman, Parsa Ghaffari, and John G. Breslin (2016) *Opinion Mining and Sentiment Polarity on Twitter and Correlation Between Events and Sentiment* IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService)

Benhardus, James, and Jugal Kalita (2013) *Streaming trend detection in twitter* International Journal of Web Based Communities 9.1: 122-139.

Bechini, Alessio, et al. (2016) *Towards a General Architecture for Social Media Data Capture from a Multi-Domain Perspective.* 2016 IEEE 30th

International Conference on Advanced Information Networking and Applications (AINA). IEEE

Barber, Pablo, and Gonzalo Rivero. (2014) *Understanding the political representativeness of Twitter users.* Social Science Computer Review: 0894439314558836.

Barber, Pablo, et al. (2015) *The critical periphery in the growth of social protests.* PloS one 10.11 (2015): e0143611.

Boecking, Benedikt, Margeret Hall, and Jeff Schneider (2015) *Event prediction with learning algorithmsA study of events surrounding the egyptian revolution of 2011 on the basis of micro blog data.* Policy and Internet 7.2: 159-184.

Borge-Holthoefer, Javier, et al. (2015) *Content and network dynamics behind Egyptian political polarization on Twitter.* Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work and Social Computing. ACM

Burnside, Grard, Dimitris Milioris, and Philippe Jacquet. (2014) *One Day in Twitter: Topic Detection Via Joint Complexity.* SNOW-DC@ WWW

Buntain, Cody, et al. (2016) *Evaluating Public Response to the Boston Marathon Bombing and Other Acts of Terrorism through Twitter.* Tenth International AAAI Conference on Web and Social Media.

Cao, Cheng, and James Caverlee (2014) *Behavioral detection of spam URL sharing: Posting patterns versus click patterns* Advances in Social Networks Analysis and Mining (ASONAM), IEEE/ACM International Conference on. IEEE

Carley, Kathleen M., et al. (2016) *Crowd sourcing disaster management: The complex nature of Twitter usage in Padang Indonesia.* Safety Science

Cavalin et al. (2016) *A scalable architecture for real-time analysis of microblogging data.* IBM Journal of Research and Development 59.2/3 (2015): 16-1.

Cavalin, Paulo, et al. (2016) *Building a Question-Answering Corpus Using Social Media and News Articles.* International Conference on Computational Processing of the Portuguese Language. Springer International Publishing

CERON, ANDREA, LUIGI CURINI, AND STEFANO M. IACUS (2016) *First and Second Level Agenda-Setting in the Twitter-Sphere. an Application to the Italian Political Debate.* Journal of Information Technology and Politics: 159-174

CHAN, C. K., M. VASARDANI, AND S. WINTER. (2014). *Leveraging Twitter to detect event names associated with a place* Journal of Spatial Science 59.1: 137-155.

CHEN, CHAO, ET AL. (2015). *A performance evaluation of machine learning-based streaming spam tweets detection.*, Transactions on Computational Social Systems 2.3 : 65-76

CHEPURNA, IULIIA, AND MASOUD MAKREHCHI (2015). *Exploiting Class Bias for Discovery of Topical Experts in Social Media* IEEE International Conference on Data Mining Workshop (ICDMW). IEEE

COLETTO, M., ET AL. (2015). *Electoral Predictions with Twitter: a Machine-Learning approach* CEUR Workshop Proceeding 2015

COMPTON, RYAN, ET AL. (2013). *Detecting future social unrest in unprocessed twitter data:emerging phenomena and big data.* Intelligence and Security Informatics (ISI), 2013 IEEE International Conference On. IEEE

DAS, T. K., D. P. ACHARJYA, AND M. R. PATRA. (2014). *Opinion mining about a product by analyzing public tweets in Twitter.* Computer Communication and Informatics (ICCCI), 2014 International Conference on. IEEE

DASS, P., LU, Y., CHOWDHURY, MD., LAMPL, D., KAMALANATHAN, J., NYGARD, K.E. (2016). *Gender differences in perceptions of genetically modified foods* Proceedings of the 31st International Conference on Computers and Their Applications, CATA 2016, Pages 183-188

DAVIS, CLAYTON A., ET AL. (2016) *OSoMe: The IUNI observatory on social media* PeerJ Computer Science 2: e87

DE VEAUX, RICHARD D., ROGER W. HOERL, AND RONALD D. SNEE (2016) *Big data and the missing links.* Statistical Analysis and Data Mining: The ASA Data Science Journal 9.6: 411-416.

Doong, Shing H. (2016). *Predicting Twitter Hashtags Popularity Level*, 49th Hawaii International Conference on System Sciences

Dos Reis, Virgile Landeiro, and Aron Culotta (2015). *Using matched samples to estimate the effects of exercise on mental health from Twitter* Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

Drovandi, C. C., Holmes, C., McGree, J., Mengersen, K., Richardson, S. and Ryan, E. (2017). Principles of experimental design for Big Data analysis Statistical Science (In Press)

Eisenstein, Jacob, et al. (2014). *Diffusion of lexical change in social media.* PloS one 9.11: e113114.

Eisenstein, Jacob (2015) *Systematic patterning in phonologicallymotivated orthographic variation.* Journal of Sociolinguistics 19.2: 161-188.

Elliott, Michael R., and Richard Valliant (2017). *Inference for nonprobability samples* Statistical Science 32.2: 249-264.

Elmongui, Hicham G., Hader Morsy, and Riham Mansour (2015). *Inference models for Twitter user's home location prediction* Computer Systems and Applications (AICCSA), 2015 IEEE/ACS 12th International Conference of. IEEE

Eom, Young-Ho, et al. (2015). *Twitter-based analysis of the dynamics of collective attention to political parties.* PloS one 10.7: e0131184.

Fink et al. (2013). *Twitter, public opinion, and the 2011 Nigerian presidential election.* Social Computing (SocialCom), 2013 International Conference on. IEEE

Frster, Thorsten, and Agnes Mainka (2015). *Metropolises in the Twittersphere: An Informetric Investigation of Informational Flows and Networks.* ISPRS International Journal of Geo-Information 4.4: 1894-1912.

Gazaz, Hatim, et al. (2016) *Geo-fingerprinting social media content.* Proceedings of the Third International ACM SIGMOD Workshop on Managing and Mining Enriched Geo-Spatial Data. ACM

Giordano, Andrea, et al. (2015) *Twitter to integrate human and Smart Objects by a Web of Things architecture.* Computer Supported Cooperative Work in Design (CSCWD), IEEE 19th International Conference on. IEEE

Gisselbrecht, T., Gallinari, P. and Lamprier, S.b. (2016) *Dynamic data capture from social media streams: A contextual bandit approach.* Proceedings of the 10th International Conference on Web and Social Media, ICWSM 2016 2016, Pages 131-140

Gisselbrecht, Thibault, et al. (2015) *Whichstreams: A dynamic approach for focused data capture from large social media.* Ninth International AAAI Conference on Web and Social Media. 2015.

Gonzlez-Bailn, Sandra, Ning Wang, and Javier Borge-Holthoefer (2014) *The emergence of roles in large-scale networks of communication.* EPJ Data Science 3.1: 1.

Gnip *http://support.gnip.com/*

Gu, Yiming, Zhen Sean Qian, and Feng Chen (2016). *From Twitter to detector: Real-time traffic incident detection using social media data* Transportation Research Part C: Emerging Technologies 67: 321-342.

Hagar, Douglas (2015)*#vote4me: the impact of Twitter on municipal campaign success.* Proceedings of the 2015 International Conference on Social Media and Society. ACM

Hanna, Alexander, et al. (2013) *Partisan alignments and political polarization online: A computational approach to understanding the French and US presidential elections.* Proceedings of the 2nd Workshop on Politics, Elections and Data. ACM

Hargittai, Eszter, and Eden Litt. (2012) *Becoming a tweep: How prior online experiences influence Twitter use.* Information, Communication and Society 15.5: 680-702.

Harvey, Morgan, and Fabio Crestani (2015). *Long time, no tweets! Time-aware personalised hashtag suggestion.* European Conference on Information Retrieval. Springer International Publishing

Hazra, Tapan Kumar, et al. (2015). *Mitigating the adversities of social media through real time tweet extraction system.* International Conference and Workshop on Computing and Communication (IEMCON). IEEE

Hogan, Bernie. (2013). *Comment on Elena Pavan/1. Considering Platforms as Actors* Sociologica 7.3 (2013): 0-0.

Hwang, Myung-Hwa, et al. (2013). *Spatiotemporal transformation of social media geostreams: a case study of twitter for flu risk analysis.* Proceedings of the 4th ACM SIGSPATIAL International Workshop on GeoStreaming. ACM, 2013.

Janetzko, Dietmar. (2014). *Predictive modeling in turbulent timesWhat Twitter reveals about the EUR/USD exchange rate.* NETNOMICS: Economic Research and Electronic Networking 15.2 (2014): 69-106.

Janses et. al (2009). *Micro-blogging as online word of mouth branding*, CHI'09 Extended Abstracts on Human Factors in Computing Systems. ACM.

Java, Akshay, et al. (2007) *Why we twitter: understanding microblogging usage and communities.* Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis. ACM, 2007

Kang, Byungkyu, John O'Donovan, and Tobias Hllerer. (2012) *Modeling topic specific credibility on twitter* Proceedings of the 2012 ACM international conference on Intelligent User Interfaces. ACM, 2012.

Joseph, Kenneth, Peter M. Landwehr, and Kathleen M. Carley (2014) *Two 1% s dont make a whole: Comparing simultaneous samples from Twitters streaming API.* International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction. Springer International Publishing

Kenett, Dror Y., et al. (2014) *Discovering social events through online attention.* PloS one 9.7: e102001.

Kergl, Dennis, Robert Roedler, and Sebastian Seeber (2014) *On the endogenesis of Twitter's Spritzer and Gardenhose sample streams.* Advances in Social Networks Analysis and Mining (ASONAM) IEEE/ACM International Conference on.

Lamy, Francois R., et al. (2016) *Those edibles hit hard: Exploration of Twitter data on cannabis edibles in the US.* Drug and alcohol dependence 164: 64-70.

Lu, Jianguo, and Dingding Li. (2013) *Bias correction in a small sample from big data.* IEEE Transactions on Knowledge and Data Engineering 25.11 (2013): 2658-2663.

Magdy, Walid, Kareem Darwish, and Ingmar Weber (2015) *#FailedRevolutions: Using Twitter to study the antecedents of ISIS support.* arXiv preprint arXiv:1503.02401

Majak, Marcin, et al. (2017). *Tweet Classification Framework for Detecting Events Related to Health Problems.*, International Conference on Computer Recognition Systems, Springer

Malik, Momin M., and Jrgen Pfeffer (2016). *A MACROSCOPIC ANALYSIS OF NEWS CONTENT IN TWITTER.* Digital Journalism: 1-25.

Marcus, Adam, et al. (2012). *Processing and visualizing the data in tweets.* ACM SIGMOD Record 40.4: 21-27.

McGregor, Shannon C., Rachel R. Mouro, and Logan Molyneux. (2017). *Twitter as a tool for and object of political and electoral activity: Considering electoral context and variance among actors.*, Journal of Information Technology and Politics: 1-14

Milioris, Dimitris, and Philippe Jacquet. (2015). *Topic detection and compressed classification in Twitter* Signal Processing Conference (EUSIPCO), 23rd European. IEEE

Missier, Paolo, et al. (2016). *Tracking Dengue Epidemics using Twitter Content Classification and Topic Modelling.* International Conference on Web Engineering. Springer International Publishing

Morgan, Jonathan Scott, Cliff Lampe, and Muhammad Zubair Shafiq (2013). *Is news sharing on Twitter ideologically biased?* Proceedings of the 2013 conference on Computer supported cooperative work. ACM

Morstatter, Fred, et al. (2013) *Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose* (2013) Seventh International AAAI Conference on Weblogs and Social Media. 2013.

MORSTATTER, FRED, JRGEN PFEFFER, AND HUAN LIU (2014) *When is it biased?: assessing the representativeness of twitter's streaming API.* Proceedings of the 23rd International Conference on World Wide Web. ACM

NAGAR, RUCHIT, ET AL. (2013) *A case study of the New York City 2012-2013 influenza season with daily geocoded Twitter data from temporal and spatiotemporal perspectives.* Journal of medical Internet research 16.10

NTZEL, JRGEN, AND FRANK ZIMMERMANN (2012) *Real-time Language Independent Sentiment Analysis in Social Network.* VIRTUAL GOODS + ODRL 2012

ORDUA MALEA, ENRIQUE, DANIEL TORRES SALINAS, AND EMILIO DELGADO LPEZ CZAR (2015) *Hyperlinks embedded in twitter as a proxy for total external inlinks to international university websites.* Journal of the Association for Information Science and Technology 66.7: 1447-1462.

OSBORNE, MILES, AND MARK DREDZE(2014) *Facebook, Twitter and Google Plus for breaking news: Is there a winner?* ICWSM. 2014.

Olteanu, Alexandra, Sarah Vieweg, and Carlos Castillo (2015) *What to expect when the unexpected happens: Social media communications across crises.* Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work and Social Computing. ACM

OUSSALAH, M., ET AL. (2013) *A software architecture for Twitter collection, search and geolocation services* Knowledge-Based Systems 37 (2013): 105-120.

PALMA, FRANCIS, ET AL. (2015) *Are restful apis well-designed? detection of their linguistic (anti) patterns.* International Conference on Service-Oriented Computing. Springer Berlin Heidelberg

PALTOGLOU, GEORGIOS (2015) *Sentimen-based event detection in Twitter.* Journal of the Association for Information Science and Technology.

PICHL, MARTIN, EVA ZANGERLE, AND GNTHER SPECHT (2014) *Combining Spotify and Twitter Data for Generating a Recent and Public Dataset for Music Recommendation.* Grundlagen von Datenbanken

PUROHIT, HEMANT, ET AL. (2013) *Emergency-relief coordination on social media: Automatically matching resource requests and offers.* First Monday 19.1

RAHIMI, AFSHIN, TREVOR COHN, AND TIMOTHY BALDWIN (2015) *Twitter user geolocation using a unified text and network prediction model.* arXiv preprint arXiv:1506.08259

REZAPOUR, REZVANEH, ET AL. (2017). *Identifying the Overlap between Election Result and Candidates Ranking Based on Hashtag-Enhanced, Lexicon-Based Sentiment Analysis.*, IEEE 11th International Conference about Semantic Computing.

ROEGIEST, ADAM, ET AL. (2016). *A Platform for Streaming Push Notifications to Mobile Assessors.* Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. ACM

SAMPSON ET AL. (2015). *Surpassing the limit: Keyword clustering to improve twitter sample coverage.* Proceedings of the 26th ACM Conference on Hypertext and Social Media

SAVESKI, MARTIN, SOPHIE CHOU, AND DEB ROY. (2016). *"Tracking the Yak: An Empirical Study of Yik Yak* Tenth International AAAI Conference on Web and Social Media

ELSEVIER SCOPUS *https://www.elsevier.com/solutions/scopus*

PENG SHI, LIANHONG DING, TIANLE ZHANG (2016). *Looking for Information Source from Online Social Network with Incomplete Observation* 9th International Symposium on Computational Intelligence and Design

SOCIAL BAKERS *https://www.socialbakers.com*

SOFEAN, MUSTAFA, AND MATTHEW SMITH. (2012). *"A real-time architecture for detection of diseases using social networks: design, implementation and evaluation.* Proceedings of the 23rd ACM conference on Hypertext and social media. ACM, 2012.

TOGIAS, KONSTANTINOS, AND ACHILLES KAMEAS. (2012). *An ontology-based representation of the Twitter REST API.* IEEE 24th International Conference on Tools with Artificial Intelligence. Vol. 1. IEEE, 2012.

TOWERS, SHERRY, ET AL. (2012). *Mass media and the contagion of fear: the case of Ebola in America.* PloS one 10.6.

TWITTER *https://support.twitter.com/*

TWITTER DEVELOPERS *https://dev.twitter.com/docs*

UDDIN, MD YUSUF S., ET AL. (2012) *On diversifying source selection in social sensing.* Networked Sensing Systems (INSS), 2012 Ninth International Conference on. IEEE

VALKANAS, GEORGE, KATAKIS, IOANNIS AND GUNOPULOS, DIMITRIOS (2014) *Mining twitter data with resource constraints.* Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 01. IEEE Computer Society

WANG, YAZHE, JAMIE CALLAN, AND BAIHUA ZHENG (2015) *Should we use the sample? Analyzing datasets sampled from Twitters stream API.* ACM Transactions on the Web (TWEB) 9.3: 13.

WANICHAYAPONG, NAPONG, WASAN PATTARA-ATIKOM, AND RATCHATA PEACHAVANISH (2014) *Road Traffic Question, Answering System Using Ontology* Joint International Semantic Technology Conference. Springer International Publishing

WEBB, STEPHEN (2016) *Twitter use in physics conferences.* Scientometrics 108.3: 1267-1286.

WELLS, CHRIS, ET AL. (2016) *Coproduction or cooptation? Real-time spin and social media response during the 2012 French and US presidential debates.* French Politics 14.2: 206-233.

WHITE, JOSHUA S., JEANNA N. MATTHEWS, AND JOHN L. STACY (2012) *Coalmine: an experience in building a system for social media analytics* SPIE Defense, Security, and Sensing. International Society for Optics and Photonics

XU ET AL. (2015) *The Impact of Sampling on Big Data Analysis of Social Media: A Case Study on Flu and Ebola.* IEEE Global Communications Conference (GLOBECOM)

YIN, ZHIJUN, ET AL. (2015) *A scalable framework to detect personal health mentions on Twitter* Journal of medical Internet research 17.6

ZAFAR, MUHAMMAD BILAL, ET AL. (2015) *Sampling content from online social networks: Comparing random vs. expert sampling of the Twitter stream.* ACM Transactions on the Web (TWEB) 9.3 (2015): 12.

Zhang, Yubao, et al. (2014) *What scale of audience a campaign can reach in what price on Twitter?* IEEE INFOCOM 2014-IEEE Conference on Computer Communications. IEEE

MOX Technical Reports, last issues

Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

15/2018 Simona, A.; Bonaventura, L.; Pugnat, T.; Dalena, B.
High order time integrators for the simulation of charged particle motion in magnetic quadrupoles

14/2018 Cuffaro, M.; Miglio, E.; Penati, M.;Viganò, M.
Mantle upwelling driven by asymmetric sea-floor spreading at northern Mid–Atlantic ridge

13/2018 Gandelli, E.; Penati, M.;Quaglini, V.;Lomiento, G.; Miglio, E.; Benzoni, G.M.
A novel OpenSees element for single curved surface sliding isolators

11/2018 Delpopolo Carciopolo L.; Bonaventura L.; Scotti A.; Formaggia L.
A conservative implicit multirate method for hyperbolic problems

12/2018 Dal Santo, N.; Deparis, S.; Manzoni, A.; Quarteroni, A.
Multi space reduced basis preconditioners for large-scale parametrized PDEs

10/2018 Menafoglio, A.; Gaetani, G.; Secchi, P.
Random Domain Decompositions for object-oriented Kriging over complex domains

09/2018 Menafoglio, A.; Grasso, M.; Secchi, P.; Colosimo, B.M.
Profile Monitoring of Probability Density Functions via Simplicial Functional PCA with application to Image Data

08/2018 Bonaventura, L.; Casella, F.; Delpopolo, L.; Ranade, A.;
A self adjusting multirate algorithm based on the TR-BDF2 method

06/2018 Antonietti, P.F.; Mazzieri, I.
High-order Discontinuous Galerkin methods for the elastodynamics equation on polygonal and polyhedral meshes

07/2018 Ieva, F.; Bitonti, D.
Network Analysis of Comorbidity Patterns in Heart Failure Patients using Administrative Data