

Algorithmics for Data Mining: Deliverable 3

Democrat Vs. Republican Tweets

Oriol Borrell

FIB - UPC Student

Barcelona, Spain

`oriol.borrell.roig@est.fib.upc.edu`

April 14, 2020

0 Abstract

We will analyze what do politicians of the Republican and Democrat Parties (in USA) tweet about. We will first see if we can extract any conclusion about particularities of each party tweet. Afterwards we will train a model that will predict, given a tweet, if it belongs to the Republican or to the Democrat party. The Code for this project can be founded here ¹. The data was extracted from here ²

1 Context

The tweets obtained are all tweet from 2019 of USA politician's. We have to take into account that Donald Trump (from the Republican party) is the president of USA since January 20, 2017. The Democrat's candidate was Hillary Clinton. The 2019 in USA was a Off-year election.

Republicans and Democrats are the two main and historically the largest political parties in the US. After every election, they hold the majority seats in the House of Representatives and the Senate as well as the highest number of Governors.

2 What do Democrat and Republican members tweet about?

Before analyze the tweets we applied the following preprocessing steps:

- Remove the newline characters
- Remove commonly used ampersand
- Remove ' from contractions such as I'm and don't
- Lowercase the string
- Remove https-links from the string
- Tokenize the string with the given pattern

Once we had the tokens, we computed the frequencies of each token in each party. We created the wordcloud shown in Figure 1.

¹Github repository of the project: <https://github.com/oriolborrellroig/ADM-Deliveries/tree/master/ThirdDelivery>

²Data used in the project: https://github.com/suneman/socialgraphs2019/tree/master/files/data_twitter

The obtained wordclouds are very interesting. We see that the Democrats apparently have a larger focus on *climate* and *health care* where the Republicans focus on *anti-abortion* and *tax-cuts*. These topics are all more interesting and polarizing than the previous results, where we saw that both parties often referred to different political personalities, committees and other political jargon. We see a clear indicator, that the phrases for both parties are political slogans, such as *endgunviolence* and *bornalive*. Twitter, therefore, gives us a valuable insight into the key-issues and focal points of each party.

3 Predicting the party of a tweet

In order to predict the party of a tweet we trained a Convolutional Neural Network (CNN). We applied the same preprocessing steps mentioned before in order to obtain the tokens. Once obtained the token, I substituted each token with their position in the top used words, or 0 if it's not a frequently used word. With this substitution I obtained continuous numerical data that I could send it to the model.

About the model, in other deliveries I briefly explained the behavior of a CNN and the different types of layers. I will assume this knowledge is already explained. The following table shows the CNN I designed:

Layer(type)	Units	Filter	Stride
Embedding	-	-	-
Conv1D	32	3	1
MaxPooling1D	-	3	2
LSTM	-	-	-
Dense	5	-	-

Table 1: CNN Layers

Once trained the model we tested with the 20% of data that we reserved for this purpose. We obtained an accuracy of 0.74.

4 Conclusions

Personally, I think that spending more time trying to build the model with different layers or parametrization, or finding a better way to convert the tokens into numeric data, could lead us to achieve a better accuracy. It's the first time I work with text data, and the project helped me to learn steps that are important to apply, and useful statistics when dealing with this type of data.

However, the results obtained in Section 2, in my opinion are very powerful, and reveal the main priorities and way of thinking of each party in some political topics.