

A Text Mining Analysis on Big Data Extracted from Social Media

Gabriella Schoier(✉), Giuseppe Borruso, and Pietro Tossut

DEAMS – Department of Economic, Business, Mathematic and Statistical Sciences "Bruno de Finetti", University of Trieste, Tigor 22, 34100 Trieste, Italy
{gabriella.schoier,giuseppe.borruso}@deams.units.it

Abstract. The aim of this paper is to analyze data derived from Social Media. In our time people and devices constantly generate data. The network is generating location and other data that keeps services running and ready to use in every moment. This rapid development in the availability and access to data has induced the need for better analysis techniques to understand the various phenomena. We consider a Text Mining and a Sentiment Analysis of data extracted from Social Networks. The application regards a Text Mining Analysis and a Sentiment Analysis on Twitter, in particular on tweets regarding Coronavirus and SARS.

Keywords: Text Mining · Sentiment analysis · Big data · SARS · Coronavirus

1 Introduction

Nowadays a huge amount of data i.e. big data are collected and stored in several Data Warehouses by different public and private organizations. The analysis of big data is becoming more and more useful.

In this paper an analysis based on a Text Mining and Sentiment Analysis semi-automated approach is presented [5]. This approach is useful in several exploratory pattern-analysis, grouping, decision-making and machine-learning situations [13], including Data Mining, Web Mining and Spatial Data Mining(see e.g. [4, 12]).

The new Industry 4.0 paradigm, with digitalization, big data analytics, and so on, is heavily influencing different aspects of human being. Big data analytics could provide opportunities to develop new knowledge to reshape our understanding of different fields and to support decision making. Even though Internet has a great impact on information search behavior, several aspects of online user are not yet clear and need further investigation. Moreover there is a growing interest in utilizing user-generated data.

Computational aspects and the visual representation are becoming attractive tools in big data analysis. This paper is related to Social Network Analysis; in more details we focused our analysis on findings relevant terms and topics related to tweets on Coronavirus and SARS with the goal of identifying the trend of international opinions and preoccupations. In this analysis the *R* language has been used.

© Springer Nature Switzerland AG 2020
O. Gervasi et al. (Eds.): ICCSA 2020, LNCS 12252, pp. 351–364, 2020.
https://doi.org/10.1007/978-3-030-58811-3_25

2 Text Mining and Sentiment Analysis of Social Media

2.1 Big Data Characteristics and Methods of Analysis

In our age there has been an increase in the accessibility of data; this has created a real revolution in the organization of these data for instance those coming from social networks. The study of big data plays an important role not only in the field of computer science, but also in the socio-economic one, for this reason the extraction of information from them is very useful (see [8, 9, 11]).

A very important characteristic of big data is the size but it is not the only one as the rapid evolution of the phenomenon has highlighted other characteristics.

In 2001 Laney identified three dimensions: volume, variety, velocity, called the three V's [6]. In support of this Gartner defines big data with this expression: "Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making".

In addition to these big data have other features which are becoming important: veracity, visibility and value. In the majority of cases big data are presented in a heterogeneous, redundant and unstructured form so traditional tools do not allow a proper analysis of this type of data, a solution is given by Text Mining.

2.2 Text Mining and Sentiment Analysis

Text Mining aims to study methods and algorithms to automatically extract information from text and to classify documents on the base of the content. If one want to give a definition, one could say that Text Mining is "the discovery by a computer of new, previously unknown, information through the automatic extraction of different written documents" [1]. Text Mining is similar to Data Mining, even if it focuses on text, which is usually unstructured.

In recent years Text Mining has increased its importance due to the development of big data platforms and Deep Learning algorithms able to analyze enormous series of unstructured data. Text Mining is often used in conjunction with Text Analytics, or Text Analysis, so sometimes they are considered as synonyms. According to this approach, text data (keywords, concepts, verbs, names, adjectives, etc.) are derived from the text extraction process and are subsequently used in the Text Analytics phase to produce useful information.

Text Mining techniques try to find the thematic information hidden in a text to facilitate the process of archiving and building a logical map of knowledge. These techniques are based on algorithms that select the relevant parts of a document.

Among the most popular there are the categorization of texts, the extraction of information, the recovery of information, the processing of analyzing natural language, clustering, the summary of text and the Sentiment Analysis also known as Opinion Mining [3]. This last method is used to extract subjective information from the content. As the term suggests, it has to do with emotion and feeling. It is used to understand a subject's emotional response in a given context.

In recent years it has become an indispensable tool for companies that want to know the so-called *brand perception* taking advantage of the interaction of users of Social Networks or more in general of the web.

Since the information available on the Internet is constantly growing, it is very easy to access texts that express opinions in sites, forums, blogs and social media.



Fig. 1. Sentiment analysis: Emoticon [7]

The subjects that produce these 'judgements' are called opinion holders; they are authors of posts or reviews and users of social media. The expressed opinion, on a characteristic of an object, has an orientation indicating whether it is positive, negative or neutral. This orientation can also be defined as sentiment orientation (see Fig. 1).

In general, Sentiment Analysis is mainly structured on three levels:

- *document*: often known as document-level sentiment classification, it classifies the whole document investigating the positivity or negativity of the opinion;
- *phrase*: the analysis is focused on phrases to determine the feeling/sentiment of each of them, it requires more precision because it is necessary to identify objective, subjective and neutral phrases;
- *target*: consists in the analysis of feelings on entities and not necessarily on the whole sentence.

The analysis of opinion detection is based on the use of opinion words recognized by the machine and classified under the aspect of the positivity or negativity of the feeling.

A further classification problem occurs when there are not opinion words but emoticons that represent the feeling of the user. For these reasons, at this stage more than in others it is indispensable human work that is able to understand better the tone of a comment, to contextualize a message and recognize whether what one wanted to communicate was, for example, a real positive feeling or, on the contrary, sarcasm.

3 The Methodology and the Application

3.1 Social Network: The Case of Twitter

Social Networks, like any other tool have strengths and weaknesses depending on its the way they are used (see [14]). The advantages are many for example [7]: *global communication*, you can, with a click, communicate in real time with the rest of the world; *the convenience of use*, they are accessible by anyone just an email address or a phone number, and a password and one is immediately connected with virtual reality; *share and publish everything*, Social Networks allow everyone to express passions and thoughts so one can find a comparison with its contacts;. *social media marketing*, companies increasingly use them to communicate as they are incredibly effective for the online promotion of a large number of products, activities and services. Since social media are free their campaigns have absolutely competitive prices compared to traditional advertising.

Some defects of Social Network usage are: *alienation* if used recklessly, Social Networks can lead to estrangement from reality; d*ependence* can compromise work productivity; *privacy*: often the subject of discussion is whether there is privacy on social networks. The most important Social Networks provide the possibility to set up profiles in order to choose what you want to make public or keep private; but one must be careful and remember that sharing content on major social networks is on default public; *fakenews*: information and news circulating on Social Networks are not always true.

Twitter is a free Social Network platform or, to be more precise, is a very popular micro-blogging service that allows one to communicate with other users through the publication of tweets, i.e. short messages (the maximum number of characters allowed until 07/11/2017 was 140, now it is 280), photos or videos and all posts are freely readable. It is an excellent environment for social analysis, as there are around 220 million active users, out of 500 million subscribed users, creating over 500 million tweets every day.

The structure of Twitter is represented by two social groups: *followers* and *following*. The platform was born as a one-sided communication tool but over time the concept of conversation has been integrated. One can now mention or answer to a user through the use of the @ symbol before the user's name. When one reads a tweet published by one of the people one follows, one has three actions that allow to interact with that tweet: *answer:* clicking on reply can replicate that tweet, sending a mention to the author of the tweet; *retweet*: clicking on retweet gives the opportunity to share that tweet with all the followers; *like*: clicking on like, the tweet is put in the collection of the favorite tweets.

Through the Twitter search system one can search for tweets based on a single word that must be preceded by the symbol "#". This can be useful for searching for tweets about a single concept.

3.2 The Objective of the Analysis

In this period, a virus is creating great agitation and its spreads frightens millions of people: it is the Cornavirus [10] (see Fig. 2). The coronavirus belongs to a large family of viruses known to cause diseases ranging from common colds to more serious diseases such as Middle Eastern Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS) [2].

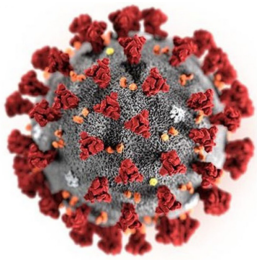

Fig. 2. Coronavirus (www.ilmessaggero.it/salute/focus/coronavirus_termini_usati_a_z_alfabeto-.html)

They are positive filament RNA viruses with similar appearance to a *corona* as seen using an electron microscope. The subfamily Orthocoronavirinae of the family Coronaviridae is classified in four genera of coronavirus (cov): Alpha-, Beta-, Delta-, and Gamma coronavirus. The genus of the betacoronavirus is further separated into five sub-genera (including Sarbecovirus).

The coronaviruses have been identified in the mid-1960s and are known to infect humans and some animals (including birds and mammals). The primary target cells are the epithelial cells of the respiratory and gastrointestinal tracts. Up to date, seven coronaviruses have been shown to be able to infect humans. The coronavirus (COV-19) is a new strain of coronavirus that has never been previously identified in humans. In particular, the one called COV-19, has never been identified before being reported in Wuhan, China, in December 2019.

On the contrary SAR means the Severe Acute Respiratory Syndrome, first detected in China at the end of 2002 (see [2]). In mid-2003, there has been a worldwide epidemic that has caused a lot of cases and deaths worldwide, including Canada and the United States. It is believed that the source of infection derived from animals infected with contact with a bat carrying the virus before being sold to a slaughter market.

SARS is caused by a coronavirus; it is much more serious than most other coronavirus infections which generally only cause flu-like symptoms. In addition, coronavirus may also cause Middle Eastern Respiratory Syndrome (MERS). SARS is transmitted from person to person through direct contact with infected persons or through droplets dispersed in the air by coughing or sneezing of an infected person [2]. SARS symptoms resemble those of other more common viral respiratory infections, but are more severe. They include fever, headache, chills and muscle pain, followed by dry coughing and sometimes difficulty in breathing. Most patients recovered in a week or

two. Some, however, have developed severe respiratory difficulties and in about 10% of cases death has occurred.

The first symptoms of COVID-19 and SARS are very similar and this has led to a medical glitch in the early stages of spreading the virus because the new disease has been cured as SARS, even if it was not.

It is interesting to check out what results the search of these two words will bring on Twitter.

3.3 The Preliminary Phase

One of the principal objective of Text Mining and Sentiment Analysis is the acquisition of information from the net and the extraction of the sentiment expressed.

In order to this the *R* language, that has made possible to extract tweets in which the keywords coronavirus and SARS appears, has been used. The preliminary step has been to register in the Social Network, to create an *Application Programming Interfaces* (API), to extract tweets and to create a *corpus*.

Twitter allows to discover in real time what users write about. One can access the social media via Web or via a mobile device. For companies, developers or users, an organized access to data is also available through the use of API. The method requires an authentication through the protocol "OAuth6" which, once authenticated, releases a token without expiration, to be used for all future connection requests.

An API is created in the developer section of the site. Access codes are divided into Consumer API keys (API key and API secret key), Access Token and Access Token Secret (Access Token and Access Token Secret).

In this way one is enabled to download tweets to *R*, using the twitter and RCurl packages. We downloaded 500 tweets in English each containing the word "coronavirus", the same operation has been carried out with the word SARS, the operation has been repeated in different days. Given the worrying situation in Italy regarding coronavirus, 500 tweets have been downloaded in Italian each containing the word "coronavirus", the same operation has been carried out with the word SARS, the operation has been repeated in different days.

At this point the data downloaded from Twitter have been transformed in a Data Frame. Before the representation through word clouds and other graphical methods and the application of sentiment Analysis, it has been necessary to use the *tm* package to create a *corpus*. At this step the *corpus* has been cleaned eliminating characters and words that are not of interest such as: punctuation, emoticon, adverbs, conjunctions and from the link to the tweet. At the end one gets a "clean" matrix.

3.4 Word Clouds and Other Graphical Analysis

After cleaning the texts one can proceed to the creation of the so-called clouds of words. As the name implies, Word Cloud is a visual representation of keywords used in a text. In general, it is similar to a list with the peculiar characteristic of assigning a font of larger dimensions to the words most cited by the users. For this operation the wordcloud function of the same package has been used; in addition, a maximum limit of 500 words has been chosen, a scale of 3 for the most relevant words up to 0.5 for the

least frequent. As far as regards the order the words more frequently are put in the middle.

The analysis have been carried out in different days. We reported the results referred to tweets of 10/02/2020 (See Figs. 3 and 4).

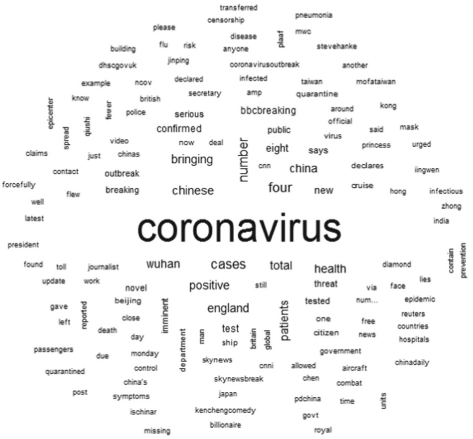

Fig. 3. Word cloud Corona virus 10-2-20 (English)

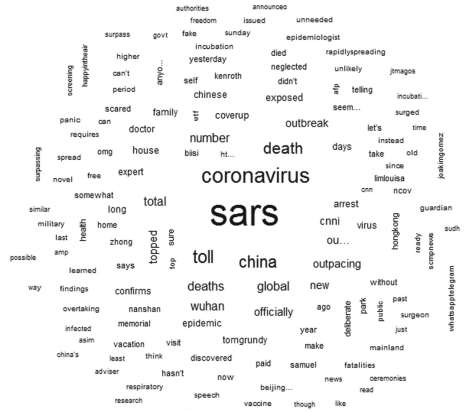

Fig. 4. Word cloud sars 10-2-20 (English)

As regards the graphical representation of the words with higher frequency the barplot command has been used. Only words with a frequency greater or equal to 5 have been represented. This choice was made taking into account the number of words that have been returned for each topic (See Figs. 5 and 6).

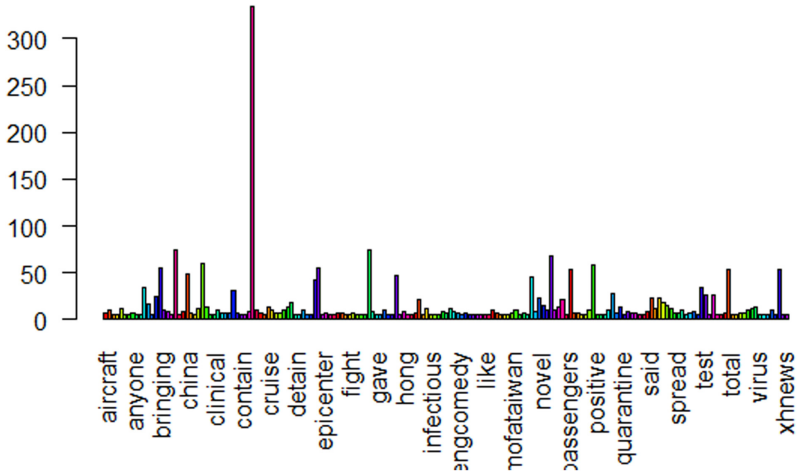

Fig. 5. Barplot Coronavirus 10-2-20 (English)

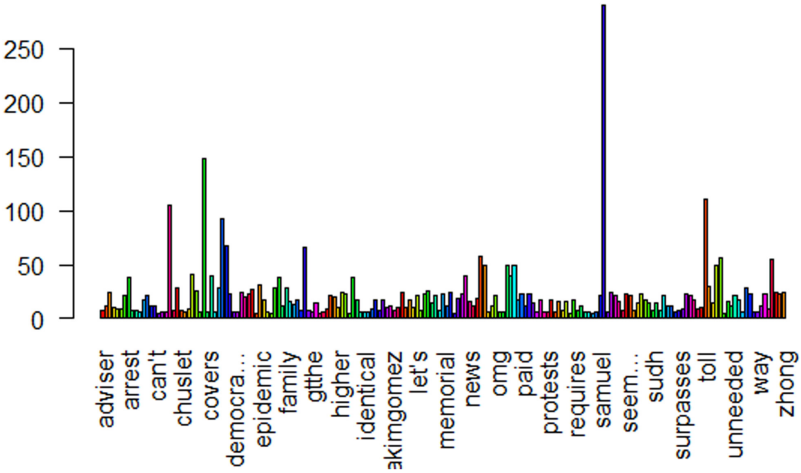

Fig. 6. Barplot SARS 10-2-20 (English)

At this point as the spread of Coronavirus in Italy has increased very much we have considered tweets both in English and in Italian. Data referred to the 11, March 2020. As one can see there are changes both in the word clouds and in the bar plots (See Figs. 7, 8, 9 and 10).

The graphical representation of the words with higher frequency has been proposed using the barplot. Only words with a frequency greater or equal to 5 have been represented. This choice was made taking into account the number of words that have been returned for each topic.

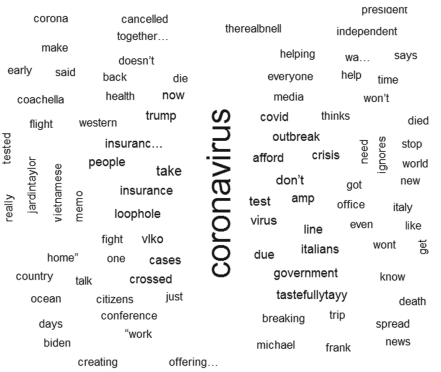

Fig. 7. Word cloud Coronavirus 11-3-20 (English)

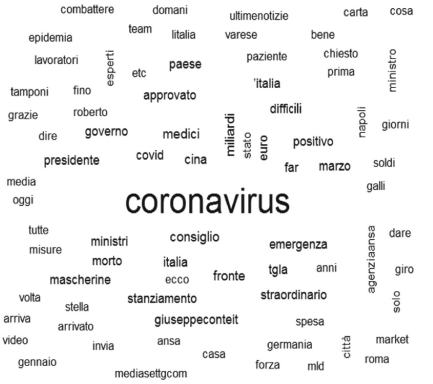

Fig. 8. Word cloud Coronavirus 11-3-20 (Italian)

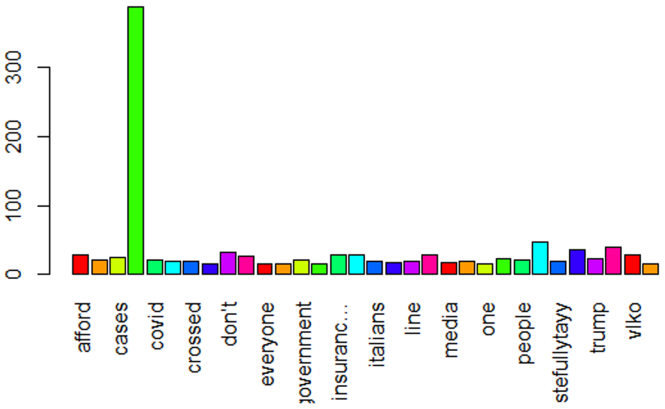

Fig. 9. Barplot Coronavirus 11-3-20 (English)

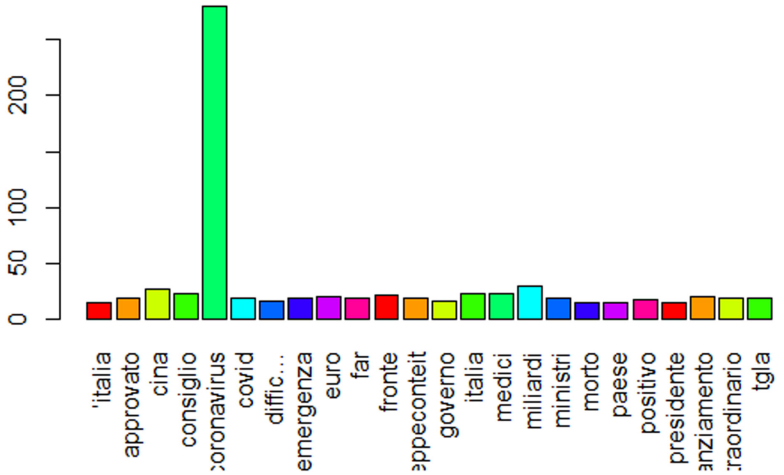

Fig. 10. Bar plot Coronavirus 11-3-20 (Italian)

3.5 A Sentiment Analysis

This phase is characterized by a first analysis carried out by the machine using *R* language and a second manual analysis.

We proceed to carry out an Opinion Analysis/Sentiment Analysis of the tweets obtained by classifying them in 8 categories: anger, anticipation (anticipation is considered as an emotion that causes different feelings thinking of unexpected events), disgust, fear, joy, sadness, surprise, trust; the following categories are provided by default by *R* and the tweets are classified according to the words used by the users.

To perform this classification, the command *get_nrc_sentiment* of the *syuzhet* package of *R* has been used and then the data obtained using the barplot command has been represented.

The reported results refer to di tweet of the 03/02/2020.

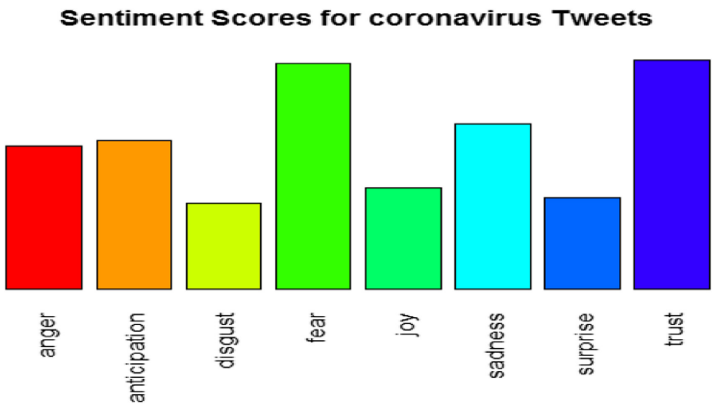

Fig. 11. Sentiment analysis Coronavirus 10-2-20 (English)

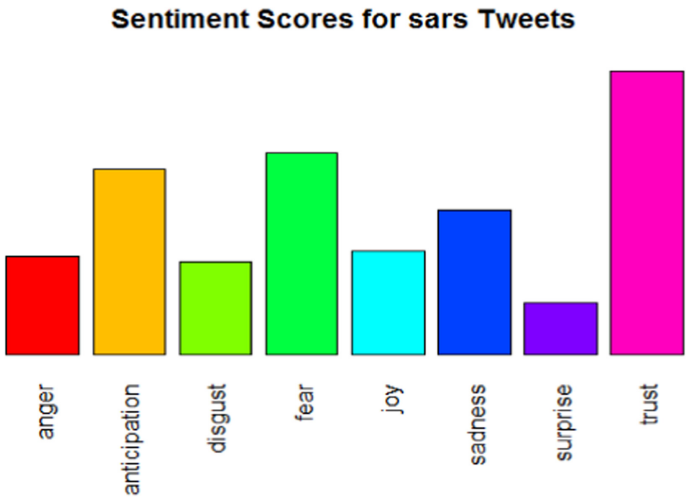

Fig. 12. Sentiment analysis Sars 10-2-20 (English)

In Figs. 11 and 12 you can see how tweets are divided by R into different groups according to the feeling that the function can determine based on the text.

Next Table 1 show the division in groups: in the first line there are the values related to coronavirus and in the second to SARS.

Table 1. Different groups for Coronavirus and SARS 10-2-20 (R analysis)

Anger	Anticipation	Disgust	Fear	Joy	Sadness	Surprise	Trust
47	72	31	89	40	51	27	82
40	76	30	88	35	45	24	80

As far as the manual analysis has been concerned, the tweets have been read one by one and classified them in three macro categories: positive, negative and non-concerning or non containing opinions or neutral. The data thus obtained have been summarized in Tables giving back, some examples for each of the categories and the retweets that are most frequently presented within the data frame obtained in the previous phases (Table 2).

Table 2. Different groups for Coronavirus and SARS 10-2-20 (manual analysis)

CORONAVIRUS 10-02-20			
	Positive	Neutral	Negative
	112	98	229
example	We appreciate India for extending support in China's battle against the coronavirus	Coronavirus: Turn off air-conditioners and open windows to reduce risk of being infected, say experts	Four further patients in England have tested positive for #coronavirus, bringing the total number of cases in the UK to eight
SARS 10-02-20			
	Positive	Neutral	Negative
	101	89	228
example	It's more deadly than SARS	Plagur Epidemics Crona 2020- Ebola 2014 Swine 2009-Bird 2005 SARS 2002	Doctor who exposed Sars cover-up under house arrest in China, family confirms

At this point as the spread of Coronavirus in Italy has increased very much we applied a Sentiment Analysis both to tweets in English language and in Italian language. Data referred to the 11, March 2020. The results are presented in Fig. 13 and in Fig. 14, as one can see the difficult situation is highlight much more than before.

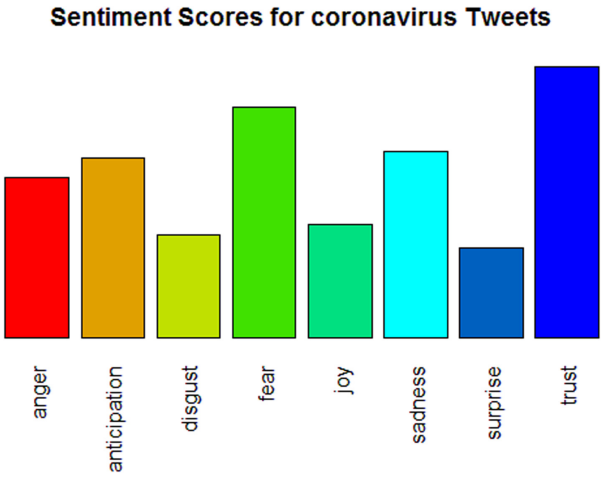

Fig. 13. Sentiment analysis Coronavirus 11-3-20 (English)

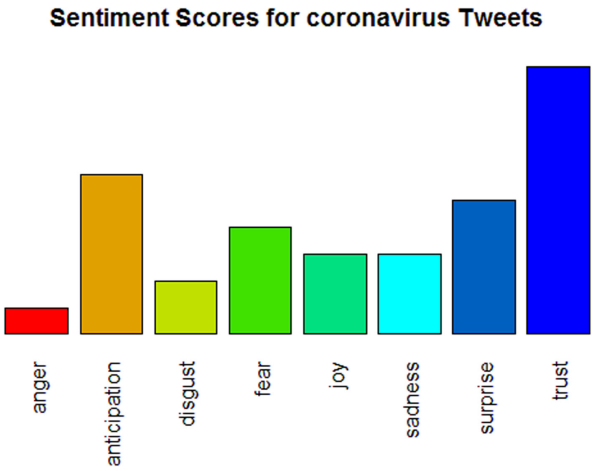

Fig. 14. Sentiment analysis Coronavirus 11-3-20 (Italian)

4 Conclusions

In this paper we have analyzed from a theoretical point of view definitions and properties of big data, we have also illustrated processes of extraction of information from the data. Processes of machine learning text mining and sentiment analysis have been analyzed in a specific area such as that of Social Network analysis. Finally, an example of these types of analysis has been given using the "R" software and its various packages in the context of Twitter. The results for the keywords coronavirus (COVID-19) and SARS have been analysed, highlighting how sentiment analysis can be useful for monitoring people's emotional situation.

References

1. Bolasco, S.: Statistica testuale e Text Mining: Alcuni Paradigmi Applicativi, Quaderni di Statistica, vol. 7 (2005)
2. Branda, L.: Sindrome respiratoria acuta (SARS) (2018). https://www.msdmanuals.com/it-it/casa/infezioni/virus-respiratori/sindrome-respiratoria-acuta-grave-SARS. Accessed March 2020
3. Inside Marketing pull information: Sentiment Analisis (2017). http://insidemarketing.it/glossario/definizione/sentiment-analysis/. Accessed March 2020
4. Jan, A.K.: Data clustering. 50 years beyond K-means. Pattern Recogn. Lett. **31**, 651–666 (2010)
5. Govoni, L.: Text mining: il processo di estrazione del testo (2019). https://lorenzogovoni.com/text-mining/. Accessed March 2020. Hearst
6. Laney, D.: 3D data management: controlling data volume, velocity and variety, application delivery strategies META GROUP (2001). https://blogs.gartner.com/. Accessed March 2020
7. Lucarini, L.: (2020). https://giacomolucarini.it/sentiment-analysis-ci-aiuta-lintelligenza-artificiale/. Accessed March 2020

8. Manhika, J., Chui, M., Brown, B., et al.: Big data: the next frontier for innovation, competition, and productivity. McKinsey Global Institute (2011). www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_datathe_next_frontier. Accessed February 2018
9. Mayer-Schonberger, V., Cukier, K.: Big Data: A Revolution That Will Transform How We Live, Work, and Think. Mariner Books, Boston (2013)
10. Ministero della salute: "FAQ - Nuovo Coronavirus COVID-19" (2020). http://www.salute.gov.it/portale/nuovocoronavirus/dettaglioFaqNuovoCoronavirus.jsp?lingua=italiano&id=228. Accessed March 2020
11. Piva, A.: Le 5 V dei Big Data: dal Volume al Valore (2019). https://blog.osservatori.net/it_it/le-5v-dei-big-data. Accessed March 2020
12. Schoier, G., Borruso, G.: A methodology for dealing with spatial big data. Int. J. Bus. Intell. Data Mining **12**(1), 1–13 (2017)
13. Software Testing Help: Data Mining Vs Machine Learning Vs Artificial Intelligence Vs Deep Learning (2019). http://www.intelligenzaartificiale.it/data-mining/. Accessed March 2020
14. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge University Press, New York (1994)