PRELIMINARY PAPER:

Can Big Data provide good quality statistics? A case study on sentiment analysis on Twitter data

Authors: Silvia Biffignandi*[1], Annamaria Bianchi*[2], Camilla Salvatore*[3]

*Università degli Studi di Bergamo
[1]silvia.biffignandi@unibg.it
[2] annamaria.bianchi@unibg.it
[3] c.salvatore@studenti.unibg.it

1. Introduction

Nowadays, economists, political scientists, managers and sociologists are interested in Big Data, that is, the huge quantity of digital data provided by people interactions, machines and processes. They can be used to answer new questions, to build new socio-economic indicators, to provide an insight on people's preferences, behaviours, political movements, and to generate competitive advantages in companies. Moreover, Big Data can offer new macroeconomic now-casting opportunities for policy-makers, providing complementary and faster information on the state of the economy and its development. In particular, the combination of data from multiple sources can provide a better overview of the economic phenomena (Baldacci et al., 2016). Furthermore, in Official Statistics the integration of Big Data with traditional data sources is a challenging opportunity for the construction of social and economic indicators. Actually, it is unlikely that Big Data will completely replace survey-based activities: they can provide complementary and specific information about a topic or they can help to asses unmeasured or partially measured socio-economic phenomena, providing new and auxiliary variables in macroeconomic models. One of the Big Data advantages in social science is that "they occur naturally without any intervention or researcher-manipulation, then, in some cases, they can be more representative of the true opinion or behaviour or they can provide further information than what could be collected with surveys" (Tourangeau et al., 2000; Japec et al., 2015). On the other hand, new risks and costs are rising. For example, social Big Data indicators "usually do not correspond to any sampling scheme and they are often representative of particular segments of the population" (Di Bella et al. 2018). Moreover, quality and ethical issues (privacy, confidentiality and transparency) as well as, the technical difficulties to deal with and to interpret this huge amount of data.

The aim of this paper is to investigate whether it is possible to produce good quality statistics on social media sentiment. To this purpose, we develop a case study focused on sentiment analysis of Twitter data, we discuss the possible sources of errors and how to get evidence of them. We used a mixed approach, concluding about the importance of the use of a mixed integrated framework for studying quality and for trying to get substantive results of good quality. Section 2 presents the literature review on the total error for Twitter and data quality paradigms in the context of Big Data. In Section 3, the framework of the analysis is presented. In Section 4 we present our data. Section 5 presents the lexicons and the sentiment analysis methodology. In Section 6 we discuss the results of the analysis and the errors that can affect it. Finally, in Section 7, the main conclusions are drawn.

2. Literature review on Total Error and Data Quality paradigms

In 1990s, the study of data quality has begun, and researchers proposed different definitions such as “fitness for use” (Wang et al., 1996), “user satisfaction” (Wayne SR, 1983) or “conformance to requirements” (Crosby, 1988), as matter of fact that quality is a multifaceted concept. The International Standardized Definition (ISO 8402:1986) is “The totality of features and characteristics of a product or service that bear on its ability to satisfy stated or implied needs”[1]. Furthermore, the data production, their analysis and the survey process are linked to different quality definitions and dimensions. Big Data is a relatively new phenomenon and, therefore, there is not aprecise definition of quality, nor indicators for quality. Moreover, due to the heterogeneity of Big Data, “specific definition of quality should be considered based on the type of Big Data and the analysis implemented” (Firmani et al., 2016). Another element to consider is that, in traditional data sources the quality at the origin is checked by the data collector, while, Big Data are found data and the quality at the origin is out of the researchers control.

Here, we focus on social media data and in particular on Twitter. The first element to consider is that, in contrast with survey, social media data could have originated from a “malicious source” (Pääkkönen et al., 2017) and suffer of a selectivity bias. The error and the selectivity bias that can affect Twitter data have been analysed respectively by Hsieh and Murphy and by Beręsewicz et al. (Hsieh & Murphy 2017;Beręsewicz et al. 2018).

Hsieh and Murphy (2017) adapted the TSE paradigm to Twitter and developed the Total Twitter Error framework. They identify three exhaustive and mutually exclusive sources of errors: query error, coverage error and interpretation error (Fig. 1).

The *query error* depends upon the misspecification of the search queries. It can be due to the key words used (irrelevant or missing) and to the inappropriate inclusion or exclusion of retweets. Edwards and Cantor (2004) argue that this type of error is like the “error of selectivity that affects the response formation process in surveys”. Researchers should formulate the search query trying to maximizethe knowledge on the topic.

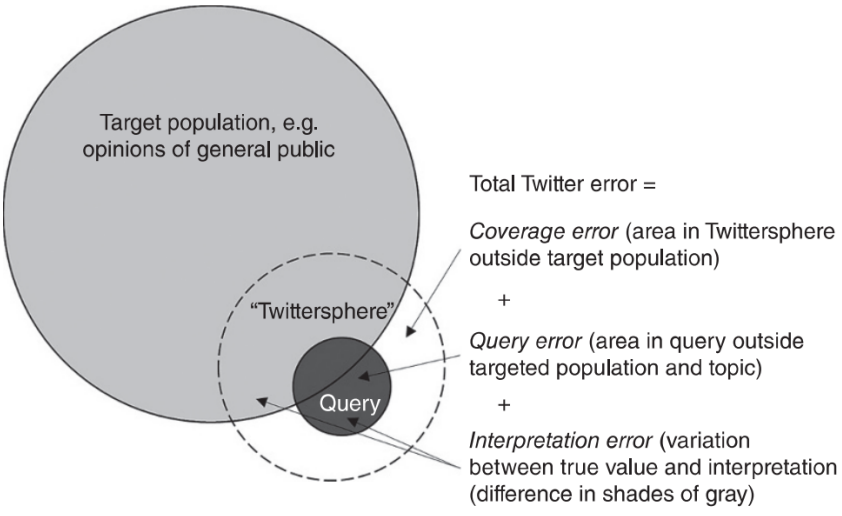

Figure 1. Total Twitter Error
Source: Hsieh, Y. P., and Murphy, J. (2017).

The *interpretation error* is due to the process of extracting insight from the text (sentiment).

[1]https://stats.oecd.org/glossary/detail.asp?ID=5150

The *coverage error* represents the difference between the target population and the units available for analysis on Twitter. In particular, depending on the research purpose, there could be a mismatch between the target population and the observed one, for example, the Italian young population does not correspond to the Italian young population on Twitter. More precisely, the population of Twitter accounts suffers from both over coverage and under coverage (Fig.2)

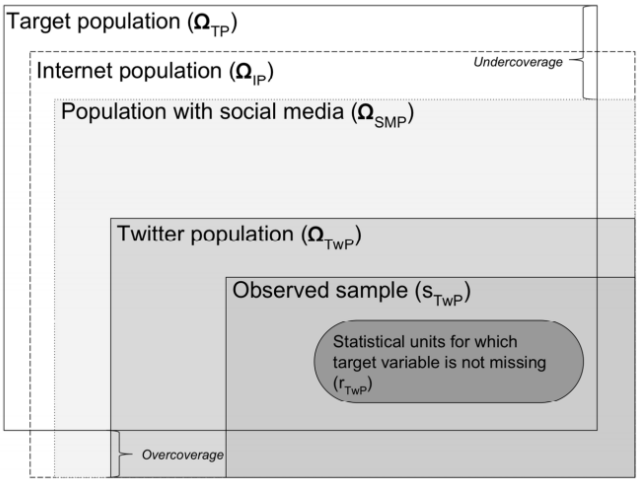

Figure 2. Over coverage and under-coverage
Source : Beręsewicz et al. 2018

The over-coverage is due to the fact that theTwitter population is composed not only by accounts that are associated to people, but also to organizations and BOTs. Thus, this over coverage can generate noise in the analysis. The under-coverage refers to the fact that the observed data do not cover all units of the target population.

3. Framework of the analysis

The way forward is to understand how to detect and check the quality and in particular which indicators can be used. The aim of this analysis is to explore whether it is possible to provide good quality statistics on social media sentiment from Twitter. For this reason, we have selected and monitored a specific event to study the evolution of the sentiment over time and to get insight on the people's opinion according to the different aspects that characterize the event.

The study is organized in three parts:

1. Data collection

It is about the data retrieval through the software RStudio. To do that, we used the *twitteR* package (Gentry, et al. 2016).

2. Text mining and sentiment analysis

We analyze and compare the structure of three lexicons and we present the dictionary-based technique to do the sentiment analysis.

3. Results and discussion

In this part we compare the results obtained by using three different lexicons and we discuss the query error, the interpretation error and the coverage error.

In Fig 3. the framework of our analysis is presented in a graphical way. A detailed description of the procedure is presented.

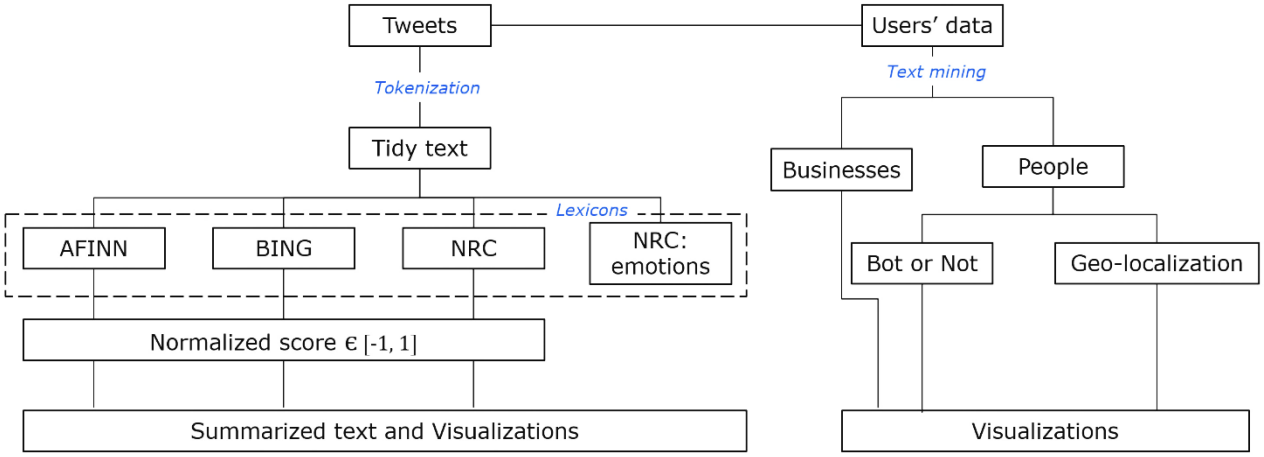

Figure 3. Framework of the mixed approach adopted in our analysis
Source: Authors' own elaboration

4. Data collection

We develop our case study using RStudio. The package we used to download data is *twitteR* (Gentry, et al., 2016). The package *twitteR* allows to download Tweets as well as users' information including: the text of theTweet, the date of publication, the source of the message, if it is a retweet or if it has been retweeted, the location (if it is geo-tagged), the screen name, the user's name, its description, if the account is verified and when it has been created, the number of followers, the profile's image URL and the location declared by the user[2].

This information is useful for the profiling of users and to assess the coverage and interpretation errors.

In this study we monitor the London Marathon (22nd April 2018) for a time span of 10 days: from the 17th to the 27th April. In the next sections, the daily volume of Tweets retrieved is presented. Since, the lexicons available are in English, we retrieved only English Tweets.

5. Text mining and sentiment analysis techniques

To implement the analysis, we work with tidy data and we apply the tidy data principles. In tidy data (Silge and Robinson, 2016, 2017):

- Each variable is a column;
- Each observation is a row;
- Each type of observational unit forms a table.

In our case of text analysis, *tidy text* format is a table with one token per row. A *token* is a meaningful unit of text, such as a word (as in this case) and the process of splitting the text into tokens is called *tokenization*. In the tokenization process we extract the word for each document by excluding punctuation and filtering the *stop words*. *Stop* words have a limited semantic meaning regardless of the document contents. Some examples are: "the, then, of, to, and".

After the tokenization, we obtain a *tibble* where the first column is the Tweet's (*document*) ID and each Tweet's word is a raw.

[2] For a complete list, please visit: [2]ftp://ftp.sam.math.ethz.ch/sfs/CRAN/web/packages/twitteR/twitteR.pdf

We refer to a *corpus* as a set of multiple similar *documents*. In our case, the totality of Tweets represents the corpus while each Tweet is a document.

In R, we use different packages to perform the analysis such as *tiditext*, *tidyverse, stringr, dplyr* and *tidyr*. We apply tidy data principles and the dictionary-based approach comparing different lexicons. In 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).

5.1. Lexicons: AFINN, Bing and NRC

We considered three unigram-based lexicons which are contained in the *sentiments* dataset of the *tidytext* package. All of them have been constructed by the authors analyzing different sources (Twitter, online reviews, etc..). They are:

- AFINN: developed by Finn Arup Nielsen (Nielsen, 2011). He manually labeled a list of 2,476 English words with a score between minus five (negative sentiment) and plus five (positive sentiment) according to their valence. Thanks to this classification, we have different "shades" in the sentiment and this can improve the results of the analysis.
- Bing[3]: developed by Bind Liu and collaborators. They categorized 6,788 English words into positive and negative categories. Then, score for positive words is 1, while for negative words it is -1.
- NRC: developed by Saif Mohammad and Peter Turney. In this case 6,468 words are classified into categories of positive, negative, anger, anticipation, disgust, fear, joy, sadness, surprise and trust. This classification is based upon the Plutchik's wheel of emotions theory (Plutchik, 1980). The NRC's dataset structure is more sophisticated: each word is classified in a binary way (Y/N) into the previous categories and each word can belong to more than one category (example: "abandon" belongs to the following categories: fear, negative and sadness).

Table 1 compares the number of positive, negative and *emotional* words of the tree lexicon.

AFINN is the smallest dataset, while Bing is the biggest one. Moreover, by calculating the ratio between positive and negative words for each lexicon we obtain an indicator of the negative or positive propensity of the lexicon. The ratio for AFINN is 0.55, for Bing is 0.41 and for NRC is 0.7. This implies that in AFINN and Bing, the number of negative words almost doubles the positive ones and thus, these lexicons have a negative propensity toward the sentiment. On the contrary, for NRC there is not a big difference in the number of positive and negative words. It is important to note that there are different aspects of the lexicon structure that can influence the results of sentiment analysis. For example, a lexicon with a very low ratio can affect the sentiment analysis negatively, while the sentiment analysis can result more precise if we use a lexicon like AFINN where the score is assigned according to the level of negativity/positivity of the words.

Moreover, it is interesting to investigate whether the lexicons share common words. AFINN and Bing share 1,315 words, in details, 870 are negative and 445 are positive. Bing and NRC have 2,484 words in common, 1,764 are negative while 720 are positive. Finally, AFINN and NRC have 1,029 words in common, 686 are negative and 343 are positive.

[3]https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

	AFINN (-5, +5)	Bing (-1, +1)	NRC (-1, +1)
N° of words	2,476	6,788	6,468*
Positives	878	2,006	2,312
Negatives	1,598	4,782	3,324
Fear			1,476
Anger			1,247
Trust			1,231
Sadness			1,191
Disgust			1,058
Anticipation			839
Joy			689
Surprise			534

Table 1. Lexicons composition. * NRC lexicon is composed by 6468 unique words that are classified into one or more categories. The final dataset is composed by 13901 elements.
Source: Authors' own elaboration

Given the different structure of the words' score assigned by the three lexicons, a normalization of the final score is needed. The normalization is implemented by following the formula proposed by Hutto et al. (2014).This is the formula that the authors used in writing the Vader library, a Python package to implement the sentiment analysis.

$$Normalizedscore = \frac{score}{\sqrt{score^2 + \alpha}} \quad (1)$$

The alpha parameter is empirically derived as to approximate the maximum expected value of sentiment words can be found in a sentence.It is set by default as $\alpha = 15$.

The range of the normalized score is:

$$Normalizedscore \in [-1, 1]$$

Moreover, the scores were converted into categories according to the following classification:

- Positive (P): $normalizedscore > 0.20$
- Negative (N):$normalizedscore < -0.20$
- Neutral (E): $-0.20 < normalizedscore < 0.20$

6. Results and discussion

In this section we investigate and discuss the three main sources of error with reference to our data.

6.1. Query error

The *query error* relates to the retrieval of the data and the observed sample. When researchers formulate the *search string*, their aim is to maximize the knowledge on the topic. Here we compare the results of two different specifications.

The first query takes into account only the hashtag and its structure follow: **"#londonmarathon OR #londonmarathon18 OR #londonmarathon2018".** Hashtags are used by users to define the topic of their posts and to allow other users to find messages on a specific topic. With this

formulation, we are sure that we are targeting Tweets that concern only the London Marathon 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, we formulated a second query: **"#londonmarathonOR #londonmarathon18 OR #londonmarathon2018 OR (london +marathon)"**. The difference in the hourly volume of Tweetsis represented in Fig 4.

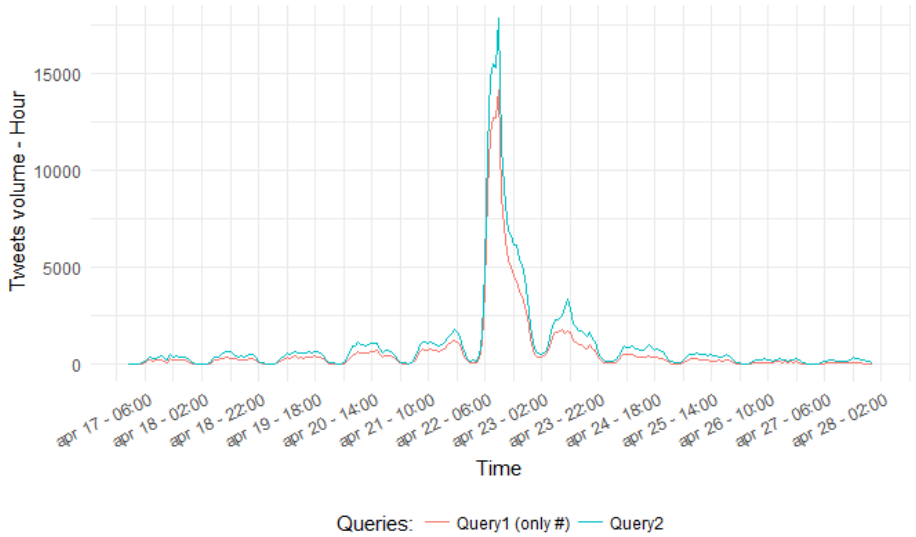

Figure 4. Volume of Tweet: Query1 versus Query2
Source: Authors' own elaboration

We can see that the highest difference is on the 22nd and 23rd April which correspond to the day of the marathon and the day after, respectively.

The differences in volume per day can be observed in the Tab. 2:

Day	Query 1	Query 2
April 17th	3,731	6,225
April 18th	4,814	8,318
April 19th	6,153	10,088
April 20th	9,645	16,066
April 21st	14,854	21,773
April 22nd	115,494	149,380
April 23rd	24,176	38,800
April 24th	7,870	15,423
April 25th	4,428	9,128
April 26th	2,307	4,779
April 27th	1,385	4,353
Total	**194,857**	**284,333**

Table 2. Volume of Tweets collecting with Query1 and Query2
Source: Authors' own elaboration on R.

For a better understanding of the difference between the queries we can assess the relative frequency distribution and the cumulative relative frequency distribution (Fig. 5 and Fig. 6).

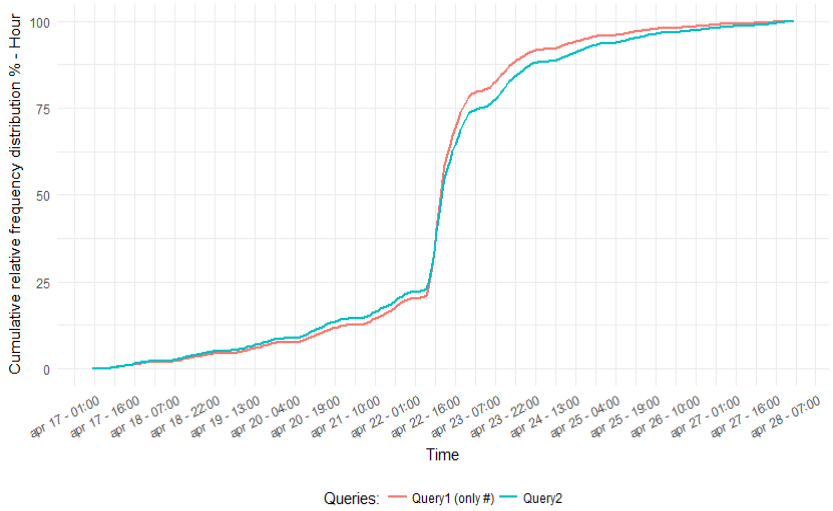

Figure 5. Cumulative relative frequency distribution - Query1 versus Query2
Source: Authors' own elaboration

The cumulative distributions show a similar pattern and, as expected, on the 22nd of April there is a strong increase. The difference between the two distributions lies in the fact that, from the day of the marathon, the cumulative percentage distribution of "Query 1" overcomes that of "Query 2". From both figures we can understand that the main difference is in the day of the marathon, when the relative frequency distribution of Query 1 is higher, and its cumulative relative frequency distribution is slightly more concentrated from the day of the event onward. Indeed, hashtags are mainly used by users to mark their messages as in relation to an event. The difference is not very relevant; thus, we expect a small difference due to the type of query.

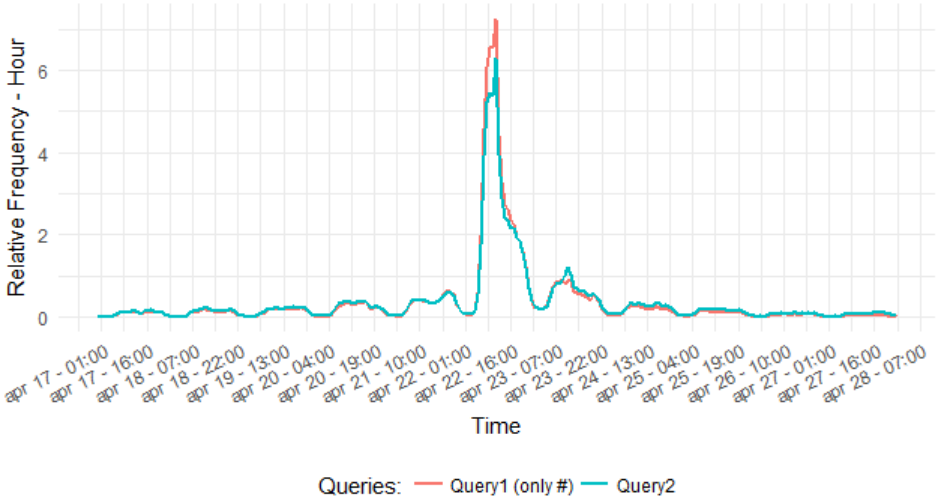

Figure 6. Relative frequency distribution: Query1 versus Query2
Source: Authors' own elaboration

We chose to analyze the data retrieved with the search specification of "Query 2" because it concerns a higher amount of data. The second element that can affect the *query error* is the inclusion or exclusion of retweets and replies.

A retweet is defined as "the repost or forward of a message posted by another user"[4]. In the retweet you can also add a comment. A reply is "a response to another person's Tweet"[5]. A retweet can have a double meaning: a person can retweet a message because he shares that opinion or, he can add a comment which is in contrast with the message. Thus, the inclusion of retweets could mislead the analysis. We also excluded replies because the sentiment of a reply refers to the topic expressed in the first message. However, for other types of investigation, such as product review and politicians' messages, it would be interesting to study separately the message's replies. In Fig. 7 it is possible to observe the difference in the volume of Tweets including and excluding retweets and replies. It is very high most of all in the day of the marathon.

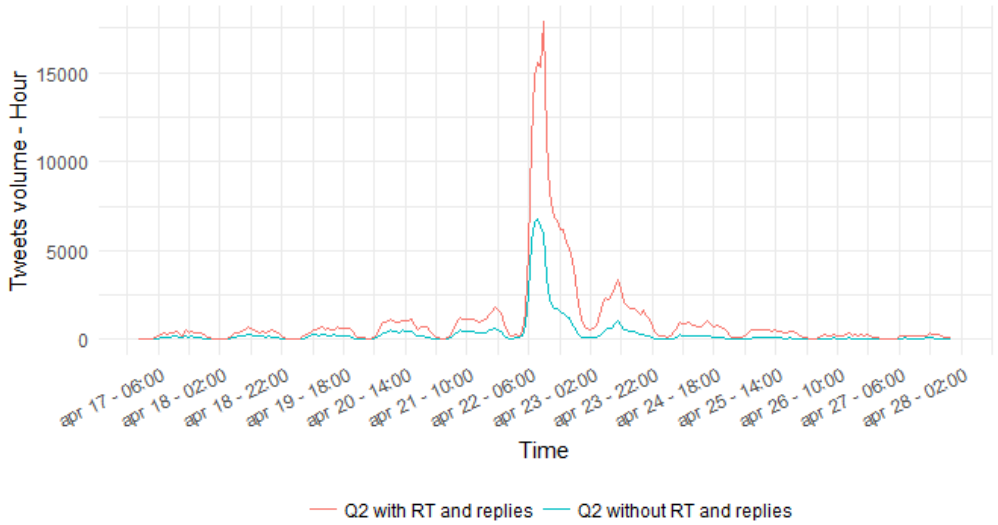

Figure 7. Tweets volume: a comparison between the exclusion and exclusion of retweet and replies
Source: Authors' own elaboration

In Fig.8 and Fig. 9, we can see that at the time of the event (22nd April in the morning), original messages (not retweet or replies) becomes more important and their relative frequency increases. This pattern is event-specific, in fact, the previous and following days are characterized by stability and relevant differences are small and not visible.

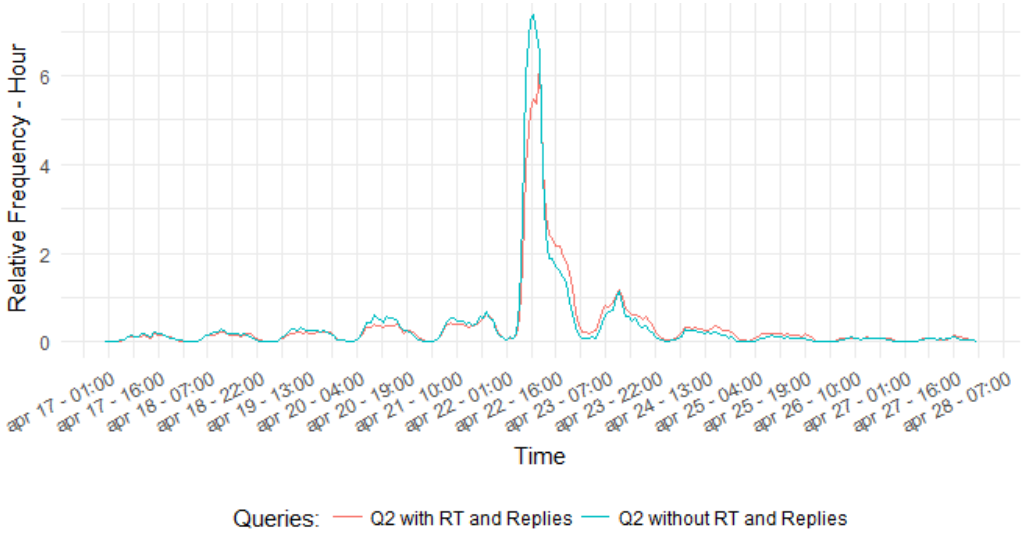

Figure 8. Relative frequency distribution: Query2 with and without retweets and replies
Source: Authors' own elaboration

[4]https://en.oxforddictionaries.com/definition/retweet
[5]https://help.twitter.com/en/using-twitter/mentions-and-replies

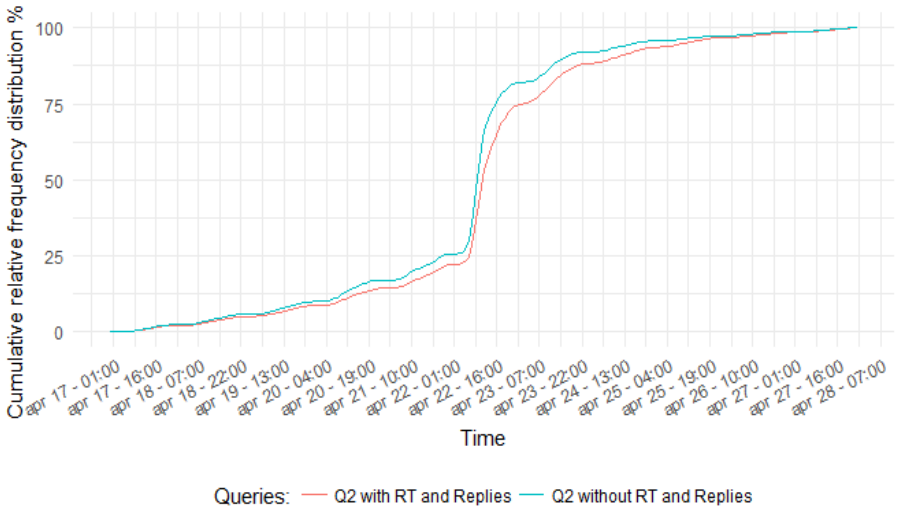

Figure 9. Cumulative relative frequency distribution, - Query2 with and without retweets and replies
Source: Authors' own elaboration

In Tab. 3, the exact number of replies and retweets per day are presented with reference to Query2.

Day	Total Volume of Tweets	N. of Retweets	N. of Replies	N. of Tweets excluding RT and Replies
April 17th	6225	3651	301	2273
April 18th	8318	4857	440	3021
April 19th	10088	5731	491	3866
April 20th	16066	8876	736	6454
April 21st	21773	12982	1001	7790
April 22nd	149380	92525	5177	51678
April 23rd	38800	28389	1088	9323
April 24th	15423	11605	475	3343
April 25th	9128	7251	249	1628
April 26th	4779	3348	178	1253
April 27th	4353	3068	164	1121
Total:	**284333**	**182283**	**10300**	**91750**

Table 3. Daily volume and composition of Tweets in Query2
Source: Authors' own elaboration

In conclusion, we decided to analyze the Tweets retrieved with the expanded query (Query2) excluding retweets and replies. So, starting from 284,333 collected Tweets (according to Query2) we selected 91,750 Tweets. The main message here is that more data does not imply a better knowledge of the topic. We exclude retweets and replies because they could bias the results as we will see in the Fig. 12. We suggest that a deeper analysis of retweets and replies should be made to understand in which way they affect the sentiment and to decide whether to include or exclude them, totally or partially.

6.2. Interpretation error

After having implemented the sentiment analysis algorithm, only 45.2% of Tweets (41,514 Tweets on 91,750) have been analyzed by all the three lexicons. In details, AFINN classified 56,646 Tweets, Bing 59,817 Tweets and NRC 58,322 Tweets. These results are due to the different composition of the three lexicons (number of words, ratio of negative/positive words and score structure). Thus, in order to provide comparable results, the analysis focuses only on the Tweets that all three lexicons have classified.

Table 4 shows the number of Tweets classified into positive, negative and neutral categories by the three lexicons.

	AFINN	BING	NRC
Positive	34430	32416	32233
Negative	6090	5938	5642
Neutral	994	3160	3639

Table 4. Number of Tweets classified into positive, negative and neutral by the three lexicons
Source: Authors' own elaboration on R.

Afterwards, we computed the concordance of the classification and the correlation of normalized scores. According to our framework, 33,279 Tweets have been classified in the same way by all the lexicons, while for only 8,235 the results are contrasting. Tab 5 shows the Goodman and Kruskal's Gammaindex of concordance which "evaluate the net proportion of concordant pairs of observation, as compared with all pairs of observations" (Sirkin, 2005). We can see that there is a high and positive association of the results.

	AFINN	BING	NRC
AFINN	1		
BING	0.884	1	
NRC	0.826	0.824	1

Table 5. Goodman and Kruskal's Gamma
Source: Authors' own elaboration on R.

Moreover, the normalized scores computed using the three lexicons are positively correlated and the correlation between AFINN and Bing is the highest one (Tab. 6).

	AFINN	BING	NRC
AFINN	1		
BING	0.8308680	1	
NRC	0.6614174	0.6729434	1

Table 6. Correlation matrix (normalized scores)
Source: Authors' own elaboration on R.

Fig. 10 shows the average hourly sentiment and it can be helpful to understand the emotional trajectory. The two straight lines represent the area where the sentiment is said to be neutral. We can observe that it is generally positive, while from the afternoon of the 22nd April and in the days after the marathon it shows a decreasing trend. This can be due to the fact that a runner collapsed during the marathon and then died. Moreover, people also complained that the marathon's day has been characterized by an unexpected above average temperature (23.2 Celsius degrees or 73.76 Fahrenheit degrees). In this situation, even if the negative/positive propensity is rather high, NCR shows a more negative pattern than the other lexicons. The more complex structure of the NCR

taking the wheel of emotions into account better catches the feelings. A deeper study of the characteristics of this lexicon will be undertaken in another study.

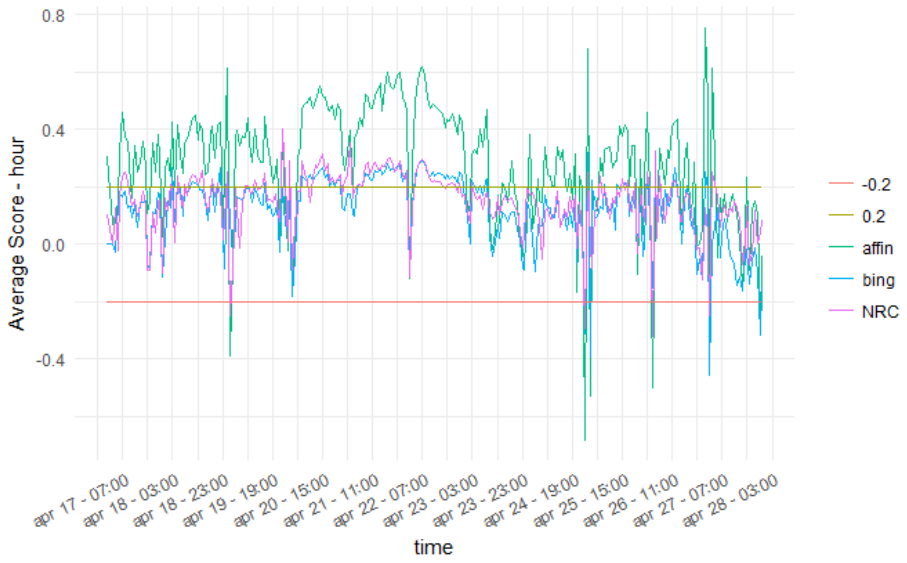

Figure 10. Hourly average score by lexicon over time (Query2 excluding RT and replies)
Source: Authors' own elaboration

This can be even more clear if we focus on the 22nd April's sentiment distribution obtained by using the NRC lexicon: in the afternoon, as the number of words connected to positive sentiment decreases, the number of words indicating a negative sentiment increases (Fig. 11, shaded area)

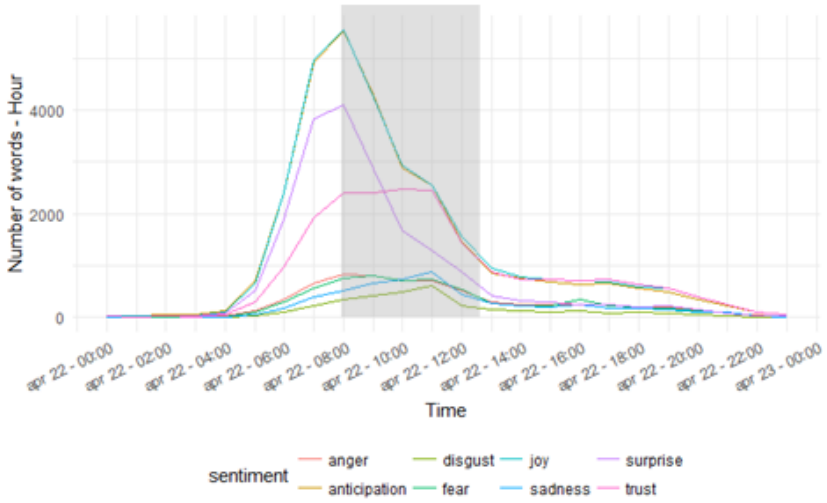

Figure 11. Hourly number of words according to the sentiment (NRC lexicon)
Source: Authors' own elaboration

To complete the overview and understand the differences among lexicons, we compare the previous sentiment trajectory with that of Query 2 including retweets and replies. Fig. 12 shows that NRC registers rather interesting differences with respect to the results of Fig.10: negative standardized average score is smaller when retweets and replies are included.

As discussed before, if we are interested in studying the public opinion, including retweetsand replies can bias the result. In fact, here, the sentiment is generally positive, and the negative peaks are due the large number of retweets. However, we suggest that retweets should be analyzed to

understand who published them (people or others), the topic of the message (advertising, news or messages) and eventually decide to include a part of them in the analysis.

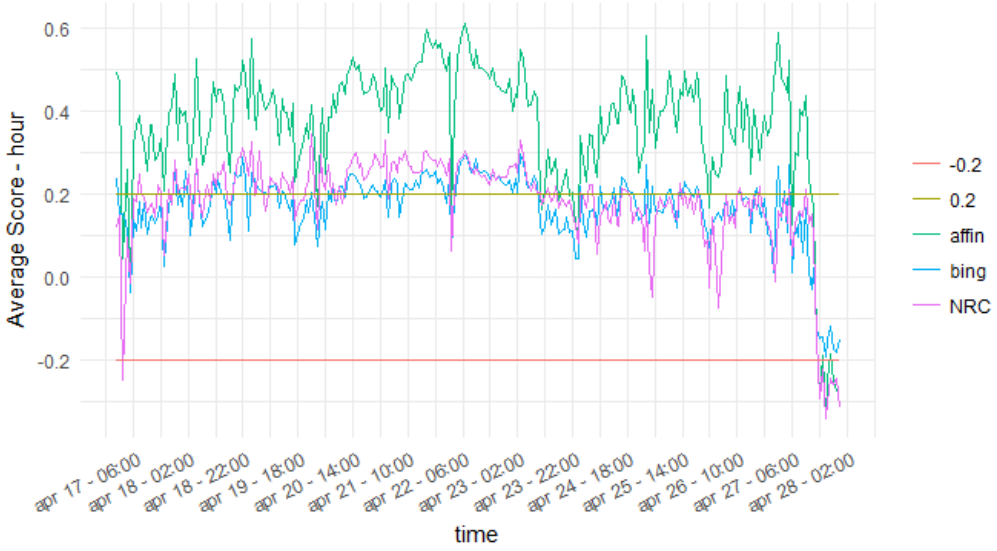

Figure 12. Hourly average score by lexicon over time (Query2 including RT and replies)
Source: Authors' own elaboration

When investigating the interpretation error, we need to take into account that the dictionary-based approach has some drawbacks. One drawback concerns the structure of the lexicons. AFINN, Bing and NRC classified respectively 61.7%, 65.19% and 63.5% of the tweets selected with Query2 excluding retweets and replies. This can depend on 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. 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 can be understood by looking at Figures 13, 14 and 15.

In the London Marathon queries, words that define diseases (cancer, autism, hospital, hospice, dementia, bloody) are mainly linked to the presence of charity companies and people that raise funds, thus, they should not be considered as negative. Other words such as breaking which is labeled as negative, can refer to "breaking news" while hottest, which is labeled as positive, is negative in this context because it refers to people's complain about the "hottest day in London". Moreover, the lexicon can contain wrongly classified words, for example in NRC feeling, winning and lovely are classified as sad words.Thus, it is evident that the lexicon can be the main source of the interpretation error.

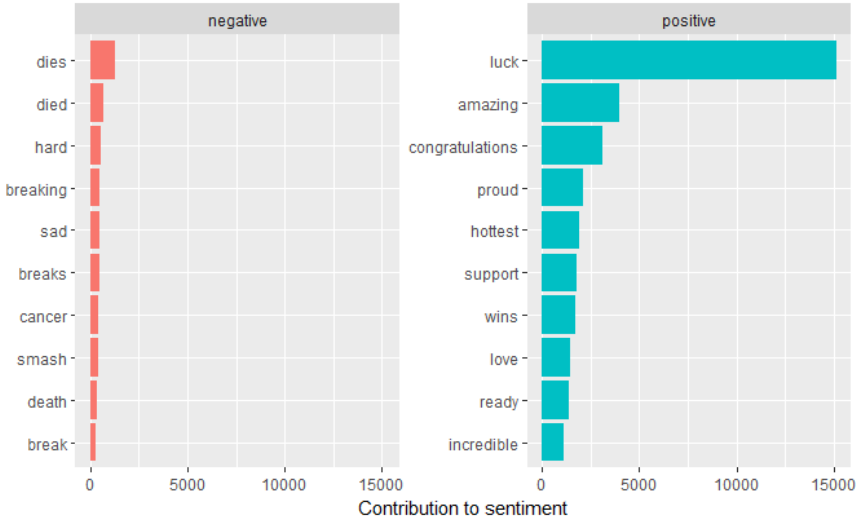

Figure 13. Contribution to sentiment of negative and positive words (BING lexicon)
Source: Authors' own elaboration

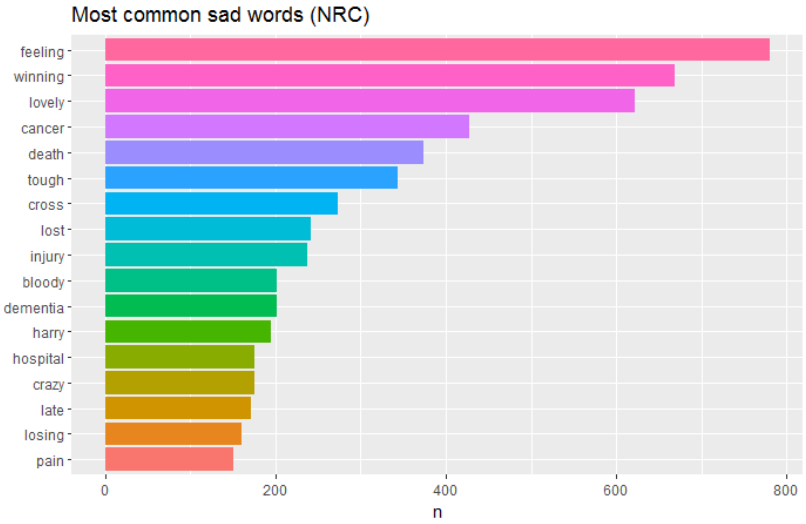

Figure 14. Most common sad words (NRC lexicon)
Source: Authors' own elaboration

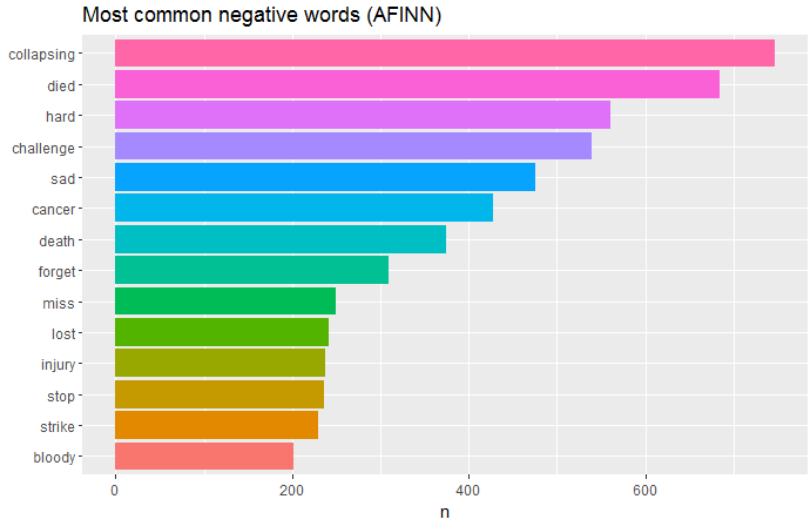

Figure 15. Most common negative words (AFINN lexicon)
Source: Authors' own elaboration

Further drawbacks are that these lexicons do not contain urban slang and abbreviation which are common in social media texts and the sarcasm, which is very difficult to detect.

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

We suggest different approaches that can be followed to improve the quality of the analysis. First, a good lexicon should rank the score according to the level of the word's sentiment like AFINN does. Second, a lexicon can be constructed by integrating Big Data sources and survey. 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 more to describe a particular situation. Third, abbreviations and slang can be included in the lexicon, even if this requires a big effort. Finally, an interesting possibility is to integrate lexicon-based approaches with machine learning approaches.

Moreover, to make the analysis more precise, we should identify to which sub-aspect the sentiment is linked (Liu, 2015). Latent direct allocation (LDA) is a method that allows to find the mixture of words associated with a topic but also the mixture of topics that describes each document. It performs very well with long texts, but with Twitter short texts it finds it more difficult to identify the exact topics. Indeed, we applied this method without significant results.

However, to have a general idea of the topics discussed, we can analyze the most common bigrams (Fig. 16). The general topic is the London marathon, but messages can refer to specific sub-aspects including for example: supporting runners, charity and fundraising, the death of Matt Campbell (a MasterChef contestant), the complains about the hottest day in London and the arrival of runners to the finish line. Indeed, we expect that the prevailing sentiment will be positive due to the general atmosphere of joy that characterizes this type of event while messages related to the death of a runner and to the "hottest day in London" are more likely to be the ones with a negative sentiment.

The main message is that, when you work with big data and in particular with text data, the quantitative analysis in not enough, but it is necessary to know the context and to provide a qualitative analysis.

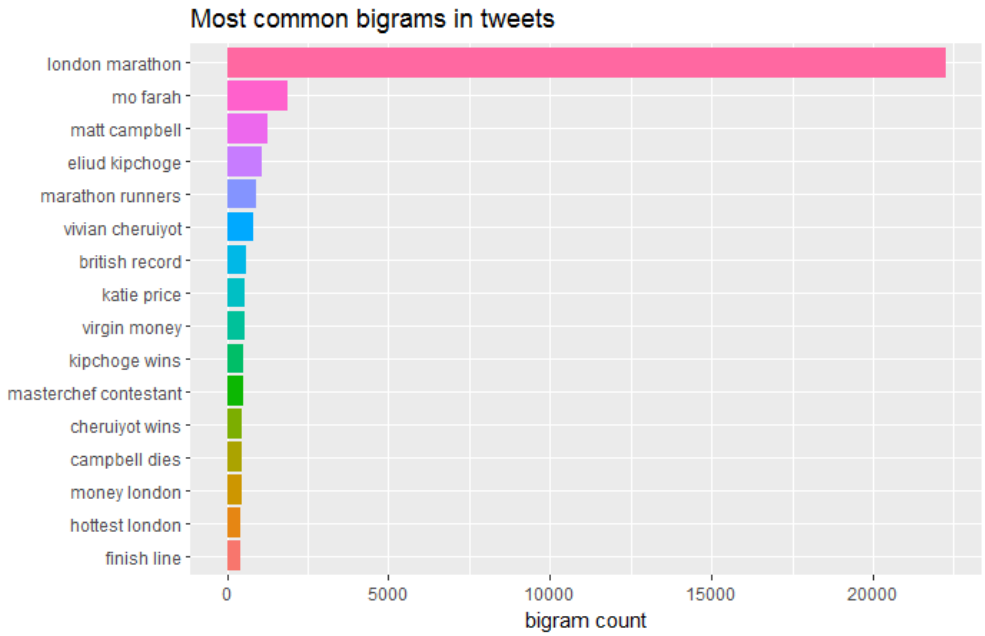

Figure 16. Most common bigrams in Tweets
Source: Authors' own elaboration

6.3. Coverage error

The *coverage error* concerns both the mismatch between the population of interest and the population available on Twitter, but also the over coverage due to the fact that the population is composed by accounts that are associated not only to people but also to organizations and to BOTs. Because of the presence of businesses and BOTs, the analysis of public opinion could contain noise or be biased.

Moreover, each user can have one or more account and each user can post more than one message from its account on the topic analyzed. Thus, each user can give different opinions on different sub-aspects that affect the sentiment. The one-to-many relationship between users and messages can be observed in Fig. 17.It is evident that users share more than one message and that the majority of messages have been posted by organizations or media companies/pages, such as: London Marathon, Virgin Money Giving, BBC and Athletics Weekly.

As a matter of fact, the 91,750 messages analyzed have been generated by 53,839 users. To understand if they are businesses or persons, we retrieved the information about their accounts. However, only data on only 44,469 accounts were available. This is because it is not possible to download data about user's private account.

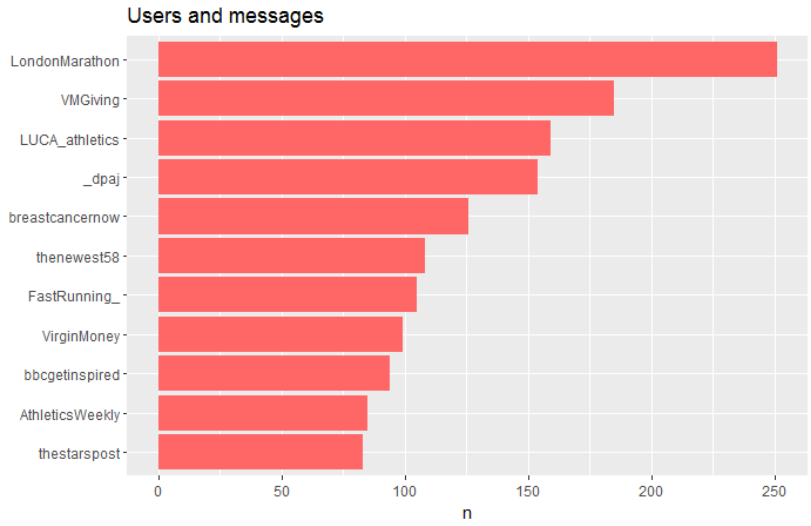

Figure 17. One-to-many relationship between users and messages
Source: Authors' own elaboration

If we consider the 41,514 messages classified from all the lexicons, only 30,712 messages were associated with public accounts. More in details, the latter have been generated by 25,286 accounts.

We tried to distinguish between people and businesses' account as well as between true accounts and BOTs. We focused on the 25,286 accounts described above. These numbers are more clear by looking at Fig. 21.

As far as businesses are concerned, we implemented text mining techniques to classify them. In particular, we retrieved user's information and we used the name and the description to check whether it is a person or an organization. In the first phase we read the data and we identify the common patterns that characterize a business. We also based our analysis on the charity organizationsthat rose funds during the event and we manually eliminated them. We labeled users as "businesses" when their name contains some specific words, such as: "*news, B&B, hotel, hostel, office, job, compass, care, healthcare, skincare, services, service, company, businesses,*

cancer, organization, society, foundation, foundation, charity, research, hospice, fundraising, hospital, hospitalcare, hospitality" etc.

Next, we also assessed the description of the user because the name analysis is not sufficient. Following the same procedure, we identified a number of common patterns, including: "*we help, we are the, we are specialist, we are founded by, we are reliant, we are raunchy, we are here to bring, we are unit, we are proud, we are fundraising, we are team, we are now open, we are delighted, we are building, we are available, we are one, we are academy, we are looking, we advise, we are revolutioning, we are in, we are part, we are dedicated, we are committed, we are working, we are here for, we are professional, we are medical, we provide, contact us, we support, our clients, our aim, our clients*" etc.

The pie graph (Fig. 18) shows the composition of the analyzed account population: the 9% of accounts are expected to be businesses (corresponding to 2,339 accounts) while the 91% are expected to represent people.

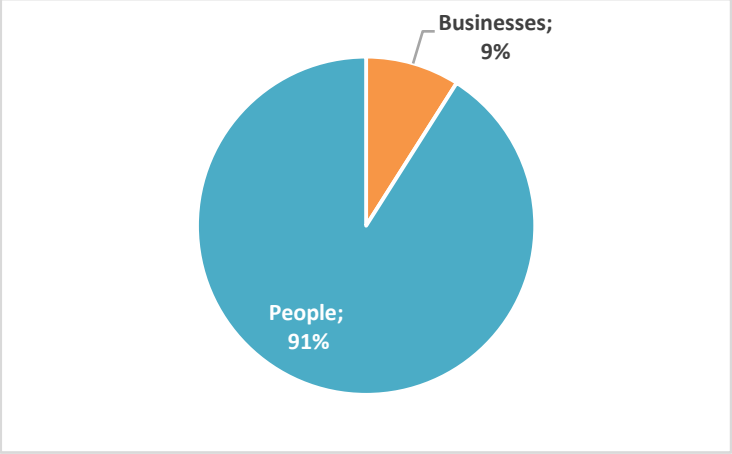

Figure 18. Businesses and people
Source: Authors' own elaboration

However, this analysis is not free from error because many accounts do not have a description and the user name is not enough to understand if it is a person or not. Moreover, also the list of words used to classify them could be incomplete.

The next step is to classify the accounts that are expected to represent people asBOTs and non-BOTs. This is done by using the "*botrnot*"[6] R package developed by Michael W. Kearney[7]. It uses a machine learning approach to classify Twitter accounts as BOTs or not. The algorithm can be implemented in two ways:

- Normal (default):it uses both users-level (bio, location, number of followers and friends, etc.) and Tweets-level (number of hashtags, mentions, capital letters, etc. in a user's most recent 100 Tweets) data to estimate the probability that users are BOT.
- FAST: This method uses only users-level data.

Since our database is large and the default algorithm can take days to download and evaluate each account, we used the FAST model. Moreover, the default model is correct 93.8% of the time,

[6] Available at: https://github.com/mkearney/botrnot
[7] Assistant Professor, School of Journalism, Informatics Institute, University of Missouri. Personal website: http://mikewk.com/

while the fast model is correct 91.9% of the time. The result of this procedure is that the 55% of accounts that are not businesses are likely to be real people (Fig. 19).

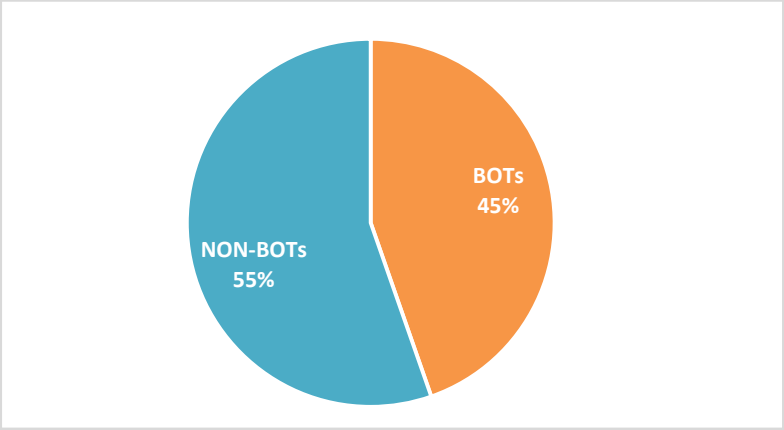

Figure 19. BOT and NOT-BOT
Source: Authors' own elaboration

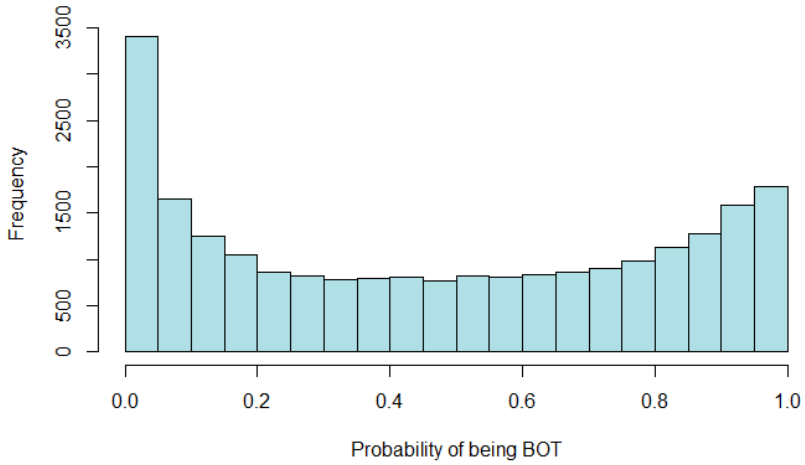

Figure 20. Histogram for the probability of being BOT
Source: Authors' own elaboration

Fig. 20 shows the histogram of the probability of being BOT. Looking at the graph, it is possible to conclude that there are almost 7,243 accounts that are most likely people (Prob< 0.2), 5,731 accounts that are most likely BOTs (Prob> 0.8), while the remaining accounts classification is uncertain, and they have a probability of being BOTs between 0.2 and 0.8.

The percentage of BOTs is quite high, and it would be interesting to compare these results with the default model which also consider Tweets-level data. Moreover, a deeper study on how the BOT's sentiment affects the sentiment analysis should be made because they could include "malicious sources" that try to influence people's sentiment.

Our findings are summarized in Fig. 21.

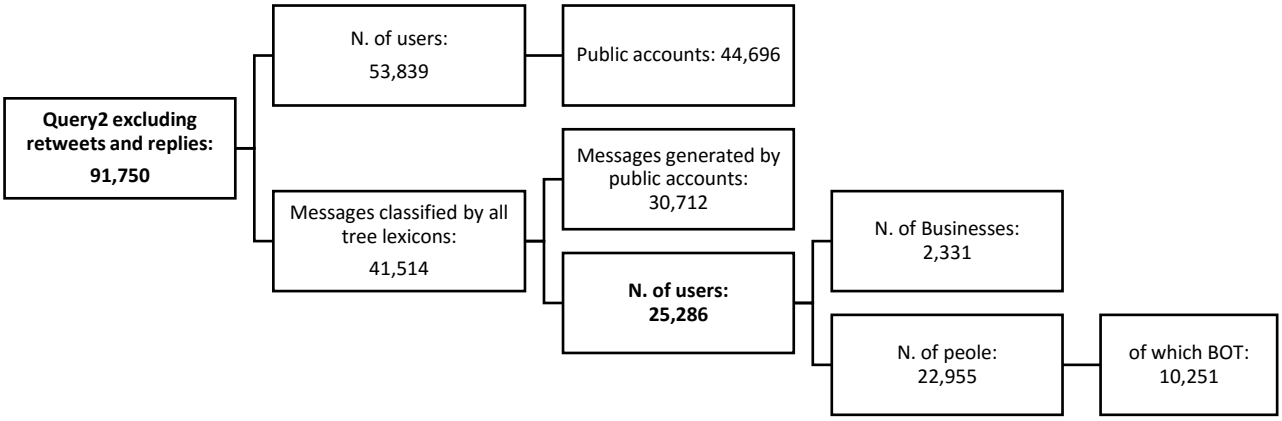

Figure 21. Structure of Accounts
Source: Authors' own elaboration

Another possibility is profiling users according to their location. Geolocalization profiling is an appealing information and would be useful for several interesting application. There are two types of geolocalization: a) the localization captured from the device: it registers actively the mobility of the device and it can be switched on or off from the user; b) the second type of localization is the geographical reference declared by the user in his account.The results of our study show that the geolocalization variable is far from being available for all the users; thus, quality need to be improved to implement statistical information in some specific applicative topics. Nevertheless, the geolocalized data look to provide some interesting information. With regards to geolocalization type a), considering the 91,750 Tweets retrieved according to Query2, only 866 Tweets (1%) are geotagged and have been generated by 635 users. If we plot the location on a map focusing only on London (Fig.22), we can observe that the distribution of points follows the marathon route and is more concentrated in the arrival area (around St James's Park).

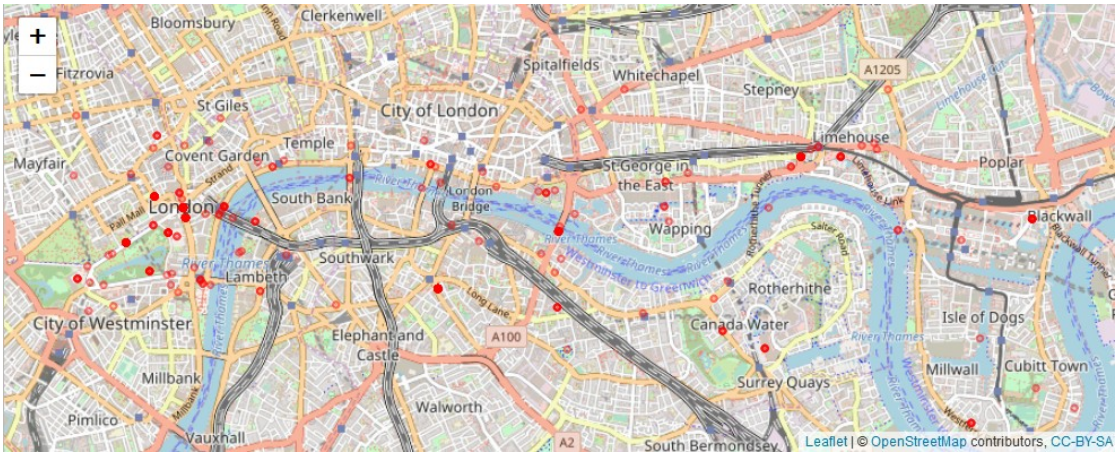

Figure 22. Geo-tagged Tweets in London
Source: Authors' own elaboration

Since the number of geotagged Tweetsis very low, we considered the geolocalization type b) and we retrieved the geographical coordinates of the users' declared location. Therefore, we used the *ggmap* package to geocode the locations. This process can be biased for different reasons. Users

can declare a false or "*fantasy*" location or they can declare more than one location at the same time. In the last two cases, the algorithm cannot find the coordinates. Moreover, before implementing it, it is necessary to clean the location eliminating all special characters. The errors depend also on the accuracy of the cleaning procedure. We retrieved the location of users previously classified as people (including BOTs) and we grouped the locations that were written in the same way. We obtained 7,679 unique locations (the same location can be written in different ways or languages). The algorithm was able to identify only 4,371 locations (56%).

The coordinates are plot in Fig. 23and Fig. 24.

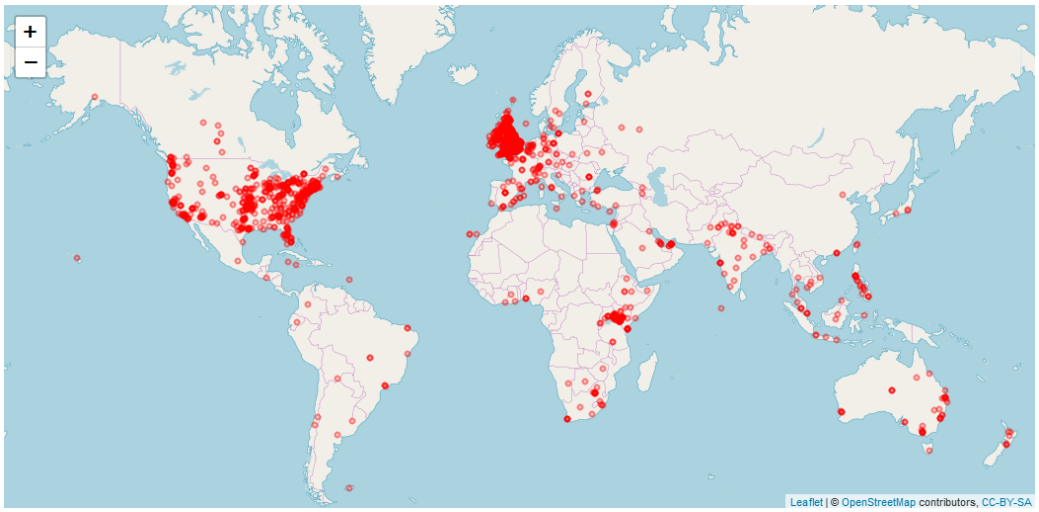

Figure 23. Users' declared location in the world
Source: Authors' own elaboration

As expected, the higher number of users is located in Anglophone countries and this depends on the fact that we retrieved only English-languageTweets. However, also European and Indian' people Tweeted on the marathon.

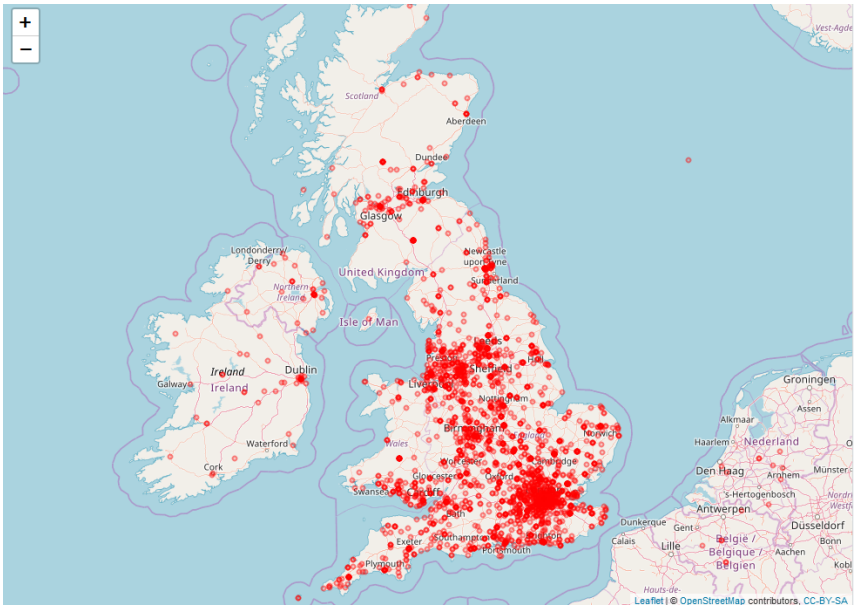

Figure 24. Users' declared location in UK and Ireland
Source: Authors' own elaboration

Finally, we can focus on the UK and observe that users are concentrated in the South-Central part of the nation, and in particular in the London, Liverpool and Manchester's surrounding.

7. Conclusions

The main message of this paper is that Big Data does not mean "*big information*". On the contrary extracting valuable and meaningful information is very hard. It is even more difficult to exactly quantify the amount of errors and the quality of the analysis.

In this sense, this paper contributes to the literature classifying possible sources of errors and quality for Twitter data.

Social media data can be an important source to monitor public events, to study citizens' sentiment and they can provide early-warning indicators to understand specific aspects of a phenomenon and take faster and suitable decisions. However, before they are actually used by policy makers, there are some issues to be solved. It is necessary to improve the quality and to develop quality and errors indicators to provide a trustworthiness measure to policymakers.

As far as the errors are concerned, the query error can be reduced formulating appropriate search string and, according to the type of analysis, including or excluding retweets and replies. Retweets should be deeply analyzed to decide whether including them totally or partially. Replies should be treated separately and for example, a network analysis could be useful to show the relationship between users and the length of the conversation. Some aspects of the interpretation errors will be solved in the next years. For example, improved and context-specific lexicons are expected to be produced. In the next future, a scenario could be the integration of Big Data and survey sources to draw up more sophisticated lexicons. Statistics Netherland made an attempt to integrate these two sources with the aim of creating a lexicon that fits the research purpose of identifying social tensions in Twitter messages[8]. In this paper, we suggest methods that can be used to evaluate the lexicons. One is considering a propensity indicator of the negativity/positivity of the lexicon expressed as the ratio between the negative and positive words. When evaluating the strength of association between scores computed by different lexicons, the correlation matrix and the Goodman and Kruskal's Gamma index of concordance can be used. Furthermore, to improve the sentiment analysis, lexicon-based and machine learning techniques could be integrated. Defining the coverage error and profiling users is the main issue. We argue that it is very important to distinguish between real people, businesses and BOT because each of them can be characterized by a different behavior on the web. For example, businesses' messages may try to attract people to their shops (or to make donations in case of fundraising) biasing the sentiment analysis. Also, BOTs that are "malicious" can significantly influence the people's sentiment. Thus, a deeper analysis is needed. Moreover, new studies to enrich the current literature are expected to be done to identify the sub-topic-aspects of the messages and to infer user's missing characteristics (Daas et al. 2016; Zhao et al. 2011).

The main message we want to share is that to make the sentiment analysis results trustable it is necessary to define the quality and the errors trough indicators; in doing this, it is fundamental to use a mixed method based on quantitative as well as on qualitative analysis.

[8]https://www.cbs.nl/en-gb/our-services/innovation/project/social-tension-indicator-based-on-social-media

References

Baldacci, E., Buono, D., Kapetanios, G. Krische, S., Marcellino, O., Mazzi, G., &Papailias, F. (2016). *Big Data and Macroeconomic Nowcasting: from data access to modelling*. Eurostat.

Crosby, P. B. (1988). *Quality is Free: The Art of Making Quality Certain*. New York: McGraw-Hill.

Daas, P. J., Burger, J., Le, Q., ten Bosch, O., & Puts, M. J. (2016). Profiling of Twitter users: a big data selectivity study. *Discussion paper 201606, Statistics Netherlands*.

Di Bella, E., Leporatti, L., & Maggino, F. (2018). Big data and social indicators: Actual trends and new perspectives. *Social Indicators Research*, 135(3), 869-878.

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., Gentry, M. J., RSQLite, S., & Artistic, R. L. (2016). *Package 'twitteR'*. R package version, 1(9).

Gilbert, C. H. 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). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf*.

Hsieh, Y. P., & Murphy, J. (2017). *Total Twitter Error*. Total Survey Error in Practice, 23-46.

Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., & Usher, A. et al. (2015). Big data in survey research: AAPOR task force report. *Public Opinion Quarterly*, *79*(4), 839-880.

Liu, B. (2015). *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press.

Maciej Beręsewicz, Risto Lehtonen, Fernando Reis, Loredana Di Consiglio, Martin Karlberg, (2018). An *overview of methods for treating selectivity in big data sources,* EUROSTAT, Luxembourg (*Forthcoming*)

Nielsen, F. Å. (2011). Afinn. *Richard PetersensPlads, Building*, *321*.

Pääkkönen, P., &Jokitulppo, J. (2017). Quality management architecture for social media data. *Journal of Big Data*, 4(1), 6.

Plutchik, R. (1980), A general psychevolutionary theory of emotion, *Academic press, New York, pp. 3–33.*

Silge, J., & Robinson, D. (2016). tidytext: Text mining and analysis using tidy data principles in R. *The Journal of Open Source Software*, 1(3).

Silge, J., & Robinson, D. (2017). *Text Mining with R: A tidy approach*. " O'Reilly Media, Inc.".

Sirkin, R. M. (2005). *Statistics for the social sciences*. Sage Publications.

Tourangeau, R., Rips, L. J., &Rasinski, K. (2000). The psychology of survey response. Cambridge University Press.

Wang, R. Y., & Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems* 12(4), pp 5–33.

Wayne SR. (1983). Quality control circle and companywide quality control. *QualProg*16(10):14–17

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E. P., Yan, H., & Li, X. (2011, April). Comparing twitter and traditional media using topic models. *In European Conference on Information Retrieval* (pp. 338-349). Springer, Berlin, Heidelberg.