Twitter for election forecasts: a Joint Machine Learning and Complex Network approach applied to an italian case study

International Conference on Computational Social Science 2015

Mauro Coletto
ISTI-CNR, Pisa - Italy
IMT Institute for Advanced Studies
Lucca - Italy
mauro.coletto@isti.cnr.it

Claudio Lucchese
ISTI - CNR
Pisa - Italy
claudio.lucchese@isti.cnr.it

Salvatore Orlando
DAIS - Unive
Venice - Italy
orlando@unive.it

Raffaele Perego
ISTI - CNR
Pisa - Italy
raffaele.perego@isti.cnr.it

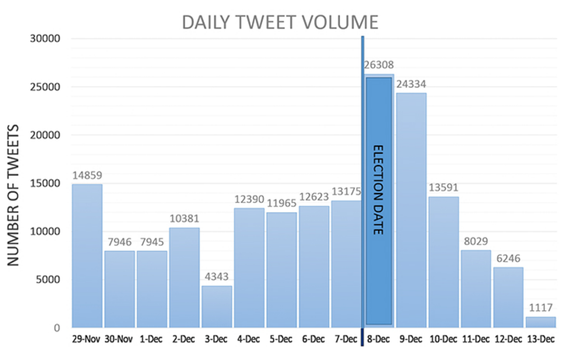
Alessandro Chessa
IMT Institute for Advanced Studies
Lucca - Italy
alessandro.chessa@imtlucca.it

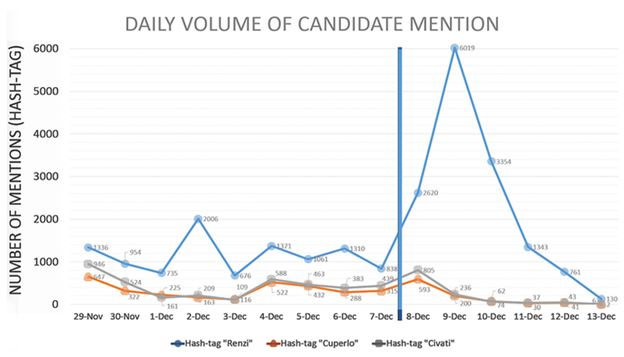
Michelangelo Puliga
IMT Institute for Advanced Studies
Lucca - Italy
michelangelo.puliga@imtlucca.it

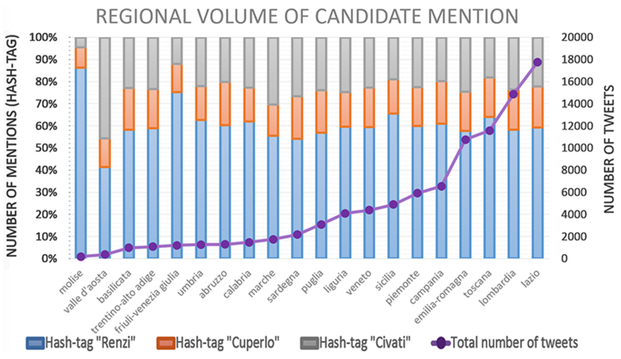
Several studies have shown how to approximately predict real-world phenomena, such as political elections, by analyzing user activities in micro-blogging platforms. This approach has proven to be interesting but with some limitations, such as the representativeness of the sample of users, and the hardness of understanding polarity in short messages. We believe that predictions based on social network analysis can be significantly improved by exploiting machine learning and complex network tools, where the latter provides valuable high-level features to support the former in learning an accurate prediction function.

DATA:

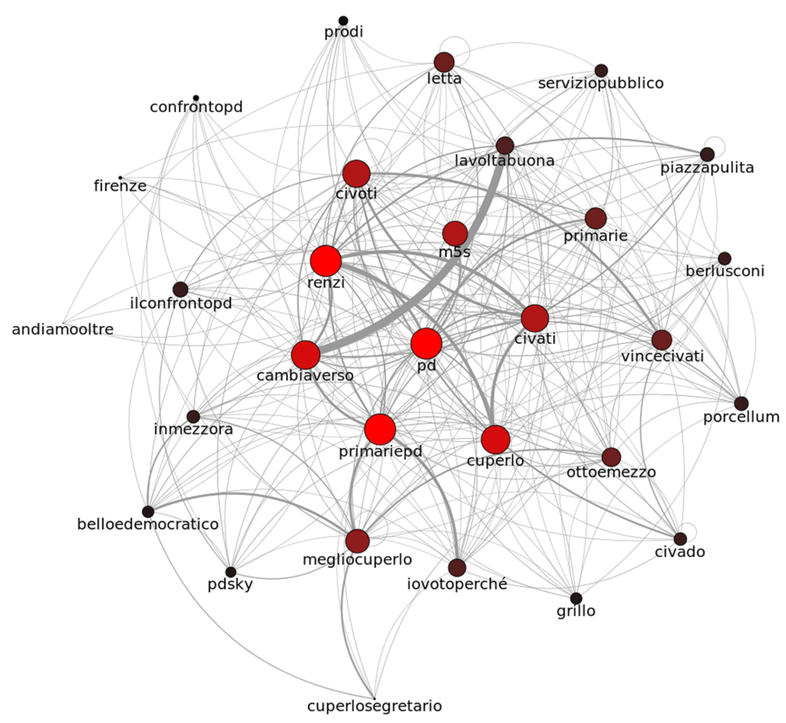
dataset	PRIMARIE "PD" (tweets)
original raw dataset	≈1.7 million
Italian data	≈416K
geo-located data	175 252
pre-electoral data	95 627



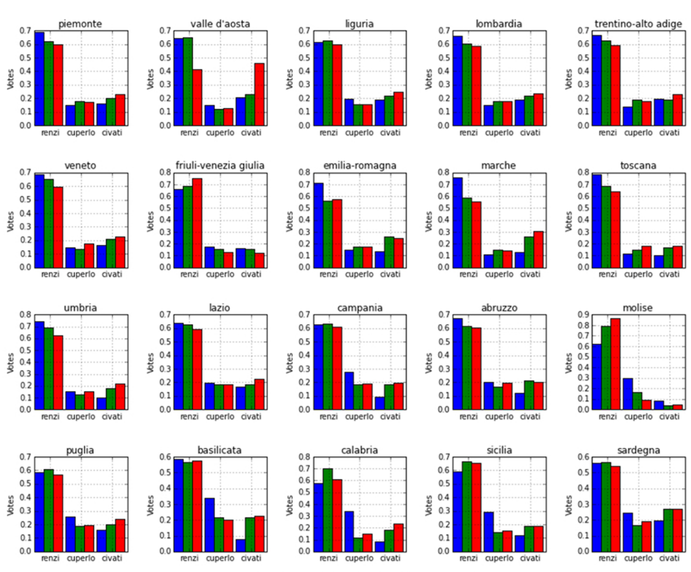




EXPERIMENTS:



Method	MSE	Δ%
tweets	0.0105	–
classified tweets	0.0062	-41%
users	0.0063	-40%
extended classified tweets	0.0156	+49%
regression	0.0044	-58%



The volumes of the tweets reflected, with a good approximation, the relative strength of the candidates. However, when dealing with social network platforms, it is well known that data can be biased toward specific classes of users (i.e., young and educated people). How these biases can change the final predictive power of counting the tweets is an open research question. In this work we investigated the echo on Twitter of the primary election of the Italian major political party: the "Partito Democratico". In this electoral campaign three candidates (Mr. Renzi, Mr. Cuperlo, and Mr. Civati) ran for the leadership. They appeared in the traditional media (TV shows and Press interviews) as well they used the new social media to create hype and discussion. Using a collection of tweets covering nine days before the elections, as well as the official electoral results data by Italian region, we were able to perform an analysis based on joint techniques from Machine Learning and Complex Networks. Our goal was to shift beyond the tweets counting paradigm by learning a function able to estimate the actual votes received by a candidate given its presence on Twitter. We included features based on complex networks measures of Twitter networks (networks of hashtags, mentions) in the machine learning process. In particular several topological measures, such as the average degree, the betweeness centrality and others were tested. We were able to establish which network features were the most effective in improving the forecast. An example of hashtag network concerning our dataset is the figure on the left, which shows the co-occurrences of the most frequent hashtags, including the three candidate hashtags : #renzi, #cuperlo, and #civati.

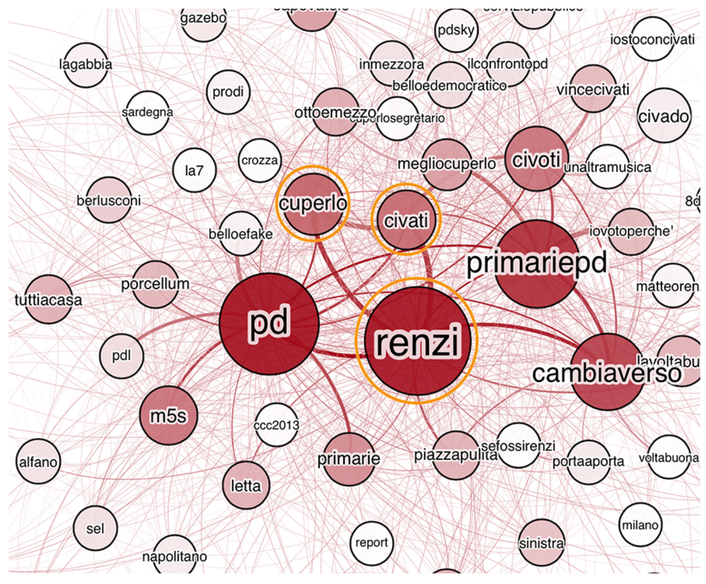
In order to evaluate the feasibility and accuracy of our machine learning approach, we build a ground-truth dataset as follows. The number of votes received by the three candidates is known in each of the 20 regions. Therefore, we were able to correlate independently twitter features of users in each region with the actual, i.e., ground-truth, electoral result. We thus transformed our initial Twitter data into a set of 60 prediction experiments, i.e., the percentage of received votes of 3 candidates in 20 regions. We report MSE error of several preliminary techniques. Five-fold cross validation is applied. To avoid overfitting we investigated simple regression methods.

We evaluated the following predictors:

– **tweets**. The predicted percentage of received votes is based on the percentage of tweets mentioning the candidate. This is the usual tweet counting approach.
– **classified tweets**. Each tweet t is assigned to the candidate c that maximizes the score $S(c|t) = \sum_{h \in t} P(c,h)/P(h)$, where h is a hashtag in t corresponding to one of the candidates. The prediction is based on the percentage of classified tweets.
– **users**. Percentage of unique users mentioning the candidate. In case more mentioned candidates by a unique user, his unit value is divided by the number of mentioned candidates.
– **extended classified tweets**. The tweets are classified as in classified tweets, with h being a hashtag in t related to a candidate, after having clustered the 1,000 most frequent hashtags in 3 groups - one per candidate - including the corresponding candidate hashtag, along with the other hashtags that co-occur most frequently.
– **regression**. We use the predictor classified tweets as a feature for fitting a simple linear regression model on a training set. The learnt regression model is then applied to the test set for each candidate separately (with five-fold cross validation). The resulting MSE is **0.0044**.
As reported in the final table, the regression method halves the error of the baseline. The final histogram on the left shows on a regional basis 3 columns per candidate: (blu) real percentage of votes received by the candidate; (green) percentage of classified users by candidate and (red) percentage of tweets mentioning the candidate hashtag (baseline).

NETWORK ANALYSIS

An additional study of the tweets leads to the introduction of predictors based on networks. We studied the network of co-occurring hashtags and the network of mentions. Due to lack of space we describe only the approach based on the first network, which provided more accurate results.
The hash-tags co-occurrence graph is undirected and weighted considering the frequency of the co-occurrences. The figure on the right shows a portion of the 100 most frequent hash-tags co-occurrences network. Color gradient follows the degree distribution of nodes, whereas the size is related to the betweenness centrality of each node.
To some extent, this graph reflects the political discussion observed in Twitter before the elections. By analyzing this graph, we aim at understanding the weight of each candidate in such discussion. We evaluated different topological measures, such as node average degree, network diameter and clustering coefficient.
An interesting measure is the betweenness centrality of the three nodes corresponding to the candidate names.
For each region, we considered the co-occurrence graph of the top 100 most frequent hash-tags, the top 200 and the full set of hash-tags. As shown in the table on the right, the error decreases with larger graphs.
Intuitively, we can say that the betweenness centrality captures the relevance of the candidates in the political discussion observed in twitter. By considering all the shortest paths, it is able to accurately evaluate hash-tags that are often used together with the candidate name, such as synonyms, substitutes, nicknames or catchphrases. If we consider an hash-tags as a topic of the discussion, betweenness centrality is likely to be able to capture the most related topics with each candidate, the importance of those topics in the graph, and, as their variety, which is also related to the users in the network.
We believe that this kind of network measures may provide some additional interesting information which is not yet fully exploited by state-of-the-art algorithms. We think that it needs further investigation and a wider evaluation on other data sets. We leave this investigation as future work.



Network feature	MSE	MAE
Bet-Centrality 100 Hash-tags	0.0217	0.1122
Bet-Centrality 300 Hash-tags	0.0192	0.1023
Bet-Centrality All Hash-tags	0.0182	0.0976