MACHINE LEARNING AND TEXT MINING TO CLASSIFY TWEETS ON A POLITICAL LEADER

Agostino Di Ciaccio, Giovanni Maria Giorgi

1. Introduction

The Social Network Twitter was created in 2006, but it has had a slow expansion in Italy starting from 2009. Twitter is now very popular and counts 255 million users, becoming the social media most used by public personalities, showmen, politicians. In Twitter, each user handles his own personal page that can be updated via text messages, with a maximum length of 140 characters, known as "*tweets*". Anyway, the user can add links to pictures, videos, or other documents. The limit on the length of each tweet is, at the same time, the strength and weakness of this social network: with 140 characters you cannot develop a speech, but you can write a sentence quickly using a smartphone.

Let us recall some of the unique aspects of this social network. A Twitter user can choose to follow another user (becoming a "*follower*"), automatically getting the communication of all his/her messages. A message may be written independently, or may be in response to someone else's tweet (i.e. it is a "*reply"*). A "*retweet*" is a message promoting in the community a message of another user without altering it in any way, stressing we fully agree with it. The *hashtags* are keywords provided by the user in the tweets; a *fake user* is, usually, a humoristic duplicate of a celebrity, finally an "*influencer*" is someone who has a large number of followers (cf. Bentivegna, 2014).

A key feature of Twitter is that it is an open system, where everyone can read the tweets of other users and participate in a discussion. Many public figures, particularly politicians and showmen, have a Twitter account and anyone can write to them directly (but it is unlikely to receive a response). Therefore, Twitter is an important showcase and an inexpensive way to communicate instantly with other users of the social network, bypassing the traditional media (TV, newspapers, radio).

In the 2014 European elections, 92% of the Italian candidates had a Twitter account. In this paper, we will see how to analyze Twitter to get the sentiment towards a political figure and describe the community connected to him, although having to handle millions of tweets.

2. Political leaders on Twitter

It is interesting to note that politicians who have the highest number of Twitter followers in the world are Obama (43 million followers), followed by the Presidents of Turkey, Argentina, Colombia, Mexico, Brazil and the Queen of Jordan. The most followed politicians in Italy are Beppe Grillo (1.48 Mln followers) and Matteo Renzi (1.15 Mln followers). Then we have, by number of followers, Vendola, Bersani, Letta, Monti, Boldrini, De Magistris, Alfano. Silvio Berlusconi is not on the list because, after an initial presence on Twitter, he decided to pull out. The number of followers changes continuously, increasing or decreasing, but this number alone is not of great interest. We must consider the fact that followers are not necessarily users who share the opinions of the politician they are following and several followers could be no longer active. It is also not true that being popular on Twitter involves being popular in the country: Vendola, with 421,000 followers, should be the most popular politician following Grillo and Renzi, but this is not true.

The analyses carried out on the tweets of politicians generally use retweets, hashtags and mentions. A user who retweets a message of a politician necessarily agree with it, hence analyzing the retweets of messages we can measure the popularity of a politician. Some hashtags are of particular relevance in the political debate: *#lavoltabuona, #sfiduciamorenzie, #vinciamonoi, #vinciamopoi, #M5S* are some examples observed in the period March-May 2014 in Italy. Analyzing the popularity of hashtags can help the evaluation of political opinions; in fact, the hashtags usually can be politically labelled. The analysis of mentions of a politician is the easiest, but also coarse, way of assessing his/her popularity. Indeed, mentions and replays do not express a clear sentiment towards the politician; thus, in order to define the opinions we need to analyze the text of the tweets.

3. The information that we can get from Twitter

If we are interested in how the network judges a politician, the basic information would be the classification of the tweets as positive, negative or neutral. Of course, an expert could classify manually the tweets reading the texts, eventually discarding some tweets (ambiguous, or linked to other documents or simply jokes). If we have hundreds of thousands of tweets, this approach is clearly unfeasible and it is necessary to look for an automatic procedure or give up the classification (the last is the most common approach).

Our analysis has focused on the tweets, written between March and May 2014, which contained the name of the premier *Renzi* or the username *@matteorenzi*. We have collected, during these three months, 1,290,965 tweets, written by 136.967 users, of which 602,663 are retweets. Overall, 72% of the users wrote no more than one message per month, while the more active users, with more than 100 messages in three months, represent only 1.5% of the users. The first large group wrote 11% of the tweets, while the small group of hyperactive users wrote as many as 43% of tweets. Each individual in a group of 56 users wrote more than 1,000 tweets in this period. This consideration should make us reflect on the difference between the sentiment of the tweets and the sentiment of the users.

Table 1 – *Most retweeted users in the period, with the number of retweets*

User	Description	Number
Matteo Renzi	Premier	64262
Matteo Salvini	Secretary of Lega Nord	8352
Gianni Kuperlo	Fake user, close to M5S	6061
La Repubblica	La Repubblica	5060
CorrieredellaSera	Corriere della Sera	5049
Andrea Scanzi	Journalist of Il Fatto Quotidiano	4987
Il Fatto Quotidiano	Il Fatto Quotidiano	4952
Francesco Manna	Blogger of Il Fatto Quotidiano	4839
Sky TG24	SKY TG 24	3893
Franco Maria Fontana	Intellectual and writer	3749
Patrizia Fiori	Quota96Scuola	3669
Fratelli d'Italia-AN	Fratelli d'Italia - Alleanza Nazionale	3622
ABATE FARIA	Blogger close to M5S	3527
Spinoza	Satirical Blog (S. Andreoli & A. Bonino)	3307
Partito Democratico	Partito Democratico	3242

What are the most retweeted users in this data? The list of the first 15 users is shown in Table 1. We also show the main hashtags and mentions in Figures 1 and 2. To make the figures readable, however, we removed "*matteorenzi*" from the hashtags, *Renzi* and *Quota96Scuola* from the mentions (*Quota96Scuola* refers to 4000 teachers who claim the right to retire). However, the number of hashtags and mentions is not very informative and of ambiguous interpretation. Therefore, when analyzing these data we should ask what the most interesting goal is.

In this paper, we identified as our main objective the understanding of the network structure of users who express opinions on a politician, identifying influencers and the relationships that bind them to each other, identifying sub-networks characterized by a particular sentiment. To achieve this goal, we must be able to classify the collected tweets with respect to the sentiment, positive or negative, on the politician. If we have more than one million of tweets, as is our case, we face a complex problem. This is the reason why all the analyses that appear in the newspapers are based on hashtags or mentions, that can be analyzed with much more ease. We must interpret the sentiment expressed by users and influencers, analyzing the text of the tweets, also taking into account that some users, such as political parties and information agencies, are very special users.

Figure 1. *The most frequent hashtags*

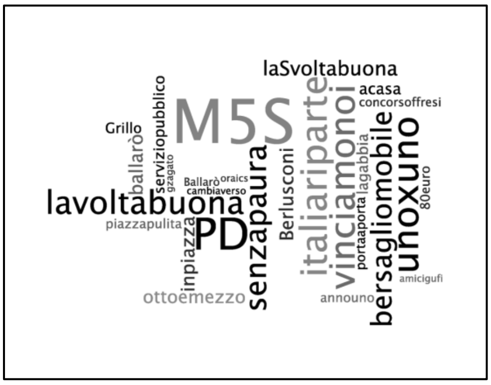
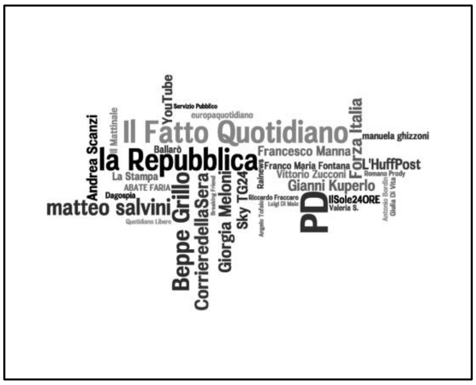


Figure 2. *The most frequent mentions*



4. Mining the sentiment from 1,200,000 tweets

The data collected consist in 1,290,965 tweets written by 136.967 users from March to May 2014. For each tweet, we recorded several information about the user who wrote it and who retweeted, if any.

We created a procedure to mine the sentiment, which requires the following steps:

1. For each month, the 500 most retweeted messages were classified manually, for a total of about 1500 distinct retweets.
2. Taking into account that each of these posts had a high frequency in the archive, the first step allowed to classify 111.490 tweets of the archive. These tweets were written (more precisely retweeted) by 38.694 different users.
3. Successively, we identified all the posts, in the three months period, written from these 38.694 users, achieving 775.686 tweets.
4. We then assumed that a user approves the sentiment of the message he is retweeting (which seems obvious), and that all his posts, at least in the short term, maintain the coherence of sentiment shown in the retweet. In this way, we were able to assign the sentiments to all 775,686 tweets.
5. Eliminating some contradictory assignments, we finally got 769,982 tweets classified and 520,983 yet to be classified.

6. To manage the unclassified tweets, we built a classification model using the archive of 769,982 messages as the data sets for training and validation. The analysis can use the typical tools of text mining (cf. Applied Analytics using SAS Enterprise Miner, 2011) and a suitable classification model. In the model choice step, the policy was not to choose the model with the lowest expected classification error. Conversely, we looked for a model that was able to classify with high probability a good percentage of the data. In our data, classification trees have proven to be the most effective. In particular, we set the parameters of the tree in order to have at least 30% of the tweets with a very high probability of correct classification. To estimate the model, we used the text of the tweet and some quantitative variables that describe the user's profile.
7. The classification model, estimated in the previous step, was then applied to the remaining 520,983 messages, identifying messages with higher probability of classification (> 0.95). In this way, we were able to classify 178,243 tweets.
8. The tweets classified by the model, were written by 59,215 different users. As done in step 3, all the messages of these users were identified and classified, for a total of 377,417 tweets.
9. The tweets classified in point 8 were joined with those already classified in point 5, for a total of 1,147,399 tweets. The remaining tweets (11%) were discarded.

It is possible to make improvements that lead to change some of the above steps. It is usual, for example, that we know a priori the sentiment of some specific users (e.g. political parties or party newspapers). Another improvement consists in assigning scores to tweets or users, i.e. a non-binary value that expresses the intensity of the sentiment (e.g. an insult is a more negative sentiment than a criticism). In this way, through the evaluation of a number of scored messages, we could obtain a more reliable estimate of the user’s sentiment. These corrections can lead to improve classification accuracy especially for users with many messages.

5. Analysis of the Network on Renzi’s Tweets

At the end of the analyses carried out in the previous paragraph, we have available a large archive of tweets classified with respect to the sentiment. In this archive, all tweets speak, good or bad, about the premier Renzi. These data allow the analysis of the relations among users in the observed community, taking account of political opinions.

Figure 3 shows a simplified view of the network corresponding to our data. To analyze the relationships between the users, we considered only the retweets, as they represent directed links among the users. More precisely, we considered all retweets that we were able to classify, i.e. 596,413 (97.3% of retweets), corresponding to 64,783 users. Since we could not represent a network with all of these users, we selected the most relevant nodes, showing the *influencers* and *assiduous followers*, defined in this way:

- The *influencers* are users who have been retweeted at least 700 times, during the observed period.
- An *assiduous follower* is a user who retweeted at least 25 times a specific influencer, during the observed period.

An influencer who has not assiduous followers is excluded from the graph. This is the case, for example, for "*Spinoza*" which is a satirical blog with many retweets but that has not assiduous followers and therefore does not appear on the graph.

The influencers (and his followers) who have expressed mainly positive opinions are represented by a gray square, conversely negative opinions are represented by black circles. The triangles represent the information agencies to which we have not assigned a sentiment; however, their position on the graph could be interpreted as an implicit political opinion. In the lower right, we see two sets of white diamonds: they correspond to two groups of people who are claiming certain rights and are addressing the current prime minister; the tweets in this case represent a form of pressure and do not express a clear political opinion.

The polygon size of the influencers represents the number of corresponding retweets. *Renzi* has the largest square with 64,262 retweets. The size of the followers shows how many messages they have retweeted. For all the influencers we reported their name, while the name of the followers is shown only if the number of retweets is large (>40). Being the tweets addressed to a prominent political figure, as expected, the network shows a political characterization due to the major parties and movements. In figure 3 we can easily identify sub-networks for the main political groups: *PD*, *M5S*, *Forza Italia*, *Lega Nord*, *Fratelli d'Italia*. Overall, 73% of tweets criticizes Renzi, while only 27% supports him.

The sub-network that refer to the political opinion of M5S (on the right) is broad and diversified and includes many bloggers (someone satirical), the *M5S* spokespersons and some news agencies/blogs. The *PD* sub-network is smaller, with a hierarchical structure and three fundamental references: *Matteo Renzi*, *YouDem TV* and the official account of *PD*. *Europaquotidiano* and *La Repubblica* are the closest news agencies. *Lega Nord* and *Fratelli D'Italia* are two sub-networks very active but isolated.

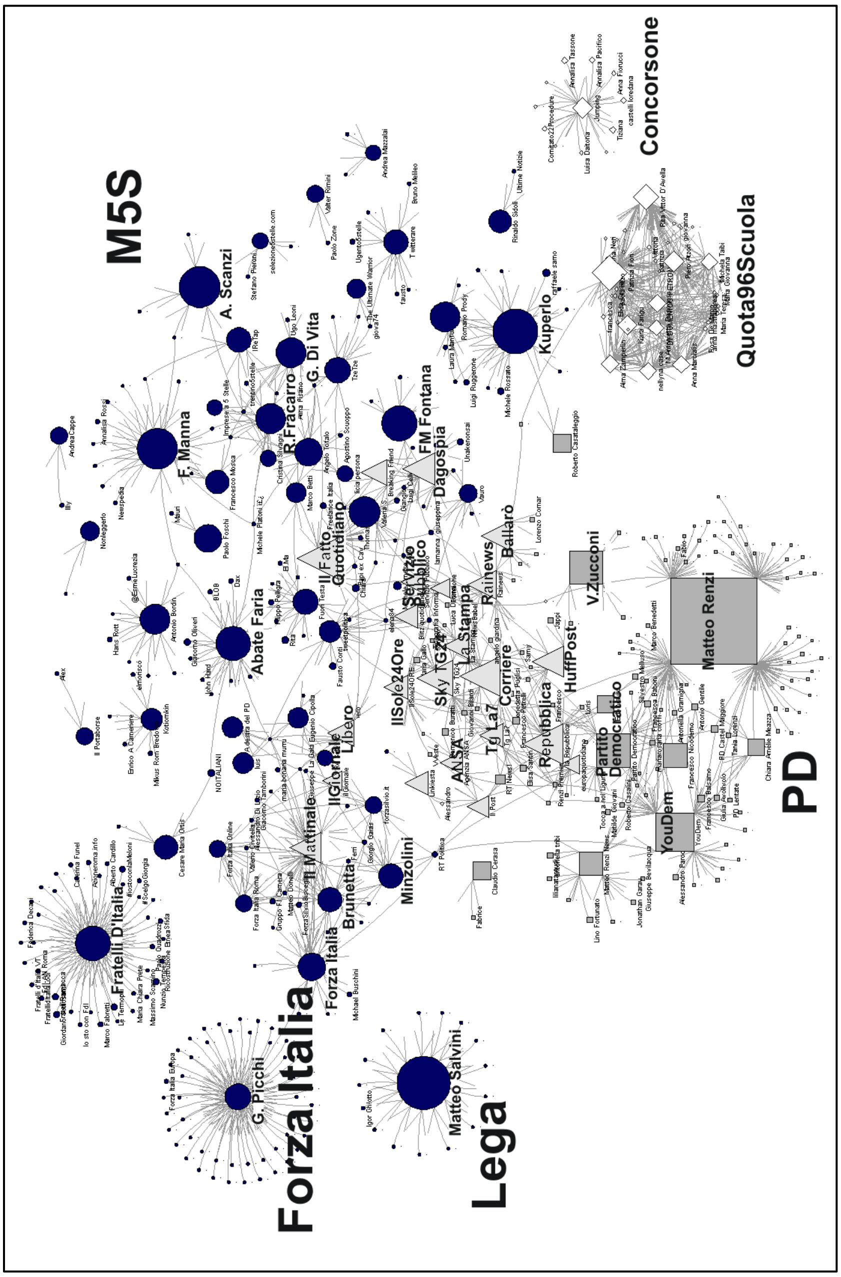


Figure 3 – *A simplified view of the full network*

The Forza Italia sub-network looks quite articulate with the presence of some very marked individualities (Picchi, Brunetta, Minzolini) and three news agencies (Il Giornale, Il Mattinale, Libero).

6. Conclusions

The paper shows how it is possible to analyze the popularity of a politician, examining millions of posts on Twitter. This was obtained using a reliable and cheap procedure, that includes text mining and statistical classification models. The information that we extracted, which also include the sentiment of users, are not typically used in the analysis of social network data. The results of our analysis also show the difference between Twitter popularity and consensus in the country.

In a deeper analysis, we can extend this approach to analyze the followers of a politician, describe the type of users in the network analyzed (always including the sentiment), and also investigate other social networks (e.g. Facebook).

Bibliographic references

SAS INSTITUTE (2011). Applied Analytics using SAS Enterprise Miner. Course notes. SAS Institute Inc., Cary, USA.

BENTIVEGNA S. (2014). La politica in 140 caratteri, Franco Angeli.

SUMMARY

Machine learning and text mining to classify tweets on a political leader

Twitter is a well-known social network. Users communicate with other users by posting short messages. These ‘tweets’ point out links among users that can be analyzed and that help to individuate “communities” who share opinions and comments. To achieve this result, we have to analyze textual data. In this paper we propose a procedure that combines machine learning techniques and text mining for the sentiment analysis on a political leader.

Agostino DI CIACCIO, Sapienza, Università di Roma, agostino.diciaccio@uniroma1.it

Giovanni Maria GIORGI, Sapienza, Università di Roma, giovanni.giorgi@uniroma1.it