Report

Elections in Brazil - 2nd Round

Introduction

In this project we build models for classifying“tweets” into positive, negative and neutral sentiment in order to predict the winner of 2014 presidential elections in Brazil.

The 2014 elections, in the first round had 11 candidates and in the second just 2 of them, Aécio Neves and Dilma Rousseff. In this project was considered just the second round of the elections.

The idea is to capture tweets which correspond to mentions of the Brazil presidential candidates, classify them into sentiment in order to predict the sentiment of tweets towards a candidate. Through this, we can see if we find a pattern of tweets in correlation to candidates, and also if we can predict the results of the election. In the other words, can we predict who will be the winner by using the Twitter data?

In this scenario, we used machine learning algorithms, such as Logistic Regression and Naive Bayes in order to classify tweets into positive and negative. And also, we used hashtags to try to find out the winner of the elections.

Collect Data

For this assignment the streaming API and rest API were used to capture data from Twitter during the election time in Brazil. We separated the data into two main sets. The first set is defined as the retrieve of data containing the hashtag #DebateNaGlobo

(Globo is the channel where the debate was held), and the second set is captured by searching for tweets of mentions of Dilma and Aécio.

Data Set 1 looking for two keywords and collecting Tweets of a predetermined Hashtag (#DebanteNaGlobo).

The algorithms executed from Friday, October 24th at 01:00am to October 26th. More than 20.000 tweets were retrieved and after removing the retweets remained 8720 unique tweets.

Data Set 2

As previously described this set is created by searching tweets that mentioned the candidates. The data started to be collected on Oct 22 until Oct 26 (the day of the election). It was collected 107431 tweets mentioning Aécio and 128646 mentioning Dilma.

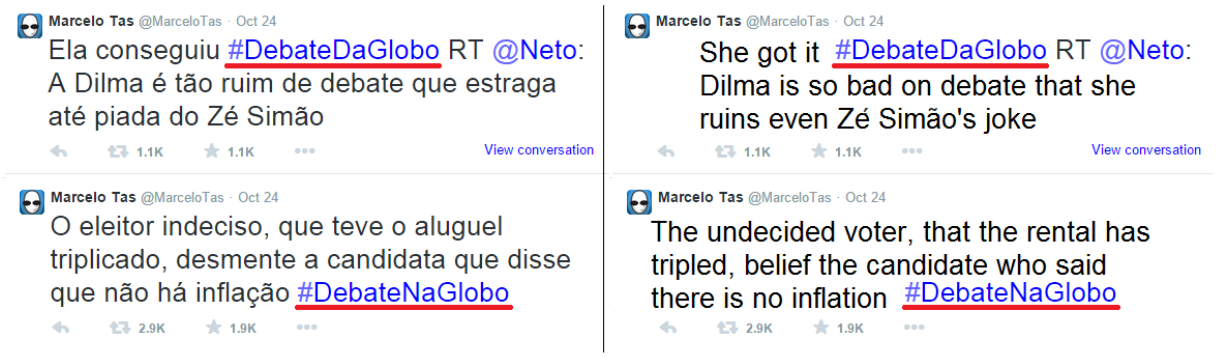
Different text encoding became a problem to face because the data was collect without a correct encoding. Find the problems and remove the wrong encoding is very important to get a meaningful dataset and convert to UTF-8 as the default encoding. After this step, we tokenized the data and replace symbols, the mentions, and URL with THIS_IS_A_SYMBOL, THIS_IS_A_MENTION , THIS_IS_A_URL, respectively. The tokenized tweets were saved in a new text document. The last step for preprocessing data was to eliminate the tweets.

After all these steps the data decreased and we got 39327 tweets for Aécio and 65509 tweets for Dilma.

Approach and methods

Dataset 1:

Analyzing the tweets on the base, and retrieving the most retweeted tweet of the dataset we discovered that users were using similar hashtags for the Debate (e.g. #Debate*OnGlobo* and #Debate*ofGlobo*, in portuguese: #DebateNaGlobo and #DebateDaGlobo (Fig. 1)), and as we used only the hashtag shared by the channel, we probably have lost an amount of tweets of persons that were giving their opinions about their candidate on twitter.



(Fig. 1 - Use of different hashtags)

After this, we used the Titan Graph Database to create a graph of mentions , and analyzing the graph we discovered another curious aspect of the tweets, that they were talking a lot about the presenter (William Bonner) of the debate. Reading some reviews of the debate, we found out that the presenter made some mistakes during the debate, resulting in a lot of tweets with ironic content referred to him (See Fig. 2).

(Fig. 2 - Ironic Content of the tweets referred to the presenter)

(Translation of the first picture - "*I command this circus!*")

Finally, we just checked the amount of tweets containing pictures, because many of them just posted ironic content not related to the elections, consequently not usefull for us.

The number of tweets containing pictures on the dataset is 1970 tweets (22.6%).

(Fig3 - Example of tweets containing pictures)

(Translation of the first picture: I Vote on the 3rd Candidate #DebateNaGlobo)

We count the number of different hashtags on the tweets, and labeled them with positive/negative/neutral labels according to the candidate.

Total Hashtags found: 351 / Neutral: 250 (71.22%)

Pos(Dilma) - 31 (8.83%) / Neg(Dilma) - 10 (2.85%)

Pos(Aecio) - 51 (14.53%) / Neg(Aecio) - 9 (2.56%)

The total number of tweets labeled "*Neutral*" demonstrate that many users use hashtags not related to the debates or related to the candidates, making the analysis difficult.

After collected the labels for each candidate, we count the number of unique tweets that we found the labeled hashtags.

Total Tweets Analyzed: 1873 (21.47% of total Tweets)

Neutral: 9 (0.48%)

Pos(Dilma) - 992 (52.96%) / Neg(Dilma) - 53 (2.83%)

Pos(Aecio) - 796 (42.50%) / Neg(Aecio) - 23 (1.23%)

If a tweet had positive and a negative label, we considered it neutral (maybe could be twitter asking about the two candidates like: "Who do you support? #aecio or #dilma ?").

After labeling the tweets we used the equation utilized in some related works to compute the preference of each candidate.

$$\frac{pos(c_1)+neg(c_2)}{pos(c_1)+neg(c_1)+pos(c_2)+neg(c_2)} \quad (1)$$

We used the equation 1, c1 is candidate for the person who support is computed when c2 is another candidate. pos(c) , neg(c) are the positive and negative number of tweets that mentioned the candidates c. In addition, neutral tweets are not used because they don't demonstrate a preference for each of candidate. [Metaxas 2011]

Computed Preference for: Aecio: 0.45 / Dilma: 0.54

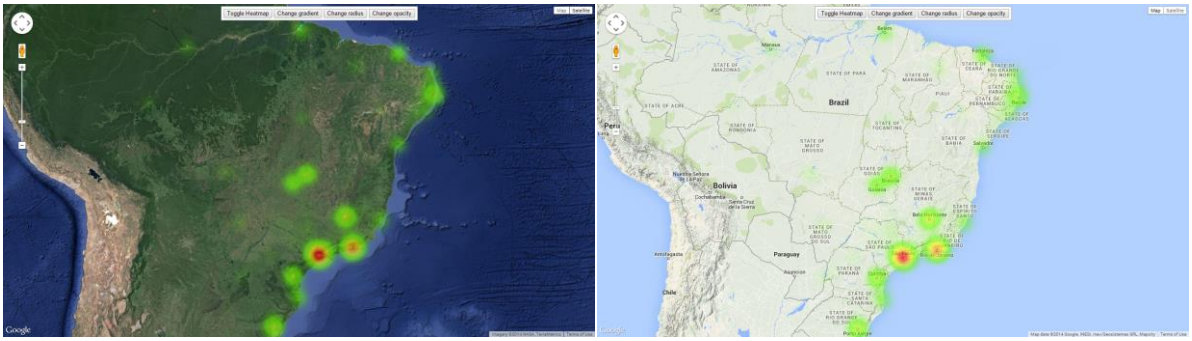
Preference in the last poll: Aecio: 0.48 / Dilma: 0.52

Election Result: Aecio: 0.484 / Dilma: 0.516

With the labels of each candidate, we analyzed the tweets and the computed preference of our dataset if 0.45 supporting Aecio and 0.54 supporting Dilma. Considering that the polls find similar results to the elections, and the last poll in Brazil interviewed 4.000 persons (0.0002% of total population), our result showed very precise to the result of elections. Maybe in a future work we could analyze more hashtags, and use the retweets, because these, like the unique ones, still express the opinion of the electors.

In a final analysis, we looked for the location of the electors, and plotted on map using Google API V3 - Heatmaps to locate our electors, and in a future work analyze

the similarity of where the people live, and the prediction of who them vote (Fig. 5).



(Fig. 5 - Location of the users)

Data Set 2

First of all, we vectorized the data using the library Sklearn from python. For each one of the candidates according to the size of the training set were defined the size of the test set. For Aécio we had a training set with 180 tweets and other with 330 and for Dilma we had 250 and 429. The size of each test set is the same as training set, but the tweets are randomly chosen from all tweets preprocessed. The number of features after vectorization if found in the tables below.

Dilma Dataset	
Trainning set	**Number of features**
250	about 1000
429	about 1400

Aécio Dataset	
Training set	**Number of features**
180	about 800
330	about 1200

The training set was built manually by labeling tweets according to human assumptions in positive (2), negative(1) and neutral(0). As described we used two sizes of data set to each candidate in order to see what would happen with the accuracy.

We used two algorithms for classification: Gaussian Naive Bayes and Logistic Regression. And also we considered other type of vectorization, the TdfidVectorizer in other to compare the results.

Experiments

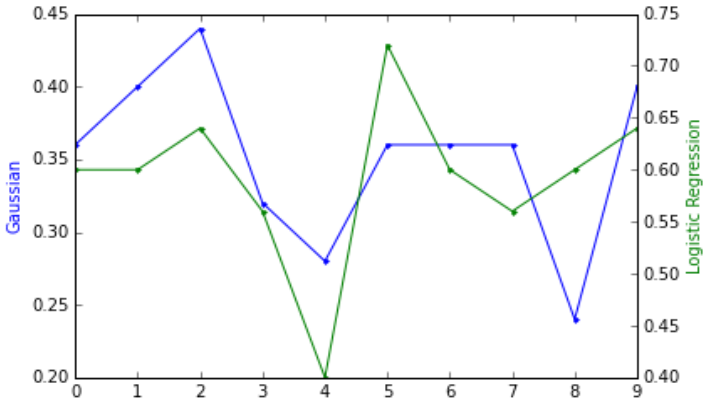
Number of the users which selected is 1669 and the number of the users who do not have a corresponding Location is 583 (34.93%).

Data Set 2

Naive Bayes, GaussianNB, Logistic Regression, and TdfidVectorizer we got the following results.

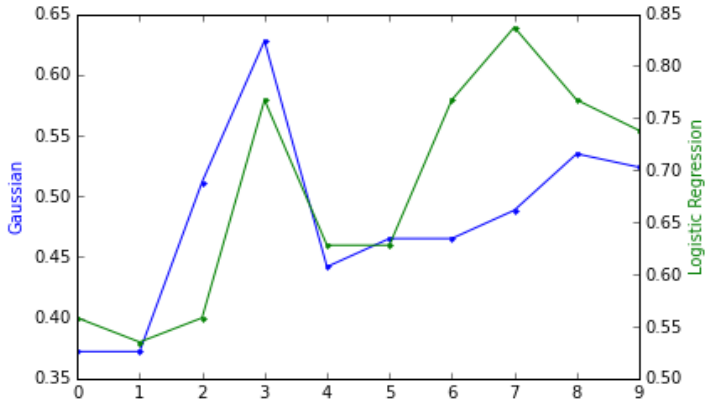
In the table below, we can see the results for the small portion of training data for the candidate Dilma with 250 tweets we had a good accuracy using TfidfVectorizer with Logistic Regression as its algorithm, but the accuracy using Gaussian Naive Bayes was poor. The graph below shows in the X axis the k-fold for cross validation, and in the Y axis it shows the accuracy for each algorithm.

Technique	Accuracy
Logistic Reg	0.61
GaussianNB	0.368
TfidfVec	0.608



The graph and table below show results using the big portion of training set that contains 429 tweets. Through them we can see that the accuracy increased using more data for training.

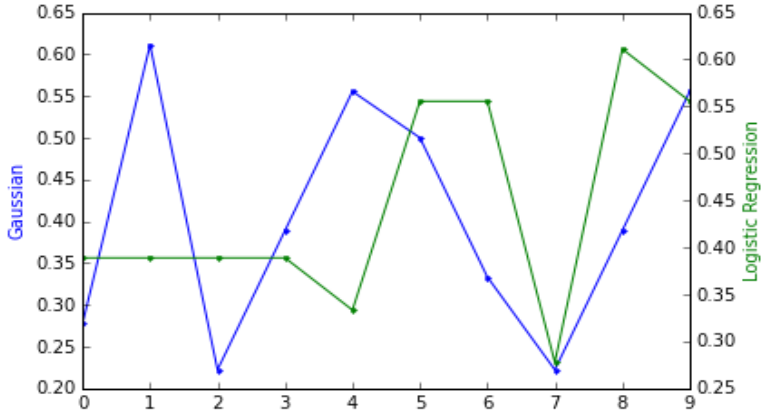
Technique	Accuracy
Logistic Reg	0.66
GaussianNB	0.48
TfidfVec	0.678



In the table below, we can see the results for the small portion of training data for the candidate Aécio with 180 tweets we had similar accuracy using all Techniques. The graph below shows in the X axis the k-fold for cross validation, and in the Y axis it shows the accuracy for each algorithm.

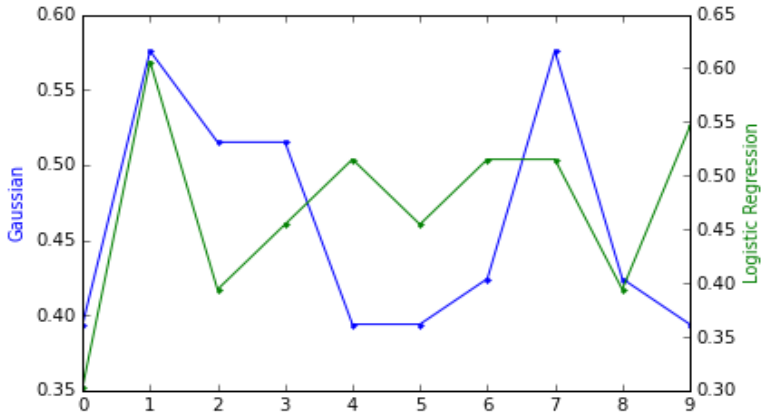
Technique	Accuracy
Logistic Reg	0.440
GaussianNB	0.422

TfidfVec	0.472



The same occurs to the big portion of the data with 330 tweets.

Technique	Accuracy
Logistic Reg	0.47
GaussianNB	0.478
TfidfVec	0.487



Related work:

[Teitler, 2009] This research did a sentiment analysis that expresses that tweets correspond to the candidates show that they tweet reflects a good sentiment overview on situation of candidate. They believe that " Tweets carry news in an even wider range of expressions, thoughts, opinions, and artifacts pertaining to news so that it is truly representative of the present moment."

[Gayo-Avello, 2012] believed that predicting an election is extremely difficult and it is almost impossible.

[Gayo-Avello, 2011] investigated the strengths and weakness of diffrent approach of extracting election data from tweeter. They said that this process needs to add some extra method to a simple sentiment analysis.

[Finn, 2012] suggested some ways to pre-processing and filter useless data.

PoliTwi is a used by the German community during the parliamentary election and express the idea of gathering information in tweeter as those data are more fresh than the real time news in the world. [Rill 2014]

Conclusions and Future Work:

Through this project we can conclude that the size of training set helps to improve the accuracy and the prepossessing step also helps in the improvement. Most of the tweets did not have sentiment, so we can say that is likely that the accuracy was low because of that. In addition, acquiring the data already encoded could have helped

to increase the accuracy because in this assignment the data was not encoded correctly. We could see that the tfidfVectorizer also influences in the accuracy and we can explore more on that in future work.

For the first data set we had a good result compared to the polls and the result of the elections, but it was not considered age, gender, social or ethnic group, they are equally represented on the dataset. And also there are several tweets have ironic content which can complicate the analysis, so we should look further into that.

For the future work we can say to increase size of training set, analyze geolocated Tweets better alternative than a random test set, also select features, collect encoded data, and consider the “mentions” to manual classification for training set could help to improve the results.

Team work

For this project all member worked together at library, helping each other with analysis, coding problems and results.

References

H.,Teitler, B.E. andSperling,J., TwitterStand: news in tweets. (2009), In GIS’09.

Gayo-Avello, D., No, you cannot predict elections with twitter. (2012), Internet Computing, IEEE, 16(6):91–94.

D. Gayo-Avello, P. T. Metaxas, and E. Mustafaraj., Limits of electoral predictions using twitter. (2011), In Proceedings of the 5th International AAAI Conference on Weblogs

and Social Media, Menlo Park, CA.

Finn, S., & Mustafaraj, E. . Real-time filtering for pulsing public opinion in social media. (2012), Proceedings of the 25th International Florida Artificial Intelligence Research Society Conference (pp. 32-37). Menlo Park, CA: The AAAI Press.

PT Metaxas, E Mustafaraj, D Gayo-Avello, How (not) to predict elections (2011), IEEE third international conference on social computing (SocialCom)

PT Metaxas, E Mustafaraj, Social media and the elections
Science (2012), 338 (6106), 472-473

S. Rill, D. Reinel, J. Scheidt, R. V. Zicari. PoliTwi: Early Detection of Emerging Political Topics on Twitter and the Impact on Concept-Level Sentiment Analysis. Knowledge-Based Systems. Elsevier; (2014), DOI: 10.1016/j.knosys.2014.05.008