Sentiment analysis on the New York Times comments
Machine Learning for Natural Language Processing 2021

Elie Ba
ENSAE
elie.ba@ensae.fr

Geneviève Toubol
ENSAE
genevieve.toubol@ensae.fr

Access to the colab notebook: https://colab.research.google.com/drive/1X1t3G45_NlPNBlrQ1xNg3cnQc5n_oJxT?usp=sharing
*Access to the repo Github:*https://github.com/E2lie/NLP-New-York-Times-comments

1 Problem Framing

Over the last few years, the US political scene has become increasingly polarised and the New York Times, one of the leading newspapers in the US and worldwide, has been at the center of it, repeatedly accused by right-wing pundits, politicians and sympathizers of embodying the East-Coast elite hostile to Donald Trump.

The aim of this notebook is to investigate how sentiment analysis on the NYT comments can reflect this high level of polarisation. Our database gathers roughly 600,000 comments belonging to the political sections of the NYT.

First, we analyse the sentiments expressed in the comments section for the period 2017-2018 adopting two angles: one is the traditional approach of sentiment analysis, *i.e.* classification of text data into a "positive" category and a "negative" one, the other approach is a political labelling of the comments into a "Republican" category and a "Democrat" one. We compare the two approaches and analyze their results. Second, we link the political-labelled comments with the articles they are responding to by putting into perspective a clustering performed on the articles with the labelling of their comments. Therefore we are able to grasp how the two major parties' advocates distinguish themselves both in terms of language elements and in terms of the topics they are interested in.

2 Training models on the Stanford Sentiment Treebank

2.1 SST reviews representation

The Stanford Sentiment Treebank is a dataset frequently used in sentiment analysis. The corpus is based on the dataset introduced by Pang and Lee (2005) and consists of 11,855 single sentences extracted from movie reviews. We use a version of the labelling that is binary. We first need to create a coherent representation of the reviews in order to classify them. To do that, we will tokenize them and clean the symbols, urls and contractions that can sometimes appear in the text. Each review becomes a list of tokens that can either be a word or a sign of punctuation.

In a second part, we transform our string vectors into numeric vectors thanks to embedding methods. We use two methods of embedding: *Word2Vec* and *TF-IDF* (Cam-Stei, 2019). It should be noted that we will also use the default embedding method as we will build a LSTM modelling.

2.2 SST reviews classification

To evaluate our representations, we use different methods of classification. First, for the two specified embedding methods, we use two classic classifiers, a *Decision Tree* and a *Random Forest*. Thanks to these classifiers, we will be able to evaluate quantitatively our representations (which we cannot do with the NYT comments database since it is unlabelled). We also use a LSTM model to classify our reviews, and we can thus compare the efficiency of a combination embedding + classifier to a neural network.

Model	Precision	Recall	F1-score
Word2Vec-Decision Tree	0.49	0.49	0.49
Word2Vec-Random Forest	0.50	0.89	0.64
TF-IDF-Decision Tree	0.54	0.56	0.55
TF-IDF-Random Forest	0.57	0.59	0.58
LSTM	0.50	1.0	0.67

As you can see, the most performing model is the LSTM, followed by the *Word2Vec & Random Forest*. We obtain a F1-score of 67% for the LSTM model, which is very encouraging about its usefulness to label our comments.

Before applying the model on the NYT comments, We will now develop our second labelling approach. Then we will be able to compare the two and draw conclusions regarding our initial objective which was reflecting the polarisation of the US political scene thanks to NLP methods.

3 Training models on an *ad hoc* database

3.1 General idea

The idea of this section is to label a portion of our dataset by detecting users having very polarised and easily identifiable opinions. We use their comments to build a labelled dataset of 15,685 observations. This labelled dataset is used to train and test a model that will then predict labels on an analysis dataset of 58,318 comments (amounts to only 10% of the remaining unlabelled data for computation capacity reasons). It is important to mention that the analysis dataset is built making sure none of the comments selected to be manually labelled are part of it.

3.2 Labelling process

In order to detect users whose comments can be manually labelled, we select keywords that we believe are clearly associated to one of the two parties. We then look at the users who use the most this term. Once we have an user, we check that his total number of comments is sufficient for two reasons: (i) we want to label the maximum comments possible at once, (ii) most importantly we do not want to do a labellisation that is overly focused on the keywords, it would entail overfitting, we rather want to spot staunch democrat/republican supporters who post a variety of comments so the algorithm can learn how they talk and about what they talk. If the user meets our conditions then all of his comments are labelled according to his political stance.

3.3 Data pre-processing

Once we have our 15,685 labelled comments (to reach this number we spotted 43 presumably Republicans users and 48 presumably Democrats) we can preprocess the data using cleaning and tokenization functions. We also use a stemmer and a phraser. Finally we remove stopwords. This latter step has to be made carefully so we do not lose words carrying meaning in the context of sentiment analysis such as negations for example. We remove words appearing only once as well as some of the words with the most appearances.

3.4 Modelling

For this political labelling we use a Word2Vec model as we believe it is important to take the contexts of the words into account when it comes to predicting a political stance. We then use a neural network (Mandav, 2018). We obtain an accuracy of 78% and a F1 score of 84%.

3.5 Comparison with the first labelling and qualitative analysis

We now use our still unlabelled analysis dataset (that we preprocess using our previously built cleaning functions) in order to predict two sorts of labels, the positive/negative labels developped in the first approach and the Republican/Democrat labels we just built.

We start by predicting the political labels. Over our 58,318 comments we end up with 45,204 Democrat-labelled comments and 13,114 Republican-labelled comments. The imbalance is a bit sharper than what we had in our manually-labelled dataset but it is therefore getting closer to the sociology of the NYT readers, 91% of whom declare themselves as Democrats (Grieco, 2020).

With the application of our SST-trained models to the NYT comments database, we have a harder time to get a proper quantitative evaluation than we had for the evaluation on the SST database itself (we now compare to the political labels, but this is only a comparison with our own previous results so there might be errors). We still get encouraging results, with F1-scores often around 60 to 70%. But it is more interesting to focus on the qualitative evaluation in this situation. With the *Word2Vec* and *TF-IDF* representations, we observe that previously Democrat-labelled comments are often labelled with positive feelings while Republicans will tend to be more negative.

These observations seem coherent with the topics to which they are related. Indeed, we linked comments and articles, and we find that the most common keywords associated with positively-received articles are quite close to those found with a Democrat cluster as developed in the last paragraph of this document. Likewise, the keywords we have associated with negatively-received articles are quite similar to Republican-attractive topics.

We now evaluate our political labelling on two aspects, adopting a qualitative approach. First, we look at the most-used words by the two camps in their comments thanks to wordclouds (*cf.* Appendix A.) and foind it is coherent with what we know about the two parties. Indeed the wordclouds reflect some of the obsessions of the two parties such as Barack Obama and immigration for the Republicans and the Russia investigation for the Democrats.

The second aspect we evaluate concerns the articles the two parties' advocates are interested in. For this part we apply our Word2Vec model on the articles and cluster them using a Kmeans algorithm. Then for each article we compute the proportion of Republican labelled comments it has received, if it is above the proportion of Republicans comments we have in our entire database (22%) we decide the article is Republican-attractive and vice-versa. The various articles' themes reflected by the clusters are sometimes clearly attached to one of the two parties, articles about gun control being much more attractive to Republicans than articles about the Russia investigation for example (*cf.* Appendix B.).

4 Conclusion

Our models trained on the SST dataset allows us, in most cases, to label the NYT comments in a positive/negative sense. With the comparison to the political labels, we find an interesting pattern that seems to associate Democrats with positive comments and Republicans with negative ones. We still encounters difficulties, especially with the LSTM model, performing on the SST database but hardly transposable to the NYT opinions because of the difference of vocabulary.

Our political labelling allows an analysis of the language elements of the two parties' advocates (with the wordclouds) as well as their topics of interests (with the clusterisation). The analysis is very general but it reflects quite well the polarisation of the American political scene.

References

Jatin Mandav. 2018. Sentiment analysis using word2vec, fasttext and universal sentence encoder in keras.

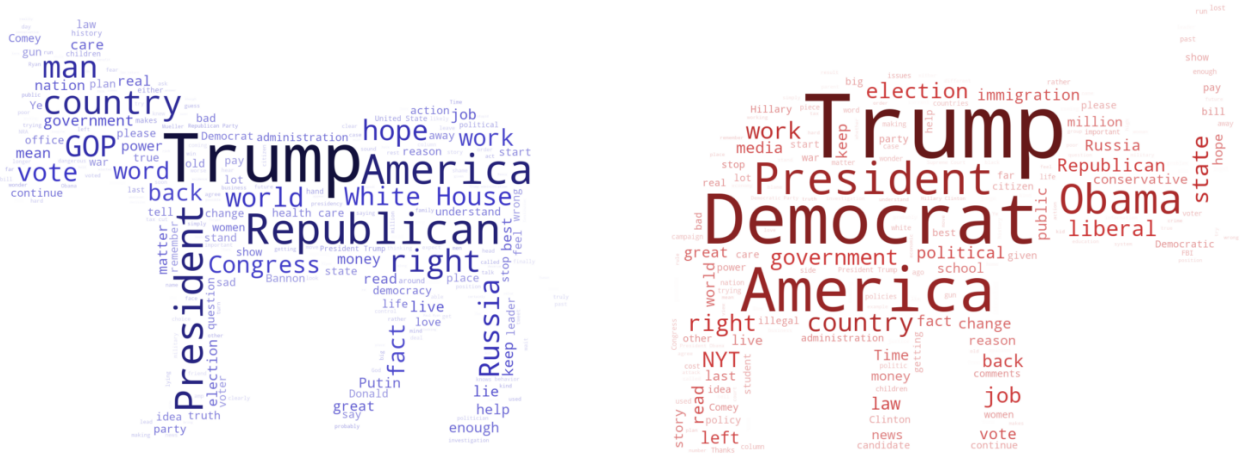
Duncan Cam-Stei. 2019. Word embedding explained, a comparison and code tutorial.

Elizabeth Grieco. 2020. Americans' main sources for political news vary by party and age.

Appendices

Appendix A.

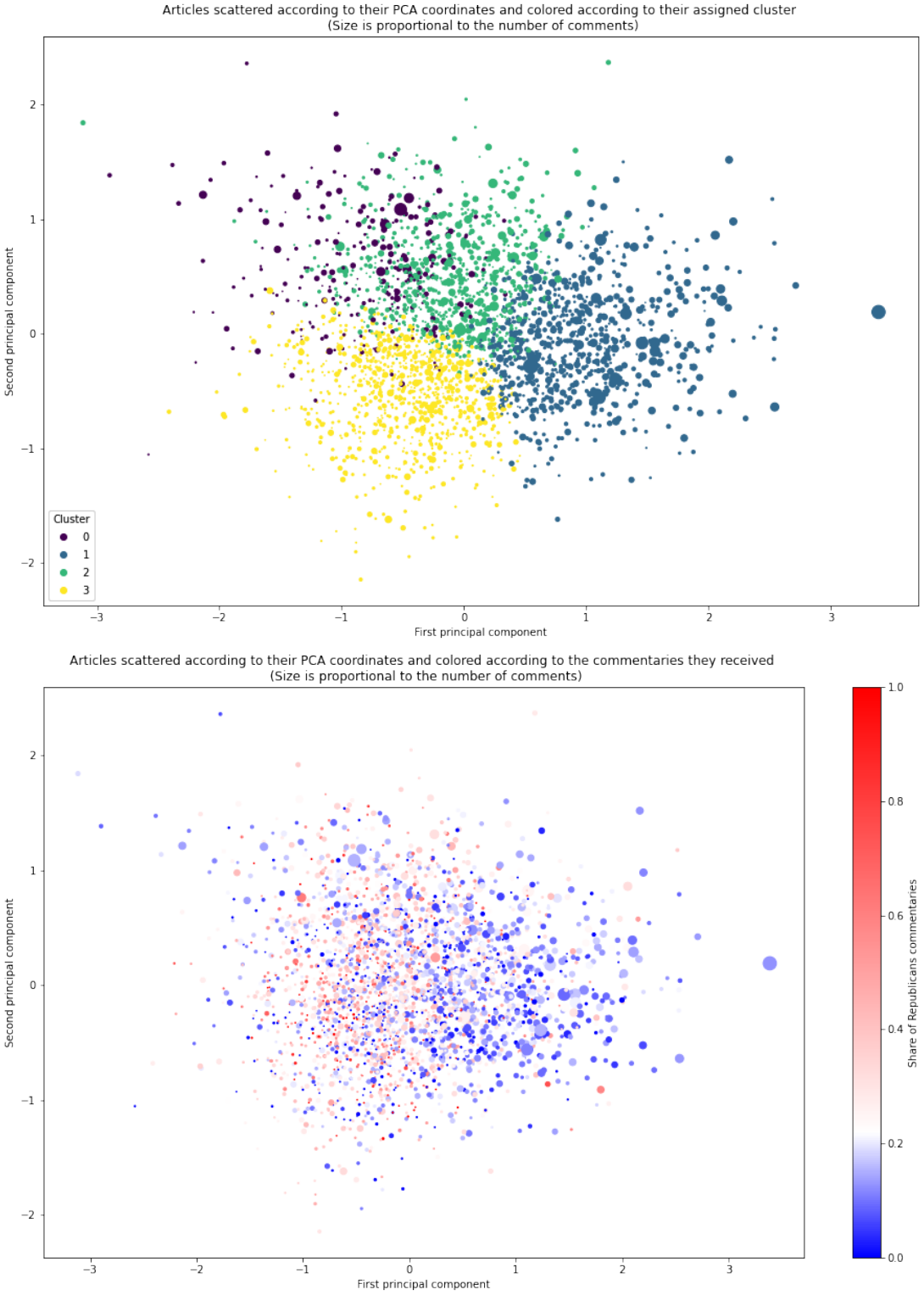
Figure 1: **Donkey *vs.* Elephant: how language elements differ between Democrats and Republicans?**



Republicans seem to talk a lot about Barack Obama which is easily explained by the fact he was one of the main target of Trump during the campaign either because of his political record or because of his alleged origins. Democrats are more focused on the investigation over Russian interference in the 2016 elections (words ”Russia”, ”Putin” and ”Comey”). The word ”lie” is also quite important for Democrats, which is coherent because we know they repeatedly accused Trump of lying, while the word ”immigration” seems to be mentioned a lot by Republicans, it was indeed one of the key points of Trump campaign. The two camps seem to be obsessed with each other since ”Republican” is one of the most frequent word for Democrats and vice-versa.

Appendix B.

Figure 2: **Interactions between articles' assigned clusters and the polarisation of the comments they received**



The first cluster is well reconstructed by the political categorization. Indeed we see that the vast majority of the articles at the right of the picture are blue, thus considered to be "Democrat-attractive". Clusters 2 and 3 appear to gather mostly Republican-attractive article despite the contrast is less sharp than for cluster 1. From the graphic it is hard to conclude anything regarding cluster 0.

Table 1: **Clusters' statistics**

	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Mean share of Republican-labelled comments	26%	19%	30%	31%
Number of articles considered Republican-attractive	121	275	515	559
Number of articles considered Democrat-attractive	135	540	320	331

The statistics of the clusters confirm what we see on the plots: Cluster 1 gathers a sharp majority of Democrat-attractive articles while clusters 2 and 3 gather a majority of Republican-attractive articles. Cluster 0 is more balanced. The next table allows to take look at the most frequent keywords among the articles of each cluster, thus giving an idea of the topics of interest of Democrats and Republicans.

Table 2: **Most frequent keywords depending on the cluster**

Cluster 0	Cluster 1	Cluster 2	Cluster 3
"Trump, Donald J" (162)	"Trump, Donald J" (702)	"Trump, Donald J" (440)	"Trump, Donald J" (235)
"US Politics and Government" (159)	"US Politics and Government" (605)	"US Politics and Government" (423)	"US Politics and Government" (218)
"Republican Party" (96)	"Russia" (129)	"Republican Party" (143)	"Women and Girls" (72)
"Health Insurance and Managed Care" (95)	"US International Relations" (125)	"Democratic Party" (105)	"Blacks" (59)
"Patient Protection and Affordable Care Act" (78)	"Russian Interference in 2016 US Elections" (122)	"Immigration and Emigration" (92)	"Colleges and Universities" (49)
"House of Representatives" (70)	"Presidential Election of 2016" (116)	"Senate" (89)	"School Shootings and Armed Attacks" (47)
"Senate" (44)	"Federal Bureau of Investigation" (101)	"United States International Relations" (79)	"Gun Control" (45)
"Federal Budget (US)" (38)	"Republican Party" (87)	"House of Representatives" (70)	"Republican Party" (44)
"Law and Legislation" (28)	"Comey, James B" (64)	"US Defense and Military Forces" (54)	"Demonstrations, Protests and Riots" (43)
"Ryan, Paul D Jr" (24)	"Mueller, Robert S III" (62)	"Politics and Government" (51)	"Parkland, Fla, Shooting (2018)" (42)
"Democratic Party" (23)	"Special Prosecutors (Independent Counsel)" (58)	"Supreme Court" (51)	"Education (K-12)" (36)
"Medicaid" (20)	"Justice Department" (52)	"Illegal Immigration" (50)	"United States" (35)

Looking at the keywords we understand why the graphic representation identified the cluster 1 as a cluster gathering Democrat-attractive articles. The most frequent keywords of the articles of this cluster reveal it gathers the articles dealing with the investigation on the Russian interferences in the 2016 elections (Mueller and Comey led the investigations). The accusation was led by Democrats and directed against Trump so it is coherent that the articles of this cluster are mostly Democrat-attractive articles.

Cluster 0 gathers more diversed articles but the dominant theme is health insurance with three keywords related to it appearing in the top 10 (excluding the first two keywords that are common to all the clusters. This cluster has a sort of intermediate position in our classification as it is not the most Democrat neither the most Republican. Thus health insurance can be viewed as a topic interesting both parties depsite the fact they have diverging opinions about it.

Clusters 2 and 3 are the most Republicans. It is therefore no surprise that cluster 2 is the cluster that gathers the most articles dealing with immigration as well as military forces which are traditionnally Republican themes. Many articles regarding mass shootings and gun control are gathered in cluster 3. We know that Republicans are usually staunch opponents of gun control so it is coherent that this theme attracts more of their responses than others. Racial justice seems also to be one of the themes of cluster 3 with the keywords "Blacks" and "Demonstrations, Protests and Riots".