Multidimensional Signal Processing
Project
A.Y. 2017/2018

Analysis of Social Media's Impact on the US Presidential Election

Candidates:

Giuseppe PIACENZA

Antonio SURIANO

Index

Goals and Dataset Presentation

The main goal of our project is to provide several studies of social media streams (with particular reference to Twitter) in order to analyze their impact on the U.S. Presidential Election of 2016. To do that, we made use of a dataset made of 6444 records each corresponding to a particular tweet written or retweeted by Donald Trump and Hillary Clinton during the early 2016 campaign trail. Every record is provided with 28 features, but only a part of them has proved to be useful for our project.

The dataset can be downloaded for free here:

https://www.kaggle.com/benhamner/clinton-trump-tweets/data

Data Preprocessing

To carry on the preprocessing of our data we exploited a Toolbox called «Text Analytics Toolbox» thanks to which we have been able to clean our data doing the following operations:

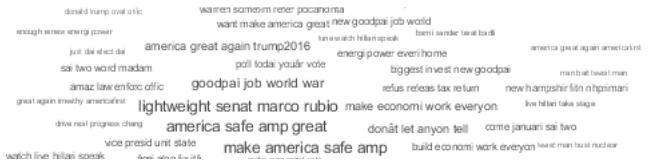
```
rng('default')
filename = "tweets.csv";
data = readtable(filename,'TextType','String');
idx = strlength(data.text) == 0;
data(idx,:) = [];
textData = data.text;
textData(1:6444);
cleanTextData = erasePunctuation(textData);
cleanTextData(1:6444);
cleanTextData = lower(cleanTextData);
cleanTextData(1:6444);
cleanDocuments = tokenizedDocument(cleanTextData);
cleanDocuments(1:6444);
cleanDocuments = removeWords(cleanDocuments,stopWords);
cleanDocuments(1:6444);
cleanDocuments = removeShortWords(cleanDocuments,2);
cleanDocuments = removeLongWords(cleanDocuments,15);
cleanDocuments(1:6444);
cleanDocuments = normalizeWords(cleanDocuments);
cleanDocuments(1:6444);
```

- ❑ **Elimination of punctuation**
- ❑ **Conversion of uppercase characters to lowercase**
- ❑ **Tokenize and normalize the document**
- ❑ **Elimination of stop words**
- ❑ **Elimination of short words**
- ❑ **Elimination of long words**

Frequent Sentences

In order to understand which are the most recurrent topics in Hillary Clinton's and Donald Trump's tweets we operated a text mining using bag of words and wordclouds.

```
bag = bagOfNgrams(cleanDocuments,'NGramLengths',4);
figure
wordcloud(bag)
title("Tweets")
tbl = topkngrams(bag,100)
```



	Ngram			Count	NgramLength
"make"	"america"	"great"	"again"	104	4

Wordcloud and ranking of the most tweeted sentences made of four words

Wordcloud and ranking of most tweeted sentences made of two words

Tf-Idf Matrix 1/2

Tf-idf stands for term frequency-inverse document frequency. It's a weighting statistical measure used to evaluate how important a word is to a document in a collection or corpus. In *term frequency*, the importance increases proportionally to the number of times a word appears in the document but in *inverse document frequency* an idf factor is incorporated which diminishes the weight of terms that occur very frequently in the document set and increases the weight of terms that occur rarely.

From an analytical point of view:

$$\mathrm{tfidf}(t, d, D) = \mathrm{tf}(t, d) \cdot \mathrm{idf}(t, D)$$

In Matlab R2018a the Tf-Idf Matrix is made as follows:

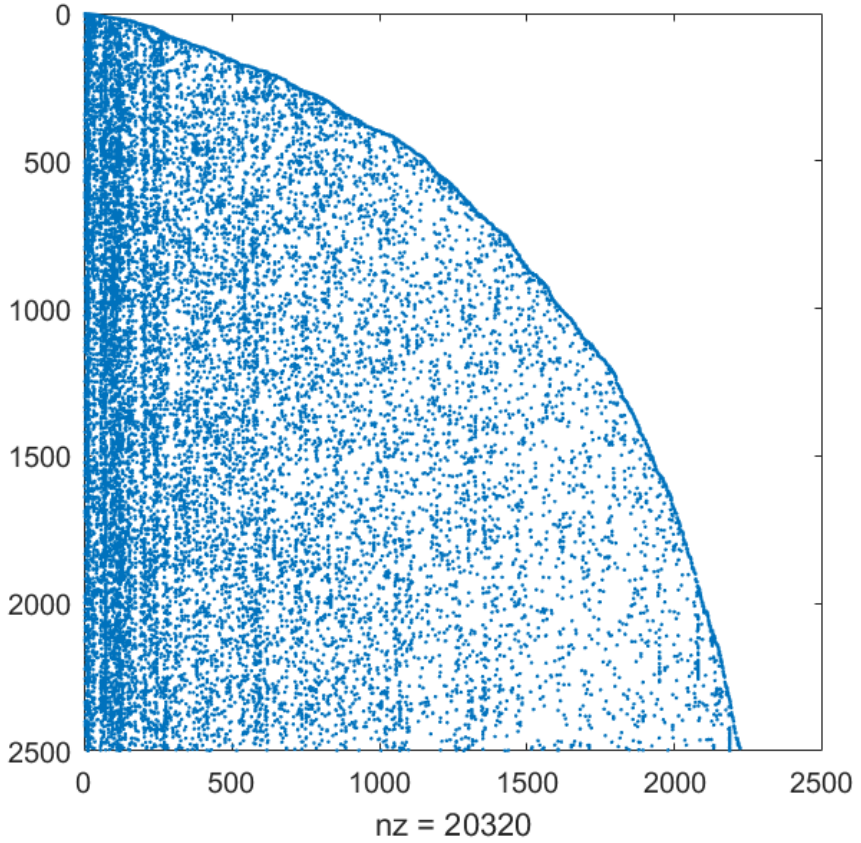
```
bag=bagOfWords(cleanDocuments);
bag=removeInfrequentWords(bag,2);
matrix=tfidf(bag);
full(matrix(1:10,1:10))
```

5.1333	3.9834	4.4271	3.8957	5.3697	2.7071	4.7279	4.3050	0	0
0	0	0	0	0	0	0	0	3.7942	4.0173
0	0	0	0	0	0	0	0	3.7942	4.0173
0	0	0	0	0	2.7071	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	3.9834	0	0	0	2.7071	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0

Tf-Idf Matrix 2/2

We used the matrix as an input of a spy() function, thanks to which we plotted its sparsity pattern

```
k=2500;
figure
spy(matrix(1:k,1:k))
```



Sparsity pattern of the Tf-Idf Matrix counting 20320 non-zero elements

LDA 1/4

A Latent Dirichlet Allocation (LDA) model is a topic model which discovers underlying topics in a collection of documents and infers word probabilities in topics.

```
numTopics = 60;
mdl = fitlda(bag,numTopics);
figure
for topicIdx = 1:4
    subplot(2,2,topicIdx)
    wordcloud(mdl,topicIdx);
    title("Topic: " + topicIdx)
end
```

LDA 2/4

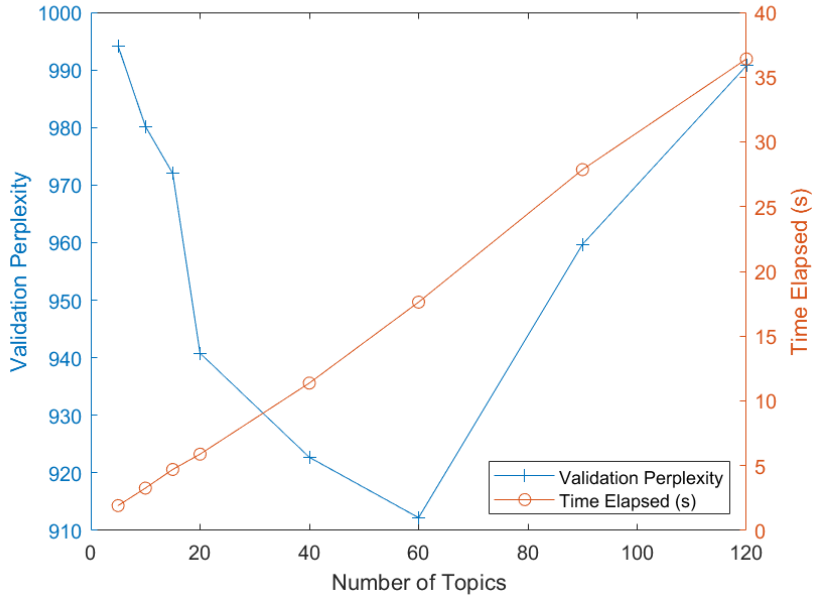
To decide on a suitable number of topics, we compared the goodness-of-fit of LDA models fit with varying numbers of topics. Thus, we could evaluate the goodness-of-fit of an LDA model by calculating the perplexity of a held-out set of documents.

The perplexity indicates how well the model describes a set of documents. A lower perplexity suggests a better fit.

The goal is to choose a number of topics that mini the perplexity is lowest compared to other numbers of topics.

LDA 3/4

The plot suggests that fitting a model with about 40–70 topics may be a good choice since the perplexity is low compared with the models with different numbers of topics.
With this solver, the elapsed time for this many topics is also reasonable.
We choosed a number of topics equal to 60 because this number of topics fits the model.



Number of topics

LDA 4/4

We choosed to select 4 main topics and to plot their wordclouds:

Topic 1

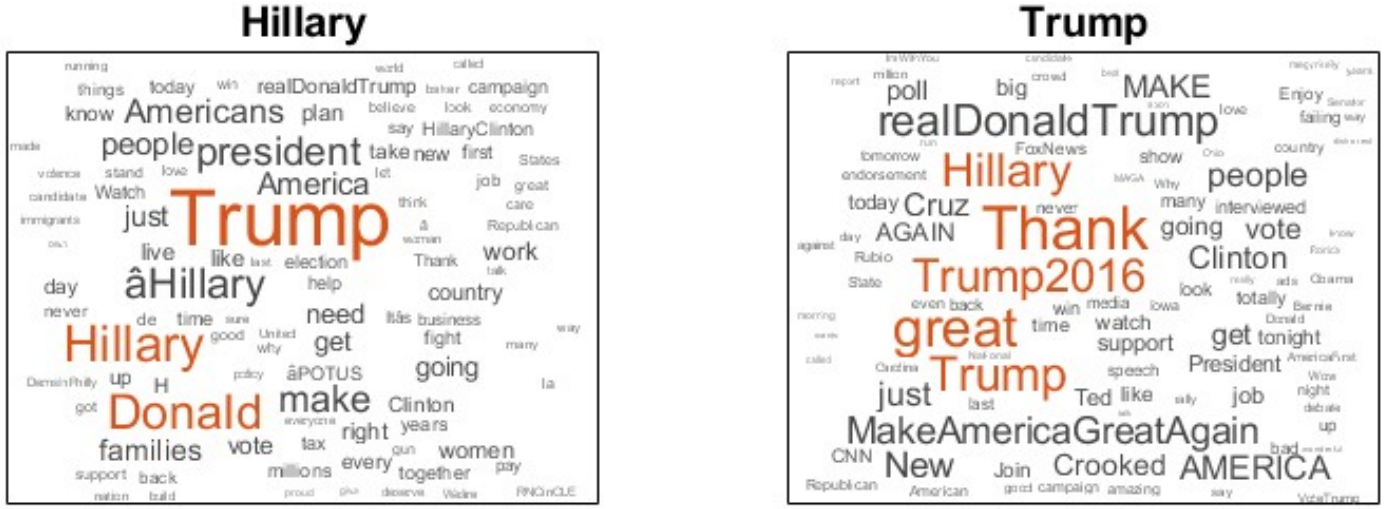
Topic 2

Topic 3

Topic 4

Most Tweeted Words by Candidates

We made a division of the data according to their handle. Then we analized the tweets to find out the most used words; then we plotted the results through *wordclouds*, respectively for Hillary Clinton and Donald Trump.



Wordclouds for Hillary Clinton and Donald Trump

Places and Languages 1/3

Using histograms, we plotted the Nations, the cities and the States where the tweets were posted and the languages in which they were written.
Due to data incompleteness, the selection was made by preparing the data discarding all the records for which the values of the attrbibute were missing. In particular, we noticed that only 6 tweets were posted in the UK and 198 in the USA.

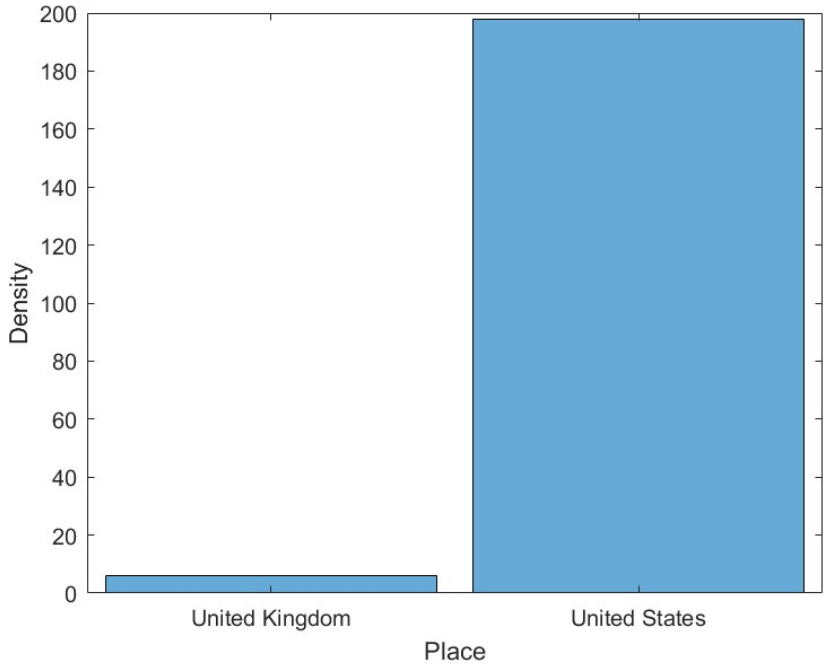


Chart for the nations

Places and Languages 2/3

For what concerns the different cities and states were the tweets were posted, we have the following results:

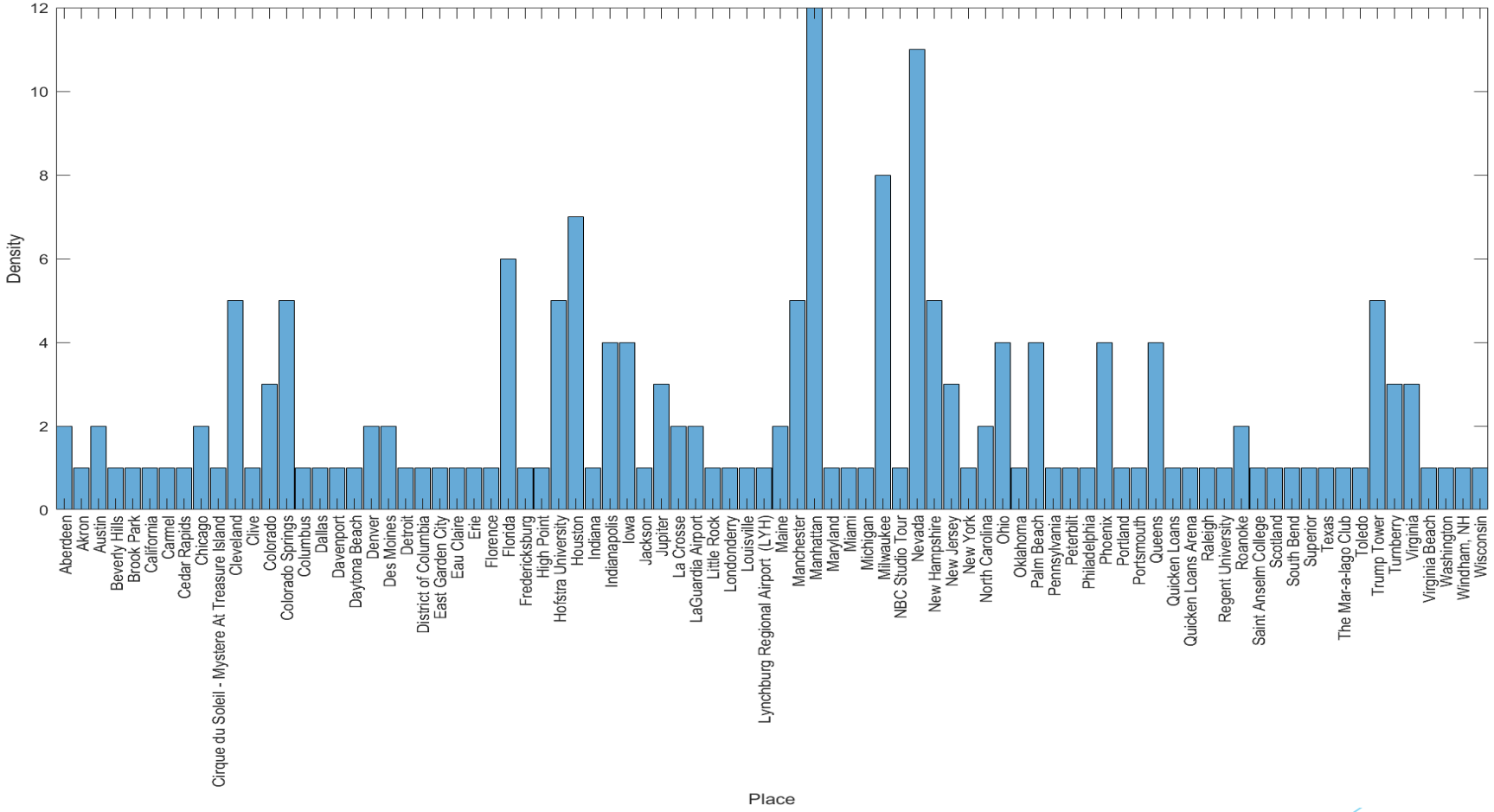


Chart for cities and States

Places and Languages 3/3

Almost all the tweets (6248) were written in English (en), but few of them (105) were written in Spanish (es). In this case we considered only the most occurring languages, neglecting the ones with only few tweets.

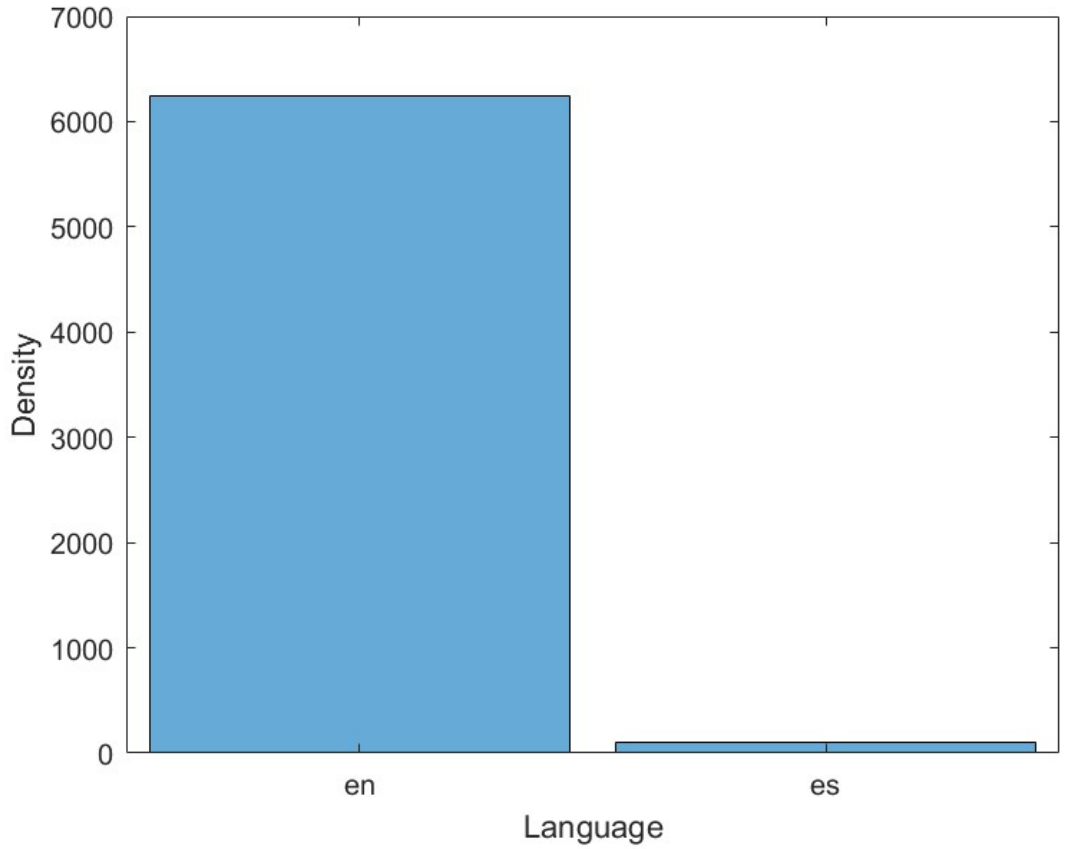
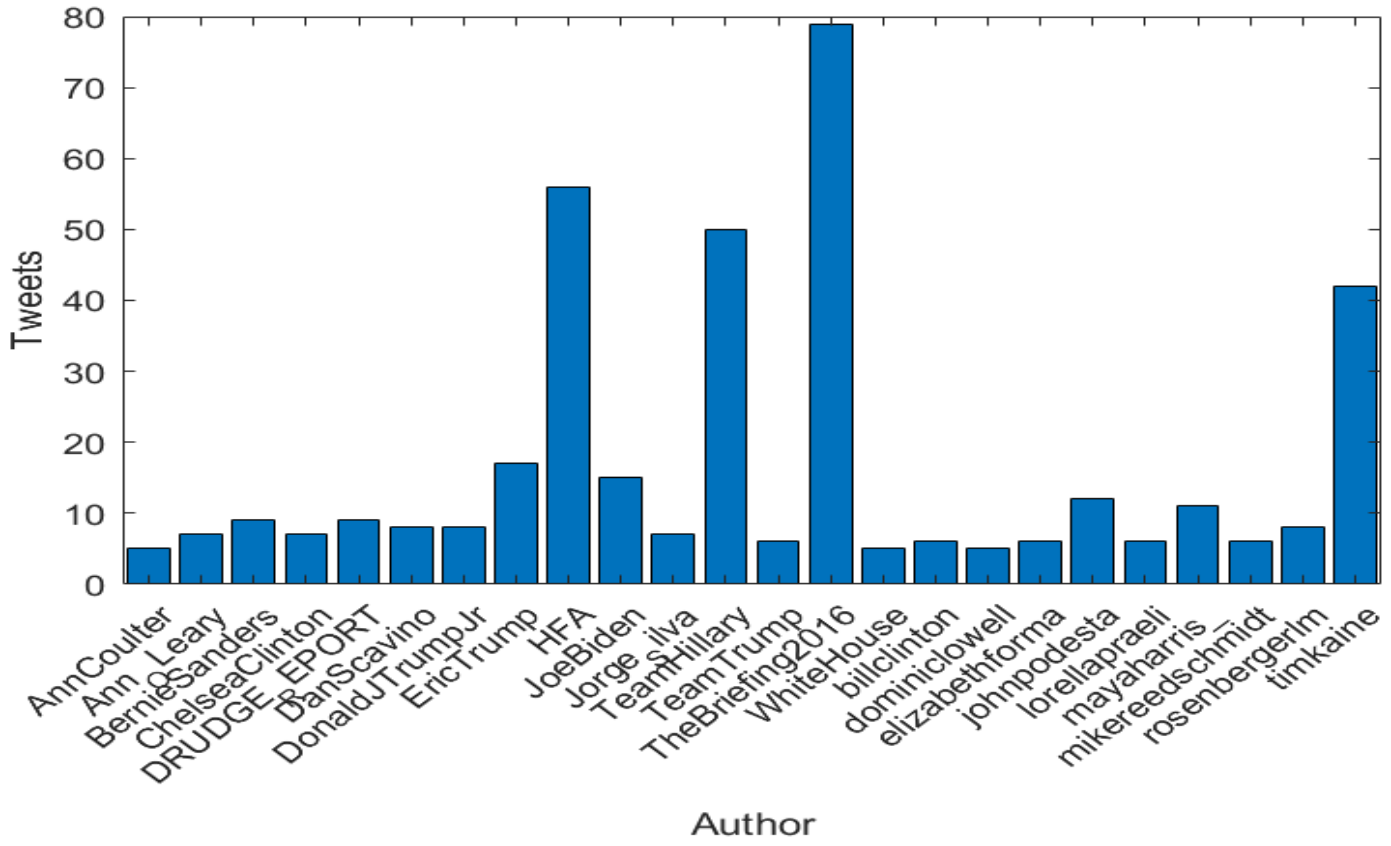


Chart for the language

Retweets 1/3

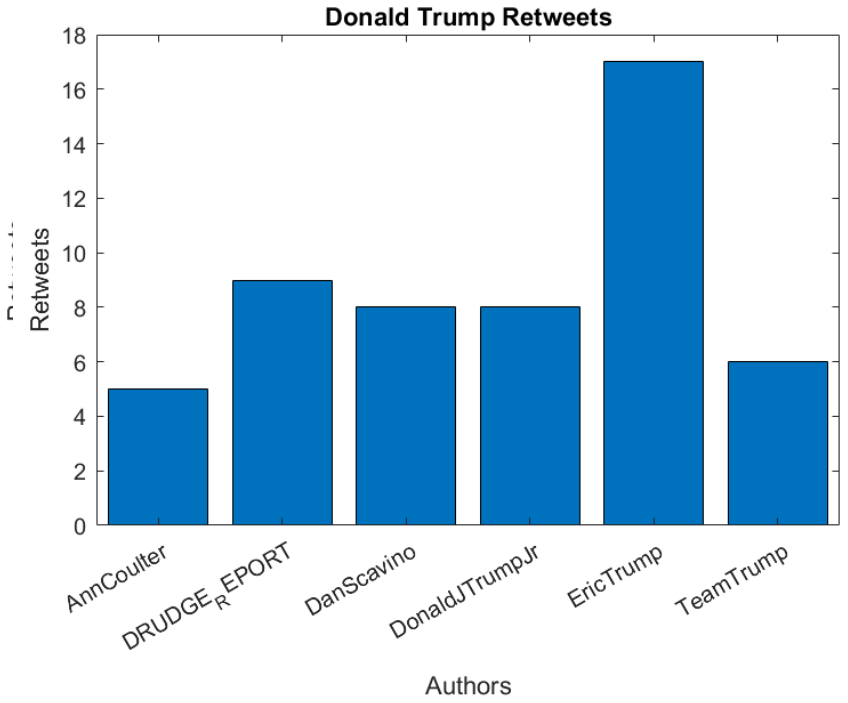
We focused on the rewteets from other authors; in this situation we filtered the authors with at most 5 retweets.
The results are shown in the following histogram.



Histogram of the retweeted authors

Retweets 2/3

In particular, Donald Trump and Hillary Clinton retweeted from different authors with different frequencies:

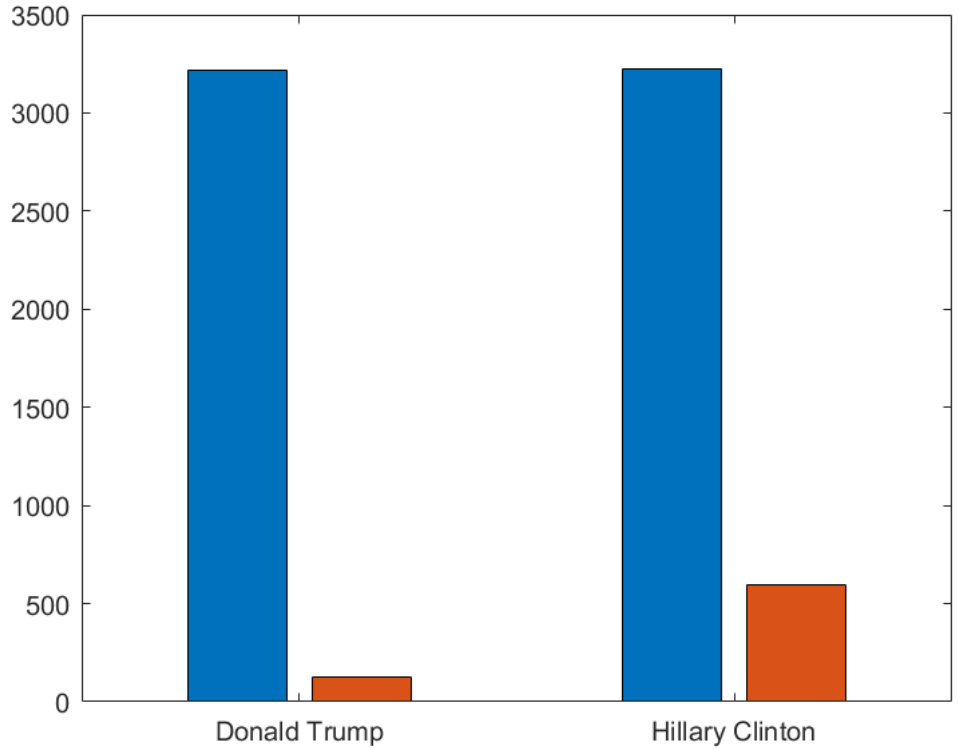


Authors retweeted by Donald Trump and Hillary Clinton

Retweets 3/3

We used two bar charts to compare the number of retweets with respect to the total number of tweets, both for Donald Trump and Hillary Clinton.

- Blue: total number of tweets;
- Red: number of retweets.

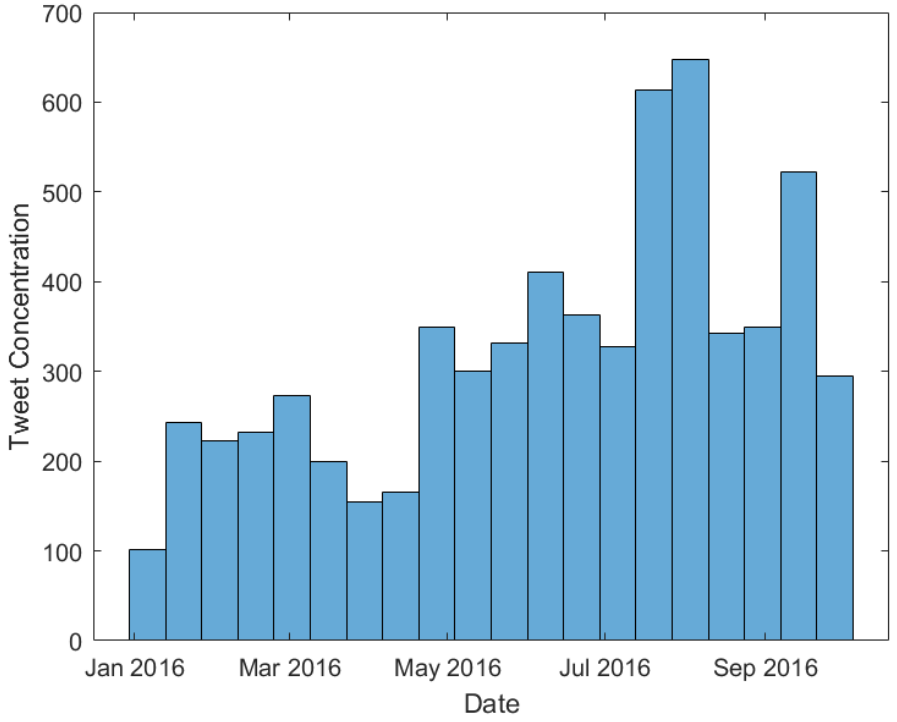


Tweets-Retweets bar charts for Donald Trump and Hillary Clinton

Date Analysis 1/2

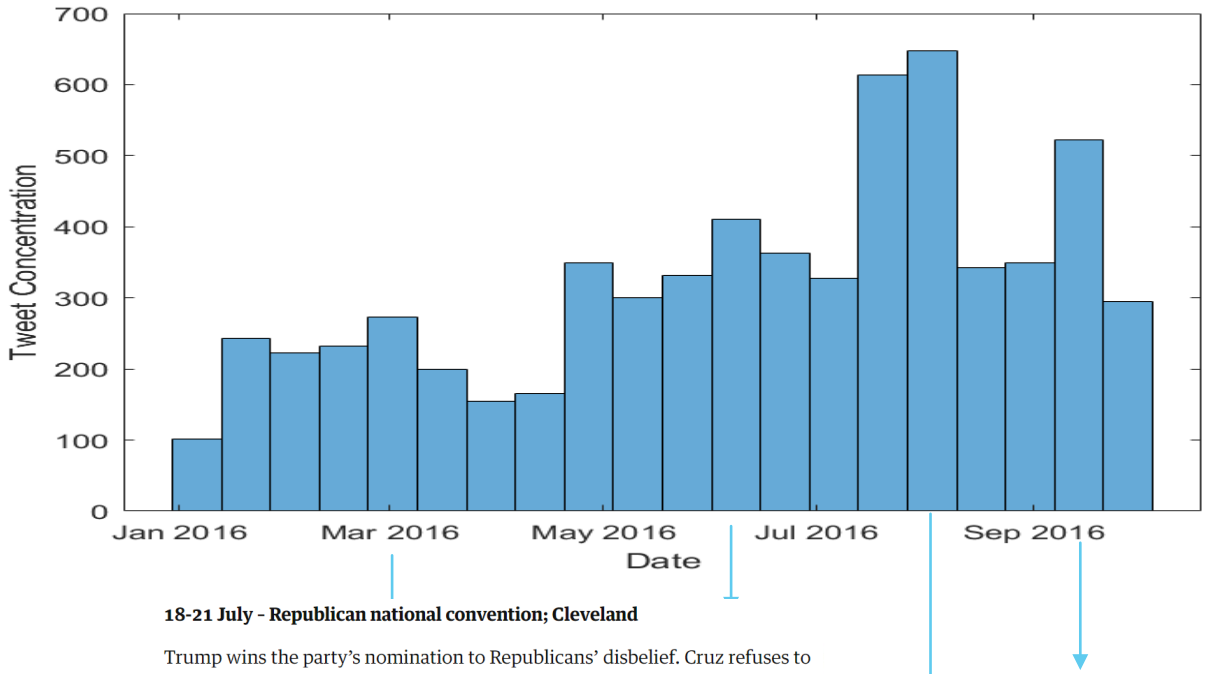
Another analysis that we carried on regards the dates in which tweets were posted. In order to do that, we generated an histogram that displays the tweet stream during the period of time in which tweets in our dataset are sampled.

```
date=data.time;
format='yyyy-mm-ddTHH:MM:SS';
t=datenum(date,format);
figure
title("Tweet Distribution");
d = datetime(t,'ConvertFrom','datenum');
h=histogram(d)
xlabel('Date')
ylabel('Tweet Density')
```



Date Histogram

Date Analysis 2/2



18-21 July - Republican national convention; Cleveland

Trump wins the party's nomination to Republicans' disbelief. Cruz refuses to

March - 26 September - First presidential debate; Hempstead, New York

https://www.theguardian.com/us-news/2016/nov/07/us-election-2016-complete-timeline-clinton-trump-president

The Rep Most watched debate in American history, with at least 80m people tuning in. Clinton comes prepared, Trump not so much. She baited him with charges of racism, sexism and tax avoidance and he took the bait - every time.

outlawe

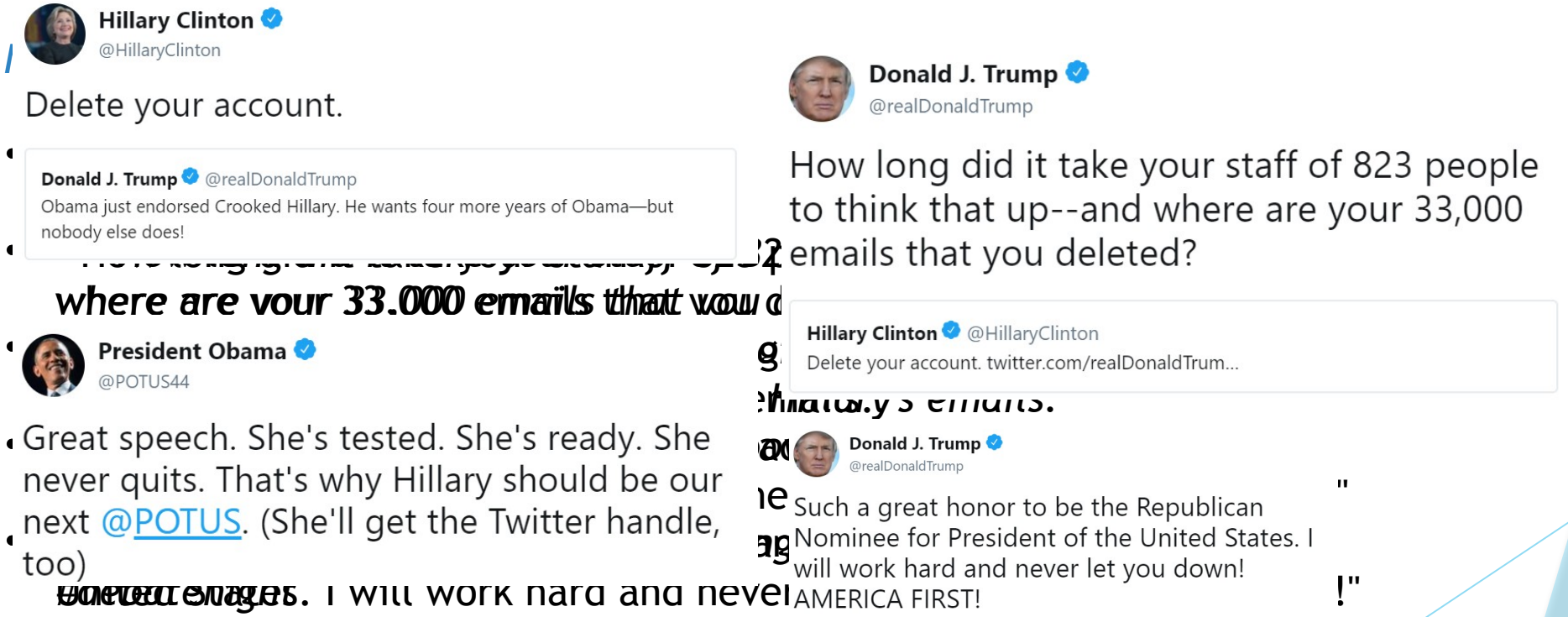
stateme

in the US:

"Standing here as my mother's daughter, and my daughter's mother, I'm so happy this day has come." She added: "When any barrier falls in America, for anyone, it clears the way for everyone."

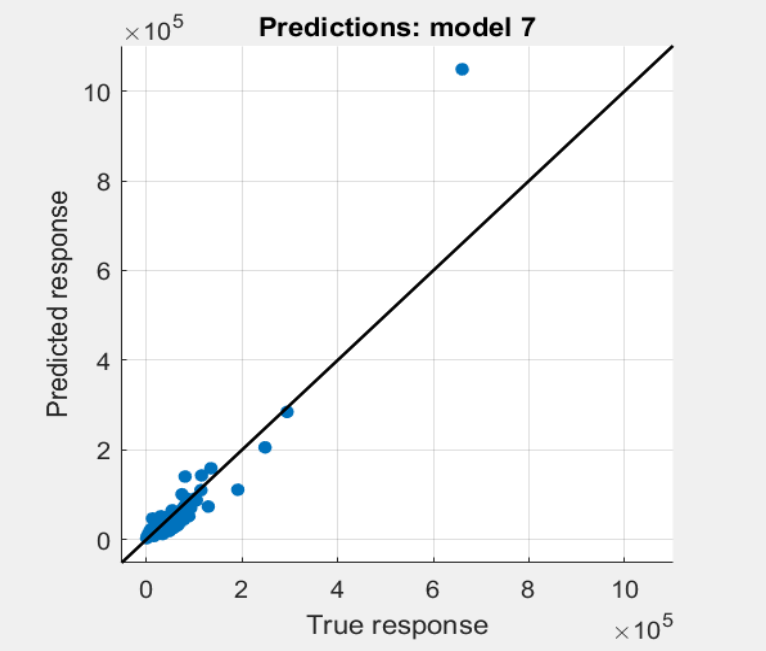
Data Ranking

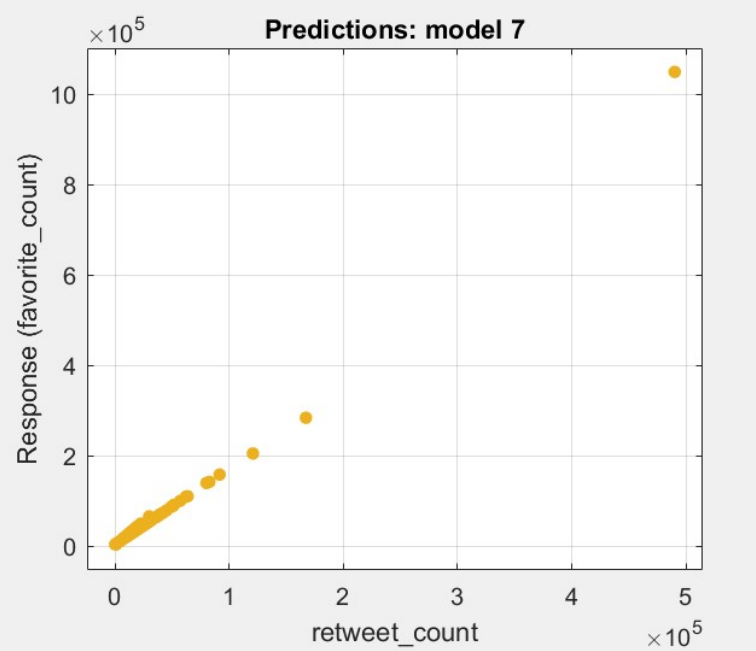
By considering the features retweet_count and favorite_count, which contain respectively the number of retweets and the number of like for each tweet it's possible to rank the data displaying the most favourite and retweeted ones.



Regression 1/5

The regression is the most used model for estimation problem. Our goal is to predict the number of likes of the tweets, starting from the number of retweets. In the following pictures we can see the distribution of the real values and the distrbution of the predicted ones, where the black line corresponds to a perfect prediction.





Real values vs Predicted values

Regression 2/5

To deeply understand if a linear regression model is suitable for our case, we computed the Pearson coefficient. It ranges from −1 to 1; a value of 1 implies that a linear equation describes the relationship between X and Y perfectly, with all data points lying on a straight line for which Y increases as X increases.
In our case the Pearson coefficient is equal to 0.9274, so the data lie on an almost straight line with positive slope and we can use the linear regression to achieve better results.
The formula for the Pearson coefficient is:

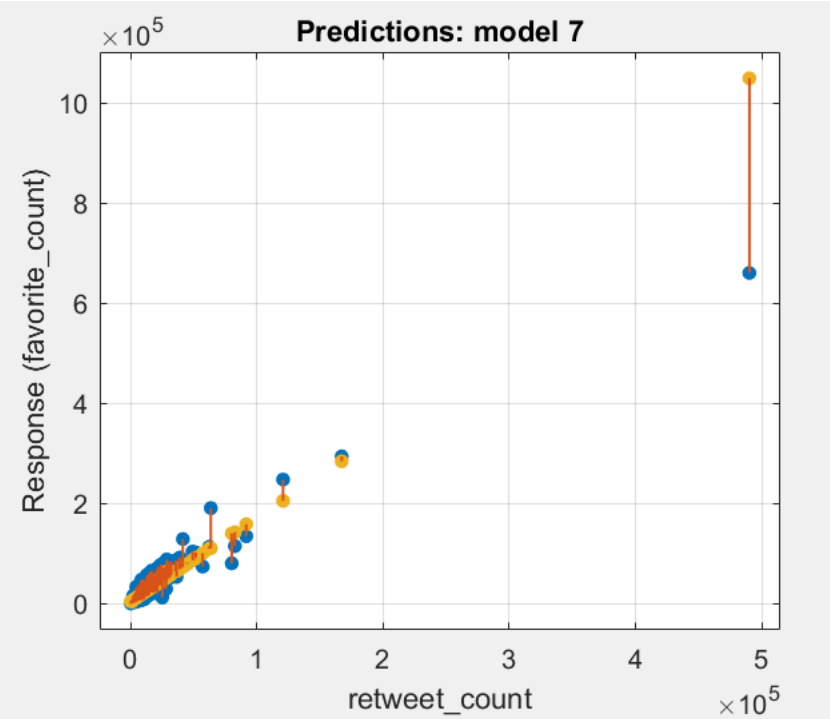
$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

Where cov(X, Y) is the covariance and σ_X and σ_Y are the standard deviations of X and Y, respectively.

```
R=data.retweet_count;
F=data.favorite_count;
covariance = cov(R,F);
pearson_coeff=covariance(2)/(std(F)*std(R))
```

Regression 3/5

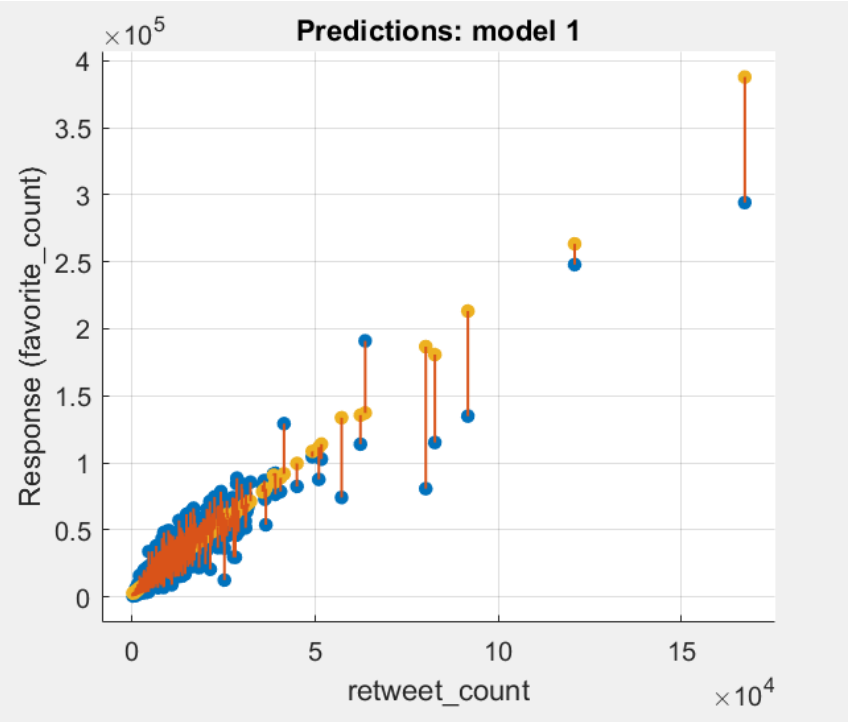
The errors between the real and the predicted values are shown in the following pictures, where the vertical lines represent the residuals.



Errors

Regression 4/5

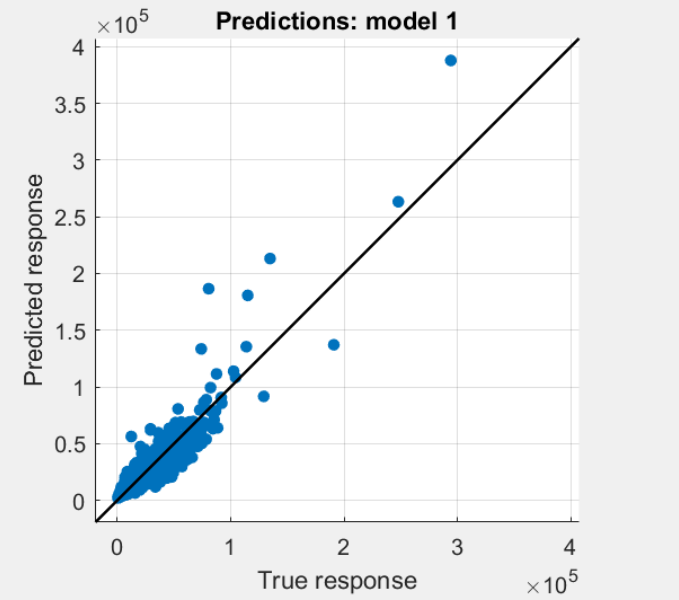
The R2 value obtained from the computation is 0.78, a quite good result.
The algorithm uses only the variable *retweet_count*.
To improve the results we can eliminate the outlier (the results are shown in the figure below), obtaining an R2 value of 0.86.

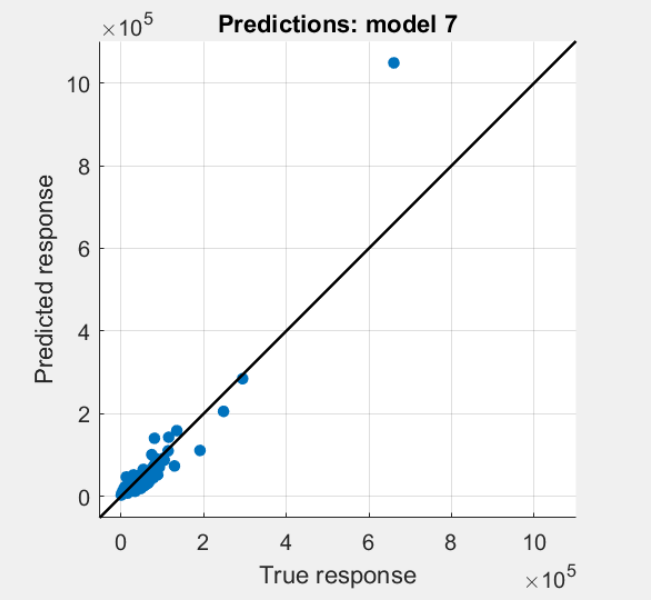


Errors without outlier

Regression 5/5

In the following figures we can see the difference between the regression model on the original dataset (on the right) and the regression model without the outlier (on the left). In both cases, the straight black line represents the ideal prediction.

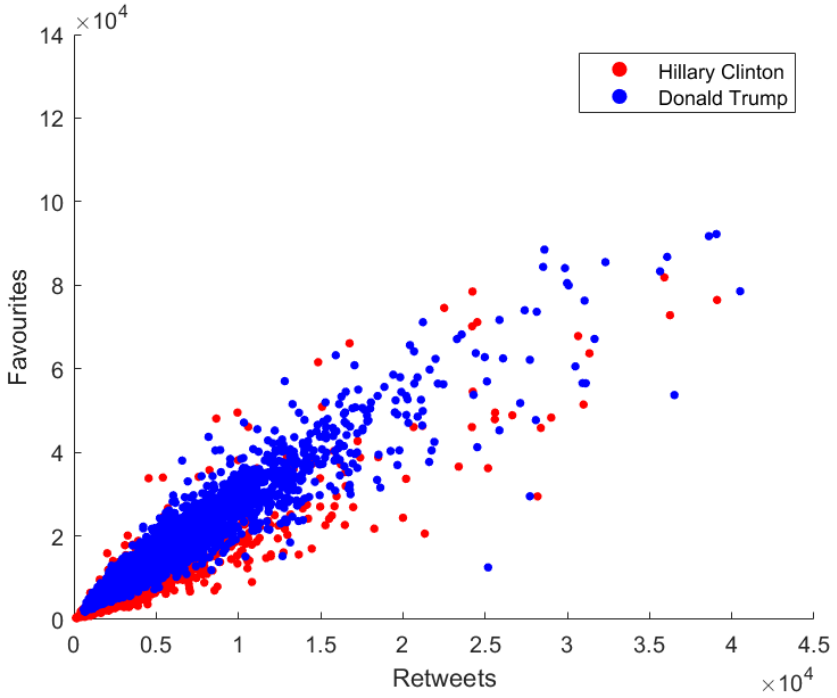




Regression model with and without outlier

Classification 1/4

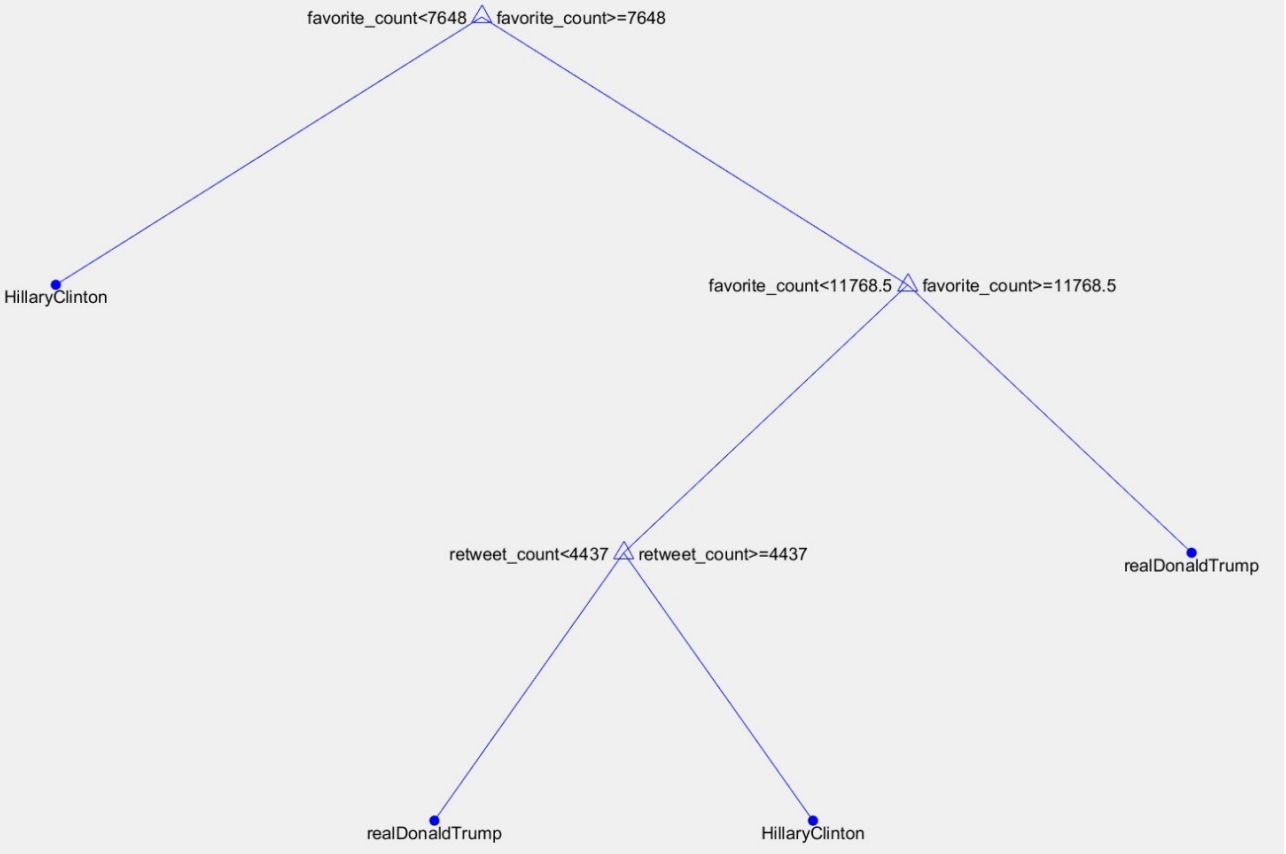
Classification is a supervised learning method for predicting value of a categorical target attribute. To train our classification model we used tenfold cross-validation.



Scatterplot

Classification 2/4

With selected only two features: *retweet_count* and *favourite_count*, and, for a good legibility, we set up *coarse tree*.



Classification Tree

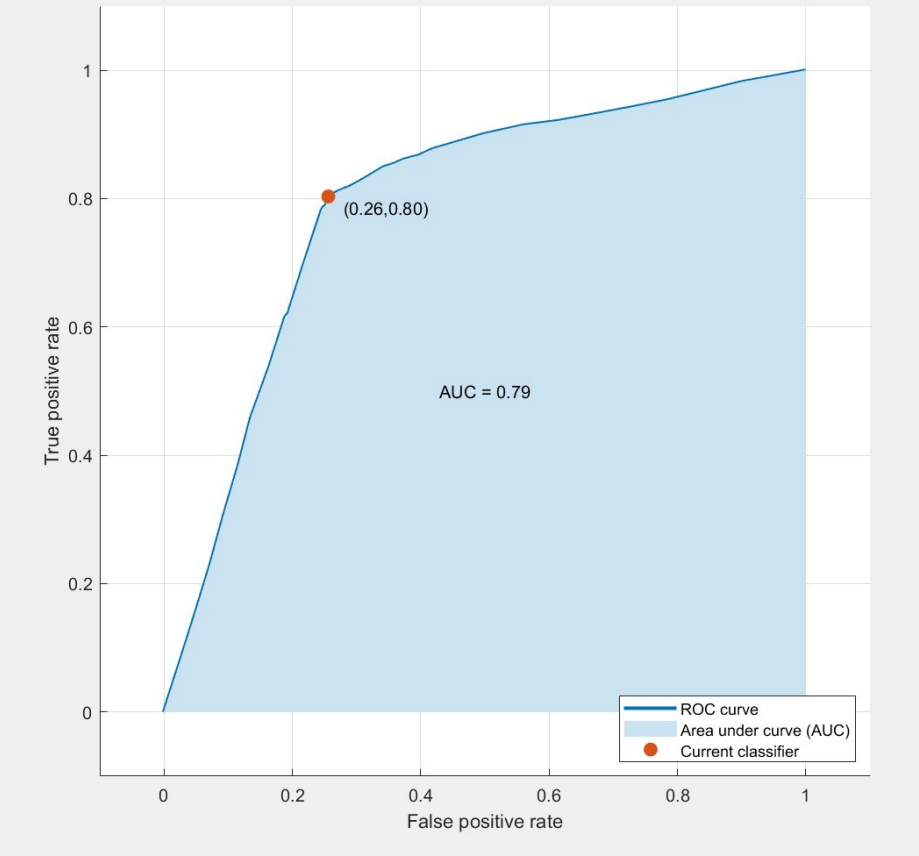
Classification 3/4

The accuracy of the classification model is 77.37%, its performance can be evaluated by considering the confusion matrix and the ROC curve.

Confusion Matrix

- Accuracy =(TP+TN)/(Tp+TN+FP+FN)= 77.37%
- Sensitivity (or recall) = TP/(TP+FN) = 80.35%
- Specificity (or TN rate) = TN/(TN+FN) = 74.40%
- Precision = TP/(TP+FP) = 75.88%

Classification 4/4



ROC Curve

Conclusions

Thanks to this analysis, it is possible to understand how social media played a crucial role on the 2016 U.S. Presidential Election. Furthermore, we can highlight the deep differences between the main themes and feelings that characterized Hillary Clinton's and Donald Trump's campaign trails.

Some future goals consist in:

- ❑ Optimize the LDA model
- ❑ Improve the classification analysis

THANKS FOR YOUR ATTENTION