Social Indicators Research (2021) 156:601–630
https://doi.org/10.1007/s11205-020-02296-w

ORIGINAL RESEARCH

Social Media and Twitter Data Quality for New Social Indicators

Camilla Salvatore[1] · **Silvia Biffignandi**[2] · **Annamaria Bianchi**[2]

Accepted: 9 February 2020 / Published online: 19 February 2020

Abstract
Social media represent an excellent opportunity for the construction of timely socio-economic indicators. Despite the many advantages of investigating social media for this purpose, however, there are also relevant statistical and quality issues. Data quality is an especially critical topic. Depending on the characteristics of the social media a researcher is using, the problems that arise related to errors are different. Thus, no one unique quality evaluation framework is suitable. In this paper, the quality of social media data is discussed considering Twitter as the reference social media. An original quality framework for Twitter data is introduced. A reformulation of the traditional quality dimensions is proposed, and the new quality aspects are discussed. The main sources of errors are identified, and examples are provided to show the process of finding evidence of these errors. The conclusion affirms the importance of using a mixed methods approach, which involves incorporating both qualitative and quantitative evaluations to assess data quality. A collection of good practices and proposed indicators for quality evaluation is provided.

Keywords Big Data · Twitter · Quality · Error

1 Introduction

Social media can be defined as a family of websites and applications that allow users to share messages and contents (images, videos, articles, etc.). These include social networking, blog/microblog, content sharing, and virtual world websites and applications. Despite the fact that social media were considered a fad in the beginning, they now occupy a key role in people's lives, reflecting several aspects of our virtual and real social life, as well as shaping our identity (Gündüz 2017).

According to the *Global Digital Report 2019,* released by *We Are Social* and *Hootsuite* (2019), the number of worldwide active users of social media, i.e. those who logged in in the reference period of 30 days, follows an increasing trend and is equal to, on average,

✉ Silvia Biffignandi
silvia.biffignandi@unibg.it

[1] Department of Economics, Management and Statistics, University of Milano-Bicocca, Milan, Italy

[2] Department of Management, Economics and Quantitative Methods, University of Bergamo, Via dei Caniana 2, 24127 Bergamo, Italy

45% of the world population. However, while in North America, Northern Europe, and East Asia, social media penetration reaches 70%, in Central Asia and Middle Africa it is still low and equal to, on average, 10%. According to the same report, at a global level, *Facebook* is the most used social media type since it has, on average, 2271 million monthly active users (MAU). Then, the most popular social media are *YouTube* (1900 million MAU), *Instagram* (1000 million MAU), the Chinese-specific social media *Qzone*, *Douyin,* and *Sina Weibo (*500 million MAU*)*, and finally there are *Reddit, Twitter,* and *LinkedIn* (300 million MAU). At local level, different patterns can be observed, and the use of some social media is clustered around specific regions. It is also important to notice that the user audience changes according to the social media type.

The rise of social media has also changed the way enterprises do business, in that social media, and particularly social networks, are now well integrated in the business strategy of both small and big enterprises.

Official Statistics is also interested in the use of social media for the construction of more timely socio-economic indicators, describing new aspects of the society. It is however clear that social media data will not replace survey-based activity: they can provide complementary, faster, and specific information about a topic or they can help to asses unmeasured or partially measured socioeconomic phenomena. In this respect, the integration of social media data with traditional data sources is a challenging opportunity for the construction of timely and new social and economic indicators, as the combination of data from multiple sources can provide a better overview of economic phenomena (Baldacci et al. 2016; Stier et al. 2019).

In terms of constructing socio-economic indicators, social media, due to their large amount of data and timeliness, have to deal with several statistical problems, such as privacy, methodological, and, especially, quality issues. In this respect, social media data, considered as *Big Data*, share some of their advantages and critical issues. Several topics, however, are specific to social media and thus require specific methodological investigation, especially with respect to quality.

The need for quality measures and for a quality framework for Big Data has been advocated by several authors (Japec et al. 2015; Di Bella et al. 2018). So far, neither a shared definition of Big Data quality nor indicators for such quality have been established. What is clear is that Big Data come from a wide range of different sources, and therefore quality should be assessed specifically according to the source and the type of analysis to be performed. The development of a quality framework is extremely important in several respects. First, it is important to be aware of possible errors inherent in the use of Big Data, both when they are used by themselves or integrated with other data sources. Second, a quality framework can warn users (particularly non-statisticians) about the type of errors they may encounter when using Big Data. Third, such a framework can advise users of good practices when using this type of data.

This paper focuses on the quality of social media data and, in particular, Twitter data. Twitter has been chosen not only because it is one of the most popular social media platforms but also because of its features, which are relevant for data analysis. The main contributions of this paper are: (1) the introduction of an original quality framework for Twitter data, with identification of the main sources of error; (2) the proposal of a set of indicators to obtain evidence of the errors; and (3) a collection of good practices that should be undertaken when using this type of data.

In the first part of this paper, some statistical considerations on this new data source are proposed. In so doing, the method of inductive reasoning is followed. First, two social media-based indicators are compared with the corresponding official indicators to

practically highlight the differences from a statistical perspective (Sect. 2). Next, in Sect. 3, the discussion is generalised, and the social media data generation process with the related statistical issues are described. In Sect. 4, a review of quality definitions and a general framework of quality evaluation are presented. In this discussion, a user perspective is adopted, and quality is broadly defined to include user-specified dimensions. Section 5 presents a case study, and Sect. 6 draws some conclusions.

2 Social Media Experimental Statistics

Studies using social media data sources are a rather recent and growing research area; however, the literature is still scarce with respect to the complex issue of using social media data in a meaningful and appropriate way. The purpose here is to show the variety of research areas which could derive interesting insights from an adequate and aware use of social media data. The intention is not to provide an extensive literature review.

The existing literature shows the application of social media data in several fields. In *political science*, social media can be used to assess the impact of fake news on voting behaviours (Allcott and Gentzkow 2017), to predict citizens' political preferences and election results (O'Connor et al. 2010; Ceron et al. 2014, 2016; Celli et al. 2016; Hürlimann et al. 2016), to study the impact of candidates' campaigns (Hong and Nadler 2012; Enli 2017), to predict emerging political trends (Rill et al. 2014), and to measure active citizenship (Sanchez et al. 2017). In *economics*, promising applications include the use of social media for stock and market volatility forecasting (Bollen et al. 2011; Ranco et al. 2015), to generate early predictions of initial claims for unemployment insurance (Antenucci et al. 2014), and to assess the reaction to public policy decisions (Ray et al. 2018). In *medicine* and *psychology*, through social media data, it is possible to predict disease outbreaks (Achrekar et al. 2011; Signorini et al. 2011) and depression (De Choudhury et al. 2013) and to understand the sentiments towards vaccination (Salathé and Khandelwal 2011). Finally, in the *social sciences,* social media data are used to measure subjective well-being (Luhmann 2017), to understand the reaction of people to natural disasters (Sakaki et al. 2010) or terrorist attacks (Monsour 2018; Wilson et al. 2017), and for predicting personality (Golbeck et al. 2011; Qiu et al. 2012). Other relevant applications include the study of online debates and news coverage (Burscher et al. 2016) and the prediction of movie success (Krauss et al. 2008).

To generate official statistics, both Eurostat and National Statistical Institutes are interested in computing social media-based indicators, mostly on an experimental basis. As will be illustrated later in the paper with special reference to Twitter data, the social media data-generating process cannot guarantee the general validity of the derived statistical information. For this reason, National Statistical Institutes in general cannot guarantee the accuracy of these indexes with reference to the overall population and, thus, they are not considered official indicators. Instead, they can be used as additional data sources to be eventually integrated with other statistical data sources. Some of these indicators are discussed next.

The *Social Tension indicator (STI)* is a social media-based indicator developed at Statistics Netherlands. The STI is based on Twitter messages whose topic is disorder or unsafety, and it provides the percentage of messages related to social tension over the total. This indicator provides similar information to that of the survey *Netherlands' Safety Monitor* (Veiligheidsmonitor-VM), which aims to measure people's feelings about the neighbourhood's safety and crime level. This survey follows sound statistical principles that assure

the quality and the reliability of the output. The STI and the VM differ in the perception of unsafety they measure. For example, the impact of terrorist attacks is more evident in the STI than in the VM, while the tension in local events, or that does not affect the collective security, is more evident in the VM. The relevance of the STI is that it can provide daily data, thus serving as an early-warning system. Table 1 compares the statistical features of the two indicators.

Daas and Puts (2014) developed the *Social Media Sentiment (SMS)* indicator, using Dutch social media public messages retrieved from 400,000 sources including Twitter, LinkedIn, Facebook, and blog news. These messages were classified as positive, neutral, or negative, and an index was computed by taking the difference between the percentage of positive and negative messages. This indicator was developed as an alternative to the Consumer Confidence Index (CCI), an official indicator which provides a measure of consumers' attitudes on whether the economy is doing better or worse. The CCI is drawn from survey data according to the procedure outlined in Table 2. Daas and Puts (2014) found a stable and strong association between the path of SMS and that of the CCI. In particular, the changes in the SMS precede by almost seven days the changes in the CCI. The main opportunity to be exploited, thus, is to provide high frequency and advance information on consumer confidence with respect to the publication of official statistics estimates.

Tables 1 and 2 compare the social media-based and corresponding survey-based indicators and show that there are substantial differences with respect to many aspects of the two sources: statistical units, frequency, and accuracy of the analysis.

Another index based on Twitter data is the *Social Mood on Economy Index*, which has been released daily by Istat since 2016. This index is not an official indicator, but it is part of Istat's experimental statistics.[1] It provides a measure of the Italian sentiment on the state of the economy and is derived from samples of public tweets in the Italian language.

3 Statistical Considerations

In order to state some introductory statistical considerations, generic social media and, in some cases, Big Data in general are considered. First of all, it is important to understand the social media data production process, which statistical units can be observed, and which populations are involved. Hereunder, a general framework is presented, and a generic social media platform to describe the entire process is considered. This framework adapts very well to social networks but, with some adjustments, it can be reproduced for other types of social media as well.

Consider a generic user that decides to join a social media platform. Such a user can be categorised as a *person* or *other*, such as a firm (profit or non-profit) or an association (formal or informal). For simplicity, the last category is called *businesses*. Both people and businesses can create multiple accounts. For instance, each person can have a professional and a personal account, and each business can have an account for each division, unit or shop. Also, the so-called 'malicious' accounts, i.e. fake accounts and Internet robots (BOTs) exist. BOTs are software programmed to share specific contents at specific moments and are used by both people and businesses. For example, businesses use them for advertising purposes. BOTs are not always malicious, but their use can be, since they

[1] https://www.istat.it/en/experimental-statistics/experiments-on-big-data.

Table 1 VM versus STI: a comparison between survey and social media indicators. *Source*: Authors' own elaboration based on: https://www.cbs.nl/en-gb/our-services/innovation/project/social-tension-indicator-based-on-social-media and https://www.cbs.nl/en-gb/our-services/methods/surveys/korte-onderzoeksbeschrijvingen/netherlands-safety-monitor

	VM—Survey	STI—Twitter (social media)
Target population	Persons aged 15 years or older living in private household	Dutch Twitter users (private and business) that publish tweets (TW) or retweets (RT) related to unsafety in the considered period
Statistical unit	Persons, households	Person and business users
Frequency	Annually (August–November)	Daily (Jan 2010–Jan 2017)—nearly real-time indicator in development
Survey/Research method	Mixed-mode design: sequential CAWI (Computer-assisted web interviewing), paper, telephone	• Retrieval of all public TW and RT by Dutch users • Identification of the most used words to describe situations of unsafety by means of interviews based on VM content • Final glossary with 350 words, including synonyms describing safety and unsafety • Identification of the number of messages related to unsafety • Elimination of messages relating to sports events and politics due to the distorting effect (high amount of negative content)
Number of respondents	It differs every year. The nationwide minimum number is 65,000	It changes every day. It can include private and business users, and each user can post multiple messages.
Weighting	It compensates for differences in personal (age, sex, ethnic background, household size, and income) and regional (police district, province, and municipality) characteristics between the sample and the population	No weighting. Lack of information about users (personal and regional). However, they could be inferred applying different techniques
Accuracy	95% confidence interval (SPSS Complex Sample Module)	Unknown
Data structure	Structured	Unstructured
User/respondent identity	It is kept anonymous	It is kept anonymous
Consent for use of data	Explicit consent	Users give consent accepting the user agreement. Sometimes they are not aware that their data can be used for research purposes

Table 2 CCI versus SMS: a comparison between survey and social media indicators. *Source*: Authors' own elaboration based on: https://www.cbs.nl/en-gb/our-services/methods/surveys/korte-onderzoeksbeschrijvingen/consumer-confidence-survey and Daas and Puts (2014)

	CCI—Survey	SMS—social media
Target population	Dutch consumers	Dutch social media users (private and business) that post public messages in the considered period
Statistical unit	Households (the head of the household or his/her partner is interviewed)	Private and business users
Frequency	Monthly	Daily, weekly, and monthly aggregates
Survey/Research method	CATI (Computer-assisted telephone interviewing) Five questions on the economic general situation, financial situation of household, and consumption perspective. Respondents indicate if the situation has improved, deteriorated or remained unchanged	• Retrieval of all public messages; • Sentiment determination (sentence level-based classification): ○ negative ○ positive ○ neutral
Number of respondents/messages	1600 per month	• Many-to-many relationship between users, accounts, social media platforms, and messages. 2.5 million messages per day • 3 billion messages (2009–2014)
Weighting, accuracy, and quality	Data are checked for internal consistency and completeness Seasonal adjustment Reweighted to ensure sequential comparability	• Messages' sentiment is aggregated to reduce error • Sentiment is adjusted including emoji
Data structure	Structured	Unstructured
User/respondent identity	It is kept anonymous	It is kept anonymous
Consent for use of data	Explicit consent	Users give consent by accepting the user agreement. Sometimes they are not aware that their data can be used for research

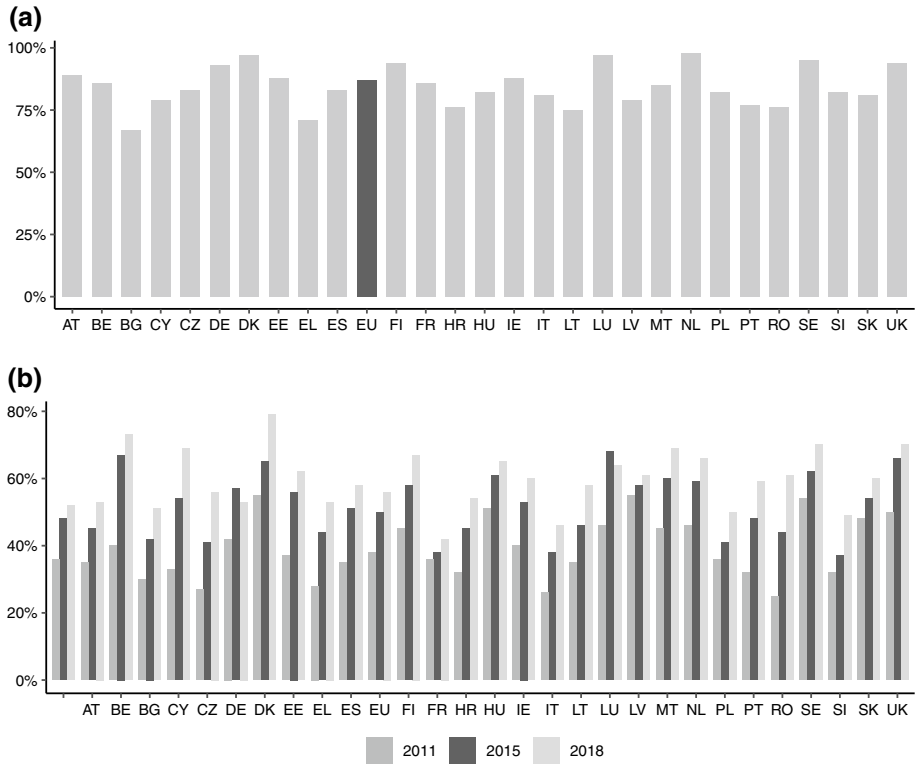

Fig. 1 **a** Internet access of households in % (2017); **b** People participating in social media in % (2011, 2015, and 2018). *Source:* Authors' own elaboration based on Eurostat data

can be used to spam and to share fake news or false negative/positive comments about a product in order to influence public opinion.

Only users who have an Internet connection can register and participate in social networking. Users can decide to join one or more social media platforms, according to their interests, and they can post or share contents with different frequencies and on different topics. As a consequence, the social media process is affected by several self-selection issues. The first depends on having an Internet connection, the second depends on the decision of the user to join a specific social media platform, and the third depends on the decision to post contents about a specific topic (Beresewicz et al. 2018).

To give an idea of the self-selection dimension, Fig. 1a shows the Internet access of households in the European Union in 2017. The average is almost 82%. Figure 1b shows the number of people participating in social media in the EU over 3 years: 2011, 2015, and 2018. The participation trend is clearly increasing, with the 2018 average about 56%. Both indicators vary across countries.

From a statistical perspective, three populations can be recognised in this process: the population of accounts Ω_A, the population of contents Ω_C, and the population of users Ω_U. Ω_A contains accounts that belong to people and to businesses. Users can create different accounts for different purposes (professional, personal, institutional, unit, division, shop, etc.). Notice that there is also the possibility that one account is managed by more people,

as in the case of business-type accounts in which a multi-user login is allowed. There is thus a one-to-many relationship between Ω_U and Ω_A. Further, each user can share more than one content with her/his accounts, so there is also a one-to-many relationship between Ω_A and Ω_C. Moreover, Ω_C contains both *original content* and *shared or comment-type content*.

There are several statistical issues related to the use of social media data as a source for social indicators. First, for the purposes of statistical inference, one should check whether Ω_U corresponds to the target population of the analysis. However, the characteristics of Ω_U are not observed. Rather, the information related to the accounts (if available) is observed, and from there, tentatively, the user's characteristics should be inferred. Further, there is usually no sampling scheme as data are usually retrieved according to the specific topic of the contents.

The second statistical consideration derives from the fact that Ω_A includes also malicious accounts. This can mislead the analysis because these accounts may act in a particular way. For example, they can share fake news, spam, or they are *fake accounts* that pretend to be other people, thus biasing the results.

The third statistical consideration deals with the link between the statistical phenomena of interest and the collected data. This link is usually indirect. For example, the opinions or sentiments of people about a certain topic are usually of interest, but, differently from surveys, the 'sentiment' variable is not observed, but rather should be extracted from a text. Further, the nature of the data should be taken into consideration, as well as the fact that a Twitter message is not a survey answer, and thus it might not correspond to what a user would have answered to a survey question (Schober et al. 2016).

Finally, there are other considerations that are not social media-specific but relate to Big Data in general. First, in some cases, it is necessary to deal with the *data deluge,* which means that the quantity of data available overcomes the capacity of storing, managing, and interpreting them. In this respect, a question of interest is 'How big is Big Data?'. The McKinsey Global Institute defines Big Data as 'datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyse; […] we assume that, as technology advances over time, the size of datasets will also increase' (Manyika et al. 2011). Even if this definition is true, sometimes data that are considered 'big' are smaller than other data that are not considered as Big Data (e.g. tweets on a specific topic in a short time frame versus census data). Second, the volatility of these data should be considered, as volatility makes it difficult to conduct analyses over time. Further, there are also considerations around privacy and consent to the use of the data. Obviously, privacy should always be granted in the whole process. The privacy and legal framework is beyond the scope of this paper and will not be assessed. Nevertheless, it is important to notice that, while the consent to the use of the data is explicit for surveys, it is not so for Big Data. In fact, the consent to the use of the data is usually indirect in Big Data settings. For example, when users join a social media platform, they subscribe to the user agreement, in which it is often stated that their personal information can be used by third parties. Even though users should be aware of this clause, this is not always the case. Moreover, a related aspect to consider is that the more users become aware that their data can be used, the more they tend to behave differently, for example by setting some privacy restrictions on the use of their data.

4 Assessing the Quality of Twitter Data

4.1 Quality Framework: A Paradigm Shift in the Big Data Era

To introduce a quality framework for social media data and Twitter data, it is useful to start from the quality framework that has been developed for surveys and assessing whether it can be applied or adapted to the analysis of Big Data. The evaluation of quality for surveys is a very large research area, which was developed in the early 1990s and includes the definition of quality in the different phases of the process: survey design, sampling scheme, data analysis, and dissemination of results. Moreover, quality is a multidimensional concept and, over time, researchers have proposed different definitions, such as 'fitness for use' (Wang and Strong 1996), 'user satisfaction' (Wayne 1983), or 'conformance to requirements' (Crosby 1988).

In general, most survey quality frameworks contain at least nine dimensions: accuracy, credibility, comparability, usability/interpretability, relevance, accessibility, timeliness/punctuality, completeness, and coherence (Eurostat 2019; OECD 2011). An important step towards a general framework to measure errors has been made with the definition of the Total Survey Error (TSE) paradigm (Biemer 2010). In the framework of this paradigm, errors are linked to the accuracy dimension, and the major sources of error to minimise TSE are identified and allocated. The TSE framework categorises errors into sampling errors (due to sampling scheme, sample size, and estimator choice) and non-sampling errors (due to specification, non-response, frame, measurement, and data processing). It is clear that this definition applies very well to survey data and to traditional data processing methods.

As previously discussed, Big Data substantially differ from survey data: they are usually unstructured, they do not correspond to any sampling scheme, they cover only a particular segment of the population, the link between the statistical phenomena of interest and the data is indirect, and both the inconsistency of the data across time and the volatility of the data sources weaken the continuity of the analysis over time. Data management and data processing techniques also differ. The data are usually stored in NoSQL databases (graph, key value, column, and document databases), and they are analysed with new techniques (machine learning, deep learning, natural language processing, multimedia processing, etc.). Moreover, in traditional data sources, quality at the origin is checked by the data collector, while quality at the origin for Big Data is out of the researchers' control, since Big Data are 'found' data. It is thus the responsibility of the analysts to be aware of the data's limitations and to take the necessary precautions to limit the effects of Big Data errors.

The diffusion of Big Data in every domain, including official statistics, has led to the necessity of setting data quality standards and best practices for Big Data. The definition of Big Data quality has thus become a subject undergoing intense study. The main point arising from recent studies is that a *general* definition of Big Data quality is meaningless, while a *source-specific* definition is more appropriate: quality should be defined for each source and according to the analyses to be performed (Cai and Zhu 2015; Batini et al. 2015; Immonen et al. 2015; Merino et al. 2016; Liu et al. 2016; Firmani et al. 2016; Japec et al. 2015).

Cai and Zhu (2015) argued that the dimensions of survey quality are general enough to be adapted to Big Data, with some adjustments, and they proposed a hierarchical definition of quality and its indicators that considers similar dimensions to those described above and from the users' perspectives. According to this framework, quality can be divided into five

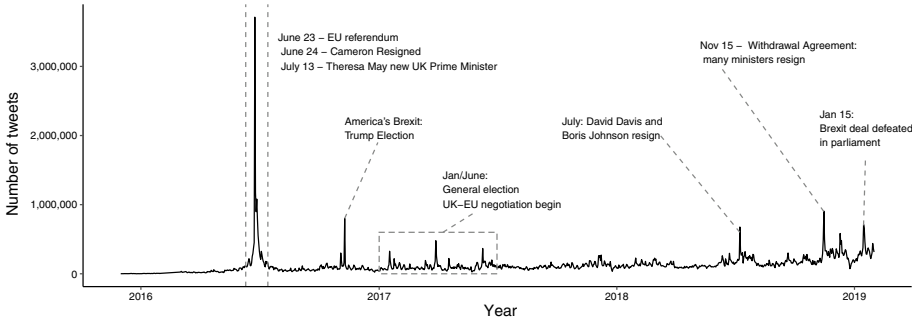

Fig. 2 '#Brexit' tweets from 1st December 2015 to 31st January 2019—Search API, Count endpoint estimates. *Source:* Authors' own elaboration

dimensions: availability, usability, reliability, relevance, and presentation of quality. Various elements and good practices are associated to each dimension according to the type of data and the data source. In Sect. 4.3, this framework is adapted to the analysis of Twitter data.

4.2 Why Twitter Data?

In the last few decades, online communication has taken an impressive jump, and Twitter grew very rapidly as well. In 2009, the tweets per day were 65 million; in 2011, they increased to 200 million[2]; and, in 2013, they reached 500 million.[3] After this, they have remained stable. This corresponds to, on average, 5700 tweets per second. The number of active users (MAU) has followed a similar path. According to Twitter's annual reports,[4] the number of active users was about 60 million in 2011; this increased to 100 million in 2012, and in the following years, it stabilised at around 300 million on average.

Currently, Twitter is one of the biggest social media companies. Thus, even though this study considers only one part of individual communication through social media, it covers an important part of it. Twitter communication is characterised by very short messages. This allows for a communication tool that catches sentiments in a reactive and synthetic way. Thus, by using statistical analyses, it is possible to perceive 'signals' that something relevant is happening and about the 'instantaneous' mood and successive evolutions.

In this respect, to demonstrate how 'signals' break out from Twitter data, Fig. 2 shows the time series of English-written tweets containing '#Brexit' from 1st December 2015 to 31st January 2019. Each peak corresponds to a specific event: the highest one matches the Brexit referendum (June 23, 2016), and then, in sequence, there are Trump election (November 8, 2016), defined as *America's Brexit*, the beginning of the UK-EU negotiation (March 29, 2017), the UK general election (June 8, 2017), and, finally, the main steps which led to the Brexit deal defeated in parliament on 15th January 2019.

It is evident that the number of tweets presents peaks as soon as something relevant happens. Sometimes, a relevant increase in the number of tweets is observed even if apparently

[2] https://blog.twitter.com/official/en_us/a/2011/200-million-tweets-per-day.html.
[3] https://blog.twitter.com/engineering/en_us/a/2013/new-tweets-per-second-record-and-how.html#.
[4] https://investor.twitterinc.com/financial-information/annual-reports/default.aspx.

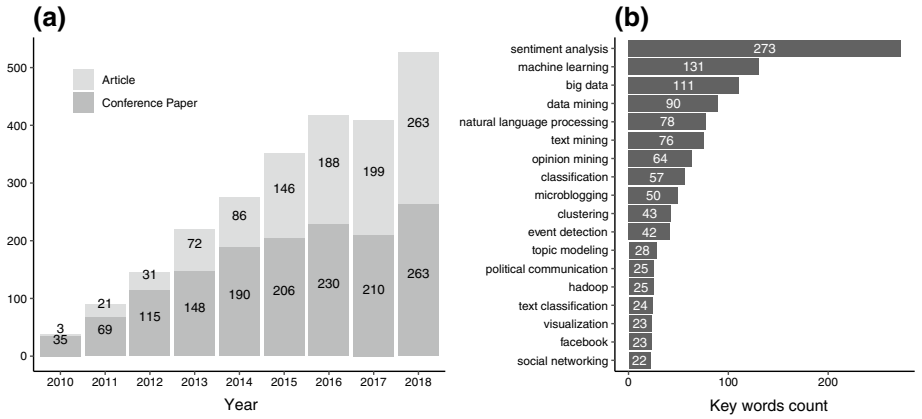

Fig. 3 **a** Internet number of articles published in Scopus database containing the term 'Twitter' in the title and 'Twitter data' in the abstract by year. **b** Most common keywords in articles (excluding generic ones such as: twitter, social media, social networks, etc.). *Source:* Authors' own elaboration based on Scopus data

no relevant event is going on. Taking the signal into account, it is possible to think about the situation and discover some interesting information.

The use of Twitter data is also the object of increasing interest by the scientific community. Figure 3a shows the number of papers and conference papers with *Twitter* in the title and *Twitter data* in the abstract contained in the Scopus database. There are 2475 articles and conference papers that match this query, and a constant increase over time can be observed. To give an idea of the topics covered, Fig. 3b shows the most common keywords in the articles. Most articles concern sentiment analysis. Other common topics are machine learning, natural language processing (NLP), and text and opinion mining.

4.3 Twitter Data Quality Framework

In this section, a data quality assessment method for Twitter data and its analysis is presented. In order to understand the development of quality dimensions for Twitter data, it is important to briefly recall its main characteristics and uses. Twitter requires users to register on the platform by creating a username and invites them to create a profile, including a brief description, name, and possibly photos for the account header and the profile headshots. Other socio-demographic information (such as gender and education) are not collected or stored in the Twitter metadata. Twitter does not require users to submit real personal information, thus allowing people to maintain their privacy. The default setting in Twitter is to make the profiles and contents public unless users modify their privacy settings.

Twitter allows users to post 280-character messages. When posting on Twitter, users tend to use shorthand, symbols, and emoticons. Hashtags (#) followed by topical keywords define the topic of posts and allow users to associate the tweet with all other tweets using the same identifying hashtags. Users can also add images and videos to their posts and can geotag them. Further, users can like, reply, or retweet (RT) other users' posts. An RT is

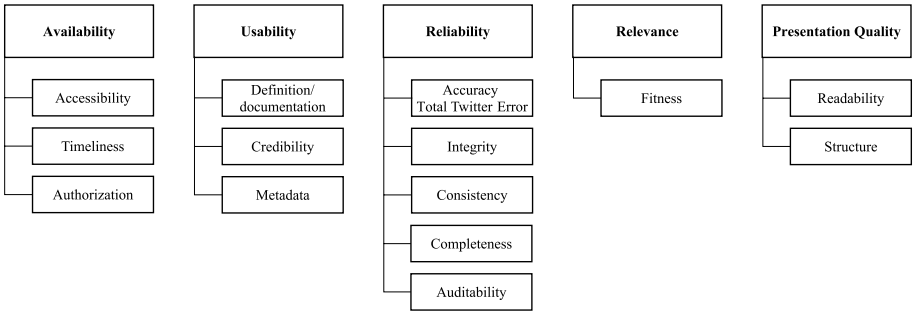

Fig. 4 Quality framework for Twitter. *Source:* Authors' own elaboration, based on Cai and Zhu (2015)

defined as 'the repost or forward of a message posted by another user.'[5] When retweeting a message, it is also possible to add a comment. Thus, an RT can have a double meaning: a person can RT a message because they share that opinion or, they can add a comment, which can also be in contrast with the original message.

Typically, researchers extract data from Twitter using keyword queries. The search query specifies a set of keywords relevant for the research, a time frame, language(s), and possibly a geographical region of interest. Results are returned at the tweet level. This means that there may be more than one tweet per user.

Next, typically a sentiment analysis is used to extract meaning from the text. Sentiment analysis is one of the most common text classification tools that allows to classify the sentiment underlying a tweet as positive, negative, or neutral. The sentiment classification can be based on three main approaches: machine learning (ML), lexicon-based, and hybrid, which combines the two previous techniques. The main idea underlying ML approaches is to improve computer skills with experience. ML methods can be supervised or unsupervised. Supervised learning means that a large number of labelled training documents are available, while with unsupervised methods, these documents are not available (Medhat et al. 2014). Notice that using ML techniques still involves subjective decisions similar to human coding. The idea underlying lexicon-based approaches is that some 'opinion-words' can be found in the text that can be ranked according to the intensity, or simply divided into positives, negatives, or neutrals. Here, dictionary-based or corpus-based approaches can be distinguished. The difference is that the latter uses statistical and semantic methods to identify the sentiment polarity, taking into account the context. Finally, hybrid methods combine ML techniques with sentiment lexicons.

Sometimes, the researcher also attempts to infer users' missing demographic and/or geographic information (Murphy et al. 2014).

In the following sections, a data quality assessment method for Twitter data and its analysis is presented. Starting from the framework proposed by Cai and Zhu (2015) for Big Data, we adapt and extend it to Twitter data, considering the quality dimensions presented in Fig. 4.

[5] https://en.oxforddictionaries.com/definition/retweet.

Table 3 Twitter APIs—Access type. *Source*: Authors' own elaboration

Access type	Description	Free/paid	Completeness
Search API: historical tweets	Standard: 7 days	Free	No
	Premium: 30 days or Full-Archive	Free (sandbox) or paid	Yes
	Enterprise: 30 days or Full-Archive	Paid	Yes
Filter real-time tweets: streaming API	Standard: *statuses/filter* endpoint	Free	No
	Enterprise: PowerTrack API (Firehose)	Paid	Yes
Sample of all public tweets: streaming API	Standard: *statuses/sample* endpoint	Free	No
	Enterprise: Decahose	Paid	10% random sample
Batch: historical tweet	Enterprise: Historical Power Track API	Paid	Yes

4.3.1 Availability

Availability refers to the ease and conditions under which data and related information can be obtained. It has three sub-dimensions: accessibility, timeliness, and authorisation. Accessibility refers to the difficulty level for users to obtain the data. For timeliness, the definition of Cai and Zhu (2015) for Big Data, which defines timeliness as the time delay from data generation and acquisition to utilisation, is adopted. Notice that this is rather different from the usual definition of timeliness/punctuality in the TSE framework (see, e.g. Biemer 2010), where timeliness is defined as adherence to schedules. In this respect, Twitter data are generally timely. The third dimension is authorisation, which refers to whether an individual or organisation has the right to use the data. There is no further investigation of this dimension in this paper. A synthesis of the proposed indicators for availability is shown in Table 7.

Accessibility

Accessibility is closely related to data openness. In contrast to most social media, Twitter data are accessible with few restrictions. Data are retrieved through Twitter APIs (Application Programming Interfaces). Table 3 provides a summary of the different access types. There are three levels of access: Standard (*free*), Premium (*pay as you go*), and Enterprise (*paid*). The idea is that upgrading the access type corresponds to more data, higher data fidelity, new search operators, metadata enrichments, and specific support services.

The Standard APIs consist of Search APIs and Streaming APIs. The Enterprise APIs include firehose, historical search, and engagement APIs, which allow for the implementation of deeper data analytics. The Premium APIs represent a middle ground between the two.

The Search APIs are based on the REST (Representational State Transfer) architecture, which is very popular among web APIs and is based on the *pull* strategy, which means that one needs to expressly request the data in order to retrieve them. In contrast, the Streaming APIs use the *push* strategy, which means that after a request, the API provides a continuous stream of data.

Access to the APIs is limited by the specific number of requests given a time window of, usually, 15 min.

Twitter data can be grouped into two categories: *tweets* and *users', followers', friends' information*. To retrieve tweets, there are two alternatives: historical or real-time data. The most-used APIs for academic research are the real-time Streaming and the Search APIs. The real-time Streaming APIs allow for retrieval of real-time data. However, there are some limitations to the amount of retrievable tweets and in some cases only a sample of tweets is returned. No details are provided on how the sample is drawn. The Search API is used to retrieve historical tweets. It provides two types of endpoint: data and count. The data endpoint allows for downloading the data, which are available at the time of the request and that match a search query. In contrast, the count endpoint provides an estimate of the number of data that matches a query when the data originally occurred and does not reflect the number of deleted tweets or any subsequent event. In order to retrieve the information o users, followers and friends, Twitter offers different endpoints, such as: *follower/list*, *friends/list,* and *user/lookup*. A comprehensive APIs list is provided on the Twitter developer portal.[6]

Further, Twitter is committed to improving and facilitating data access. For this purpose, it provides several guides and tutorials, and it promotes the exchange of knowledge and codes through the developer community.

Researchers can access data directly, using different programming languages, such as Ruby, PHP, Java, Scala, Phyton, and R. There are also ready-to-use packages, such as Tweepy or TwitterAPI (Python), and TwitteR or StreamR (R), which do not require deep coding knowledge. The issues of such packages are the slowness of the process and the problems in handling exceptions, such as automatic disconnection while obtaining streaming data. For example, a way to use Twitter Streaming APIs more efficiently is to work directly in the Hadoop Ecosystem. Within the Hadoop framework, Apache Flume can be used to retrieve data from the Internet and directly store them into the HDFS (Hadoop Distributed File System). In contrast to the previous methods, Flume is a distributed, very reliable, and fault tolerant service, which solves the previous issues (Farhan et al. 2018).

Finally, researchers can also request data from specialised companies, such as Topsy, Radian6, Teezir, Coosto, and Crimson Hexagon (Mishori et al. 2014; Dyar et al. 2014; Hogenboom et al. 2013; Daas et al. 2015; Raynauld and Greenberg 2014).

It is extremely important to underline that the type of access affects the results of the analysis. Indeed, the number of data available changes according to the access type, and, in case a sample of tweets is returned, there is no information on how the returned sample is constructed. In this respect, many studies have been conducted in order to compare different access types and to develop new sampling methods for Twitter data (Rafail 2018; Sampson et al. 2015; Valkanas et al. 2014; Driscoll and Walker 2014; Morstatter et al. 2014; Li et al. 2013). Table 4 summarises some of them.

Timeliness

There are several time dimensions and indicators to consider around the issue of timeliness.

The first is the difference between the publication of a tweet and its availability to be downloaded. The corresponding indicator can be called *availability timeliness*. For example, this difference is equal to 30 s for the Search API.

[6] Twitter API reference index available at: https://developer.twitter.com/en/docs/api-reference-index.html.

Table 4 Relevant studies on the accessibility of Twitter data. *Source*: Authors' own elaboration

Reference	Results
Hino and Fahey (2019)	They used public Twitter APIs to build a data set which is representative of the full data set retrieved through Twitter Firehose for Japanese-language tweets (target population: Twitter users who tweet in Japanese). The novelty of this approach is that it allows post hoc searching. The issue is that this approach cannot be replicated for all languages, such as for English tweets
Tromble et al. (2017)	They showed the potential biases in results when using different access types (Streaming and Search). They found significant differences, especially when the amount of data is larger than 1% of all tweets on Twitter at a given time
Joseph et al. (2014)	They compared many samples obtained from Twitter Streaming API through different connections about the same query and at the same time. They found a correspondence of 96% of tweets between samples, proving that this is not a suitable sampling procedure
Morstatter et al. (2013)	They performed several comparative analyses (top hashtag, Latent Dirichlet Allocation (LDA), network, and geographic measures) of tweets obtained with the Streaming API (sample data) and Firehose (full data). The bigger the tweets sample from the Streaming API, the more the analyses results were similar. Further, they repeated the comparisons using random samples from the Firehose data set. They found a bias in the way the Streaming API provides data

Tweets are not stored forever; if the user decides to eliminate a tweet or an account, all the information will be deleted. Therefore, the second aspect to consider is the time difference between data generation and data request. The proposed corresponding indicator is the *data request timeliness*. This indicator is strictly related to the *consistency* dimension, as discussed in Sect. 4.3.3.

Finally, the third aspect is the time difference between the data request and the data delivery, which varies according to the access type. By upgrading the type of access, more requests per minute can be submitted, as indicated on Twitter API documentation. The proposed corresponding indicator can be called *data delivery timeliness*.

4.3.2 Usability

Usability refers to whether the data are useful and meet users' needs. Usability has three sub-dimensions: credibility, definition/documentation, and metadata. Credibility refers to the objective and subjective components of the believability of a source or message. Definition/documentation consists of data specification (including data name, definition, ranges of valid values, standard formats, etc.). Metadata should be provided by the data producer to describe different aspects of the data sets, to reduce possible misunderstanding or inconsistencies.

Credibility of online social media is a subject of broad and current interest (Alrubaian et al. 2019). Several authors propose a credibility evaluation framework at user, tweets and topic levels (Yang et al. 2019; Verma et al. 2019; Gupta et al. 2019).

It is necessary that the data provider makes documentation available to facilitate the access and manipulation of data. Twitter data are provided in JSON format (JavaScript Object Notation), which is a well-known semi-structured form. As far as the definition/documentation and the metadata sub-dimensions are concerned, Twitter makes available *data dictionaries* that allow for an understanding of the data structure and easy recognition

Fig. 5 Basic JSON structure of Twitter data. *Source:* Authors' own elaboration

```
{
 "created_at": "Mon Feb 02 11:42:46 +0000 2019",
 "id_str": "111111111111111111",
 "text":"Twitter message",
 "user": {
 },
 "place": {
 },
 "entities": {
 },
 "extended_entities": {
 }
}
```

of Twitter objects and attributes. The objects are Tweets, Users, Entities, Extended entities, and Geo. Each object has some attributes; for example, some of the tweets' attributes are *created_at*, *id_str*, and *text*. A very basic example of a tweet's JSON structure is presented in Fig. 5. Finally, by upgrading the access type, usability is improved since extra support services are provided, and metadata are enriched.[7]

4.3.3 Reliability

Reliability refers to whether one can trust the data. Reliability is composed of five sub-dimensions: accuracy, consistency, completeness, integrity, and auditability. Accuracy refers to the difference between the measure of a social indicator provided by Twitter and some hypothetical true value. Consistency refers to whether the logical relationship between the correlated data is correct and complete. Completeness means that all components of a single datum (with multiple components) are valid. Integrity refers to the correctness of all the characteristics of the data. Auditability means that auditors can evaluate data accuracy and integrity within rational time and manpower limits during the data use phase. In the following sections, the concepts of accuracy, consistency, and completeness for Twitter data are proposed. A synthesis of the proposed indicators for reliability is shown in Table 7.

Accuracy

This dimension is strictly linked to the concept of 'errors,' which can affect both data and the analysis. In surveys, the sources and nature of the errors are accounted for by the TSE framework. Hsieh and Murphy (2017) adapted the TSE paradigm to Twitter and developed the *Total Twitter Error* framework. They identified three exhaustive and mutually exclusive sources of errors: query error, interpretation error, and coverage error.

Query error Twitter-based analyses usually start with the definition of a search query to extract information about a selected topic. Researchers formulate the search query trying to maximise topic coverage, and they can decide whether to retrieve tweets in real-time or from the historical archive through the APIs. The query error refers to the misspecification of the search query. It can be due to the keywords used (irrelevant or missing), to the inappropriate inclusion or exclusion of RTs, or to the inclusion/exclusion of other constraints

[7] https://developer.twitter.com/en/docs/tweets/enrichments/overview.html.

(geo-localisation, languages, and time frame). This type of error is similar to the 'error of selectivity that affects the response formation process in surveys' (Edwards and Cantor 2004, pp. 218–219).

A first consideration concerns the trade-off between the timing in submitting the data request, which aims at minimising the data loss, and the formulation of a search query so as to maximise the topic coverage. As a matter of fact, the observation of a social phenomenon allows researchers to identify the right keywords, hashtags, and other elements to build a proper search query. However, this process entails some time and, in the meanwhile, some tweets could be eliminated. Further, the quantity of tweets retrieved changes according to the type of access. Standard access has a limited number of operators, and the returned data are based on relevance and not on completeness.[8] This implies that some data are missing from the results, while completeness is assured only with Premium and Enterprise access types. Moreover, Premium and Enterprise access provides further search operators so as to refine the search query to obtain more topic-specific results. The advice to researchers is to formulate and compare the amount of data available for different queries and the proposed indicator for the query error is the difference in the amount of data retrieved using different queries.

Interpretation error This is due to the process of extracting insights from the text or to the process of inferring users' missing characteristics. Many studies focus on the opinions and sentiments of people about a certain topic. However, in contrast to surveys, the 'sentiment' variable is not observed and needs to be extracted from a text. Interpretation error can be defined as the extent to which the true meaning or value differs from that extracted by the researcher.

The interpretation error is influenced by textual errors. First, misspelled words cannot be recognised and elaborated by algorithms, and this affects the results of the analysis. In this respect, the percentage of misspelled words as an indicator of the accuracy of tweets at the origin is proposed. Second, abbreviations and slang words are difficult to evaluate by machines. In this context, text mining techniques represent a fundamental tool to identify and correct errors before the implementation of any analysis.

There are many techniques to extract insights from the text and to infer users' missing characteristics. Our aim is not to discuss the features or the precision of all the techniques, but to offer an overview of the elements that should be assessed to reduce the error. The main message to convey is that, when working with textual data, quantitative analyses are not enough; it is also necessary to do qualitative evaluations and to consider the context. In the case of an automatic sentiment classifier, evaluation metrics for ML models should be reported. Further, Kiefer (2016) suggested that the quality of the analysis will be higher if the input data and the training data are similar, i.e. using tweets on the same topic as training data will produce more accurate results that using tweets on a different topic. He also proposed the Cosine similarity or the Greedy String Tiling as possible indicators of the similarity between two documents. For dictionary-based approaches, the characteristics of the dictionaries should be reported. The authors argue that dictionary features can influence the results. First, it is useful to evaluate the ratio between positive and negative words for each lexicon in order to obtain an indicator of the negative or positive propensity of the lexicon. Second, a good lexicon should rank the score according to the level of the word's sentiment (Nielsen 2011; Hutto and Gilbert 2014). Moreover, the lexicon can

[8] https://developer.twitter.com/en/docs/tweets/search/overview/standard.html.

be constructed by integrating Big Data sources and survey data. For example, it can be constructed by identifying the most popular words used by people in tweets related to a specific topic, but also by asking people which words they use the most to describe a specific situation. Third, abbreviations and slang words can be included in the lexicon, even if this requires some effort. Finally, an interesting possibility is to integrate lexicon-based approaches with ML approaches. Sarcasm detection remains a problem both in dictionary-based and in ML approaches (Bamman and Smith 2015; Joshi et al. 2017).

In addition to the methodology-specific errors, linguistics-related issues should also be considered. Language is mutable across time and space. Even in the same country, there may be region-specific linguistic patterns and different dialects. In this respect, many studies argue that the inclusion of socio-demographic factors in the analysis (age, gender, location, etc.) can improve text classification (Hovy 2015; Johannsen et al. 2015; Jørgensen et al. 2015).

Coverage error Coverage error represents the difference between the target population and the units available for analysis on Twitter. Depending on the research purpose, there could be a mismatch between the target population and the observed population. For example, the actual young Italian population does not correspond to the young Italian population on Twitter. More precisely, the population of Twitter accounts suffers from both over-coverage and under- coverage. Under-coverage is related to the selectivity issues outlined in Sect. 3. For example, people not having access to the Internet are excluded from the retrieved data. This exclusion may introduce biases, as people having Internet access may be different from people without Internet access in many respects. Over-coverage is related to erroneous inclusions in the retrieved data. For the purposes of building social indicators, non-individual users (businesses and BOTs) are generally considered erroneous inclusions. In this respect, Twitter is committed to monitoring suspicious account activities and fighting spam and malicious automation.[9]

The coverage error is difficult to quantify. As a possible indicator of over-coverage, it is proposed to consider the percentage of people, businesses, and BOTs in the retrieved data. For under-coverage, it is proposed to consider the statistics on Internet penetration.

Consistency

Twitter data are mutable, especially with reference to user information such as username, declared location, and description. Twitter updates these data in the archive on a regular basis and, usually, the Search APIs return data updated at the time of the request. However, if the search query is based on operators that refer to mutable data (from, to, @, is:verified, etc.), it might happen that data are returned according to the last update of the Twitter database, which might not correspond to how they really appear on Twitter. Moreover, tweets can be deleted in any moment, and, thus, according to the request time, the output might change.

The proposed indicator for consistency is the *data loss*, given by the comparison of the count endpoint estimates and the number of data returned by the data endpoint.

Completeness

Completeness of data and the number of metadata available depend on the data access. In the Standard Search API, the returned data are based on relevance and not on completeness. This implies that some data are missing from the results. On the other hand, completeness is assured with Premium and Enterprise access types. Further, users can

[9] https://blog.twitter.com/en_us/topics/company/2018/how-twitter-is-fighting-spam-and-malicious-automation.html.

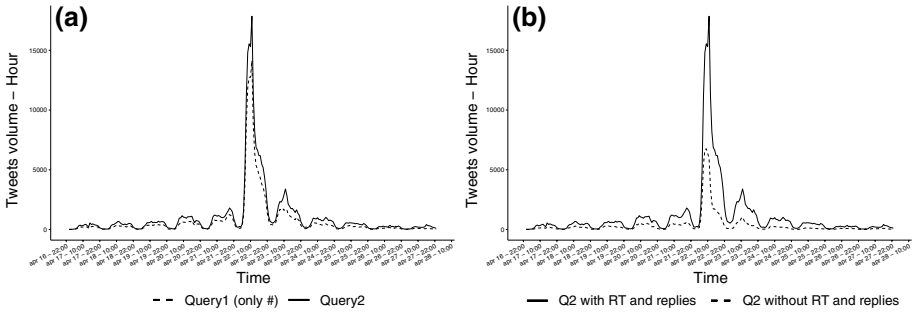

Fig. 6 **a** Query 1 versus Query 2; **b** query 2 with and without retweets and replies. *Source:* Authors' own elaboration

decide whether to disclose their information. Thus, the proposed indicators for completeness are the number of missing values and the type of access level. Data completeness has an impact on data accuracy as well, because through metadata it is possible to infer users' missing characteristics and evaluate the coverage error.

4.3.4 Relevance and Presentation Quality

Relevance is used to describe the degree of correlation between data content and users' expectations or demands. Its dimension is fitness, which refers to the amount of accessed data used by users and the degree to which the data produced match users' needs.

Presentation quality refers to a valid description method for the data, which allows users to fully understand the information. Its sub-dimensions are readability and structure. Readability is defined as the ability of data content to be correctly explained according to known or well-defined terms, attributes, units, codes, abbreviations, or other information. Structure refers to the level of difficulty in transforming semi-structured or unstructured data to structured data. These dimensions will not be discussed further in the following sections.

5 A Case Study: The London Marathon

In this section, the focus is on the reliability dimension. A case study is presented to show how to obtain evidence of errors. The tweets examined relate to the London Marathon held on 22nd April 2018 (Biffignandi et al. 2018).

To gain insight into the *query error*, a comparison of the results obtained from two different queries using the Search API with Standard access through the R package *twitteR* (Gentry 2016) is proposed. In the first query (Query 1), only the following hashtags are considered:

```
#londonmarathon OR #londonmarathon18 OR #londonmara-
thon2018
```

With this formulation, the tweets targeted concern the London Marathon only according to what the users have declared. However, this can lead to the exclusion of all tweets that do not contain the hashtag but that concern the topic. Thus, a second query (Query 2) is formulated:

Table 5 Lexicons' composition. *Source*: Authors' own elaboration

	AFINN (−5, +5)	Bing (−1, +1)	NRC* (−1, +1)
No. of words	2476	6788	6458
Positives	878	2006	2312
Negatives	1598	4782	3324
Fear			1476
Anger			1247
Trust			1231
Sadness			1191
Disgust			1058
Anticipation			839
Joy			689
Surprise			534

*NRC lexicon is composed of 6468 unique words that are classified into one or more categories

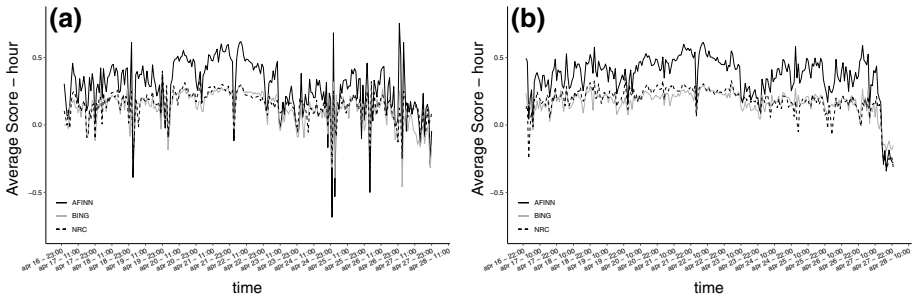

Fig. 7 **a** Sentiment over time excluding retweets and replies; **b** sentiment over time including retweets and replies. *Source:* Authors' own elaboration

```
#londonmarathon OR #londonmarathon18 OR #londonmara-
thon2018 OR (london + marathon)
```

The difference in the hourly volume of tweets is illustrated in Fig. 6a. The highest difference is on the 22nd and 23rd April, which correspond to the day of the marathon and the day after, respectively. The second element that affects the query error is the inclusion or exclusion of RTs and replies. Figure 6b shows the difference in the volume of tweets including and excluding RTs. The difference is very high especially on the day of the marathon.

To show how this choice is relevant in determining the analysis results, a comparison of the differences in the estimated sentiments, including or excluding RTs and replies, is proposed. Considering the second query, a dictionary-based sentiment analysis is implemented using three unigram-based lexicons contained in the *sentiments* dataset of the *tidytext* R package: AFINN, Bing, and NRC. In this type of dictionary-based approach, the total sentiment of the tweet is obtained by adding up the individual sentiment scores for each word in the tweet (Silge and Robinson 2016, 2017). The sentiment score for each tweet's word is obtained through the matching with a lexicon. AFINN contains a list of 2476 English words with a score between minus five and plus five in order to account for the different

shades in the sentiment (Nielsen 2011). Bing contains 6788 English words classified into positive (+1) or negative (−1) categories. NRC contains 6458 words classified into categories of positive and negative, but also into the emotions of anger, anticipation, disgust, fear, joy, sadness, surprise, and trust, according to Plutchik's wheel of emotions theory (Plutchik 1980). Lexicons are described in Table 5.

The results of the sentiment analysis are reported in Fig. 7. It is clear that the overall sentiment becomes more positive when RTs and replies are included.

Thus, a first recommendation is that researchers decide carefully which query to use and whether to include RTs or replies. It is argued that, in this specific case, Query 2 is more appropriate because it allows for the retrieval of more tweets, maintaining the topic well defined. The advice to researchers here is that RTs and replies should not be roughly included in the analysis. On the contrary, replies should be assessed with regard to the *parent* tweets and as part of a conversation. If the RT includes a comment, the tweet's sentiment should be calculated considering the comment, which can agree or disagree with the original message. If the RT does not include a comment, it can be simply included in the analysis.

As an indicator of the *interpretation error*, it is useful to evaluate the ratio between positive and negative words for each lexicon in order to obtain an indicator of the negative or positive propensity of the lexicon. The ratio between positive and negative words is 0.55 for AFINN, 0.41 for Bing, and 0.7 for NRC. This implies that the number of negative words almost doubles and more than doubles the positive ones in AFINN and Bing, respectively. Thus, these lexicons have a negative propensity towards the sentiment. On the contrary, there is not a big difference in the number of positive and negative words for NRC. It is important to note that there are different aspects of the lexicon structure that can influence the results of the sentiment analysis. For example, a lexicon with a very low ratio can affect the sentiment analysis negatively, while the sentiment analysis can be more precise if a lexicon such as AFINN, where the score is assigned according to the level of negativity/positivity of the words, is used.

When investigating the interpretation error, it should be considered that the dictionary-based approach has some drawbacks. One drawback concerns the structure of the lexicons. AFINN, Bing, and NRC represented 61.7%, 65.19%, and 63.5%, respectively, of the tweets selected with Query 2, excluding RTs and replies. This result is due to two reasons: the first is that some messages do not contain opinion words; the second is that the words included in the lexicons do not match the words included in the messages. In this respect, the dictionary-based approach requires powerful linguistic resources, which are not always available.

Another drawback is that these lexicons are not context-specific, and this might lead to errors. For example, in general, words related to diseases (cancer, autism, hospital, hospice, dementia) are classified as negative. In the London Marathon case, however, these kinds of words are mainly linked to the presence of charity companies and people that raise funds and, thus, they should not be considered as negative.

Other words such as 'breaking', which is labelled as negative, can refer to 'breaking news', while 'hottest', which is labelled as positive, is negative in this context because it refers to people's complaints about the 'hottest day in London'. Moreover, the lexicon can contain wrongly classified words. For example, in the NRC lexicon, the words 'feeling', 'winning', and 'lovely' are classified as sad words. Thus, it is evident that the lexicon itself can be the main source of the interpretation error.

Further drawbacks are that these lexicons do not contain urban slang and abbreviations, which are common in social media texts, as well as sarcasm, which is very difficult to

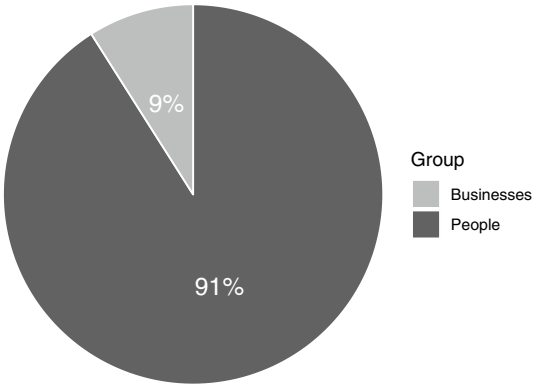

Fig. 8 Businesses and people. *Source*: Authors' own elaboration

detect. Finally, this method is based on unigrams and, hence, it does not consider the qualifiers before a word.

Different approaches might be implemented to improve the quality of the analysis. First, a good lexicon should rank the score according to the level of the word's sentiment, as AFINN does. Second, a lexicon can be constructed by integrating Big Data sources with survey data. For example, a lexicon could be defined by identifying the most popular words used by people in tweets related to a specific topic but also by asking people which words they use the most to describe a particular situation. Third, abbreviations and slang can be included in the lexicon, even if this requires some effort. Finally, an interesting possibility is to integrate lexicon-based approaches with ML approaches.

As for the *coverage error*, there is a clear one-to-many relationship between users and messages: the 91,750 messages analysed were generated by 53,839 users. If the interest is in the opinions of people, it is necessary to identify their messages among those of all users. To understand if they are businesses or persons, the information was retrieved about their accounts. However, only the information on 44,469 accounts was available. Restricting the analysis to the 41,514 messages classified by all the lexicons, only 30,712 messages were associated with these accounts (for which information on users is available). In turned out that the latter were generated by 25,286 accounts.

To compute an indicator of over-coverage, focusing on the 25,286 accounts described above, the effort was made to distinguish between people's and businesses' accounts. In order to identify businesses' accounts, text mining techniques to classify them were implemented. Using the name and the description provided in the user's information to check whether the account referred to a person or an organisation, the common patterns that characterise a business, including (manually) charity organisations that raised funds during the event, were identified. Next, we labelled users as 'businesses' when their name contained some specific words, such as *news, B&B, hotel, hostel, organisation, society, foundation, charity, research, hospice, fundraising, hospital*, etc., or their description contained patterns such as *we help, we are the, we are specialist, we are founded by, we are reliant, we are fundraising, we are team, we are now open, we are professional, we are medical, we provide, contact us, our clients, our aim*, etc.

The results are reported in the pie graph (Fig. 8), which shows the composition of the analysed account's population: 9% of the accounts are expected to be businesses (corresponding to 2339 accounts), while 91% are expected to represent people.

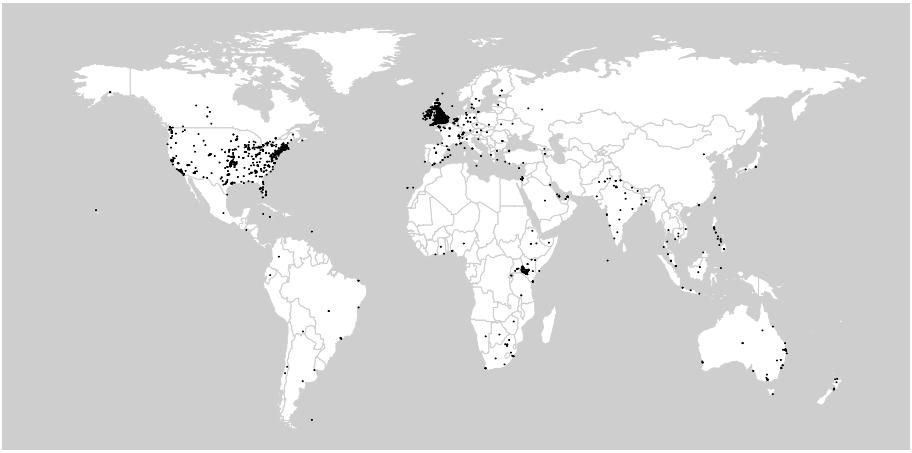

Fig. 9 Users' declared location in the world *Source*: Authors' own elaboration

Table 6 Data comparisons. *Source*: Authors' own elaboration

Day	No. Tweets[a]	Tweets retrieved[b]	Tweets available April 2019[C]	Loss	% of data loss
April 17th	3803	3731	2342	1389	37.22
April 18th	5055	4814	2940	1874	38.92
April 19th	6236	6153	3782	2371	38.53
April 20th	9833	9645	5999	3646	37.80
April 21st	14,968	14,854	9068	5786	38.95
April 22nd	116,185	115,494	72,580	42,914	37.15
April 23rd	24,954	24,176	14,777	9399	38.87
April 24th	8257	7870	4845	3025	38.43
April 25th	4443	4428	2494	1934	43.67
April 26th	2309	2307	1438	869	37.66
Total	196,043	193,457	120,265	73,207	38.00

[a]Tweets originally posted (count endpoint); [b]Tweets retrieved day by day with the standard Search API; [c]Tweets available 1 year later (tweet ID search)

Under-coverage is even more difficult to evaluate. It requires looking at some statistics about Internet and social media penetration. Moreover, to assess both over- and under-coverage, it is necessary to take into account the localisation of users. This information is not always provided, or else it is not provided correctly in many cases. Moreover, the localisation is not a proxy for citizenship. For this purpose, with reference to the users classified as *people*, the geographical coordinates of their declared location were retrieved. The coordinates are plotted in Fig. 9. As expected, the higher number of users are located in Anglophone countries, and this is because we retrieved only English language tweets. However, many users also declared themselves to be located in Europe.

Table 7 Summary of good practices and proposed indicators for Twitter quality evaluation (availability and reliability dimensions). *Source*: Authors' own elaboration

Quality dimensions	Sub dimensions	Critical points	Good practices	Proposed indicators
Availability	Accessibility	The number of retrieved data changes according to the access type No information on how the returned sample is constructed are available	Preferring Premium and Enterprise access levels	Qualitative: Type of access level
	Timeliness	Different time dimension to consider Trade-off between the timing in submitting the data request and the formulation of a search query	Preferring Premium and Enterprise access levels	Quantitative: Availability timeliness; Data request timeliness; Data delivery timeliness.
Reliability	Accuracy	*Query error*		
		The formulation of a search query aims to maximize the topic coverage Researchers should also carefully decide whether to include RT and replies in the analysis	Formulating and comparing the amount of data available for different queries	Quantitative: Difference in the amount of data retrieved using different queries.
		Interpretation error		
		Process of extracting insights from the text or to the process of inferring users missing characteristics In textual analysis, lack of context, slang and textual errors are relevant issues	Providing evaluation metrics for machine learning models, and carefully describing the dictionary used for the analysis Assessing replies with regard to the "parent" tweet and the RT with regard to the comments if any	Quantitative: Percentage of misspelled/unknown words; Evaluation metrics for ML models; Dictionary characteristics;
		Coverage error		
		Twitter data suffers both from under- and over-coverage. Under-coverage is difficult to evaluate	Classifying users in people, business and BOTs Looking at Internet penetration statistics	Quantitative: Percentage of people; Percentage of businesses; Percentage of BOTs; Internet penetration.

Table 7 (continued)

Quality dimensions	Sub dimensions	Critical points	Good practices	Proposed indicators
	Consistency	Twitter data are mutable, and they are not stored forever	Formulating the search query using non-mutable operators Retrieving data when the event occurs	Quantitative: Data loss
	Completeness	It improves with Premium and Enterprise access levels Users decide whether to disclose their information or not	Preferring Premium and Enterprise access levels	Quantitative: Number of missing values Qualitative: Type of access level

Next, consistency was studied by considering the data loss over time. It was investigated how many tweets retrieved with Query 1 between the 17th and 26th of April 2018 were still available after 1 year. In order to do so, a tweets' ID search was implemented, as tweets' IDs uniquely identify tweets and do not change over time. Table 6 shows, in order, the estimate of originally posted tweets according to the count endpoint, the number of tweets retrieved with the Standard Search APIs day by day, the number of tweets available and the data loss after 1 year from the marathon. About 38% of the data are no longer available.

Indicators of consistency were obtained by the data loss after 1 year and by comparing the count endpoint estimates with the number of data retrieved by the data endpoint.

6 Conclusions

Social media are an important and massive mode of communication between individuals. The statistical analysis of these data represents a promising source of information; however, several challenges and methodological problems need to be investigated for their effective use as statistical indicators. The quality issue is one of the main concerns, and this needs to be studied both for social media in general and for specific social media platforms.

In this paper, the authors propose a comprehensive quality framework for analysing Twitter data. The results show how errors arise in the different phases of the process: data collection and data analysis. Original and specific Twitter quality issues are described and appropriate quality indicators are experimentally computed. Table 7 summarises a collection of good practices and proposed indicators for the two most discussed dimensions, availability and reliability, that should be considered when using this type of data. This quality framework is a useful scheme for conducting evaluations of retrieved data. In this respect, a recommendation is to prepare quality reports to provide information on each quality dimension. The design of data retrieval could also benefit from referring to this quality framework. Moreover, some dimensions (like accessibility and relevance) are qualitative and difficult to quantify. Mixing quantitative and qualitative indicators is therefore recommended.

This study also entails some limitations. First, only experiments on Twitter data are undertaken in this paper. Even though, Twitter is a prominent social media platform, other platforms could be explored to complete a more general quality framework. Second, this paper investigates only one error and dimension at a time. Further developments could include investigating how the different quality dimensions and sources of error interact. Data quality dimensions are interrelated with each other and changes in one quality dimension impact on other dimensions as well. Some examples follow. Improving the accessibility dimension will improve timeliness (by ameliorating data delivery timeliness), completeness (more data are available), and usability (extra support services provided, and metadata enriched). Improving timeliness will improve consistency, but worsen accuracy. Indeed, it has been showed that there exists a trade-off between data request timeliness, which aims at minimising the data loss, and the formulation of a search query so as to maximise the topic coverage. Timeliness, consistency, topic- and population-coverage also have a positive impact on the relevance dimension. Credibility of messages impacts the accuracy of the analysis. Such relationships should be deeper investigated and researchers should be aware of these interactions and trade-offs when performing their analysis.

Third, new emerging quality aspects, such as access constraints, need to be further investigated. Indeed, social network data belong to private companies, and, in this respect the issue of digital divide between data rich and data poor emerges. As Boyd and Crawford (2012) recognised, in the system there are data producers, data collectors, and data analysers. New relationships need to be structured among these agents. The digital divide is now mainly due to data rather than to skills or technology, which can be acquired. Usually, the access to full-fidelity data is costly, and it turns out that only a few 'well-resourced' universities and companies can afford it. As a matter of fact, social media companies can decide whether, at which cost, to whom, and which data to make available through the Application Programming Interfaces (Lomborg and Bechmann 2014). In this respect, the data producers of official statistics using social media or other Big Data sources will have to face agreements with private companies; thus, a new paradigm is arising.

Some final advices concern how social media Twitter data should be considered in the statistical analysis context. These data are generated from a large number of users that belong to a specific subpopulation group (social media users), and no statistical selection criteria with respect to the total population can be applied. Therefore, it is difficult to ensure representativeness. The value of these data is that they help in catching signals. From an adequate and aware use of social media data, interesting insights could be derived to be considered in the context of more general indicators.

Acknowledgements Research supported by the grants of the University of the authors, various years.

References

Achrekar, H., Gandhe, A., Lazarus, R., Yu, S. H., & Liu, B. (2011). Predicting flu trends using twitter data. Paper presented at the 2011 *IEEE Conference on computer communications workshops*, *INFOCOM WKSHPS 2011* (pp. 702–707). https://doi.org/10.1109/infcomw.2011.5928903

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives, 31*(2), 211–236.

Alrubaian, M., Al-Qurishi, M., Alamri, A., Al-Rakhami, M., Hassan, M. M., & Fortino, G. (2019). Credibility in online social networks: A survey. *IEEE Access*, 2019, 7, art. no. 8572695, 2828–2855.

Antenucci, D., Cafarella, M., Levenstein, M., Ré, C., & Shapiro, M. D. (2014). *Using social media to measure labor market flows*. National Bureau of Economic Research, Working Paper 20010. https://doi.org/10.3386/w20010.

Baldacci, E., Buono, D., Kapetanios, G., Krische, S., Marcellino, O., Mazzi, G., et al. (2016). *Big Data and macroeconomic nowcasting: From data access to modelling*. Brussels: Eurostat.

Bamman, D., & Smith, N. A. (2015, April). Contextualized sarcasm detection on Twitter. In *Ninth international AAAI conference on web and social media*.

Batini, C., Rula, A., Scannapieco, M., & Viscusi, G. (2015). From data quality to Big Data quality. *Journal of Database Management (JDM), 26*(1), 60–82.

Beresewicz, M., Lehtonen, R. T., Reis, F., Di Consiglio, L., & Karlberg, M. (2018). *An overview of methods for treating selectivity in Big Data sources*. Brussels: Eurostat.

Biemer, P. P. (2010). Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly, 74*(5), 817–848.

Biffignandi, S., Bianchi, A., & Salvatore, C. (2018). *Can Big Data provide good quality statistics? A case study on sentiment analysis on Twitter data.* Presented at the "International Total Survey Error Workshop", June 2018, Duke University, North Carolina.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science, 2*(1), 1–8.

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

Burscher, B., Vliegenthart, R., & Vreese, C. H. D. (2016). Frames Beyond words: Applying cluster and sentiment analysis to news coverage of the nuclear power issue. *Social Science Computer Review, 34*(5), 530–545.

Cai, L., & Zhu, Y. (2015). The challenges of data quality and data quality assessment in the Big Data era. *Data Science Journal, 14,* 2.

Celli, F., Stepanov, E., Poesio, M., & Riccardi, G. (2016). Predicting Brexit: Classifying agreement is better than sentiment and pollsters. In *Proceedings of the workshop on computation modeling of People's opinions, personality and emotions in social media*, *110–118 Osaka December 12 2016*.

Ceron, A., Curini, L., & Iacus, S. M. (2016). *Politics and Big Data: Nowcasting and forecasting elections with social media*. London: Routledge.

Ceron, A., Curini, L., Iacus, S. M., & Porro, G. (2014). Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New Media & Society, 16*(2), 340–358.

Crosby, P. B. (1988). *Quality is free: The art of making quality certain*. New York: McGraw-Hill.

Daas, P. J. H., & Puts, M. J. H. (2014). Social media sentiment and consumer confidence, *European Central Bank Statistics Paper Series*, No. 5.

Daas, P. J., Puts, M. J., Buelens, B., & van den Hurk, P. A. (2015). Big Data as a source for official statistics. *Journal of Official Statistics, 31*(2), 249–262.

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. In *Seventh international AAAI conference on weblogs and social media*.

Di Bella, E., Leporatti, L., & Maggino, F. (2018). Big data and social indicators: Actual trends and new perspectives. *Social Indicators Research, 135,* 869–878.

Driscoll, K., & Walker, S. (2014). Big Data, big questions| working within a black box: Transparency in the collection and production of big twitter data. *International Journal of Communication, 8,* 20.

Dyar, O. J., Castro-Sánchez, E., & Holmes, A. H. (2014). What makes people talk about antibiotics on social media? A retrospective analysis of Twitter use. *Journal of Antimicrobial Chemotherapy, 69*(9), 2568–2572.

Edwards, W. S., & Cantor, D. (2004). Toward a response model in establishment surveys. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys*. Hoboken: Wiley.

Enli, G. (2017). Twitter as arena for the authentic outsider: Exploring the social media campaigns of Trump and Clinton in the 2016 US presidential election. *European Journal of Communication, 32*(1), 50–61.

Eurostat. (2019). *Quality assurance framework of the european statistical system*. Eurostat report.

Farhan, M. N., Habib, M. A., & Ali, M. A. (2018). A study and performance comparison of mapreduce and apache spark on Twitter data on hadoop cluster. *International Journal of Information Technology and Computer Science (IJITCS), 10*(7), 61–70.

Firmani, D., Mecella, M., Scannapieco, M., & Batini, C. (2016). On the meaningfulness of "Big Data quality". *Data Science and Engineering, 1*(1), 6–20.

Gentry, J. (2016). Package 'twitteR'. *CRAN repository.* https://cran.r-project.org/web/packages/twitteR/twitteR.pdf. Accessed 18 Feb 2020.

Golbeck, J., Robles, C., Edmondson, M., & Turner, K. (2011). Predicting personality from twitter. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing* (pp. 149–156). Washington, DC: IEEE.

Gündüz, U. (2017). The effect of social media on identity construction. *Mediterranean Journal of Social Sciences, 8*(5), 85–92.

Gupta, P., Pathak, V., Goyal, N., Singh, J., Varshney, V., & Kumar, S. (2019). Content credibility check on Twitter. *Communications in Computer and Information Science, 899,* 197–212.

Hino, A., & Fahey, R. A. (2019). Representing the Twittersphere: Archiving a representative sample of Twitter data under resource constraints. *International Journal of Information Management, 48,* 175–184.

Hogenboom, A., Bal, D., Frasincar, F., Bal, M., de Jong, F., & Kaymak, U. (2013). Exploiting emoticons in sentiment analysis. In *Proceedings of the 28th annual ACM symposium on applied computing* (pp. 703–710). New York: ACM.

Hong, S., & Nadler, D. (2012). Which candidates do the public discuss online in an election campaign? The use of social media by 2012 presidential candidates and its impact on candidate salience. *Government information quarterly, 29*(4), 455–461.

Hovy, D. (2015). Demographic factors improve classification performance. In *Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing* (Vol. 1: Long Papers) (Vol. 1, pp. 752–762).

Hsieh, Y. P., & Murphy, J. (2017). Total twitter error: Decomposing public opinion measurement on twitter from a total survey error perspective. In P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. Lyberg, N. C. Tucker, & B. T. West (Eds.), *Total survey error in practice: Improving quality in The Era of Big Data, Wiley Series in Survey Methodology* (1st ed., pp. 23–46). Hoboken, New Jersey: Wiley.

Hürlimann, M., Davis, B., Cortis, K., Freitas, A., Handschuh, S., & Fernández, S. (2016). A Twitter sentiment gold standard for the Brexit Referendum. In *SEMANTICS* (pp. 193–196).

Hutto, C. J. & Gilbert, E. E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international conference on weblogs and social media (ICWSM-14), Ann Arbor, MI, June 2014*.

Immonen, A., Pääkkönen, P., & Ovaska, E. (2015). Evaluating the quality of social media data in Big Data architecture. *IEEE Access, 3,* 2028–2043.

Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., et al. (2015). Big Data in survey research: AAPOR task force report. *Public Opinion Quarterly, 79,* 839–880.

Johannsen, A., Hovy, D., & Søgaard, A. (2015). Cross-lingual syntactic variation over age and gender. In *Proceedings of the nineteenth conference on computational natural language learning* (pp. 103–112).

Jørgensen, A., Hovy, D., & Søgaard, A. (2015). Challenges of studying and processing dialects in social media. In *Proceedings of the workshop on noisy user-generated text* (pp. 9–18).

Joseph, K., Landwehr, P. M., & Carley, K. M. (2014). Two 1% s don't make a whole: Comparing simultaneous samples from Twitter's streaming API. In *International conference on social computing, behavioral-cultural modeling, and prediction* (pp. 75–83). Cham: Springer.

Joshi, A., Bhattacharyya, P., & Carman, M. J. (2017). Automatic sarcasm detection: A survey. *ACM Computing Surveys (CSUR), 50*(5), 73.

Kiefer, C. (2016). Assessing the quality of unstructured data: An initial overview. In *LWDA* (pp. 62–73).

Krauss, J., Nann, S., Simon, D., Gloor, P. A., & Fischbach, K. (2008). Predicting movie success and academy awards through sentiment and social network analysis. In *ECIS* (pp. 2026–2037).

Li, R., Wang, S., & Chang, K. C. C. (2013). Towards social data platform: Automatic topic-focused monitor for twitter stream. *Proceedings of the VLDB Endowment, 6*(14), 1966–1977.

Liu, J., Li, J., Li, W., & Wu, J. (2016). Rethinking Big Data: A review on the data quality and usage issues. *ISPRS Journal of Photogrammetry and Remote Sensing, 115,* 134–142.

Lomborg, S., & Bechmann, A. (2014). Using APIs for data collection on social media. *The Information Society, 30*(4), 256–265.

Luhmann, M. (2017). Using Big Data to study subjective well-being. *Current Opinion in Behavioral Sciences, 18,* 28–33.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big Data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute, Report.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal, 5*(4), 1093–1113.

Merino, J., Caballero, I., Rivas, B., Serrano, M., & Piattini, M. (2016). A data quality in use model for Big Data. *Future Generation Computer Systems, 63,* 123–130.

Mishori, R., Singh, L. O., Levy, B., & Newport, C. (2014). Mapping physician Twitter networks: Describing how they work as a first step in understanding connectivity, information flow, and message diffusion. *Journal of medical Internet research, 16*(4), e107.

Monsour, S. (2018). Social media analysis of user's responses to terroris using sentiment analysis and text mining. *Procedia Computer Science, 140,* 95–103.

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* (pp. 555–556). ACM.

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 *ICWSM*.

Murphy, J., Link, M. W., Childs, J. H., Tesfaye, C. L., Dean, E., Stern, M., et al. (2014). Social media in public opinion research: Executive summary of the AAPOR task force on emerging technologies in public opinion research. *Public Opinion Quarterly, 78*(4), 788–794.

Nielsen, F. Å. (2011). Afinn. Richard Petersens Plads, Building, 321.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. In *Fourth international AAAI conference on weblogs and social media*.

OECD. (2011). *Quality framework and guidelines for OECD statistical activities*. Version 2011/1. STD/QFS(2011)1.

Plutchik, R. (1980). *A general psychoevolutionary theory of emotion* (pp. 3–33). New York: Academic Press.
Qiu, L., Lin, H., Ramsay, J., & Yang, F. (2012). You are what you tweet: Personality expression and perception on Twitter. *Journal of Research in Personality, 46*(6), 710–718.
Rafail, P. (2018). Nonprobability sampling and Twitter: Strategies for semibounded and bounded populations. *Social Science Computer Review, 36*(2), 195–211.
Ranco, G., Aleksovski, D., Caldarelli, G., Grcar, M., & Mozetic, I. (2015). Price effects of Twitter sentiment on stock price returns. *PLOS One, 10*(9), e0138441.
Ray, P., Chkrabarti, A., Ganguli, B., & Das, P. K. (2018). Demonetization and its aftermath: an analysis based on Twitter sentiments. *Sadana, 43,* 186.
Raynauld, V., & Greenberg, J. (2014). Tweet, click, vote: Twitter and the 2010 Ottawa municipal election. *Journal of Information Technology & Politics, 11*(4), 412–434.
Rill, S., Reinel, D., Scheidt, J., & Zicari, R. V. (2014). Politwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis. *Knowledge-Based Systems, 69,* 24–33.
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* (pp. 851–860). New York: ACM.
Salathé, M., & Khandelwal, S. (2011). Assessing vaccination sentiments with online social media: Implications for infectious disease dynamics and control. *PLoS Computational Biology, 7*(10), e1002199.
Sampson, J., Morstatter, F., Maciejewski, R., & Liu, H. (2015, August). Surpassing the limit: Keyword clustering to improve twitter sample coverage. In *Proceedings of the 26th ACM conference on hypertext and social media* (pp. 237–245). New York: ACM.
Sanchez, C. R., Craglia, M., & Bregt, A. K. (2017). New data sources for social indicators: The case study of contacting politician by Twitter. *International Journal of Digital Earth, 10*(8), 829–845.
Schober, M. F., Pasek, J., Guggenheim, L., Lampe, C., & Conrad, F. G. (2016). Social media analyses for social measurement. *Public Opinion Quarterly, 80*(1), 180–211.
Signorini, A., Segre, A. M., & Polgreen, P. M. (2011). The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic. *PLoS ONE, 6*(5), e19467.
Silge, J., & Robinson, D. (2016). Tidytext: Text mining and analysis using tidy data principles in r. *The Journal of Open Source Software, 1*(3), 37.
Silge, J., & Robinson, D. (2017). *Text mining with R: A tidy approach*. London: O'Reilly Media Inc.
Stier, S., Breuer, J., Siegers, P., & Thorson, K. (2019). Integrating survey data and digital trace data: Key issues in developing an emerging field. *Social Science Computer Review*. https://doi.org/10.1177/0894439319843669.
Tromble, R., Storz, A., & Stockmann, D. (2017). We don't know what we don't know: When and how the use of Twitter's public APIs biases scientific inference. In *SSRN*.
Valkanas, G., Katakis, I., Gunopulos, D., & Stefanidis, A. (2014). Mining twitter data with resource constraints. In *Proceedings of the 2014 IEEE/WIC/ACM international joint conferences on web intelligence (WI) and intelligent agent technologies (IAT)*—(Vol. 01, pp. 157–164). IEEE Computer Society.
Verma, P. K., Sharma, V., & Agarwal, S. (2019). Credibility investigation for tweets and its users. In *Proceedings of the 3rd international conference on computing methodologies and communication, ICCMC 2019* (art. no. 8819809, pp. 925–928).
Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems, 12*(4), 5–33.
Wayne, S. R. (1983). Quality control circle and companywide quality control. *Quality Program, 16*(10), 14–17.
We are social and Hootsuite. (2019). *Global digital report 2019*. https://wearesocial.com/global-digital-report-2019. Accessed 18 Feb 2020.
Wilson, T., Spiro, E.S., Stanek, S. A., & Starbird K. (2017). Language limitations in rumor research? Comparing French and English tweets sent during the 2015 Paris attacks. In *Proceedings of the 14th ISCRAM conference*, *Albi, France May 2017*.
Yang, J., Yu, M., Qin, H., Lu, M., & Yang, C. (2019). A Twitter data credibility framework—Hurricane Harvey as a use case. *ISPRS International Journal of Geo-Information, 8*(3), 111.