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. Afterwords 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 tweets from 2019 created by different 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 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 strings

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

[1] Github repository of the project: `https://github.com/oriolborrellroig/ADM-Deliveries/tree/master/ThirdDelivery`

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

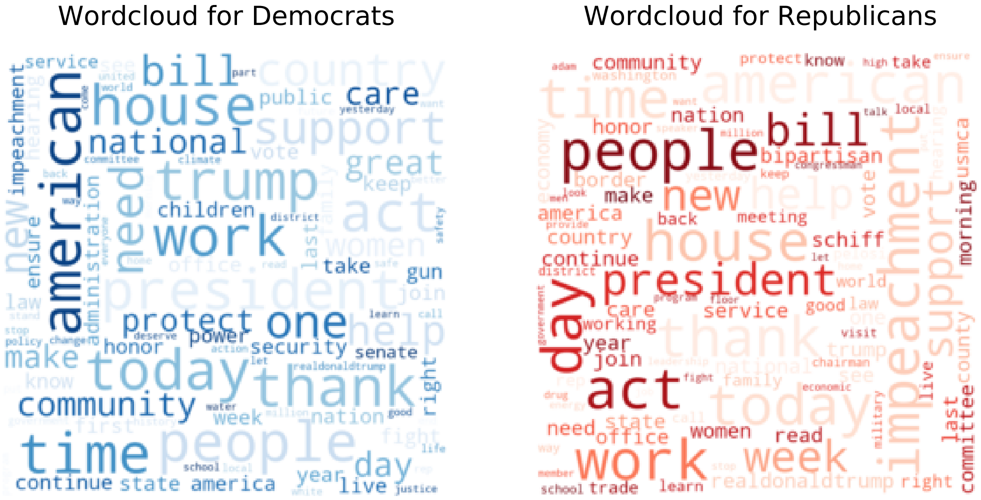

Figure 1: Wordcloud of the top used tokens of each party

Analyzing the results, we observe that the most used words of each party are fairly common words in the field of politics. We cannot extract any particularity neither from the Democrats nor Republicans.

Doing a bit of research I founded a common concept in *Information Retival* called *Term Frequency–Inverse Document Frequency, TFIDF*[3]. TFIDF is a way to compute the importance a word is in a document. As shown in Equation 1 is computed using two statics, the *Term Frequency(tf)*, and the *Inverse document frequency(idf)*:

$$tf(t,d) = 0.5 + 0.5 \times \frac{f_{t,d}}{max\{f_{t',d} : t' \in d\}}$$
$$idf(t,D) = log\frac{N}{|d \in D : t \in d|} \tag{1}$$
$$tfidf(t,d,D) = tf(t,d) \times idf(t,D)$$

Where $f_{t,d}$ is the frequency of the token t in the document d, and N is the total number of documents in the collection D.

We computed this static for each word and created another time the wordcloud of each party. The results are shown in Figure 2

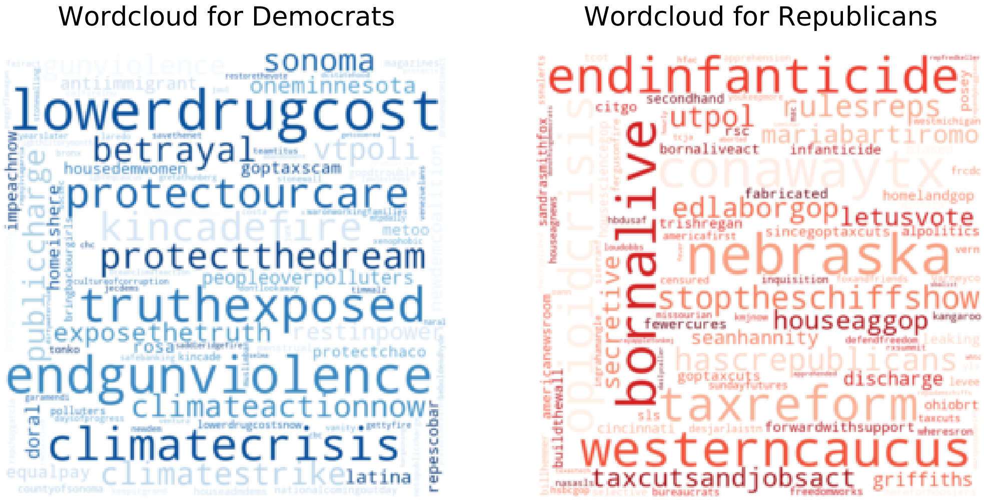

Figure 2: Wordcloud using TFIDF

[3] TFIDF's Wikipedia page: `https://en.wikipedia.org/wiki/Tf-idf`

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 Newral 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 a top used words ranking, or 0 if it's not a frequently used word. With this substitution I obtained continues numerical data that I could send-it to te 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 a accuracy of 0.74.

4 Conclusions

Personally, I think that spending more time trying to build the model with different layers or parametrization, or founding a better wey to convert the tokens into numeric data, could lead us to achieve a better accuracy. Is the first time I work with text data, and the project helped me to learn steps that are important to apply, and useful statics 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 think of each party in some political topics.