Brazilians Divided: Political Protests as Told by Twitter

Cássia de Souza Carvalho[1], Fabrício Olivetti de França[1,3](✉), Denise Hideko Goya[1,3], and Claudio Luis de Camargo Penteado[2,3]

[1] Center of Mathematics, Computing and Cognition (CMCC),
Federal University of ABC (UFABC), Santo André, SP, Brazil
cassia.carvalho@aluno.ufabc.edu.br,
{folivetti,denise.goya}@ufabc.edu.br
[2] Center of Engineering, Modeling and Applied Social Sciences (CECS),
Federal University of ABC (UFABC), São Bernardo do Campo, Brazil
claudio.penteado@ufabc.edu.br
[3] Nuvem Research Strategic Unit, Santo André, Brazil

Abstract. After a fierce presidential election campaign in 2014, the reelected president Dilma Rousseff became a target of protests in 2015 asking for her impeachment. This sentiment of dissatisfaction was fomented by the tight results between the two favorite runners-up and the accusations of corruption in the media. Two main protests in March were organized and largely reported with the use of Social Networks like Twitter: one pro-government and other against it, separated by two days. In this work, we apply two supervised learning algorithms to automatically classify tweets during the protests and to perform an exploratory analysis to acquire insights of their inner divisions and their dynamics. Furthermore, we can identify a slightly different behavior from both parts: while the pro-government users criticized the opposing arguments prior the event, the group against the government generated attacked during different times, as a response to supporters of government.

1 Introduction

In democratic elections, whenever the results are tight, the competing sides tend to express a negative sentiment towards each other, inciting a polarization among people. When this sentiment is accompanied by doubts about the legitimacy of voting system, it may influence a wave of protests and calls for a change of rules.

This situation occurred in the Brazilian presidential election of 2014, in which the two main candidates, Dilma Rousseff, representing the Workers' Party, and Party, and Aécio Neves, representing the Brazilian Social Democracy Party, obtained a result of 51.64 % and 48.36 % of votes respectively. These results, together with the spread of news about internal corruption in one of the largest semi-public multinational corporation, influenced the people from the opposing side to organize a series of protests.

A. Hameurlain et al. (Eds.): TLDKS XXVII, LNCS 9860, pp. 1–18, 2016.
DOI: 10.1007/978-3-662-53416-8_1

These protests occurred inside their homes, on the streets [21] and throughout the two main social networks: Facebook[1] and Twitter[2]. These Social Networks played an important role for the organization and discussions of such protests.

With the widespread use of the Social Networks, it is possible to extract different information about these events. For the government and opposition sides, it is important to know who are the main actors of these events, the overall sentiments, the demands and the different parts that gathered for a common goal.

In this paper, we apply two classification algorithms [2] to determine the overall sentiment of the protesters on the events that occurred during the period of 13th and 15th of March 2015. The first event (13th of March) was organized by pro-government groups, while the second (15th of March) was organized by groups against government. We explore what information we can infer from the classes by plotting the temporal relations. Despite the usual literature on Sentiment Mining [9], we will label the sentiments pro or against the government.

The paper is organized as follows: In Sect. 2 we contextualize these two political protests to better understand the overall sentiment of both sides. In Sect. 3 we explain the two classification algorithms used in this work: Naive Bayes [12] and Support Vector Machine [17], as well briefly summarize some works found in the literature of twitter sentiment analysis, particularly focusing on political context. In Sect. 4 we explain the methodology and apply these two algorithms in our collected dataset and to analyze the information that can be extracted from the results. Finally, in Sect. 5 we conclude this paper with some insights for future work.

2 Brazilian Political Protests

After a polarized campaign between the two candidates, the president Dilma Rousseff was re-elected as President of Brazil by a small margin of votes, 3,459,963 (roughly 3.28 % of the electors). The presidential campaign of 2014 was marked by intense debates between the candidates since the first round, motivating supporters and militants to produce favorable information for their candidates in the Internet Social Networks.

Disagreeing with the loss of the candidate Aécio Neves, their supporters and groups opposed to the Workers' Party manifested their unhappiness on the Internet, maintaining an intense online political mobilization. As a result from this articulation, groups against the government organized via digital media (Facebook, Twitter, WhatsApp[3]) a protest that was known as *Panelaço* (pan beating). During the initial statement of president Dilma Rousseff in national broadcast on 8th of March 2015, the protesters beat pans and swore the president and her party.

On 15th of March 2015 took place the first and largest manifestation against Dilma Rousseff, in several different cities, asking for her impeachment.

[1] https://www.facebook.com.

[2] https://www.twitter.com.

[3] https://web.whatsapp.com/.

These manifestations united on Brazilian streets millions of people, dissatisfied with the current management of the country, inflation of prices and corruption reports, chiefly in Petrobras.

On the other hand, supporters of the government decided for a counterattack. A mobilization was organized by union and social movements on 13th of March 2015. Besides occupying the streets, the political debate also occurred on the Internet.

The government supporters accused the traditional mass media of diminishing the importance of pro-government protests on news, while giving a wide coverage on protests of opposition, notably Rede Globo TV Channel, the most popular and influential media group in Brazil.

Virtual militants and connected citizen have continued the political debate in cyberspace. After the mobilization studied in this paper, there were two others great protests against the Workers' Party, on 12th of April 2015 and 17th of May 2015 (this last one with a smaller adhesion).

3 Supervised Learning

In Machine Learning, Supervised Learning [18] refers to the set of algorithms and methods that learns a function $y = f(x)$ where x is the object of study and y is a predicted value. This is performed by feeding the algorithm with a set X of object examples, associated with the expected output given by a set Y. The algorithm creates a mapping from the observed data, being capable of inferring any new object, already observed or not.

There are many algorithms created for this task, with different characteristics and capable of handling different types of variables. In this work, we will use two well-known techniques: Naive Bayes [12], a technique known for its good trade-off of performance and simplicity; and Support Vector Machine [17], a state-of-the-art algorithm for many classification problems and datasets, but with the need of more specific adjustments.

In the following sub-sections we will briefly explain these techniques.

3.1 Naive Bayes

Naive Bayes is a non-parametric probabilistic algorithm, often used for classification of categorical data [3] and text mining [6]. This algorithm assumes that the variables describing the objects of study are independent from each other regarding their classification, thus making use of the Bayes Theorem. With this strong assumption, we can use the Bayes Theorem described as:

$$p(c|X) = \frac{p(c)p(X|c)}{p(X)}, \tag{1}$$

where X is the feature set describing the object and c is the class to which it belongs.

From a training data, it is easy to estimate $p(c)$ as the proportion of objects classified as c. The estimation of $p(X|c)$ and $p(X)$ makes use of the independence assumption as:

$$p(X) = p(x_1) \cdot p(x_2) \cdots p(x_n), \tag{2}$$

and

$$p(X|c) = p(x_1|c) \cdot p(x_2|c) \cdots p(x_n|c). \tag{3}$$

After estimating all of these probabilities, a new object can be classified by finding the class c which gives the maximum probability given the features of the object.

3.2 Support Vector Machine

The Support Vector Machine (SVM) is a technique that extends the linear regression model to alleviate two problems: (i) the assumption that the data is linearly separable and; (ii) the over-fitting of the training data.

For the first problem, the first and simpler assumption during the classification task is that the objects are linearly separable, i.e., the objects of different classes can be separated with a simple line equation. But in practice, this assumption rarely holds, so a new set of features should be crafted or learned as a non-linear combination of the original features set. With this transformation, it is expected that the new features set resides on a linearly separable space, but this adds the cost of transforming to every new object to be classified. In SVM, the idea of a Kernel function was introduced to alleviate this problem [5,15].

A Kernel function $k(x, y)$ takes as input two objects described by their original features set and calculates the distance between them in a different space chosen by the function being used. This calculation is performed without explicitly transforming the feature space, thus having an efficient computational cost. The main Kernel functions used on the literature are Linear Kernel, Polynomial Kernel and RBF Kernel, the last two non-linear.

The second problem, regarding the over-fitting, is alleviated by changing the objective-function of the separation line. In Linear Regression, the objective is to find the separation line which gives the minimum error regarding the training data. In SVM, the objective-function is the maximization of the margin enveloping the separation line. In other words, the algorithm seeks a separation line that has a maximum distance from the closest points of each class.

By maximizing this margin, not only the classification error for the training data is minimized, but also it keeps some space for generalization of unseen data.

3.3 Related Work

It is well know the usage of SVM and Naive Bayes as text classifiers, and recently applied to Twitter corpora and other micro-blogging platforms [1,8,14]. In particular, we briefly summarize some studies that utilized tweets as a source of public opinion manifestations.

In the context of political sentiment mining on Social Networks, Spaiser et al. [16] applied statistical and machine learning techniques to almost $700,000$ tweets, being able to observe how they had contributed to weaken Russian protest movements.

Livne et al. [10] collected tweets from US House and Senate candidates, applied text mining using a bag-of-words model, conducted graph analysis to estimate co-alliances and divergence among candidates and generated a predictive model for a certain candidate win or lose the election.

Lotan et al. [11] analyzed the Tunisian and Egyptian Revolutions as told by Twitter, identifying the main actors of the online manifestations and flow of information.

Turkmen et al. [19] collected and labeled tweets during recent Turkey protests and used SVM and Random Forest classifier to predict political tendencies in the messages.

4 Experiments

In this section, we give a complete description of data acquisition, methodology and analysis of a real-life event on the Twitter Social Network.

4.1 Methodology

During the period of 12th to 16th of March 2015, we collected the tweets with hashtags related to both protests (see Table 1) by using the Twitter API[4] with the streaming interface that continuously collects tweets in real time. After the data collection, we ended up with 274,645 tweets from 101,452 different users.

We added the tweets published on 13th of March of 2015 in one dataset (PROGOV) and those published on 15th of March of 2015 in another dataset (CONGOV). From these two datasets we extracted the bag-of-words model, transforming the features by using tf-idf (frequency inverse document frequency) [4].

For the classification task, we randomly picked 100 tweets from each dataset, 50 for each sentiment [5], and fitted this data using both classification algorithms. After that, another 100 tweets were chosen at random and classified using these models. If the classification accuracy (percentage of correct classification) were below 70 %, these 100 tweets were added to the training data, and the process repeated until the accuracy levels reached 70 % or more on the random data. This threshold is a compromise of the reported accuracy of the literature [1,8,14] that range between as low as 60 % and as high as 85 %.

After that, we classified the entire dataset and performed some exploratory analysis to extract information about the protests dynamics. A summary of the datasets characteristics is depicted in Table 2.

[4] https://dev.twitter.com/.

[5] We are aware that this dataset is possibly unbalanced, but to know the exact balance would imply a large quantity of manual classification.

Table 1. Hashtags used during the data collecting stage.

Hashtag	Meaning
#13Marco	Date of the protest supporting the government
#AcordaBrasil	Wake-up Brazil
#DilmaNaoMeRepresenta	Dilma (elected president) does not represent me
#DilmaVaiada	Dilma booed
#ForaDilma	Go away Dilma
#ForaPT	Go away PT (Workers' Party)
#ImpeachmentDilma	Impeachment of president Dilma
#PetrobrasEhBrasil13	Petrobras (Brazilian oil company) belongs to Brazil (supporters of the gov.)
#PronunciamentoDaDilma	Speech of president Dilma
#SouPetrobras	I am Petrobras (supporters)
#TodosContraOGolpe	All against the coup d'état
#VamosVaiarDilmaNaTV	Let us shout down Dilma on TV
#VemPraRua15DeMarco	Let us go to the streets on March, 15th
#br45ilnocorrupt	No corruption in Brazil (with a pun with the code 45 of the opposition party)
#globogolpista	Coup-backer Globo (Globo is one of the largest TV Station in Brazil)
#protestos	protests

Table 2. Summary of studied datasets.

Dataset	# of tweets	Unique words
PROGOV	84, 821	36, 070
CONGOV	189, 824	60, 684

In the next subsections we will present just the main results in order to preserve clarity and brevity of this paper. The full set of results with the corresponding IPython Notebooks will be made available at https://github.com/folivetti/POLITICS.

4.2 Classification Results

After sampling 100 tweets from the datasets and manually labeling them as PRO or CON, as in pro-government and against it respectively, we trained the Naive Bayes and SVM algorithms with these sampled tweets, and applied the classification process for the entire data set. After this first step, we sampled another batch of 100 tweets from the classified results of each algorithms.

In order to use a diversified set, without a bias towards one class, we have used the Reservoir Sampling technique [20] that samples items with equal probability from a large set. The algorithm is briefly described in Algorithm 1.

Algorithm 1. Reservoir Sampling.

```
input  : Data stream $D$, number of samples $k$.
output: Sampled data $S$

$S \leftarrow \varnothing$
for $sample \in D$ do
    if $sample.index <= k$ then
        $S.append(sample)$
    else if $r\ U(0, k) < k$ then
        $S[r] \leftarrow sample$
```

The algorithm starts by inserting the first k samples into the sampled data set. After that point, every subsequent data can replace a given sample, chosen randomly by an uniform distribution ($r\ U(0, k)$), with probability $1/k$.

After the sampling process, we manually verified the classes of data to estimate the accuracy of both classifiers.

As we can see from the Truth Tables in Tables 3 and 4, both classifiers had similar results, with an accuracy around 90 %. Although this may not be statistically significant for the whole dataset, the intention of this work is to perform a practical analysis of the protests data with the minimal human effort.

Table 3. Truth table for the classification results of Naive Bayes.

		Actual values		
		PRO	**CON**	**Total**
Predicted values	**PRO**	45	4	49
	CON	6	45	51
	Total	51	49	100

Table 4. Truth table for the classification results of SVM.

		Actual values		
		PRO	**CON**	**Total**
Predicted values	**PRO**	45	9	54
	CON	2	44	46
	Total	47	53	100

4.3 Distribution of Classes

It is expected that classes are biased by the theme of the day, i.e., PRO tweets mainly occur in the PROGOV dataset, and CON tweets in the CONGOV dataset. However, our question is how imbalanced the datasets actually are, and if there is a difference on the distributions for each day.

To answer such questions, Figs. 1 and 2 show the distributions for each day and for each classifier. As we can see, regarding the classifiers, they agree on the distribution of topics on both datasets, having a very similar distribution of classes. Also, those Figures confirm that the distribution is biased towards the central theme of each protest, on March 13th the majority are supporting the government while on March 15th, the majority is against it.

We observe that on March 13th the opposing group was less active than on March 15th. This indicates that the people against the government concentrated their efforts on the protest of March 15th and did not pay attention to this pro-government manifestation. On the other hand, the group supporting the government was considerably active on both days of protests, trying to contest the claims of the other group.

Furthermore, the Figures show that the absolute number of tweets supporting the government is about constant throughout the days, with a number of around $80,000$ tweets, while the number of people against the government steps up from around $20,000$ to about $150,000$, almost 7 times more. This indicates a more consistent pattern of activists supporting the government.

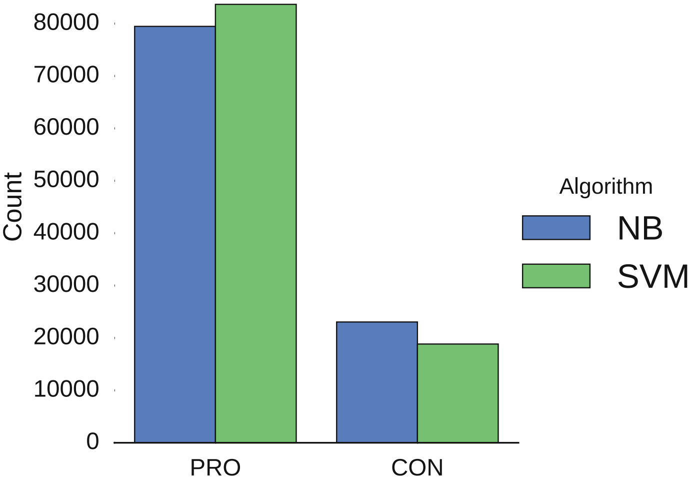

Fig. 1. Distribution of classes for March 13th.

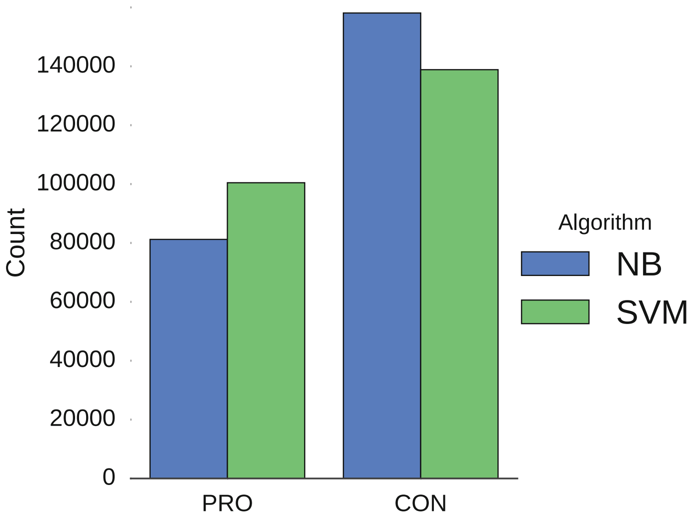

Fig. 2. Distribution of classes for March 15th.

4.4 Distribution of Words

After verifying the distribution of each class, it is also interesting to extract what people of each group are saying. For this matter we have extracted the Top 3 words used on the tweets for each class and on each type of protest.

The Figs. 3 and 4 show the results of these distributions. It is important to notice that both algorithms rendered the same set of words, so the results are grouped together on the bar plot depicted with the confidence intervals. The meaning of these words are explained on Table 5.

As we can see on March 13th, the majority of the tweets focused on the accusations against Globo TV Channel harming the democracy. In Brazilian history, Globo is often associated with the support of the military coup of 1964 [7] and the election of the only Brazilian president to suffer an impeachment [13]. The second and third more frequent words are associated with calling the people on the streets and stating they will not participate on the next protest against the government. The people against the govern limited themselves on calling people for the protests and asking the president to step out on her own.

On March 15th, the people supporting the government kept a similar behavior from the previous day, but additionally, they started a campaign claiming for democracy, stating that the people should accept the results from the past election as this is a democracy. The group against the govern intensified the use of the hashtag asking the president Dilma to step out together with the use of a similar hashtag related to her political party. The term *vemprarua* is perceived to have been used by both sides since this word is a more general term for calling people to the streets, without specifying the reason.

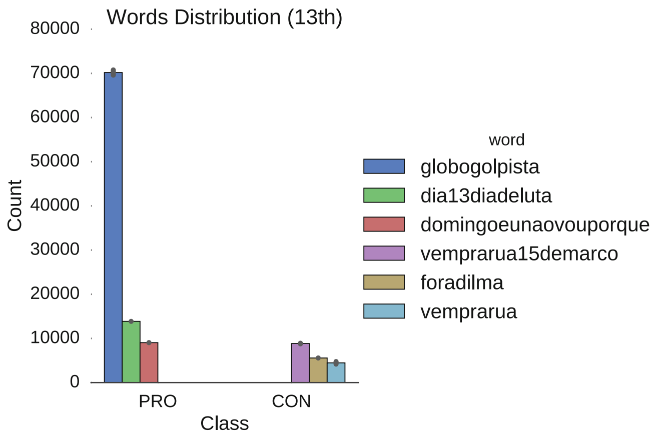

Fig. 3. Words distribution for March 13th.

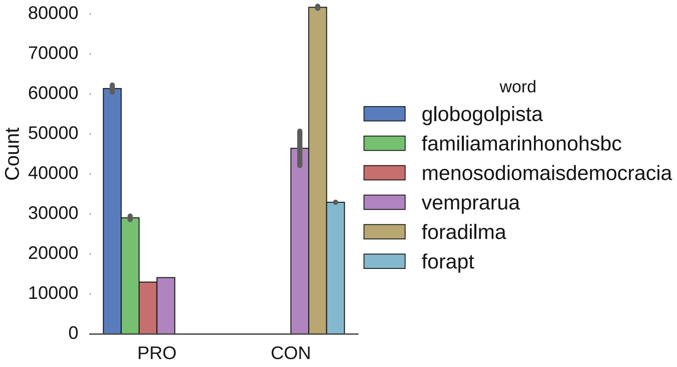

Fig. 4. Words distribution for March 15th.

4.5 Most Active Users

Another practical result of interest from these datasets is the identification of the most active users for each class. The identification of such actors may reveal the organizations and real motivation behind both manifestations. Even if they are not the *leaders* of such events, they represent a step towards finding such connections.

Initially, we analyzed the distribution of activity of all users in each day of protests. In Figs. 5 and 6 it is shown that the majority of users posted few tweets about the protests, while there were very few users responsible for about 800 tweets on March 13th and more than 1400 tweets on March 15th. This is similar to a power law distribution, indicating that few users are more active and possibly more influential than others. The next step was to identify those very active users and their role in the protests.

Table 5. Explanation of each hashtag.

Hashtag	Explanation
dia13diadeluta	Used to call the people for March 13th event
domingoeunaovouporque	Stating that they will not participate on March 15th
familiamarinhohsbc	Related to the accusations against Globo TV Station (accused of supporting the movement against the government) and HSBC bank
foradilma	Asking for Dilma Rousseff to step out of presidency
forapt	Asking for the Workers' Party to step out
globogolpista	Claiming Globo TV Station is trying a coup
menosodiomaisdemocracia	Asking for less hate and more democracy
vemprarua	Calling people to the streets, used for both events
vemprarua15demarco	Calling people to the streets on March 15th

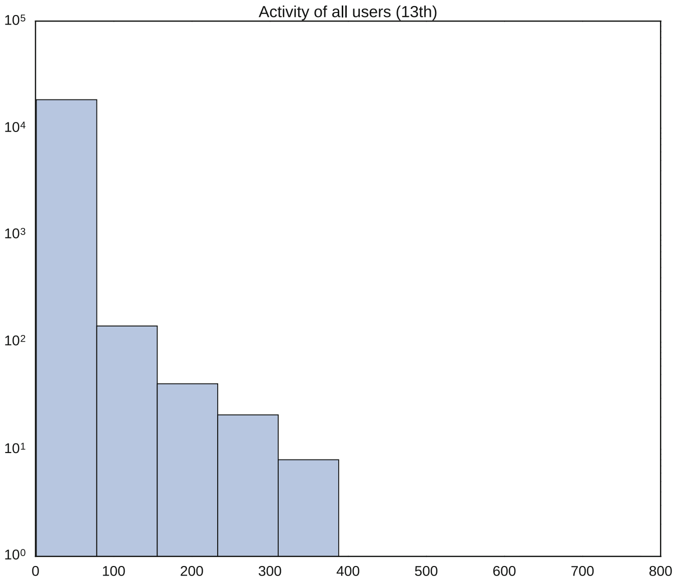

Fig. 5. Distribution of tweets from all users on March 13th, logarithmic scale for y axis.

In Figs. 7 and 8 we depict the distribution of the six most active users with confidence intervals. Regarding March 13th, the most active users for each group were *Larissa Alves* (/laripr), a twitter account of a person who actively tweets about the accomplishments of the current government, the suspicious and accusations of the opposing parties, and *Br45il No Corrupt* (/br45ilnocorrupt), an account with a pun on the number 45 corresponding to the opposing political party, replacing the letters 'A' and 'S' from *Brasil*. This account was specially

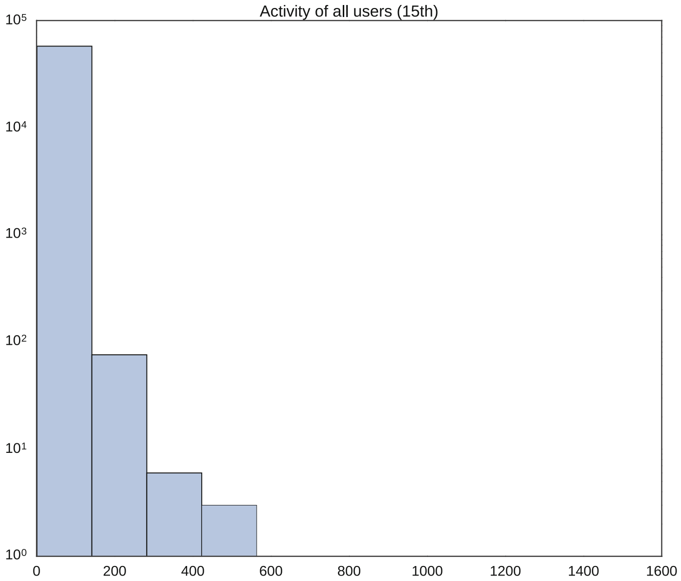

Fig. 6. Distribution of tweets from all users on March 15th, logarithmic scale for y axis.

created for accusing the Workers' Party of being corrupt and feed the discussions around the protests. This account was created by the non-profit organization of the same name that, while do not explicitly enlist a direct connection with the opposing party, it manifested support to them.

The account *#Dia13DiadeLuta* (/AdaByronKing) is an account related to a group of political activists against rumours, *#ForaDilma* (/jonhpaul11) was a common user that changed his name during the event to support the group against the government. There is no known connection with political parties but it is assumed that they have such support. The account *Revista Eletrônica* (/e_editora) refers to a self-claimed independent journalist media while *JoaoG* (/JGZZZO) seems to be a fake account created as a retweeting robot, also known as bot. These bots are computer programs created to share the messages of specific users, often used to fake the real impact of an opinion. The user is considered suspect of being a bot whenever they have more than 10 thousand tweets, consisting mostly of retweets, if they have many retweets in different languages, or have no tweet at all (i.e., retweet a message and delete some time later).

On March 15th, some of the tweets of the account *Br45sil No Corrupt* are probably incorrectly classified by one of the algorithms, generating a lower confidence. This misclassification occurred by a sequence of tweets without the common words used against the government. One example is the tweet literally translated to *Tomorrow we will be* 1 *million on the streets* that, without the

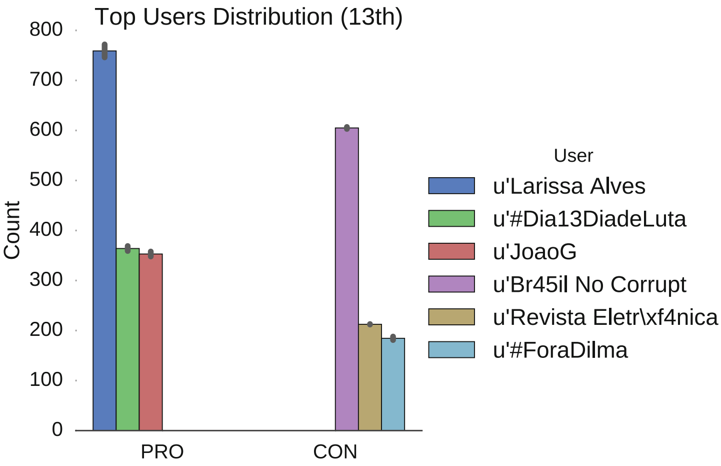

Fig. 7. Distribution of tweets from the six most active users on March 13th.

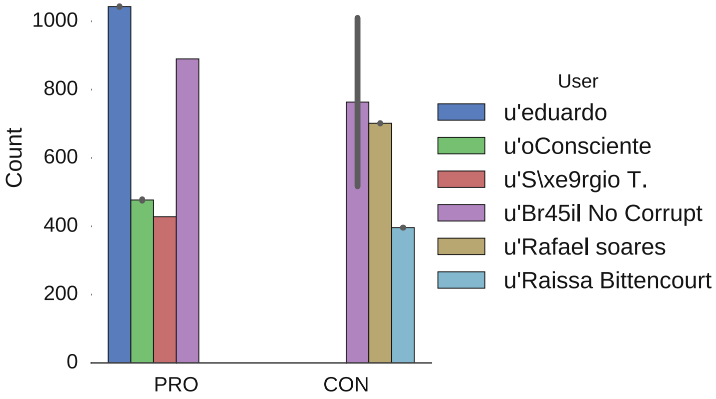

Fig. 8. Distribution of tweets from the six most active users on March 15th.

date of the tweet and the user that created the content, the correct classification cannot be inferred.

The user *Rafael Soares* (/KatycatBrasill), after manual inspection, seems to be an account created as a fan account for singer Katy Perry as a disguise for being another retweeting bot. This account has a long history of retweeting contents of different opinions in different languages. The user *Raissa Bittencourt* (/raissabittenco3) was a fake account and it is not active anymore, created probably with the purpose of retweeting opinions against the government. The user *eduardo* (/eduardonino) is a political activist supporting the government but aligned with more leftist parties. Finally, the user *oConsciente* (/oconsciente) is a political activist supporting the Workers' Party.

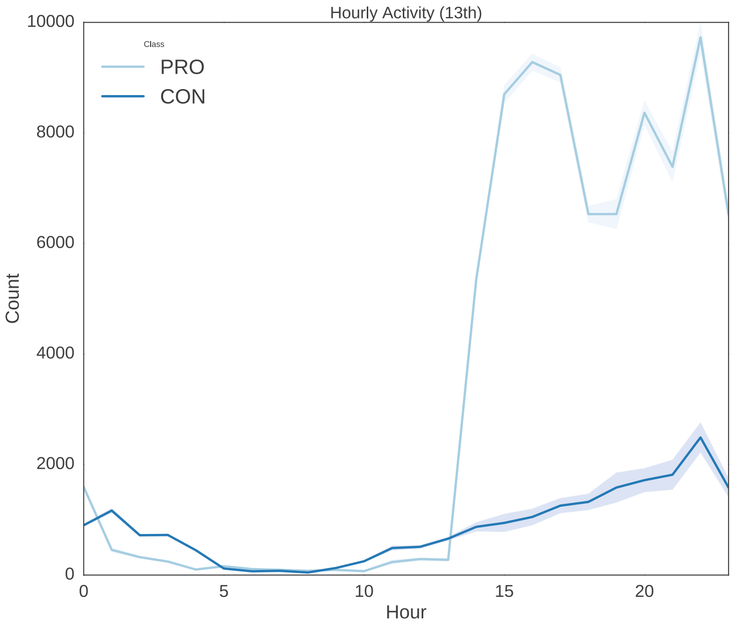

Fig. 9. Hourly distribution of classes on March 13th.

These results could find some interesting actors (i.e., *Br45il No Corrupt* and *oConsciente*) that are indicative of the organizations behind each group. But, also, it revealed the use of bots by both sides in order to inflate the importance of their claims.

4.6 Hourly Activity

Next we verify the hourly activity throughout both days of protests, first grouped by class and then by the top users. In Figs. 9 and 10 we can see the activities for each group on each day. We note that the protests took place during the afternoon of the corresponding days, thus the main activity was comprised from noon to midnight on both days. As it should be expected, the group supporting the government was more active than the group against it on March 13th, while on March 15th occurred the opposite.

However, the behaviors are different, as seen in these Figures. The first is regarding the behavior of the CON group during March 13th, as they kept a low profile in the morning but started raising their activity after 10 a.m., reaching its peak at around 11 p.m. of the Friday night. This pattern seems reasonable as a kind of attack against the supporters group, when their manifestation started. Since this is the day preceding the weekend, the working time might have prevented most of the users of tweeting before 6 p.m.

During the events of March 15th, we observe an intensified activity of the supporters group early in the morning. They seem to have organized themselves

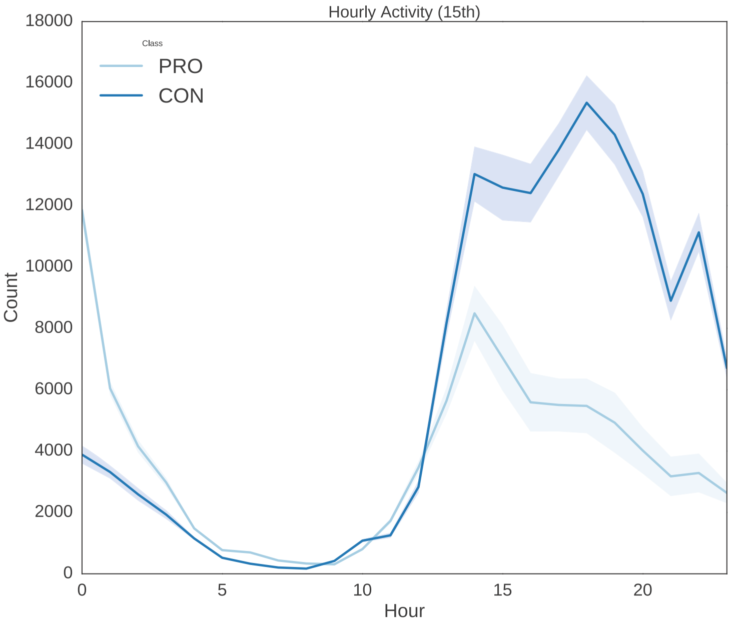

Fig. 10. Hourly distribution of classes on March 15th.

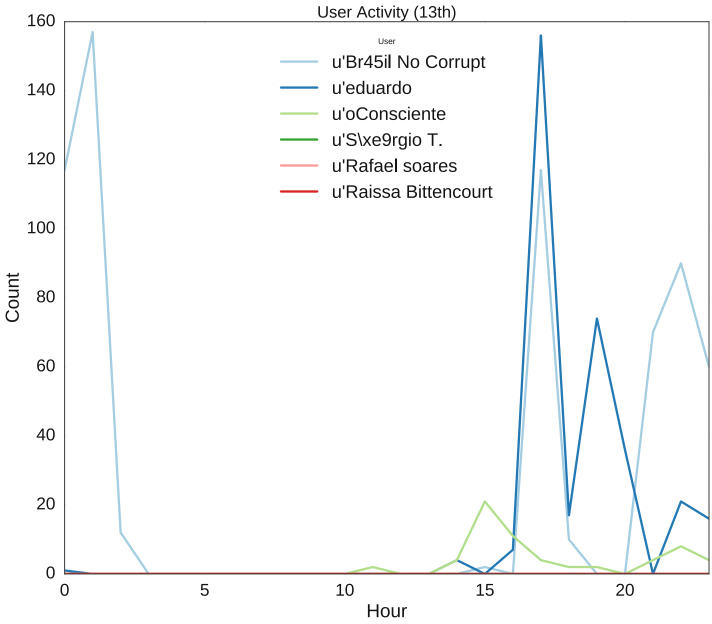

Fig. 11. Hourly user activity on March 13th.

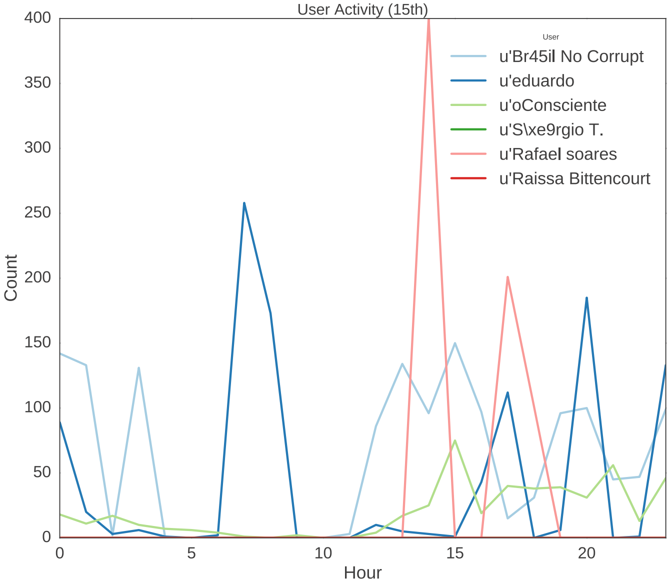

Fig. 12. Hourly user activity on March 15th.

to try attacking the protesters prior the event. Right after the start of the event, the supporters were also very active, trying to compensate for the rising of people tweeting against the government and, after that, followed the same trend of the protesters.

In Figs. 11 and 12, we depict the hourly activity of some of the top users from a previous analysis throughout each day. On March 13th, the users followed a similar behavior of the tweets by class, being more active during the afternoon. The users *eduardo* and *Br45il No Corrupt* were responsible for the most activities, having similar peaks at 5 p.m., at 7 p.m. and a final one at 9 p.m.. The events of 5 p.m. were about the presence of artists on the protest against the government, with a decay of such announcements on 7 p.m. and raising again at 9 p.m..

During March 15th, the activity of users did not match exactly the class hourly behavior, having several peaks throughout the day. The main peaks occurred on 6 a.m. by *eduardo* calling the people for a tweeting event against the protest. After that, at 2 p.m. the user *Rafael soares* chained a tweeting activity to raise the hashtags against the government on the trending topics. These users followed the same behavior later at 5 p.m. and, by 8 p.m., the user *eduardo* raised again a protest against the media trying to coup the government.

5 Conclusion

In this paper we show how we applied two algorithms for supervised learning, Naive Bayes and Support Vector Machine, in order to analyze the events of two opposing protests on the streets of Brazil, as told by Twitter users, as a consequence of the disputed presidential elections in 2014. These algorithms were trained using a very small sample of the data set in order to quickly estimate the numbers of both events.

The events were first separated in two datasets, being March 13th regarding the protest supporting the government and March 15th the protests from the group against the government. Both datasets were classified by the two algorithms on its entirety, and the distribution of the analyzed quantities were grouped together when convenient.

Ideally, to improve accuracy, a large set of labeled data should be available during the training process, so that the learning algorithms could face distinct examples that should pertain to the same class. But, in practice, we cannot always afford to manually separate a sufficient amount of data for this task, and not even verify the accuracy results. These experiments show that, even if you cannot guarantee high accuracy, some interesting information can still be extracted for using on a broader study.

The results showed that the activists supporting the government, although being a minor number, were more active throughout the weekend comprising both protests. They actively tried to reduce the importance of the protests against the government by accusing the organizations that supposedly were behind the event. On the other hand, the groups leading the protest against the government concentrated their efforts during the peak of the events, as an attempt of minimizing the importance of the other group and spread their goals.

Another interesting information found in these datasets was the use of retweeting robots from both groups to inflate the numbers of tweeters supporting each event. This not only may affect the perceivable intensity of the movements, but can also help to attract new people for both sides through the Twitter trending topics.

From this point, we have paths to follow for future research. On the Computer Science side, we will try to automatize the process of manual labeling for the training process or minimizing such efforts. We intend to do that by means of semi-supervised learning and the use of Topic Modeling. On the Data Science side, we will apply this procedure into a much larger data set containing all the events that happened during the presidential elections, and that motivated the current events.

Acknowledgment. This research was funded by FAPESP process number 2014/06331-1.

References

1. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R.: Sentiment analysis of Twitter data. In: Proceedings of Workshop on Languages in Social Media, pp. 30–38. Association for Computational Linguistics (2011)
2. Aggarwal, C.C., Zhai, C.: A survey of text classification algorithms. In: Aggarwal, C.C., Zhai, C. (eds.) Mining Text Data, pp. 163–222. Springer, New York (2012)

3. Agresti, A., Kateri, M.: Categorical Data Analysis. Springer, Berlin (2011)
4. Aizawa, A.: An information-theoretic perspective of TF-IDF measures. Inf. Process. Manag. **39**(1), 45–65 (2003)
5. Amari, S.I., Wu, S.: Improving support vector machine classifiers by modifying kernel functions. Neural Netw. **12**(6), 783–789 (1999)
6. Berry, M.W., Castellanos, M.: Survey of text mining. Comput. Rev. **45**(9), 548 (2004)
7. Chong, A., Ferrara, E.L.: Television and divorce: evidence from Brazilian novelas. J. Eur. Econ. Assoc. **7**(2–3), 458–468 (2009)
8. Kouloumpis, E., Wilson, T., Moore, J.: Twitter sentiment analysis: the good the bad and the OMG! In: ICWSM vol. 11, pp. 538–541 (2011)
9. Liu, B., Zhang, L.: A survey of opinion mining and sentiment analysis. In: Aggarwal, C.C., Zhai, C. (eds.) Mining Text Data, pp. 415–463. Springer, New York (2012)
10. Livne, A., Simmons, M.P., Adar, E., Adamic, L.A.: The party is over here: structure and content in the 2010 election. In: ICWSM 2011 (2011)
11. Lotan, G., Graeff, E., Ananny, M., Gaffney, D., Pearce, I., Boyd, D.: The Arab spring— the revolutions were tweeted: information flows during the 2011 Tunisian and Egyptian revolutions. Int. J. Commun. **5**, 31 (2011). http://ijoc.org/index.php/ijoc/article/view/1246
12. McCallum, A., Nigam, K., et al.: A comparison of event models for Naive Bayes text classification. In: AAAI-1998 Workshop on Learning for Text Categorization, vol. 752, pp. 41–48. Citeseer (1998)
13. Miguel, L.F.: Mídia e eleições: a campanha de 1998 na rede globo. Dados [online] **42**(2) (1999). http://dx.doi.org/10.1590/S0011-52581999000200002
14. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: LREC, vol. 10, pp. 1320–1326 (2010)
15. Roth, V., Steinhage, V.: Nonlinear discriminant analysis using kernel functions. In: Advances in Neural Information Processing Systems. Citeseer (1999)
16. Spaiser, V., Chadefaux, T., Donnay, K., Russmann, F., Helbing, D.: Social Media and Regime Change: The Strategic Use of Twitter in the 2011–2012 Russian Protests (2014). Available at SSRN
17. Suykens, J.A., Vandewalle, J.: Least squares support vector machine classifiers. Neural Process. Lett. **9**(3), 293–300 (1999)
18. Thrun, S., Pratt, L.: Learning to Learn. Springer Science & Business Media, New York (2012)
19. Turkmen, A., Cemgil, A.: Political interest and tendency prediction from microblog data. In: 2014 22nd Signal Processing and Communications Applications Conference (SIU), pp. 1327–1330, April 2014
20. Vitter, J.S.: Random sampling with a reservoir. ACM Trans. Math. Softw. (TOMS) **11**(1), 37–57 (1985)
21. Watts, J.: Brazil: hundreds of thousands of protesters call for rousseff impeachment. The Guardian (2015). http://www.theguardian.com/world/2015/mar/15/brazil-protesters-rouseff-impeachment-petrobas