STATISTICAL ANALYSIS OF SOCIAL NETWORKS

Agostino Di Ciaccio, Giovanni Maria Giorgi

1. Introduction

The marked increase in the use of social networks, especially among younger age groups, offers a new opportunity for statistical surveys. According to a recent study in Italy, of young people between the ages of 18 and 30, who grew up during the boom of social networks and web 2.0 (14% of the population), 91% are enrolled in a social network, 55% in a forum, 34% constantly follow one or more bloggers and 17% have their own blog.

The advent of smartphones and tablets will tend to accentuate this phenomenon: in Asia, even now, 60% of the people that frequent social networks already use a primary tool. According to our estimates, in Italy about 55% of the messages on Twitter are sent and received on a mobile phone, 54.8% of young people between 14-29 years of age have a smartphone (2012) and services that require the use of a mobile phone, for example real-time information about train delays in certain sections, are already being offered on Twitter.

Many companies are already engaged in extracting information from the social networks: to back up the launch of a new product or carry out a political poll, for example.

The potentialities of analysis are considerable: you can succeed in analysing millions of posts with costs and time extremely low compared to a traditional survey. However, analysing this information calls for special techniques that combine textual analysis with advanced statistical techniques and suitable software tools. In fact, through the social networks it is possible to "listen" to the opinions expressed by thousands or even millions of people concerning a wide range of subjects. However, these opinions are expressed textually, with language and procedures typical of the social network being used.

The BuzzMetrics application by Nielsen, which has been on the US market for 10 years now, is meeting with success on the international market and its strongpoint is the large number of sources that information can be extracted from, over 180 million blogs and 100 thousand forums throughout the world. Another example is Sysomos, a Canadian company founded in 2005 as the result of an

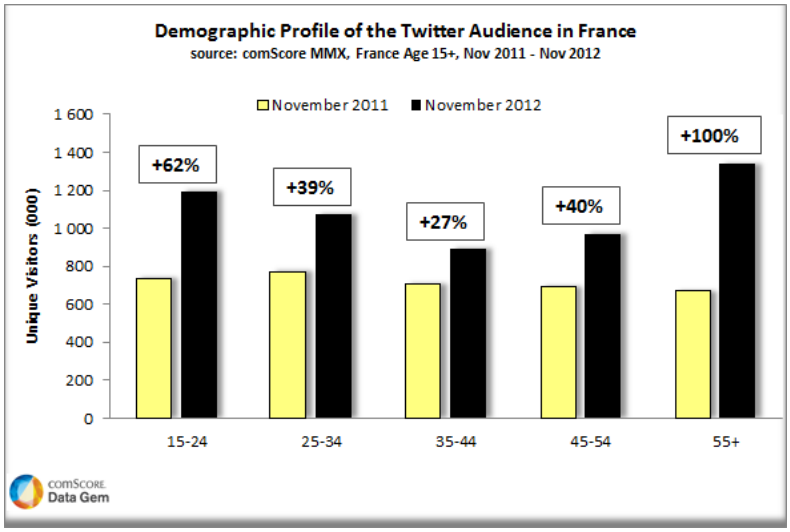
advanced research programme of the University of Toronto. Thanks to an analysis of the language combined with data mining techniques, Sysomos identifies the important subjects, the problems and the sentiments of the discussions and it activates automatic spam filtering.

In Italy an interesting example is *Voices from the Blogs*, an initiative that started off as a research project of the Milan University in October 2010 and now provides services to authorities, firms and news agencies. It uses advanced statistical techniques developed by G. King and D. Hopkins of Harvard University. Also in Italy, we have *Cogito* by Expert Systems, based on a database of millions of concepts and relations but the technology is not so suitable for analysing very short texts. The company *Blogmeter* has a similar approach.

A list of the most common applications of surveys carried out on the social networks is: web brand reputation, brand protection, analysis of the competition, market research, monitoring of social phenomena, opinion surveys, analysis and evaluation of services.

It could be particularly interested to identify the "opinion leaders", that is to say the people or organisations that can influence the online world. many people listen to them and, above all, act on their advice.

Figure 1 - *Users of Twitter in France (2011-2012).*



Unlike a traditional survey, which involves a maximum of 1-2 thousand contacts, by means of the social networks we can analyse even millions of posts. For instance, about 140 million tweets are entered on Twitter every day

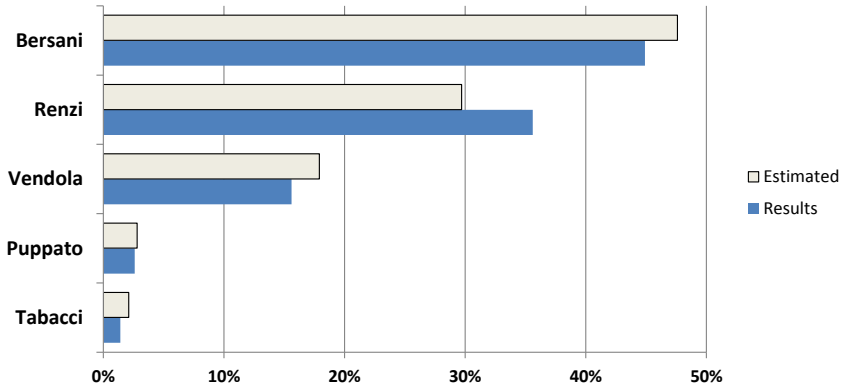
(throughout the world, 2011 survey) and the phenomenon is increasingly markedly, thanks to the fact that it does not require the use of a computer, and is even spreading to the not-so-young groups (especially women). This would make studies carried out on this social network more reliable.

A recent survey on Twitter, made in France by comScore (figure 1), shows that within the twelve months from 2011 to 2012 the users of Twitter increased by 53%, reaching 5.5 million users in November 2012 and making France the seventh largest market for Twitter. The number of visitors over 55 years of age doubled compared to the previous year and represent the most important segment of the French public, with 1.3 million users, followed by the 15-24 years group, which represents 1.2 million users (62%).

2. Advantages and disadvantages of the use of social networks

Political polls, prior to an election, are an ideal testing ground for evaluating the reliability of the use of social networks. By now almost all political elections are preceded by surveys carried out on Twitter and on blogs. This was the case, for example, in the US presidential elections, with very satisfactory results, and is now being done in Italy also. Take, for example, the primary elections of the centre-left coalition in Italy on 25 November 2012. We show in figure 2 the analysis made two days before the elections, using Sentiment Analysis on Twitter, published in the *Corriere della Sera* newspaper, the results of which were, on the whole, satisfactory.

Figure 2.- *Comparison of sentiment analysis with the real results of the primary elections of the PD (democratic party)*



Obviously, the tools used and the ability to use them properly are, as usual, vital for obtaining good results and the advantages and disadvantage related to this type of survey must always be taken into account.
Among the advantages we can mention:

- "Real-time" survey: an analysis can be made within a few days.
- Possible retrospective analysis: by analysing the tweets related to a given period of time it is possible to analyse the sentiment corresponding to events and actions that can be taken (advertising campaigns, promotion works and so on).
- Georeferencing: in some cases information about the place of origin of the tweets can be obtained.
- No questionnaires, and low costs: since no questionnaires have to be completed there are considerable savings in carrying out the survey.

Some of the disadvantages are:

- The sample observed may be distorted, especially if the target population is one that makes little use of internet.
- We cannot ask, we can only "listen" to what is being said on the social networks, and so we cannot put the questions we think most important.
- The analysis calls for expert and reliable researchers, familiar with the phenomenon being investigated, the social networks, the language used, the software, and the statistical techniques that have to be used.

3. The characteristics of Twitter

In addition to the text message a great deal of other information is present: 128 fields, which can come to as much as 250 if a retweet (a large number of fields is empty). Certain relevant information is present in the fields usually compiled, for example:

- *source*: indicates the method by which the tweet was sent
- *user.created_at*: date the account was created
- *user.description*: a string in which the user describes his or her account.
- *user.favourites_count*: the number of tweets the user entered as favourites since the time of registration
- *user.followers_count*: the number of followers
- *user.friends_count*: the number of users of which the reference user is a follower
- *user.name*: the name of the user
- *user.statuses_count*: the number of tweets written by the user

Taking all the tweets into consideration, we find that about 29% are retweets while 26% are replies. On the other hand, taking only tweets on political subjects into account, we find that as many as 42% are retweets and only 16% replies. Note that if we only consider "volume data", such as the number of followers or the number of retweets, we get conflicting and not very significant information. It can be seen that the number of times a user is retweeted is not necessarily related to the number of followers.

4. Sentiment Analysis methodology and software

The term Sentiment Analysis means techniques that can automatically extract, analyse and classify opinions expressed, on the basis of a written text, usually present on the WEB.

In this article we only take into account the Twitter social network and only distinguish between positive opinions and negative opinions. It must also be said that quite often it is difficult to classify a tweet as a positive opinion or a negative opinion. This is due to the fact that, because of their brevity, the texts are not very structured and are often ironical or allusive, with links to other documents. Users generally use a language rich in metaphors and references. The texts often do not contain complete sentences and are often closely connected with the latest news or, in any case, with very recent news.

In some cases it is also possible to define neutral opinions, even though for some subjects, like politics for instance, this type of message is not very frequent.

A commercial software available for this purpose is SAS Sentiment Analysis, which offers the user three different methods of analysis:

- a supervised classification statistical model,
- a set of rules for defining sentiment,
- a hybrid system, combining the preceding options.

The statistical model basically consists of the estimation of a feedforward neural network with a binary target (but, unfortunately, the documentation does not explain this aspect). Estimation of the model is done starting with a set of "training" documents (training corpus) already classified as positive or negative. The user cannot interfere with the model used, which makes it very simple but not very flexible for an expert user.

As an alternative or, better still, in addition, a sophisticated language can be used, capable of defining a set of very complex textual rules for finding the positivity or negativity of the texts being reviewed. On the other hand, the characteristics of the tweets and their brevity, makes the use of rules in classifying

them ineffective. Included in the program we have a dictionary of terms, a dictionary of ontology, a list of synonyms, a list of stop-words (terms we do not consider useful in the analysis). It is also possible to add, to the set of rules, words identified as positive or negative in the statistical model. This list can also be manually edited subsequently. The model created "imports" rules from the statistical model but does not create a true hybrid model, which can instead be explicitly created when both the statistical model and the set of rules have been defined. In short, the classification of the sentiment of a text is obtained by combining, by means of predefined weights, the answer obtained from the statistical model and from the set of rules.

In our application on the tweets, the set of rules we defined was able to classify only a small percentage of tweets and therefore the solution is in any case dominated by the classification of the statistical model.

Instead of that software, Enterprise Miner by SAS can be used, where there is a Text Mining module with which it is possible to construct a complete and sophisticated analysis using machine learning methods and models: Neural Networks, Support vector machine, Gradient boosting, Naïve Bayes, Ensemble methods.

After a small percentage of tweets for each day have been classified and the training data-set then created, the model is constructed as follows:

1) filtering and cleaning the texts;
2) tokenizing the texts and constructing the documents x words frequency matrix;
3) analysing and filtering the most relevant terms;
4) singular value decomposition of the frequency matrix;
5) splitting of the texts into training, validation and testing;
6) applying the supervised statistical classification models, tuning the models;
7) comparing the models and selecting the model with the best performance;
8) making a new estimation of the selected model, using all the available text;
9) assessing the sentiment on all the available texts.

With an analysis of this kind, a percentage of between 70% and 80% of correct classification can be achieved on new tweets. What is more, since we are not interested in the classification of the single tweets but only in an overall estimate of the number, this procedure can obtain very reliable estimates, with an error very small in the forecast of the number of positive and negative tweets.

Lastly, the SAS Sentiment Analysis module is easy to apply and fast and very effective in analysing medium sized texts, thanks to ontological and grammatical dictionaries. Enterprise Miner Text Mining is more flexible and more effective for short texts like tweets but requires a good knowledge of the statistical tools used

(Neural Networks, Support vector machine, classification trees). Non-commercial software can also be used with success to obtain these analyses but it requires more work and a certain amount of computer skills.

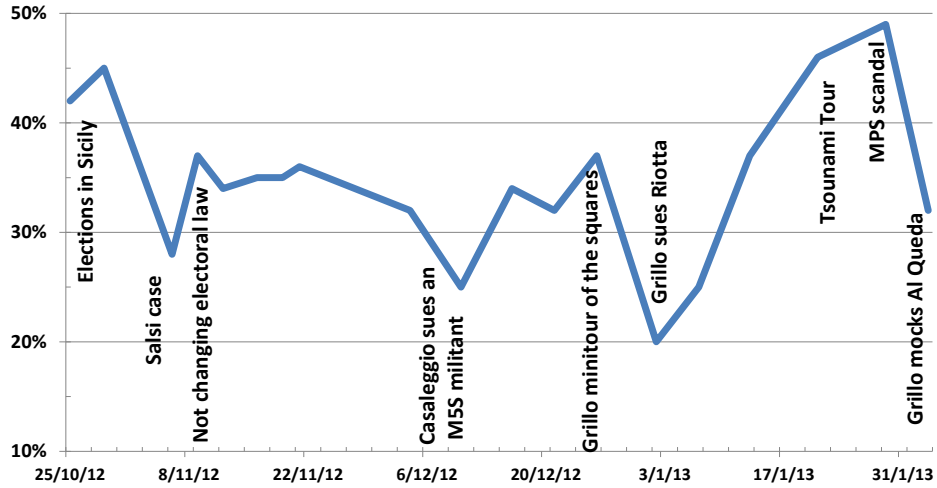
5. Results of analysis of the tweets concerning Beppe Grillo

As an example of application, we analysed the sentiment about Beppe Grillo expressed in the tweets. We fixed ourselves the aim of classifying the tweets as positive or negative.

We extracted 29646 tweets by searching for the hashtag #Grillo on 20 different days between 25 October 2012 and 3 February 2013. The first 100 tweets were classified manually for each of the 20 days.

There were many aspects that could have been taken into account in the analysis but that we neglected in this example. We did not consider the presence of retweets, of links to other pages, we did not consider information about the authors of the tweets or evaluate the presence of opinion leaders.

Figure 3. *Pattern of positive sentiment about Beppe Grillo during the period 25/10/2012 – 3/2/2013*



In figure 3 we show the results of the analysis made with SAS Enterprise Miner. We also included in the graph the presence of particular events that occurred during the period and that had great influence on discussion on Twitter. Note that the graph actually represents the evolution of the discussion and does not

give the number of supporters of Grillo. In the graph we show the percentage of favourable tweets, for each day analysed. It can, however, be noted that the average of positive tweets (around 33%) is very close to an estimate of the votes obtained by Grillo in the elections of February in the youngest age group (18-24).

6. Conclusions

The analysis we showed, made as part of a degree thesis (Claudia Proia 2013), shows the potential inherent in the statistical analysis of textual data taken from the WEB. It would, in fact, be unthinkable to conduct a similar survey using traditional tools without having enormous resources. In our opinion, the results of a sentiment analysis can be considered very interesting if one takes into account the necessary statistical methodologies and the characteristics of the reference group. The analysis can be conducted in more depth than shown in the previous paragraphs, including the analysis of: followers, opinion leaders, geolocation data, other information from blogs and from other information channels. Taking these data into account certainly will make the analysis more complex but also more complete.

References

SAS Sentiment Analysis Studio: Building models. Course notes (2011). SAS Institute Inc., Cary, USA.

PROIA C. 2013. Sentiment Analysis for the statistical classification of tweets. *Graduate thesis*. Department of Statistics, Sapienza, University of Rome.

SUMMARY

Analysis of information, expressed in a textual and unstructured manner on the web, is becoming increasingly common in web-marketing or political polls. These analyses require an advanced statistical methodology that combines text-mining with machine learning. This article analyses the potential of this type of analysis and gives an example of sentiment analysis application to a political poll.

Agostino DI CIACCIO, Full Professor, "Sapienza" University of Rome,
agostino.diciaccio@uniroma1.it
Giovanni Maria GIORGI, Full Professor, "Sapienza" University of Rome,
giovanni.giorgi@uniroma1.it