Can Social Network Influence Voters?

Gilberto Flores, Ana Lorena, Claudio Penteado, Carlos Kamienski

Universidade Federal do ABC (UFABC)
Avenida dos Estados, 5001 – 09.210-580 – Santo André – SP – Brasil
{gilberto.pochet, cak,claudio.penteado}@ufabc.edu.br,
aclorena@unifesp.br

Abstract. *Online Social Networks (OSN) has been used for different activities of human life in the last years. In the same way, political campaigns have increasingly exploring this new digital space so that candidates and their supporters can post their ideas and proposals directly to prospective voters. This paper presents a text mining analysis of the campaign for mayor of Sao Paulo (Brazil) looking at messages and comments of candidates, supporters and their friends in Facebook. We chose public data from 500 friends, and their friends, of the two best ranked candidates in the opinion polls for 75 days. Our results reveal very interesting behavior of supporters and voters, also using two friends of each candidate whom were most active in posting endorsement messages, for further comparison. We consider this an important evidence of the spread of influence in a social network.*

1. Introduction

The very concept of friend is changing and frequently people have thousands of “friends” and share sensitive information with all of them that otherwise would be kept private inside an OSN. The same movement happens in political campaigns, where candidates have they public pages with friends and/or followers and they post many activities held there. In this new world, candidates and their supporters can post their ideas and proposals directly to prospective voters and follow their reaction in real time. Ideas, thoughts, expertise and knowledge are every day posted in social networks by millions of people and having access to this valuable information can make a difference in understanding better the world. Candidates, for example, should have interest in understanding their prospective voters. However, first of all we need to have access to the information in a timely manner, because otherwise we cannot benefit from it [Borgatti 2003].

In this paper we present a text mining analysis of the campaign for mayor of São Paulo (Brazil) looking at messages and comments of the two best ranked candidates just in the first round of the elections, their supporters and their friends in Facebook. We crawled public data from 500 friends, and of their second level friends, of the two best ranked candidates in the opinion polls for 75 days. Please notice that we only used public data, since privacy is a major concern of many users of Facebook. Having this in mind, we collected information from users with open profiles, timeline option activated and that have messages in their walls, only reaching 500 friend/users for each candidate whom apply for those conditions. We developed a crawler for automating the information extraction using the Selenium tool[1]. The collected data was then pre-

[1] http://seleniumhq.org

processed using information retrieval and text mining techniques. Afterwards we applied the K-means machine learning algorithm for grouping the main terms posted by users about the candidates. In the end, for obtaining insights from the clusters of words we analyzed individually a set of messages, looked up information in the Internet and faced it with the view of an expert (a social scientist who is co-author of this work).

Our results provide interesting insights about how information and influence is spread through a social network for election purposes. We observed similar patterns of behavior by analyzing the messages and comments of 500 friends of two candidates and comparing with the friends of their two most active friends.

The remainder of this paper is organized as follows. Section 2 presents related work and section 3 explains the methodology we adopted. The main results are presented in Section 4 and discussed in Section 5. Finally, section 5 presents some concluding remarks.

2. Related Work

There are some relevant previous works that influenced this paper. [Catanese 2011] presents results of crawling Facebook for finding out how to map a network of friends of friends, extracting users ID and their friend list, then grouping them and observing their relationships. A different approach was taken in [Ellison 2007], aimed at observing how Facebook can create social capitals for the users, and how its connections can give to some people influence on their relationships. [Smets et al. 2008] employs machine learning and text mining techniques in online social networks for commercial purposes such as avoiding vandalism. Finally, [Gerrish & Blei 2011] used a different machine learning technique (called Topic Model) and focuses in the political area and extraction of key information for futures elections. None of them is similar to our work, that presents a new methodology and new results.

3. Methodology

3.1. Data Collection and Pre-processing

For extracting the information from the users in Facebook, we employed the Selenium tool. Selenium helps to replicate and store human actions inside a web page. Its main goal is to support the automatic extraction of information from Facebook users' wall [seleniumhq]. For analyzing text structures with Machine Learning techniques such as k-means, we first need to use Text Mining. The following Text mining techniques are used: Tokenizing**,** Stop List, Stemming and Bag of Words.

Our machine learning technique (K-means) works by choosing k random centers from data and iteratively adapting those centers in order to reflect the intrinsic group structure on data [Teknomo 2007]. For choosing the appropriate number of clusters k in data, we compared the sum of the squared error (SSE) [Peeples]. This method gives us a guide to choose the number of clusters present on data, by observing where those "elbows" in the SSE graphics are located. Figure 1 shows an elbow for six clusters, such that we can infer that data is consistent with six groups.

3.2. Text Mining Methodology

The experimental methodology adopted in this work is divided into three main steps, for understanding the influence spreading process in an OSN. The first step we developed a Facebook crawler based on Java and Selenium. Our tool is extracting all the messages and comments referring to the candidate, using as reference their surnames (as both are publicly known) appear, i.e., “Russomanno” and “Serra”. This process is made for 500 friends of each candidate from July, 1st 2012 to September, 24th 2012. Once this process is finished, we select two friends per candidate who were more active in posting messages/comments among the 500 friends. For each of them, our program starts the same Facebook crawling process over again collecting their messages and comments. With these measures we were able to compare the behavior patterns of the candidate’s first level connections (friends) with their second level ones (friends of their friends).

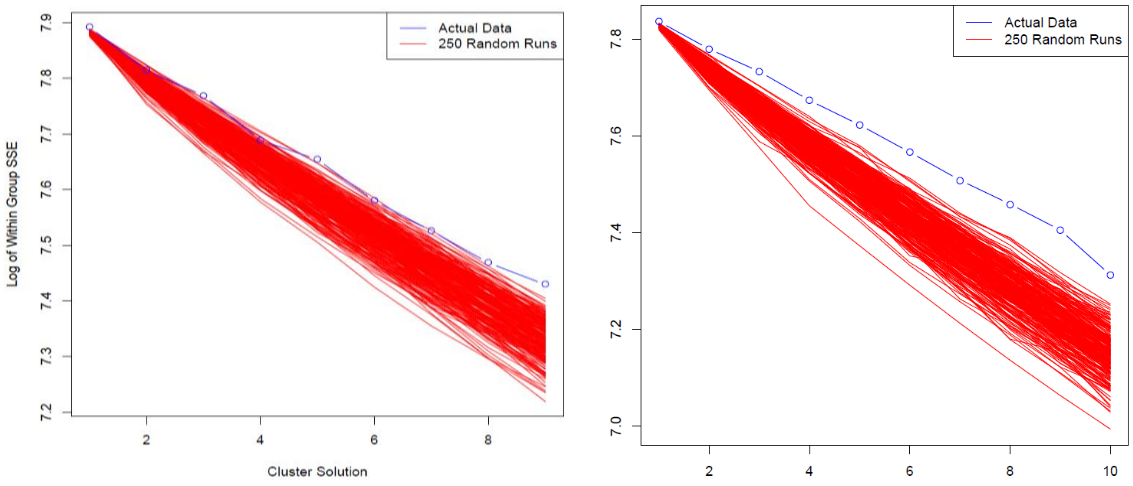
The second step is text mining, where data are pre-processed and converted into a structured format. In this step are applied text pre-processing steps described in Section 2 (tokenizing, stop list, stemming, creation of a bag of words and matrix of terms and weighting each term). The last step is the execution of the K-means clustering algorithm and the choice of the most appropriate number of clusters for each case, based on the presence of an elbow in the SSE results, as shown in Figure 1. K-means was run for both candidates and their two selected friends, summing up six Facebook users. Candidates are identified by their surnames (Russomanno and Serra) and their friends by Russomano-Friend-1, Russomano-Friend-2, Serra-Friend-1 and Serra-Friend-2.

An additional non-computational step is the analysis of the group of words within each cluster for both candidates and their two friends, summing up six sets of results from the K-means processing.

4. Results

4.1. Number of Clusters

Figure 1 determined the best number of clusters for each analysis. In both pictures, we can observe that a small elbow is formed between two and four clusters. After analyzing the results with two, three and four cluster for all users, we decided that option is three clusters and for one user (Serra-Friend-2) the best option is two clusters.



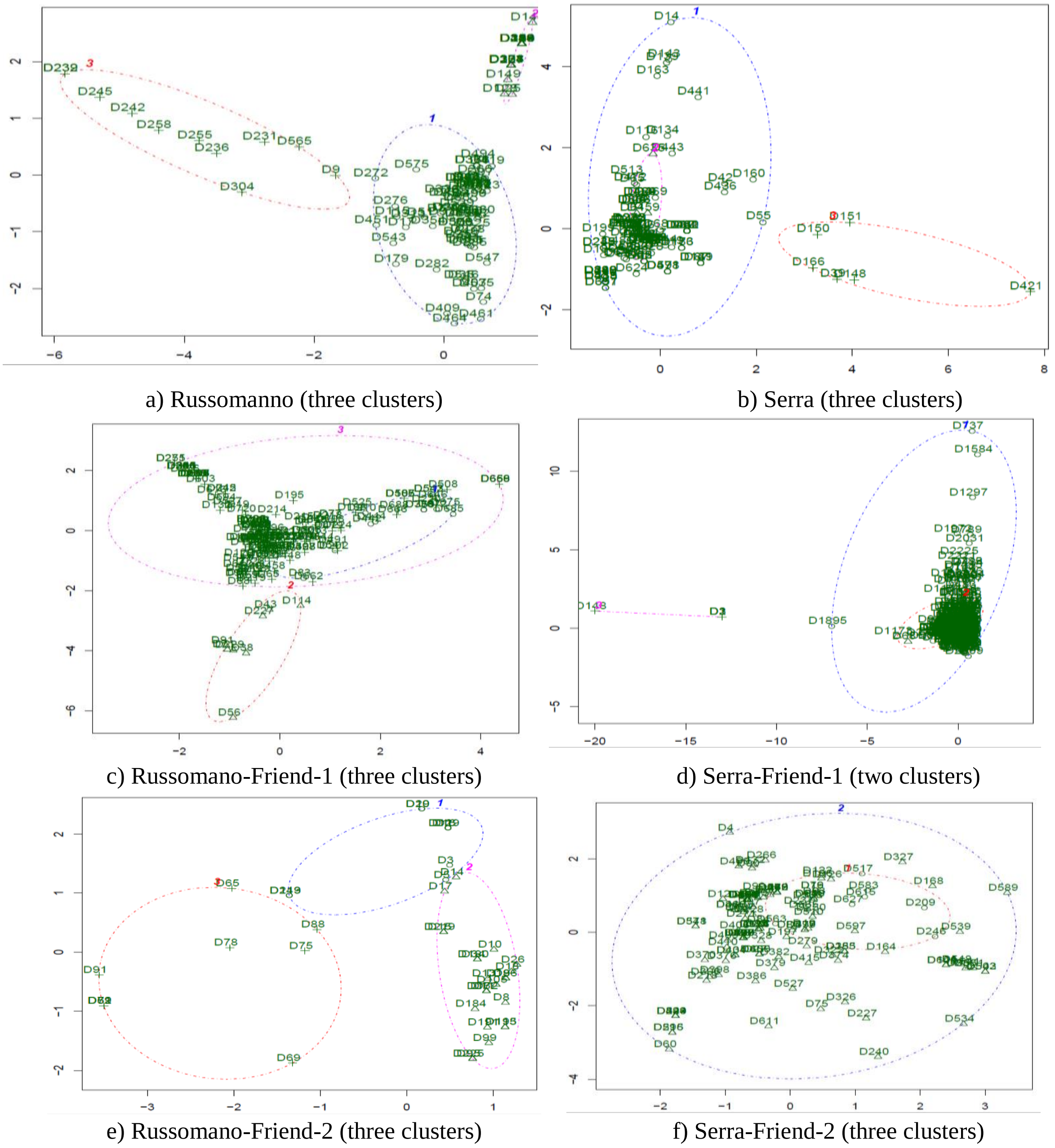
a) Russamono **b) Serra**

Figure 1 – SSE elbow for candidates Russomanno and Serra

We selected the most significant group of words for our analysis, as far as their frequency is concerned, and we call them as “Main Words”. Afterwards, we analyzed some sample messages to obtain the right understanding of their real meaning and usage for these Main Words. As we know to interpret each group of Main Words is common to use an expert, this person has the knowledge to find a deep meaning for the sequence of words.

4.2. Clustering contents

Figure 2(a) presents the three clusters for Russomanno friends. The Main Words for each cluster are the following: 1.vote, mayor, São Paulo, candidate, go, car parade, talk, people. 2. God (Assemblies God Church), good, day (car parade on Saturday). 3. city hall, city, sp, course.



a) Russomanno (three clusters)

b) Serra (three clusters)

c) Russomano-Friend-1 (three clusters)

d) Serra-Friend-1 (two clusters)

e) Russomano-Friend-2 (three clusters)

f) Serra-Friend-2 (three clusters)

Figure 2 – Clusters for candidate Serra and Russomanno and respective friends

The first cluster contains words that refer to how the candidate approaches his prospective voters. For example the term “car parade”, translated from the word “carreata” is a common campaign strategy in Brazil, where the candidate and his supporters present themselves to the population. The second cluster reveals a much more interesting behavior of the candidate Russomanno, who is frequently alleged of using religion in his campaign. Even though he denies that evangelical churches are promoting him, our analysis suggests the opposite. For example, many users refer to “the day”, which seems to be an important car parade on a Saturday supported by the Assemblies of God Church. The third cluster is about the intentions of the candidate and his supporter of making his course to the city hall.

The main words for Russomano-Friend-1´s friends (443 friends), presented in Figure 2(c), are the following: 1.Sao Paulo, mayor, investigation, new, history. 2. Take, come, turn, vote. 3. vote, good, day, city hall, friend, church. The first cluster gives the idea of achieving a new history for São Paulo and indeed Russomanno is quite a novel candidate. The second cluster expresses the ideas of voting and the chance for the candidate to be elected. And the third cluster is about religion and refers to the “day” which also points out to the important car parade on a Saturday.

For Russomano-Friend-2 the Main Words are: 1. city hall, city, Sao Paulo. 2. team, city, people, vote. 3.course, sp, mayor. Clusters 1 and 3 deal with the candidate motivation to be the São Paulo mayor. And for cluster 2 the most important is the advertising through the words “team”, “people” and “vote”. It is also worth mentioning that the word “team” is used with high frequency when Facebook messages come from Twitter. However, since this user has only 53 available friends we did not find a large number of messages to extract strong ideas from the clusters.

Figures 2(b), (d) and (f) contain cluster information for candidate Serra and users Serra-Friend-1 and Serra-Friend-2. The first difference from candidate Russomanno is the smaller size of the dataset collected from Facebook. Another problem is the excessive noise in the text, since the word “Serra” is a common surname in Brazil, but also means saw or mountain range in Portuguese, which generates useless information in some cases. The Main Words for each cluster are the following: 1. Sao Paulo, Haddad, candidate, vote, government, Brazil. 2. Sao Paulo, speak. 3. Soninha, investigations, points, Haddad.

The first cluster tells us about the candidate and his intention to be the mayor of São Paulo. Also, a very important piece of information is the word Haddad, a rival candidate and third ranked in opinion polls. The second cluster does not provide any insightful result, just a generic mention of São Paulo. The third cluster also is very meaningful because it mentions another candidate, Soninha Francine, who is frequently associated to Serra in some Internet rumors. Also, it mentions Haddad again and mentions the variations, up and down, of percentage points in opinion polls.

The analysis of Serra-Friend-1 user contains information taken from 824 friends, and the Main Words are: 1. city, Sao Paulo, vote, friends, candidate, truth, mayor. 2. commemorate. The second cluster of the candidate Serra is referring to the idea of being the mayor of São Paulo, but the friends of Serra-Friend-1 posted messages with the word truth. We found messages and comments of users that do not believe that candidates tell the truth. Also, it refers to the fact that Serra was a former elected mayor of the city and he left his mandate in the middle for running for governor of the state. The second cluster has the word “commemorate”, used in some messages for the activity of planting trees on São Paulo. The third cluster is not presented here because it contained too much noise with the word “Serra” used for other meanings.

The last results are user Serra-Friend-2 (470 friends), which give us the following Main Words: 1. vote, lies, sp, time. 2. vote, candidate, Soninha, Haddad, marijuana, campaign, questions. The first cluster discusses the idea of voting and how candidates in general lie in São Paulo, and the word time is used frequently for expressing the idea of “it is time of change and vote”. The second cluster again reveals information about the rivals of the candidate Serra that appeared before in the messages of the friends of Serra. Also, these users focus on the issue of legalizing the use of marijuana, which seems to be an idea of Serra's party, and the word "question" express how suspicious most users are with this proposal.

5. Discussion

Initially our research has a main problem to solve, how a political expert group or person can analyze a political campaign, and understand the spread of influence through the candidates' social network (Facebook), in short time and with scarce resources. Our way to innovate and find a solution was to use a combination of elements (Facebook, automatization, machine learning, politics), each element with their difficulty, our subject OSN (Facebook) was used easily thanks to the automatization with Selenium otherwise is not common to get text user information, other [Catanese 2011] works just use limited information from Facebook API and others works [Andranik Tumasjan, 2010] prefer to use a OSN more open as Twitter. Also we are saving a lot of time for the analysis, avoiding analyzing manually each text from each user in a political campaign. We observed similar patterns of behavior by analyzing the messages and comments of 500 friends of two candidates and comparing with the friends of his two most active friends. Furthermore, we confirmed how a social network can be used for influence ideas easily, cheaply and without missing the focus.

6. Conclusion

We consider this point an important evidence of the spread of influence in a social network. A very important finding when comparing the messages posted by Russomanno's friends with the friends of his two most active friends reveals a similar behavior. This encourages us to think that his two friends do have an influence on their respective friends, and that this behavior of supporters and voters can be spread down to further levels of the social network.

References

Borgatti, S. P., Cross, R. (2003), "A Relational View of Information Seeking and Learning in Social Networks", Management Science, 49(4), pp. 432-445, April 2003.

Catanese, S. A. et al. (2011), "Crawling Facebook for Social Network Analysis Purposes", ACM WIMS'11, May 2011.

Ellison, N. B., Steinfield, C., Lampe, C. (2007), "The Benefits of Facebook "Friends:" Social Capital and College Students' Use of Online Social Network Sites".

Smets, K., Goethals, B., Verdonk, B. (2008), "Automatic Vandalism Detection in Wikipedia: Towards a Machine Learning Approch", AAAI Workshop, 2008.

Peeples, M. A., "R Script for K-Means Cluster Analysis", http://www.mattpeeples.net/kmeans.html

Teknomo, K. (2007). K-Means Clustering Tutorial. Retrieved from http://staff.pradnya.ac.id/luqman/pages/K%20mean%20Clustering1.pdf

seleniumhq. (n.d.). Retrieved April 2012 from http://seleniumhq.org

Gerrish, S. M., Blei, D. M. (2011). "Predicting Legislative Roll Calls from Text", International Conference on Machine Learning, 2011.

Tumasjan, A, et. al (2010), "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment", ICWSM 2010.