Social indicators and Big Data: a case study on social interactions and active citizenship

**Keywords:** Big Data, Social Media, Sentiment Analysis, Network Analysis, Topic Modeling, Quality, Socio-economic Indicators, Active Citizenship.

# Introduction

The potentialities of Big Data new data source are relevant: They 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 [[1]](#unod). 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 socioeconomic phenomena. At international level, the discussion about social indicators and in particular quality of life, well-being and beyond-GDP activities is under constant debate. The measurement of the quality of life and well-being from an individual level perspective has become very important and social media represents a promising data source to study new topics [[2]](#due), [[3]](#tre), [[4]](#quattro). Within the ESS, the “Quality of life indicators framework” has been developed to measure the quality of life considering not only the GDP, but also other complementary and subjective aspects. From Big Data and, in particular from social media analysis we obtain dynamic indicators showing changes over time and the reaction of people to particular events. On the other hand, new issues 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” [[5]](#cinque).

The purpose of this paper is to use Twitter data to study social interactions and to provide an indicator of active citizenship. This is an on-going research, composed by two phases. The first one, which is already concluded [[6]](#sei), focuses on evaluating the overall quality of an analysis based on social media. 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 as well as the

 users’ behaviour. The second phase focuses on the development of an active citizenship indicator: “contact with politicians” based on the framework proposed by Sánchez et al. [[7]](#sette), [[8]](#otto). Then, we perform an in depth analysis to study the network relationship among users and topics discussed. More information are provided in the next section.

# Methods

We start the discussion with the methods used in the first phase which set the foundation for the second phase.

1. **First phase**

For the evaluation of quality, we apply the Twitter-TSE paradigm developed by Hsieh and Murphy to identify the three main sources of errors: the query error, the interpretation error and the coverage error [[9]](#nove). In summary, the query error depends upon the misspecification of the search queries, the interpretation error is due to the process of extracting insight from the text (sentiment, topic etc..) and the coverage error which represents the difference between the target population and the units available for analysis on Twitter [[9]](#nove), [[10]](#dieci).

We organized the first phase in three part:

***1. Data collection:*** data retrieval through the software RStudio. To do that, we used the *twitteR* package [[11]](#undici). We collect tweets about the London marathon for a time span of 10 days.

***2. Text mining and sentiment analysis:*** we compare the structure of three lexicons and we apply text mining techniques to distinguish between people, businesses and BOT. We also analyse the localization of people.

***3. Results and discussion:*** we compare the results obtained by using three different lexicons and discuss the query error, the interpretation error and the coverage error.

 In Fig 2. the framework of our analysis is presented in a graphical way.



Figure 2. Framework of the mixed approach adopted in our analysis

*Source*: Authors’ own elaboration

1. **Second phase**

The results obtained provided us the basis for moving toward the evaluation of a social indicator of active citizenship. The evaluation of the quality and the errors that can affect the analysis is very important with social media and, in general big data. Bias and a lot of noise is often present in this type of data. Initially, we follow the framework proposed by Sánchez et al. [[7]](#sette), [[8]](#otto) to develop this second phase, so we collected tweets about the messages sent by users to politicians and institution official accounts. The purpose is to compare these results with “contact to politicians” indicator developed by the European Social Survey. We are still analysing these tweets, thus we do not present the conclusive results here (they will be available before the conference in a paper), but a description of the general framework of out analyses follows:

***1. Data collection.*** It is made through the Twitter API and RStudio.

***2. Text mining, topic modelling and sentiment analysis.*** We create a specific and context-specific lexicon to reduce the interpretation error. We apply text mining techniques to distinguish between messages in which users request something and other types of messages. Then, we try to identify the topic of the message and we apply the sentiment analysis. As in the first phase, we distinguish between people and other types of users.

***3. Network analysis for topics and users*** [[12]](#dodici), [[13]](#tredici), [[14]](#quattordici), [[15]](#quindici).

***4. Results and discussion.*** In this part we assess the results obtained by comparing the official indicator with the social media one, we discuss about the network relationship and we try to evaluate the errors and the overall quality of the analysis.

# Results

The results in this section concern mainly the first phase. Comparing the sentiment analysis results with the information available on the press, we can conclude that they reflects the mood of the events occurred during the London Marathon. For example, the sentiment changes in relation of the collapse and the death of a runner. From this analysis we noticed that hashtags are mainly used by users to mark their messages as in relation to an event and mainly the day in which the event occurs rather than the previous and following days. The inclusion or exclusion of retweet and replies can significantly affect the analysis of public opinion if we pull together all the messages. Retweets and replies should be analysed separately. Moreover, we observed that the original messages (not retweet or replies) becomes more important and their relative frequency increases the day of the event. The interpretation error depends mainly upon the lexicon used and the text mining techniques applied. First, a good lexicon should rank the score according to the level of the word’s sentiment. Second, a lexicon can be constructed by integrating Big Data sources and surveys. 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. For what it concerns the coverage error, we found that the 9% of accounts belong to businesses (mainly charity organization that tried to rose funds) and among the accounts that should belong to real people, the 45% are likely to be BOTs. Geolocalization profiling is an appealing information. 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. Plotting the location of geo-tagged tweets, we noticed that the distribution of points follows the marathon route, while if we consider the localization declared in the account information, we saw that users are concentrated in the South-Central part of the nation, and in particular in the London, Liverpool and Manchester’s surrounding.

From the second phase we expect to obtain a social-media based indicator of “contacts with politicians” which cannot replace the official indicator but that can provide further information about the changes over time or that occur during particular events (elections, debates, enacting of legislation etc…). Moreover, through the topic modelling and network analysis we could understand the content of the conversation and the relationship between the users and the topics.

# Conclusions

The main message of the first phase 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 quantify exactly the amount of errors and the quality of the analysis. 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 analysed to decide whether including them totally or partially. Replies should be treated separately and network analysis is useful to show the relationship between users and the length of the conversation: this is what we are doing in the second phase. Some aspects of the interpretation error can be solved. For example, we develop an improved and context-specific lexicon that fit our research purpose. A scenario is the integration of Big Data and survey sources to draw up more sophisticated lexicons. From the results of the first phase, we can suggest some 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 has proved to be useful. Furthermore, to improve the sentiment analysis, lexicon-based and machine learning techniques can 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 behaviour on the web and we expect that this will be even more evident analysing the contact between users and politicians. 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 analysis results. Moreover, the identification of the topic and the sub-topic-aspects of the messages is very important. The main message we want to share is that to make the social-media based indicators 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. The second phase of the research is trying to develop a social-media based indicator, evaluating its reliability and main sources of errors and the overall quality of the analysis. A network analyss is implemented to understand the connections among topics and users.

# References

[1] 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.

[2] Shek, D. T., & Wu, F. K. (2018). The social indicators movement: Progress, paradigms, puzzles, promise and potential research directions. Social Indicators Research, 135(3), 975-990.

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

[4] 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.

[5] 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.

[6] Biffignandi, S., Bianchi, A., & Salvatore, C. Can Big Data provide good quality statistics? A case study on sentiment analysis on Twitter data. *International Total Survey Error Workshop 2018, Duke University - North Carolina.*

[7] Rosales Sánchez, C., Craglia, M., & Bregt, A. K. (2017). New data sources for social indicators: the case study of contacting politicians by Twitter. *International Journal of Digital Earth*, *10*(8), 829-845.

[8] Sánchez, C. R. (2018). International comparison of active citizenship by using Twitter data, the case of England and the Netherlands. *First Monday*, *23*(6).

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

[10] Beręsewicz M., Lehtonen R., Reis F., Di Consiglio L., Karlberg M., (2018). An *overview of methods for treating selectivity in big data sources,* EUROSTAT, Luxembourg.

[11] Gentry, J., Gentry, M. J., RSQLite, S., & Artistic, R. L. (2016). Package ‘twitteR’. R package version, 1(9).

[12] Ausserhofer, J., & Maireder, A. (2013). National politics on Twitter: Structures and topics of a networked public sphere. *Information, Communication & Society*, *16*(3), 291-314.

[13] Grandjean, M. (2016). A social network analysis of Twitter: Mapping the digital humanities community. *Cogent Arts & Humanities*, *3*(1), 1171458.

[14] O'Malley, A. J., & Onnela, J. P. (2014). Topics in social network analysis and network science. *arXiv preprint arXiv:1404.0067*.

[15] 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.