Abstract—In our society we are continually invested by a stream of information (opinions, preferences, comments, etc.). This shows how Twitter users react to news or events that they attend or take part in real time and with interest. In this context it becomes essential to have the appropriate tools in order to be able to analyze and extract data and information hidden in their large number of tweets. Social networks are a source of information with no rivals in terms of amount and variety of information that can be extracted from them. We propose an approach to analyze, with the help of automated tools, comments and opinions taken from social media in a real time environment. We developed a software system in R based on the Bayesian approach for text categorization. We aim of identifying sentiments expressed by the tweets posted on the Twitter social platform. The analysis of sentiment spread on social networks allows to identify the free thoughts, expressed authentically. In particular, we analyze the sentiments related to U.S President popularity by also visualizing tweets on a map. This allows to make an additional analysis of the real time reactions of people by associating the reaction of the single person who posted the tweet to his real time position in Unites States. In particular, we provide a visualization based on the geographical analysis of the sentiments of the users who posted the tweets.

Keywords—Twitter; Russiagate; sentiment analysis; U.S. President popularity.

I. INTRODUCTION

The advent of the Internet, the large amount of available data and the next document digitalization brought people to have an excessive quantity of daily information. Some problems raised in terms of storage, preservation, communication, compression, and extraction of information. It is easy to understand the content (or their opinions) of a few documents. Nevertheless, the text understanding is not an easy task when the available documents are too many, causing information overload also known as infobesity or infoxication. For these reasons, it is necessary to adopt automated tools to perform topics analysis favoring their understanding and the support to the decision making [1].

Digitalized sources have several characteristics such as heterogeneity and inconsistency in texts, which pose serious limits when trying to extract the meaning from sentences [2]–[4].

The real-time analyses support both companies in verifying the effectiveness of their marketing strategies and national or government agencies in evaluating political preferences and mood, or in supporting the earthquake and disease information diffusion. These analyses can also be applied to other similar real-time decision-making environments, such as crisis management and emergency medical assistance [5]–[7], and not only to make decisions but also to simply describe the emotions of people, like suffering and hope, in the disaster-hit areas [8]. These analyses are allowed by the diffusion of data mining and machine learning techniques [9]–[11] due to the new devices, new software [12], [13], networking [14], [15] and new technologies [16]–[20] for new challenges such as to predict health states [21]–[23] or to building a virtual assistant health coach [24].

In textual format data, the automatic recognition process plays an important role. Information Extraction and Knowledge Discovery are key research areas focused on data analysis coming from different points of view connected to sources’ heterogeneity [25]. For example, the information extraction in web data is complex for their dynamic nature [26], [27]. Here, there are changes for every new transaction and web usage, as well as for every field of applications. Therefore, one of the significant tasks for web applications is to collect users’ opinions from different
Among social media, Twitter emerged as the most popular micro-blogging platform, where information proliferates rapidly and posted information and actions cause instantaneous responses from users [42]. Several researchers are discussing how to overcome the limitations of microblog-texts to recognize relevant tweets and extract, annotate and exploit domain-independent, semantic information [25], [43], [44]. Some other researchers are investigating how to exploit machine learning techniques like deep learning [9] in various cases, e.g., in the case of companies that want to know their customers better and opportunistically manage the relationships with them, in the case of Arabic sentiment analysis on Twitter [45]. In the case of environments, such as in customer relationship management needed for the company to know their customer more closely [46]. In terms of Arabic sentiment analysis on Twitter [45] or recommending not only items of interest to the user, but also the conditions are enhancing user experiences with those items [47] are also investigated. Other researchers used naïve Bayes as a text classification algorithms to improve the accuracy of email and Twitter classification [48]. Twitter more than other social networks favors the sentiment analysis due to its “public” nature and offers some APIs \(^1\) that allow displaying the tweets of users, to get the followers and following, to search for tweets by content, etc. The message shortness due to the limited number of characters on Twitter eases the analysis. Authors in [49] highlight the reasons to choose Twitter.

In this paper, we developed a software able to analyze a short text (i.e., the tweet in our case) and to identify the sentiment expressed by the tweets posted on the Twitter social platform. According to the analysis we try to capture the orientation (e.g., judgment, thought, feeling or opinion) of the person who wrote the message. The software exploiting the language R is based on a Bayesian approach to classify the text. After a pre-processing (i.e., filtering of useless text and words), the developed software classifies the tweets expressing their sentiment (in positive, negative or neutral) and in an aggregated form. The software has been validated by identifying the minimum training set of data to be used to train the classifier. It is important to identify the minimum data set, which gives the best performance since the training phase is time-consuming. Then, the classifier has been used to classify real tweets on the work of U.S. President Donald Trump. The output accuracy (or error) of the software is evaluated and compared to the data provided by the Real Clear Politics website. It is worthy to note that this software allows accessing to the free opinions (by Twitter), expressed almost in real time to a large audience. Moreover, we investigate the possibility to analyze the expressed sentiment on a geographical basis, evaluating the position of the users who posted the tweets.

The paper is organized as follows. In section II, we describe the most popular methods to perform the sentiment analysis. In Section III, we provide the organization of the performed sentiment analysis. In Section IV, we describe the software developed to perform the analysis. In Section V, we apply the developed software to a real use case such as the popularity of the U.S. President. Finally, in Section VI conclusions are drawn.

II. MATERIAL AND METHOD

A. Related Works

Research on political communication has demonstrated that The role of the Internet as a source of political information produces the best sphere for public expression [50]. Due to its instantaneous response and its usage easiness, twitter more than other social media spread rapidly. These characteristics are ideal for promoting political viewpoints, particularly during contentious election campaigns. Political impact on social media has been examined through political discourses, and the political discourses on Twitter have been explored during the past several years by researchers. The emphasis on these studies was analyzing online networks of candidates and active users to forecast the results of the elections [51].

Different studies have investigated information behaviors of likeminded publics in the polarized political networks. Most of them focused on the strategic application of Twitter, highlighting the behaviors of the candidates while encouraging them to adopt Twitter with helpful tips for effective applications [52], [53]. Twitter can also be used to identify political preferences [54] and for day-by-day monitoring of electoral campaigns [55]–[57]. It is also possible to estimate the level of disaffection across people by counting negative tweets about politics in general, and this approach has shown that peaks in disaffection can correlate with important political news [58]. Other research studies explored how Twitter is applied within the electoral context to forecast electoral results [55], [59] by discovering

---

\(^1\) https://dev.twitter.com/
candidates’ patterns of political practice [60]–[62]. These investigations stressed the behaviors of the candidates, paying less attention to the behavior and flocking of the political public. In this context a Twitter sample may not be representative of the electorate, the general sentiment dictionaries used may not be optimal for politics, and replies to political messages may not be captured by keyword searches [63]. It follows that sentiment analysis needs to be sophisticated to make election predictions credible [63]. Twitter has also been used to study divisions within electorates [64] and by journalists to add direct quotes from politicians to stories. Moreover, Twitter sometimes helps to generate news in addition to reflecting it [65].

Miyoung Chong [42] filters tweets containing some hashtags to perform data analysis on a smaller set of data. He stresses the concept of hemophilia or the process by which people tend to create relationships with people with characteristics similar to theirs, such as, for example, gender, ethnicity, educational level, age, and social position. The filter function is used to improve the data analysis for all those textual components that are not useful for the classification activity, including the hashtags [25]. Moreover, the independence between the words of the analyzed tweets (bag of words) and the tweets themselves has been considered. Holly Thayer [66] proposes new characteristics of Twitter propaganda that create a model for determining if a message includes propagandist material. Instead, we offer a system that makes an ex-post analysis of the media effects of a given event, and therefore how this impacts for example on the popularity of the President of the United States Donald Trump.

The technique used by both Miyoung Chong and Holly Thayer is defined as cluster analysis and is one of the most used in the field of unsupervised machine learning [67]. In this work, we propose supervised learning where the proposed system preliminarily performs training receiving tweets labeled as either positive or negative. Thus, the system has learned and acquired knowledge to subsequent label data. The dataset of tweets in [42], [66] is much smaller than in our work. This is because it is restricted based on groupings as homogeneous as possible based on predefined parameters. Once the groups have been obtained, it is necessary to look inside them to see how the elements are similar in terms of the topics covered. In [42] the author uses a targeted and unique sample to investigate the distributive power of the political orientation towards Trump on the Twitter platform in a given day in the U.S. geographic scope. In this work we also extrapolated from the tweets information about their geolocation, it is therefore highlighted the possibility of obtaining the sentiment expressed at the level of geographical macro area (Federal States) an aspect not taken into consideration by [42], [66]. This last aspect can be very relevant especially in electoral propaganda in a country like the USA of a federal nature.

B. Sentiment Analysis Organization

In this paper, we developed a supervised sentiment analysis system able to detect and classify short texts from Twitter, based on the language R. We selected the R language due to its flexibility and its capacity in managing a large dataset [49]. Moreover, the R environment allows the computing of basic functions and complex mathematical operations and algorithms as well as statistical processing and graph generation. It has numerous functions and libraries dedicated to advanced mathematical evaluations and statistical analysis enabling the user to create new functions.

As for every supervised learning problem, the algorithm needs to be trained from labeled examples to generalize new data. We considered a dataset with 240,000 tweets: 2/3 of total tweets (i.e., 160,000) are used for training the system, while the rest 1/3 tweets (i.e., 80,000) are used for testing phase. The file used for the training is of type CSV and was downloaded from www.kaggle.com46. It contains 1.6M pre-classified tweets, but we restricted the analysis to a subset of the complete data, which contains some information in addition to the text of the tweet, such as the tweet identifier, the creation date or the user name writing the tweet. As said, considered tweets are pre-classified. It means that the classifier has learned the rules to assign a sentiment to texts and is therefore ready to apply these same rules to any other dataset. Each entry of the file is a tweet, and a value identifies its sentiment: from “0”, which means negative sentiment, up to “4”, which means positive sentiment. The loadFileData.R script is used to read the input file that contains the tweets. It reads and divides into two datasets tweets labeled as either positive or negative, both for the training phase and for the testing phase. In the following sections, we describe the four phases used to analyze data on sentiments: the Preprocessing and Data Filtering Phase (section C), Training Phase (section D), Classification Phase (section E), and Validation Phase (section F).

C. Preprocessing and Data Filtering

To reduce the computational errors, in the filtering phase it is necessary to reduce the number of features by eliminating all those textual components that are not useful for the classification activity. To this end, different types of characters, punctuation, brackets, and sequences in the tweets, not useful for classification, are intercepted through regular expressions. For example, hashtags (#topic), URLs (http ...), @user (i.e., a particular user mentioned in the tweet), RT (the abbreviation indicating a retweet) are identified and deleted. Moreover, emoticons searched and replaced. The emoticons that indicate positive feeling are replaced with the “posemotic” string while the emoticons that indicate negative sentiment are replaced by the “negemotic” string. In Fig. 1 examples of posemotic and negemotic emotications are reported.

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Posemotic</th>
<th>Negemotic</th>
</tr>
</thead>
<tbody>
<tr>
<td>🌞 ☀️ ⛅️ 🌦️</td>
<td>🌟 🌆 🌌 ☀️</td>
<td>🌞 ☀️ ⛅️ 🌦️</td>
</tr>
<tr>
<td>🌙 🌙 🌙 🌙</td>
<td>🌙 ☀️</td>
<td>🌙 🌙 🌙 🌙</td>
</tr>
</tbody>
</table>

Fig. 1 Example of relevant emoticons

Finally, the activity of reducing features has provided for identifying words that have the same letter repeated several times and then transformed, such as “i liiiike youuuu” becomes “i liiike youu.”
D. Training Phase

In supervised machine learning techniques, a training phase is needed to create a knowledge base, where the system can learn useful information to make the classifications on another data subset [68]. In this paper, we adopted a Bayesian classifier. We report it in Fig. 2 the sequence diagram of the training phase, in which tweets are filtered during the data filtering phase, and then two matrices called term-document matrix (tdm) are created: one for negative tweets and one for positive tweets. A document-term matrix or term-document matrix is a mathematical matrix that describes the frequency of terms that occur in a collection of documents.

![Fig. 2 Sequence Diagram of the Training phase](image)

In a tdm, rows correspond to terms in the documents and columns correspond to documents in the collection. Then, the element $E_{ij}$ of the tdm matrix is 1 if the term in row $i$ is present in the document $j$, and 0 otherwise. Tdm matrix gives us information about the frequency of a term within a set of documents. Finally, the tdm matrix allows performing some data filtering operations, such as the deletion of numeric characters, punctuation or stop-words and length control of terms.

E. Classification Phase

A Bayesian classifier is used to classify the tweets. It divides the tweets into its relative tokens, and for each one, it calculates the probability that the tweet expresses positive or negative sentiment. For each term, the classifier provides the probability that the tweet has a positive polarity or negative polarity. Table 1 shows some terms with the relative frequency of appearance in positive and negative tweets.

<table>
<thead>
<tr>
<th>Term</th>
<th>Positive Tweet Frequency</th>
<th>Negative Tweet Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bad</td>
<td>1434</td>
<td>5228</td>
</tr>
<tr>
<td>Day</td>
<td>14045</td>
<td>10955</td>
</tr>
<tr>
<td>Good</td>
<td>16162</td>
<td>7570</td>
</tr>
<tr>
<td>Just</td>
<td>15599</td>
<td>15738</td>
</tr>
<tr>
<td>Love</td>
<td>10939</td>
<td>4038</td>
</tr>
<tr>
<td>Miss</td>
<td>306</td>
<td>6943</td>
</tr>
</tbody>
</table>

In Fig. 3 an example of the output generated by the classify function is reported. Percentages and absolute values of classification are shown in addition to the elapsed time required for classification.

![Fig. 3 Example of output of the classify function](image)

F. Validation Phase

The validation phase aims to evaluate the goodness of the classifier in terms of accuracy and error. The accuracy is defined as the ratio between the numbers of corrected classified tweets concerning the total analyzed tweets. Of course, the error is the complement to it, i.e. error = 1 – accuracy. To carry out this validation, the training set size is increased, and it is checked how the classifier performance varies.

After each training, a test phase follows where 20,000 tweets are classified (10,000 positives and 10,000 negatives). It starts with a training set of 2,000 tweets (1,000 positive and 1,000 negative) up to 160,000 tweets (80,000 positive and 80,000 negative). It is necessary that the tweets used for the training are different from those used for the test. In Fig. 4 the results of the accuracy are shown according to the size of the training set. We report the accuracy for the positive and negative class as well as its mean. Results show that the accuracy tends asymptotically to the value of 71% after 40,000 tweets.
After evaluating the performance of the classifier on a pre-classified data set, we use the developed software on real data from the Twitter server. We used the “retweet” package to retrieve the tweets from the Twitter server. Moreover, it is necessary to create a new application on https://apps.twitter.com, defining its name, the “Consumer Key,” the “Consumer Secret Key” and some other information to use the search_tweets function to access the Twitter server. In this function, we set some parameters, such as the string to be searched among the tweets, the number of tweets to read and the language in which the tweets are written. Collected tweets are stored on the .csv file for the subsequent classification. The “tweets” object that is returned by the search_tweets function contains some useful information, such as the nickname of the user who wrote the tweet, the time the tweet was written, and the geographical coordinates of the device used to write the tweet.

H. Visualizing Tweets with Their Geolocalization in R

Starting from the coordinates of the device used to write the tweet, we can visualize the location of the user on a map through the script. In Fig. 5 we reported an example of tweet positioning in the U.S. map. In figure 1, we also reported the analysis of the tweets by showing if the tweet is positive in green or negative in red. It is worthy to note that not all tweets contain geolocalization information. In Fig. 5, we analyzed 10,000 tweets, but we reported only 8%, containing device coordinates.

In addition to displaying the coordinates of the tweets on a map, we can go to retrieve the country of the coordinates of the tweet. The information on the country from where the tweet has been sent can be useful to evaluate the sentiment analysis about one topic (e.g., the popularity of the U.S. President in our case) based on the position of the user to be more effective in analyzing real-time people reactions. To this aim, we used the “revgeo” library enabling us to recover the country of the tweet’s user by his lat-lon.

III. RESULTS AND DISCUSSION

In a society strongly driven by instinctive choices (i.e., not rational), it becomes increasingly important for politicians to know how they can gain more votes from potential people. Sentiment analysis tools can analyze in real time the reactions to certain debates and discussions, going to identify what are the issues that are most interesting for the population. In this section, we analyze a real use case using the developed Bayesian classifier. We investigate the general sentiment of some Twitter users about the work of the President of the United States Donald Trump, trying to carry out some indications on it.

A. Reference Poll

To evaluate the performance of the classifier, it is necessary to have statistical data for comparison. Since we have no survey, which refers to people’s feelings about the U.S. President work, we use statistical surveys on the work of President Trump in a given period. On the Internet, it is easy to find surveys of this type. Thus, we discovered tweets that contain the word “trump”, “Trump” or “TRUMP.” Then, we submitted them to the classifier. Finally, we compared the surveys published on the website www.realclearpolitics.com. On Real Clear Politics there are surveys and statistics related to the U.S. political world. We referred to the period from 15th to 31st August 2018, when 42.2% of the interviewees (on average) approved the work of Trump while 54.1% disapproved it.

B. Classification

We adopted the script miningTweet.R to get a certain number of tweets containing the selected strings. We retrieved about 15,000 tweets per day for 5 days. Then, we fed the classifier with these tweets and the corresponding output (i.e. the behavior of positive, negative or neutral sentiment) in the different days is reported in Fig. 6. Comparing the output data of the classifier with those provided by RealClear Politics website shows that the percentages evaluated by the classifier in terms of approval and disapproval are similar to the results of the website. Fig. 6 shows that the output of the classifier provides 43.2% of positive sentiment, 47.1% of negative sentiment and about 9% of neutral sentiment. Further efforts can be done on tweets expressing neutral sentiment in order to have the same percentages.

2 https://www.realclearpolitics.com/epolls/other/president_trump_job_approval-6179.html
Fig. 6 collected tweets in five days with positive, negative or neutral sentiment

C. Visualization and Geographic Localization

In this section, we report some examples showing the expressed sentiment on a map despite the low percentage of information regarding the location of tweets (lower than 10%). We used the ggmap library for these plots, through the get.map function to retrieve the map and geom_point function to draw the points on the retrieved map Fig. 7.

This library allows the selection of a single Country but also of a particular city or an area to perform analyses on narrow geographical areas. In Fig. 8, we report an example of zooming the U.S. west coast, showing the geolocalized tweets and the related associated classification, as positive (in green) and negative (in red) ones.

Based on the information on the latitude and longitude of the tweet, it is possible to identify the country from where the tweet has been sent. We have aggregated the results by country and analyzed the difference between positive and negative geolocalized tweets on a country base.

In Fig. 9, we show the results of the sentiment distribution organized per country. Results are normalized to the maximum value given by country aggregate process. It is reported a negative (positive) sentiment if most of tweets of that state are negative (positive).
IV. CONCLUSIONS

Social networks are a source of information with no rivals in terms of amount and variety of information that can be extracted from them. In particular, in this paper, we focus the attention on the possibility of identifying and collecting the free thoughts and opinions expressed by the people through social media. In this way, we can give as feedback to common people or to whom it may concern, e.g., politicians or technicians, the opportunity to read and interpret the average thought on a given topic. Here, we propose an approach to analyze, with the help of automated tools, comments, and opinions taken from social media in a real-time environment. The tool identifies sentiments in terms of positive, negative, neutral polarity expressed by the tweets posted on the U.S. President popularity. We perform the training phase with 40,000 tweets, and the classifier provides an average accuracy of about 70%. Then, the classifier has been used to classify real tweets on the work of U.S. President Donald Trump. The accuracy of the tool is similar to data provided by the Real Clear Politics website. The slight deviation (about 9% of neutral tweets evaluated by the classifier respect to about 4% of the Real Clear Politics website) is due to the reference period, which is not the same but successive of some days. Finally, we provided a visualization based on the geographical analysis of the sentiments of the users who posted the tweets.

REFERENCES


