Content-based Classification of Political Inclinations of Twitter Users

Marco Di Giovanni Marco Brambilla Stefano Ceri Florian Daniel Giorgia Ramponi

Politecnico di Milano. Dipartimento di Elettronica, Informazione e Bioingegneria. P.za L. da Vinci 32. Milano, Italy
[firstname.lastname]@polimi.it

***Abstract*—Social networks are huge continuous sources of information that can be used to analyze people's behavior and thoughts. Our goal is to extract such information and predict political inclinations of users. In particular, this paper investigates the importance of syntactic features of texts written by users in the process.**

Our hypothesis is that people belonging to the same political party write in similar ways, thus they can be classified properly on the basis of the words that they use. We analyze tweets because Twitter is commonly used in Italy for discussing about politics; moreover, it provides an official API that can be easily exploited for data extraction.

Many classifiers were applied to different kinds of features and NLP vectorization methods in order to obtain the best method capable of confirming our hypothesis. To evaluate their accuracy, a set of current Italian deputies with consistent activity in Twitter has been selected as ground truth, and we have then predicted their political party. Using the results of our analysis, we also got interesting insights into current Italian politics.

***Index Terms*—computational social science, political inclination forecast, Twitter analysis, natural language processing**

I. Introduction

A. Problem definition

Studying political inclination of people in social networks is becoming an increasingly interesting topic after the last USA election. Similar analyses can also be performed in other countries, in case one can assume that social networks are constantly used for political discussions and propaganda.

Being able to estimate political inclination of people by looking at what they share on social media can be a useful tool to predict election results and can be compared to statistics obtained through classical methods such as surveys.

Much information can be extracted from social networks, ranging from permanent connections (friend and follow relationships) to temporary interactions (like, comment, repost) to general information (geographical location, profession, education) and content posted (texts, images, links). Social networks also allow us to get this wide amount of data updated in real time, so that the analysis can be always up-to-date and changes in behaviour can be easily detected with a relatively small delay. We focus on written texts since our goal is to understand the importance of them for classification purposes and similarity of users, as opposed to classical methods that analyze the social networks as a network itself for similar tasks.

This work was partially supported by the ERC Advanced Grant 693174, Data-Driven Genomic Computing.

In Italy, the most used Social Network is Facebook, followed by YouTube. However, since our analysis is focused on syntactic proprieties, we chose Twitter, known for the possibility to share short written messages. The main advantage of Twitter is the public API that allows one to easily extract every information needed, with few limits in terms of frequency. Twitter is also widely used to discuss political issues in Italy, where politicians and supporters perform low-cost propaganda and share their opinion continuously, consolidating our choice.

Our main hypothesis is that tweets contain enough information to understand the political inclination of people. To test this hypothesis, a not easily collectible ground truth is needed, since individuals rarely share their political inclination (thus, the secretiveness of the vote). Hence, to first evaluate different prediction methods, we chose to perform the analysis on politicians, whose political inclination is obviously known. Future work will focus on the analysis of non politicians to understand if the hypothesis is still valid, meaning that users write in the same way if they vote to the same party.

An important aspect to consider is that since we are not using the social network structure, we are not confined to classify people connected in some way to others. Any user that has at least a moderate activity on Twitter can be classified, ignoring its social connections. Of course, incorporating the social network structure could be done to improve the prediction accuracy, but can also limit the prediction power since the account must be in some way connected to others, requisite not needed in the context of this work.

Unlike classical tools such as surveys, algorithms based on social network analysis are faster and can be used on a larger scale with a relatively small effort. Large quantities of data can be collected daily obtaining a wider and more heterogeneous set of people analyzed. On the contrary, bias can be a problem, since the Twitter community not always is homogeneously distributed with respect to voters. It should be taken into account that people belonging to different parties can represent different classes in the society, thus can be more or less inclined to use social networks as a political instrument. The analysis of the gap between the real distribution of voters and the one predicted can give interesting insight on the voters of different parties.

For research purposes, we chose to split the wide Italian

978-1-5386-5035-6/18/$31.00 ©2018 IEEE

Party	Deputies	Fraction
MOVIMENTO 5 STELLE	221	35.1%
LEGA - SALVINI PREMIER	125	19.8%
PARTITO DEMOCRATICO	111	17.6%
FORZA ITALIA - BERLUSCONI PRES.	105	16.7%
Other minor parties	68	10.8%

TABLE I
ITALIAN DEPUTIES CHAMBER: DISTRIBUTION OF DEPUTIES AMONG THE FOUR MAJOR PARTIES AFTER THE ELECTION ON MARCH 4, 2018.

political situation in four different categories: "MOVIMENTO 5 STELLE", "LEGA - SALVINI PREMIER", "PARTITO DEMOCRATICO" and "FORZA ITALIA - BERLUSCONI PRESIDENTE". Smaller parties have been discarded since the fraction of population that voted them was small enough not to be relevant for our purposes.

The paper structure is the following: in section II, the related work is exposed; in section III the methods to perform the classification are described; the evaluation technique and results are shown in section IV; in section V, we perform a further analysis of the dataset, given the results of the classification; we conclude in Section VI.

B. Italian political situation

Tweets used for this work are collected in August 2018. After the election of March 2018, the Italian government is composed of a coalition between "MOVIMENTO 5 STELLE" and "LEGA - SALVINI PREMIER", with respectively 32,7% and 17,4% voters at the election in March 2018. The other 2 main important parties are "PARTITO DEMOCRATICO" and "FORZA ITALIA - BERLUSCONI PRESIDENTE", with respectively 18,7% and 14,0% voters at the election. Thus, the four parties selected represent the 82,8% of the total voters. Chamber of deputies is composed by 630 members, subdivided between parties with relation to the percentage of voters. In table I, the actual numbers are reported.

II. RELATED WORK

In recent years, many works focused on analysis of social network accounts in order to obtain information about political inclination. Often, analyses are made around election days, to obtain insights and predictions of the results.

In [11], the concept of wisdom of the crowds introduced in [12] is applied twice to forecast 2010 UK election results using data from social media. Using an ARIMA model they claim to exceed the predictive power of classical surveys.

A quantitative analysis of Tweets is performed in [9] to prove that social media can be a reliable tool about political behavior, applying this technique to competitive races of 2010 and 2012 US congressional elections.

Moreover, in [18], they state that volume of tweets is not always enough to capture public opinion and they propose a better but not perfect model able to obtain more accurate results about 2012 American republican presidential election.

Interesting results are obtained observing the bias of pools and Twitter for Donald Trump and Hillary Clinton in 2016 U.S. election, suggesting to not underestimate the effect that different forecasting methods can have on the predictions based on the nature of the method itself (an heterogeneous sample of the voters is not easy to collect) [1].

An improved analysis is performed on Brexit data, classifying through SVM the leave/remain intention of users. In [6], they confirm that this kind of analysis of political topics using social media data can substitute Internet pools and telephone calls, being not only more accurate, but also faster and cheaper.

However, in [13], the limits of Twitter are exposed, revealing the scarce robustness of this approaches. They apply algorithms that obtained good results for one election forecast to other elections, showing that results are not always as good as stated before. They conclude suggesting to investigate impact of different lexicons and the application of machine learning techniques for this task.

Similar analysis has also been performed about German federal election 2009, demonstrating that Twitter can be used as a source to perform political forecasts, since it is widely used for political deliberation and it mirror the offline political sentiment [19].

An interesting analysis of prediction of political inclination of Twitter users comparing results coming from contents (defined by hashtags used) and networks structure is performed in [8], showing advantages and disadvantages of both the techniques.

Some examples of prediction using syntactic features are the forecast of box-office revenues for movies using tweets about a set of popular movies [2] and the knowledge extraction algorithm proposed in [4], [5] .

Sentiment analysis is also one of the most used techniques, applied to correlate significant events in social, political, cultural end economic sphere with moods extracted from tweets posted in the meantime [3], [14].

Interesting research in the field about regarding political echo chamber must be cited, finding huge differences between Democrats and Republican behavior on Twitter through also network analysis techniques [7].

To the best of our knowledge, syntactic analysis has not been yet applied to classify deputies through Twitter.

III. METHODS

The analysis is composed of the following steps: creation of the dataset (subsection III-A), selection of appropriate syntactic features (subsection III-B), selection of a text embedding method (subsection III-C) and selection of a multiclass classifier (subsection III-D).

A. Dataset

The dataset consists in Twitter data from Italian deputies' accounts.

Firstly, names of the 630 Italian deputies and their corresponding parties are collected from the official website of the Italian parliament [1]. Deputies belonging to small parties are discarded due to their relative small importance in the actual

[1] http://www.camera.it/leg18/1

political situation, obtaining 562 names out of 630 deputies. The 4 main Italian parties selected are: "MOVIMENTO 5 STELLE", "LEGA - SALVINI PREMIER", "PARTITO DEMOCRATICO" and "FORZA ITALIA - BERLUSCONI PRESIDENTE".

We, then, automatically associate at each deputy his/her official twitter account. Using the twitter API to search for users, the names collected are used as inputs. We often obtain more than one account for the each query, due to homonymy issues. Accounts that don't contain in their bio one of the words selected (and their corresponding variations) about politics are discarded: 'deputato' (deputy), 'camera' (chamber), 'parlamento' (parliament), 'partito' (party), 'legislatura' (legislature), 'pd', 'lega', 'movimento' (movement), 'stelle' (stars), 'forza italia', 'salvini', 'berlusconi'. Accounts with less than 100 tweets are also rejected, since our analysis relies on a statistically relevant number of written words. Finally, if more than one account still corresponds to a given name of a deputy, the right one is manually selected. After this cleaning procedure, 188 twitter accounts corresponding to Italian deputies belonging to one of the 4 main Italian parties are collected, subdivided as follows: "MOVIMENTO 5 STELLE": 64, "PARTITO DEMOCRATICO":51, "FORZA ITALIA - BERLUSCONI PRESIDENTE": 39, "LEGA - SALVINI PREMIER":34. We are aware that this procedure doesn't find every account belonging to an actual Italian deputy, however we are still able to obtain a large enough dataset to perform our analysis. A more accurate analysis can be done by manually searching for politicians' accounts, but we believe that the great part of accounts that we are missing by automatizing the procedure will not be relevant to the analysis since they will not be active enough.

For each account found, we select the last 200 tweets (one API call per user), excluding retweets, and we merge the texts into a single large document d_i. URLs, mentions and every not alphanumerical character are removed to clean the text from non useful features.

The total number of tweets collected is 30643, since not every account tweeted at least 200 tweets since their registration on the social network. We remark that for this analysis no starting date has been selected, since we assume that the political inclination of actual deputies has not changed much recently.

Since our hypothesis is that deputies belonging to the same party write in the same way, the large documents obtained should contain enough information to understand the users' political inclination, so to classify accounts into the correct political party.

B. Selection of syntactic features

To select which kind of syntactic feature is best to classify users, for each user u we tag every word w_u of the document d_u, using a standard tagset [2]. This step is followed by a lemmatization step to reduce inflectional forms of a word to a common base form, performed by a NLP python library "TreeTagger" [17] trained using an Italian dataset.

Thus, for each user u, from each original document d_u, we obtain 5 different lists of words:

1) list of every word w_u, ignoring the tags
2) list of nouns n_u
3) list of verbs v_u
4) list of adjectives a_u
5) list of adverbs ad_u

We perform this selection to understand if there is a set of words that influences particularly positively or negatively the classification accuracy. This analysis allows us to answer the question if the information of the political inclination is present in the nouns used, the verbs or in every single word tweeted.

We obtain 40554 different words, 7838 nouns, 2469 verbs, 3118 adjectives and 490 adverbs, discarding what the tagger classifies as "unknown". Other tags are neglected since we think no useful information is contained in those set of words (articles, conjunctions, ...)

C. Vectorization

To perform any kind of classification task, lists of words w_u (or list of every other feature selected before) must be embedded into vectors.

Some standard vectorization methods are tried to better understand which one is the best embedding technique for our task.

1) *Count Vectorizer* (CV): converts a collection of text documents into a matrix of token counts;
$$CV(w,u) = f_{w,u}$$
represents the number of times that user u used the word w
2) *Hashing Vectorizer* (HV): converts a collection of text documents into a matrix of token occurrences, using the hashing trick to find the map between the token string name and the feature integer index;
3) *Term Frequency Vectorizer* (TF): converts a collection of text documents into a matrix of term frequencies;
$$TF(w,u) = \frac{f_{w,u}}{\sum_{w' \in d_u} f_{w',u}}$$
represents the frequency that the word w is used by the user u;
4) *Term Frequency - Inverse Document Frequency Vectorizer* (TF-IDF): converts a collection of text documents into a matrix of term frequencies weighted by document frequency;
$$TFIDF(w,u,U) = TF(w,u)IDF(w,U)$$
where U is the set of users,
$$IDF(w,U) = \log \frac{|U|}{|u \in U : w \in d_u|}$$

[2] http://sslmit.unibo.it/ baroni/collocazioni/itwac.tagset.txt

represents the logarithm of the fraction of the total number of users and the number of users that used the word w.

HV, TF and TF-IDF can be performed with L1 or L2 norm, obtaining a total of 7 different techniques [15].

No stop words are removed in this step and no limits are selected for the matrices, that have dimensions: $n \times N_w$ where n is the number of deputies and N_w is the number of words (or nouns, verbs, adjectives, adverbs).

D. Classification

Finally, some standard multiclass classifiers are selected to perform the learning procedure:

1) *Multinomial Logistic Regression*, a generalization of Logistic Regression to Multiclass Problems (4 classes), tuning the regularization parameter;
2) *K-neighbors Classifier*, tuning K (the number of neighbors to consider);
3) *Decision Tree*, tuning the depths of the trees;
4) *Random Forest*, tuning depths and number of trees;
5) *Support Vector Classifier*: support vector machines applied for classification purposes, investigating kernel type and appropriate hyper parameters;
6) *MultiLayer Perceptron Classifier*: feed forward fully connected neural network, tuning simple architectural parameters.

For each one of the features selected and vectorization techniques, the classifiers are trained and the results and collected, fine tuning the necessary hyper parameters.

IV. Evaluation and results

To evaluate the performance of the different methods, k-fold cross validation is performed. The dataset is divided in k subsets and each one of them is iteratively selected as test set, while the others are used to train the models. This technique, then, averages the performances to get a more precise evaluation of the model, since generalization proprieties are considered. This is particularly useful since our dataset consists in only 188 users. We chose $k = 5$ for the whole analysis.

Each method, consisting in a combination of features choice, vectorizer and classifier, is trained, and compared with the other methods. Different metric scores are chosen to obtain accurate insight into the quality of predictions, possibly enabling the observation of biases or other kinds of misclassification issues.

1) $accuracy = \frac{tp+tn}{tp+tn+fp+fn}$
2) $precision = \frac{tp}{tp+fp}$
3) $recall = \frac{tp}{tp+fn}$
4) $f_1 score = 2\frac{precision \times recall}{precision+recall}$

where tp is the number of true positives, tn is the number of true negatives, fp is the number of false positives and fn is the number of false negatives.

Since this is a multiclass classification problem, precision, recall and f1 score are different for each class, and the final value is the average, considering the unbalancement of the number of politicians per party. Thus, for each party p, true positives are deputies belonging to p and actually predicted correctly, false positives are deputies wrongly predicted to belong to p, etc.

features	vectorizer	accuracy	precision	recall	f1 score
nouns	tf-idf L2	0.89	0.91	0.87	0.87
every word	tf L2	0.86	0.86	0.84	0.84
every word	hv L2	0.86	0.87	0.84	0.84
every word	tf-idf L2	0.86	0.87	0.84	0.84
nouns	cv	0.85	0.85	0.84	0.84

TABLE II
EVALUATION METRICS FOR THE FIVE BEST METHODS (FEATURES-VECTORIZER-CLASSIFIER COMBINATIONS) AVERAGED OVER THE FOUR PARTIES CONSIDERED.

A. Results

In this section, results are exposed for different selections of methods, sorted by average accuracy on 5-fold cross validation.

The highest value of accuracy is obtained using only nouns, vectorized with TF-IDF (L2 norm). Both Multinomial Logistic Regression and simple Multilayer Perceptron Classifier obtain an accuracy of 0.89, with similar values of precision, recall and f1 score (see table II).

As reported, classifiers are able to deal with political parties of different sizes, since precision and recall are high and similar.

Similar but lower results are obtained using every tweeted word, obtaining 0.86 accuracy for both Hashing Vectorizer, Term Frequency Vectorizer and TF-IDF with L2 norm. Thus, we can state that cleaning the tweets removing every word that is not a noun increases the performance of the classification. In fact, features like adjectives, verbs or adverbs obtain at best an accuracy respectively of 0.75, 0.65 and 0.50, meaning that they don't contain enough information to perform this kind of classification. These words are not used in a different way by politicians belonging to different parties, on the contrary to nouns.

As expected, TF-IDF is the best vectorizer since it can weight words taking into consideration also if they appear in tweets of other deputies, giving more importance to specific words and penalizing more common words.

K-Neighbors Classifier, Decision Trees, Random Forest and SVC don't perform well enough for this task, often obtaining very low scores for every vectorizer and features selected. Probably a more rigorous fine tuning of parameters can lead to better results, but it is not the scope of this paper.

This analysis proves that politicians belonging to the same party tend to write in the same way. Precisely, the main feature that differentiate between parties are the nouns used. It is important to take into account also the presence of words in other tweets to perform analysis (using TF-IDF vectorizer), and a simple Multinomial Logistic Regression can be trained to obtain good results. We prefer to use the latter classifier since the algorithm is more easily interpretable with respect to a Multilayer Perceptron, with no loss of precision.

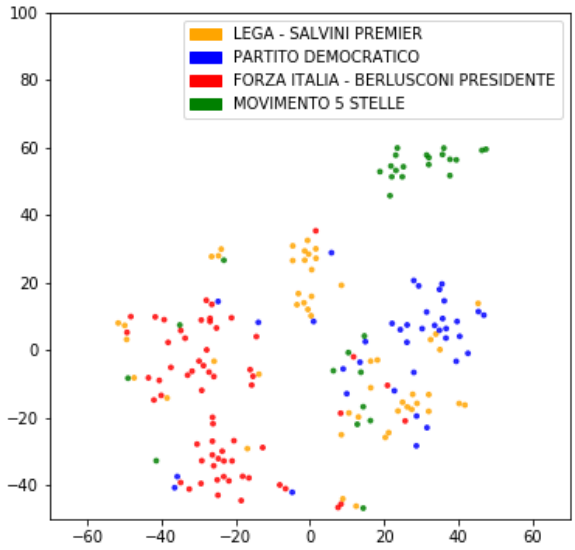

Fig. 1. t-SNE 2d projection of TF-IDF vectors calculated using nouns. Each color represents a different party. Cluster can be easily detected

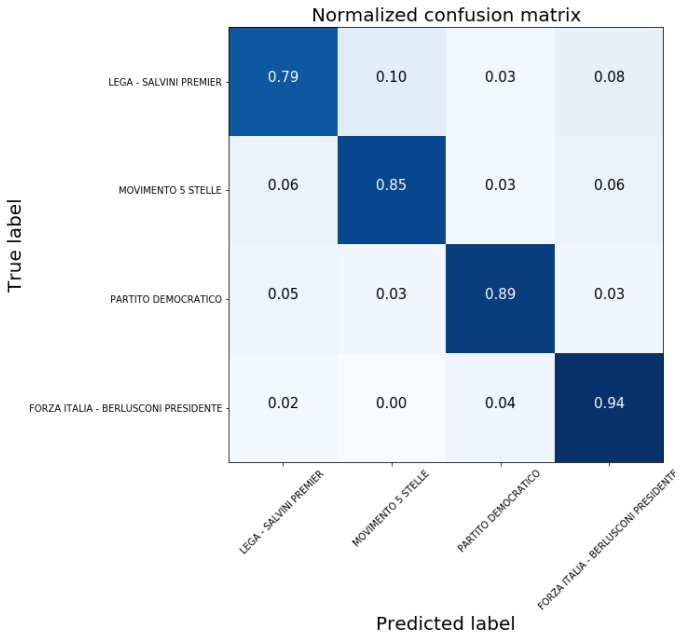

Fig. 2. Confusion matrix for the predictions of the best classifier (nouns with TF-IDF L2 norm, Logistic Regression). The values reported are the mean with relation to the 5-fold cross validation results

V. Further analysis

After finding a good processing pipeline (method) that can classify politicians given their tweeted texts, we continue our analysis inspecting the gathered dataset.

Firstly we perform a TF-IDF transformation with L2 norm on the whole dataset of politicians nouns, to obtain a set of vectors in a about 7900 dimensional space.

In figure 1, the projection of the highly dimensional vectors into a 2 dimensional space has been done using t-distributed stochastic neighbor embedding (t-SNE) as a visualization technique [20]. We can easily notice how three out of four parties are very defined, while politicians belonging ”LEGA - SALVINI PREMIER” are spread almost randomly in the surface. This suggests that that party will be harder to predict since it doesn’t have a specific dictionary of ”preferred” words, as the other parties.

We can verify this looking at figure 2, a normalized confusion matrix that shows insights on the misclassification errors. The party with fewer true positives is in fact ”LEGA - SALVINI PREMIER”, which true deputies are often classified as belonging to ”MOVIMENTO 5 STELLE” (0.10), or as belonging to ”FORZA ITALIA - BERLUSCONI PRESIDENTE” (0.08), suggesting some syntactic relationship between these parties.

However, it is also important to notice that a coefficient like silhouette score [16] calculated using L1 or L2 metric, applied to the dataset vectorized with TF-IDF with L2 norm, has low value (0.01), indicating that this kind of high dimensional vectors does not form compact and separate clusters, since, of course, users are not using a complete set of different nouns one with respect to the other. The great part of them is in common, while just a few terms are decisive for classification purposes.

We now shift our focus on decisive nouns present in the tweets. In table III, nouns, whose coefficient of Multinomial Logistic Regression is higher/lower, are listed for every party. For example, the word ”centrodestra” (centre-right) is the most significant noun for people belonging to ”FORZA ITALIA - BERLUSCONI PRESIDENTE”, meaning that a higher value of TF-IDF of this noun in the tweets of an account will bring the classifier to lean with that party as the most probable one for it, while the noun ”lega”, being in the last place, will have the opposite role. As expected, nouns of the parties, like ”movimento” (movement) and ”stella” (star) for ”MOVIMENTO 5 STELLE” are in the first positions, while they are in the last positions for others parties, suggesting that politicians tend to talk mostly about their parties. Interesting is also the presence of nouns like ”cittadino” (citizen) and ”gente” (people) for populist parties in the first positions, while for other parties they are in last positions. Finally, words like ”nord” (north) and ”sud” (south) can suggest a particular focus of the selected party with relation to the Italian geographical region, obtaining a hint on where the political interests of the parties are.

Finally, a simple topic detection algorithm is applied to the data to get further insights into what the tweets are about. We selected LSA method [10], approximatively decomposing the TD-IDF matrix X (number of deputies times number of nouns) obtained before into the product of three matrices, U (number of deputies times number of topics), S (a diagonal

	FI	Lega	M5S	PD
1	centre-right	lega	star	minister
2	president	people	movement	commitment
3	south	gazebo	citizen	suburbs
4	govern	north	change	thing
5	retired	right	spokesman	comparison
7834	thing	courtroom	family	citizen
7835	change	law	centre-right	people
7836	citizen	star	left	star
7837	star	president	minister	movement
7838	lega	govern	lega	centre-left

TABLE III
MOST RELEVANT WORDS (TRANSLATED FROM ITALIAN) PER PARTY AS INDICATED BY THE COEFFICIENTS OF THE LINEAR REGRESSION MODEL. THE UPPER HALF OF THE TABLE REPORTS NOUNS THAT SUGGEST THE BELONGING TO THE PARTY, THE LOWER HALF THE OPPOSITE

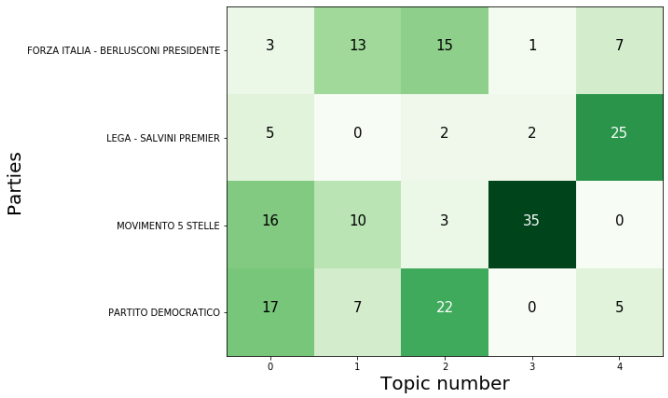

Fig. 3. Automatic detection of 5 topics from corpus of tweets by political party

matrix of length number of topics with sorted eigenvalues) and V (number of nouns times number of topics). The S matrix describes how much the topics are important, while U contains information on how the deputies are related to the topics, and V groups the nouns into different topics. Thus, observing this decomposition we can obtain information about how different parties are related to different interests.

We chose a number of topics of 5, and we decompose the TF-IDF matrix as described above. Inspecting matrix U, we can chose the most relevant topic per deputy. In figure 3 we show the results. Interesting how topic 4 is dominated by "LEGA - SALVINI PREMIER", while "MOVIMENTO 5 STELLE" is more focused on topic 3. The other two parties are more balanced between 2 topics. Analyzing which nouns characterize the topics through matrix V we notice that topic 4 is composed of "moschea" (mosque), "immigrato" (immigrant), "festa" (party), "gazebo" (gazebo), while topic 3 by "cittadino" (citizen), "video" (video), "appuntamento" (appointment), reflecting as expected the political inclination of those parties. Of course most of the words that characterize the topics are politics related, such as "legge" (law), "camera" (chamber), "ministro" (minister), "governo" (government), since the main topic is of course politics, but still we are able to identify subtopics highly related to the most characterizing ideas of the parties.

VI. CONCLUSION

In this paper, we investigated how natural language processing tools can be applied to obtain insights about the political situation in Italy. Of course, once a good language-specific word tagger is obtained, the same analysis can be repeated for any country, with the only requisite that a social network is constantly used to political discussions and propaganda. Results show that our hypothesis is true: deputies belonging to the same party use the same words (in particular, nouns) when tweeting. This fact can be used to classify of accounts obtaining good results, once the right vectorization has been selected.

Once the texts are converted into vectors, any kind of analysis can be performed to obtain meaningful insights of the political situation, ranging from the most/least important words for each party to the visual projection of the vectors into bi-dimensional spaces, e.g. for inspecting the cohesion of the parties. Outliers could be easily detected and inspected; misclassification could hint to the fact that the ideas of the misclassified user are not coherent with the dominant ones in the classified party – this analysis should be conducted with care and is subject to restrictions, as it could violate the right to individual privacy.

Future works will focus on a deeper analysis of content analysis for politics, with the objective of using this method for knowledge extraction (i.e. extracting the accounts which make most use of a given vocabulary). We will also use content-based methods for community detection; we also expect that mixed methods, using both content and network analysis, may be more effective than current methods, which typically highlight just the use of social connections between accounts.

Our classification algorithm will also be tested on Italian accounts of non politicians, in order to obtain prediction of elections by classifying users' vocabulary; this could be powerful tool when compared with expensive and often biased classical methods for political surveys. Awareness of the vocabulary being used within a party could also give insights to politicians about terms that are most expected by the people whom they address.

REFERENCES

[1] David Anuta, Josh Churchin, and Jiebo Luo. Election bias: Comparing polls and twitter in the 2016 U.S. election. *CoRR*, 2017.

[2] Sitaram Asur and Bernardo Huberman. Predicting the future with social media. 1, 03 2010.

[3] Johan Bollen, Huina Mao, and Alberto Pepe. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *CoRR*, 2011.

[4] Marco Brambilla, Stefano Ceri, Florian Daniel, Marco Di Giovanni, Andrea Mauri, and Giorgia Ramponi. Iterative knowledge extraction from social networks. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, pages 1359–1364, Republic and Canton of Geneva, Switzerland, 2018. International World Wide Web Conferences Steering Committee.

[5] Marco Brambilla, Stefano Ceri, Emanuele Della Valle, Riccardo Volonterio, and Felix Xavier Acero Salazar. Extracting emerging knowledge from social media. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pages 795–804, Republic and Canton of Geneva, Switzerland, 2017. International World Wide Web Conferences Steering Committee.

[6] Julio Cesar Amador Diaz Lopez, Sofia Collignon-Delmar, Kenneth Benoit, and Akitaka Matsuo. Predicting the brexit vote by tracking and classifying public opinion using twitter data. 8, 01 2017.

[7] Elanor Colleoni, Alessandro Rozza, and Adam Arvidsson. Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data. 64, 03 2014.

[8] M. D. Conover, B. Goncalves, J. Ratkiewicz, A. Flammini, and F. Menczer. Predicting the political alignment of twitter users. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 192–199, Oct 2011.

[9] Joseph DiGrazia, Karissa Mckelvey, Johan Bollen, and Fabio Rojas. More tweets, more votes: Social media as a quantitative indicator of political behavior. 02 2013.

[10] Susan T. Dumais. Latent semantic analysis. *Annual Review of Information Science and Technology*, 38(1):188–230.

[11] Fabio Franch. (wisdom of the crowds)2: 2010 uk election prediction with social media. *Journal of Information Technology & Politics*, 10(1):57–71, 2013.

[12] F. Galton. Vox populi. *Nature*, 75(1949):7, 1907.

[13] Daniel Gayo-Avello, Panagiotis Takis Metaxas, and Eni Mustafaraj. Limits of electoral predictions using twitter. In *ICWSM*, 2011.

[14] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. From tweets to polls: Linking text sentiment to public opinion time series, 01 2010.

[15] Anand Rajaraman and Jeffrey David Ullman. *Mining of Massive Datasets*. Cambridge University Press, New York, NY, USA, 2011.

[16] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53 – 65, 1987.

[17] Helmut Schmid. Improvements in part-of-speech tagging with an application to german. In *In Proceedings of the ACL SIGDAT-Workshop*, pages 47–50, 1995.

[18] Lei Shi. Predicting us primary elections with twitter. 2012.

[19] Andranik Tumasjan, Timm Oliver Sprenger, Philipp Sandner, and Isabell Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment, 01 2010.

[20] L.J.P. van der Maaten and G.E. Hinton. Visualizing high-dimensional data using t-sne. 2008.