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

Accepted for Poster Presentation at the 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

ABSTRACT

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.

1. BEYOND COUNTING TWEETS

The use of the Twitter micro-blogging platform as a tool to predict the outcomes of political elections is not new in literature [3, 2]. Approaches based on the analysis of the tweets volumes and polarity were able to correctly guess the final outcomes in a statistically significant number of cases, also with a correct geographical resolution [1]. 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. The traditional media and Twitter/Facebook influence each other and, for every single TV show there is a specific peak in the tweet volumes on Twitter.

Using a collection of tweets covering about ten 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 therefore assume *by design* that tweet volumes, even if accurately collected, are not proportional to future votes. To accomplish this goal, we included *features* based complex networks measures of Twitter networks (networks of users, hashtags, retweets etc.) in the machine learning process. In particular several topological measures, such as the average degree, the diameter of the network, the clustering coefficient 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 shown in Fig. 1, which shows the co-occurrences of the most frequent hashtags, including the three *candidate hashtags*: `#renzi`, `#cuperlo`, and `#civati`.

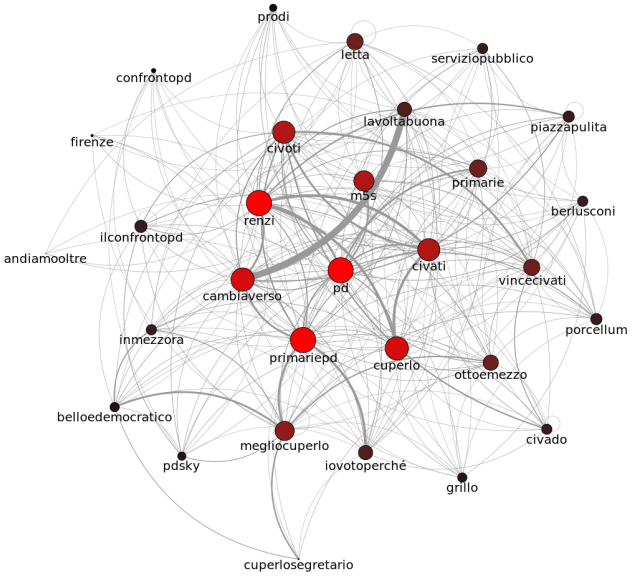

Figure 1: Hashtag correlation analysis

In a future research we will investigate not only the basic data from the networks related to the Twitter data but also

how to use the information about population and demography (for instance number of educated people by geographical location, census data of the population, age classes) to start addressing the problem of unbiasing the Twitter data. We also plan to study complex network features able to measure the impact of media campaigns, the diffusion of electoral messages, and the topologies of interaction networks witnessing a large rate of conversion from posting tweets about a candidate to actually voting the candidate.

2. ELECTORAL PREDICTIONS

Our study was conducted on a dataset of ∼1.7 million tweets collected through Twitter API by querying a list of keywords related to the election and the candidates. The election took place on December 8^{th} 2013, and the dataset covers about 10 days before and 5 days after the election day. We considered only the tweets in Italian language, geo-located with cities (∼168,000) or regions (∼175,000). In fact, only about 8,000 tweets provide GPS information, whereas the remaining tweets were geo-located by matching the user profile location with the Italian cities and regions. We finally filtered 95,627 geo-located tweets across the 20 regions of the country, taking into consideration only the tweets published before the election day.

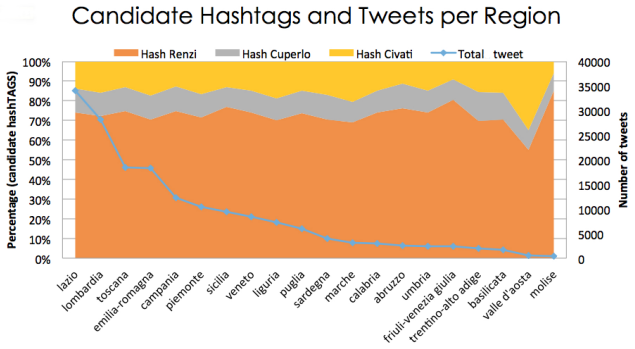

Figure 2: Hashtag analysis by region

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. Fig.2 shows the distribution of tweets and votes across the 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.

Table 1: Election Prediction

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%

– **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 (see Fig.1).

– **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 Table 1, the regression method halves the error of the baseline. Fig. 3 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).

3. RESEARCH CHALLENGES

Our proposal opens new opportunities and research challenges. The naïve approach of correlating simple social media networks measures, e.g., tweets volume, is not sufficient to provide accurate estimation of real world phenomena. We believe that machine learning methods are capable of devising more accurate models, by exploiting social media features in a non trivial way. We aim at exploiting network properties to support machine learning algorithms. Network properties provide global information about the topology and evolution of a network, and thus information about the behavior of their users. Such additional information may dramatically improve the predictive power of machine learned models.

The proposed approach renews and opens up to new research challenges. The application of machine learning methods is harmed by the lack of positive training instances, e.g., elections are not very frequent. Therefore, we need machine learning methods able to generalize well and minimize misprediction risk with a very small number of positive examples. The dynamism of social network data and their size require new network analysis tools that take into account the network evolution and that provide accurate of streaming analysis.

Preliminary results are promising, and we believe that our study can be successfully applied to other use cases by tackling the aforementioned research challenges.

4. REFERENCES

[1] G. Caldarelli, A. Chessa, F. Pammolli, G. Pompa, M. Puliga, M. Riccaboni, and G. Riotta. A multi-level geographical study of italian political elections from twitter data. *PloS one*, 9(5):e95809, 2014.

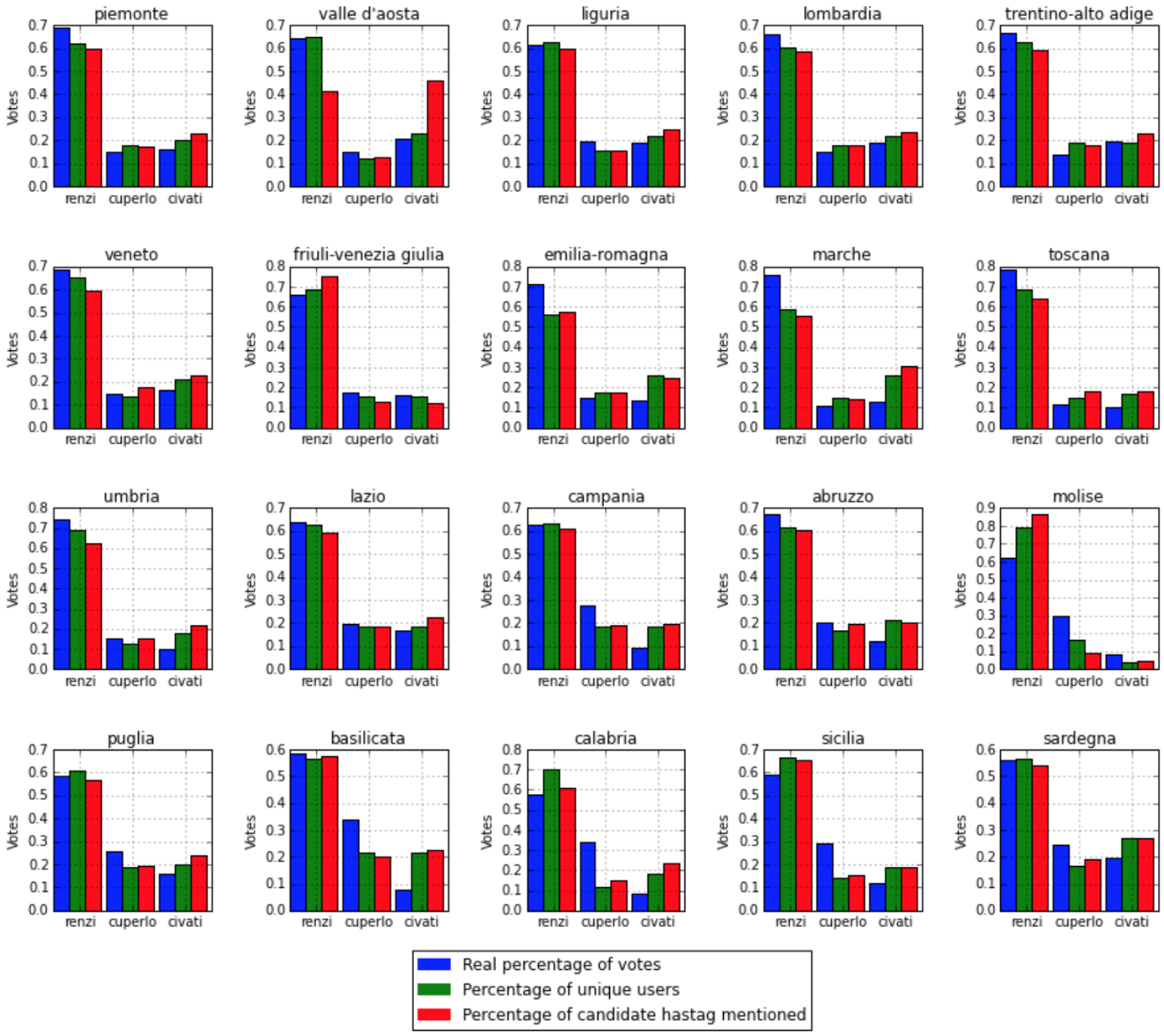

Figure 3: Real percentages of votes, baseline and predictions (classified users) by region

[2] M. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer, and A. Flammini. Political polarization on twitter. In *ICWSM*, 2011.

[3] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, 11:122–129, 2010.