Natural Language Processing Project: Analyzing the tweets by Hilary Clinton and Donald Trump During the US President Election in 2016

Instructed by Professor Mariano Rico (Universidad Politécnica de Madrid)
Delivered by Soo Min Jeong (EIT Digital Master - Data Science)

1. Scope

This project is designed to implement data analysis mainly focusing on natural language processing (NLP) on the dataset of tweets by Hilary Clinton and Donal Trump during the US president election in 2016.
Twitter announced permanent suspension on Donald Trump's account due to the risk of further incitement of violence.[1] Politicians including the president of the US deliver their political message on Twitter. Since the political communication on social networks is getting more prevalent, this project expects to answer some questions in how politicians utilize this unconventional media.
The source code is archived here: https://github.com/soomin-jeong/nlp-trump-clinton-tweets

2. Description of Dataset

The dataset is originally from Kaggle, called 'Hilary Clinton and Donald Trump Tweets'.[2] The source is also provided in the 'README.md' file. The dataset provides 6,434 tweets. The main entities of the dataset are:

- handle: the username of the tweet (*HilaryClinton*, *realDonaldTrump*)
- text: the contents of the tweet
- is_retweet: whether the tweet was retweeted or organic (written by the owner of the account)
- in_reply_to_user_id: if this tweet was a replying tweet to a user
- retweet_count: how many times it was retweeted
- favorite_count: how many times it was marked favorite

3. Preprocessing (preprocess_data.R)

Before applying the NLP approaches, the dataset is preprocessed in this section. Other analysis scripts source the preprocess script in the beginning by calling `source('preprocess_data.R')`

a. Removing unnecessary characters

As a tweet may include a username, a hyperlink to image, website, another tweet, special characters including the hashtags, the preprocessing script was developed to handle them.

b. Filtering organic tweets

An organic tweet is a tweet that the owner of the account posted. The retweeted tweets and replies are excluded here.

c. Dividing the tweets by the user

To compare the tweets by the user, the tweets are divided by the author in advance.

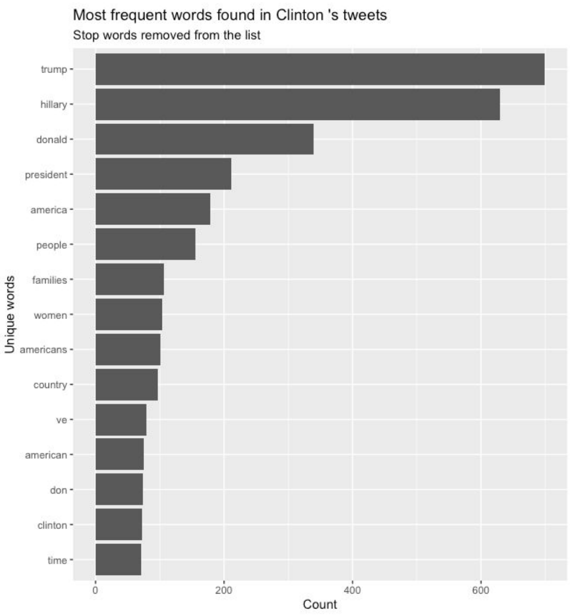
4. Data Analysis with Natural Language Processing

a. Descriptive Analysis (descriptive_analysis.R)

[1] https://blog.twitter.com/en_us/topics/company/2020/suspension.html
[2] https://www.kaggle.com/benhamner/clinton-trump-tweets

This section answers basic questions such as: *How many tweets did they make?, How many retweets did they get?, Which tweets got the most retweets for each candidate?, Which tweets got the most favorites? Which words did they mention the most?*

The last question was answered in a bar chart. Apparently they both mentioned each other the most frequently. Clinton mentioned ‘president’ more frequently than Trump and focused on mentioning the keywords: ‘trump’, ‘hilary’, ‘donald’. In the meantime, Trump has a relatively broader distribution on his topics.



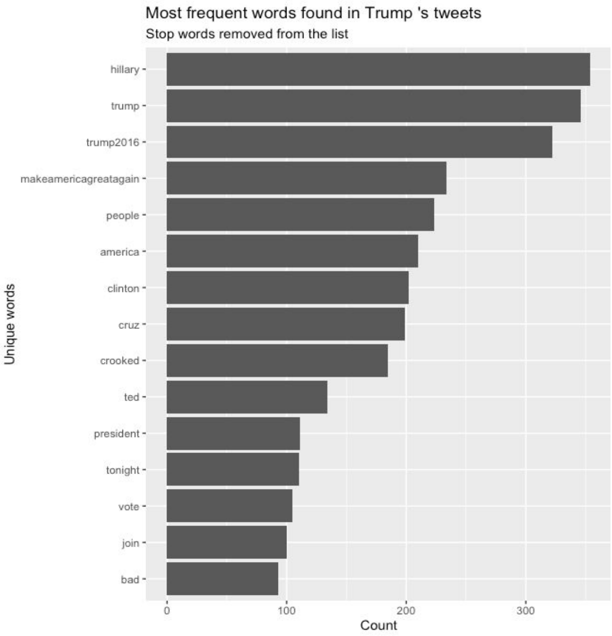
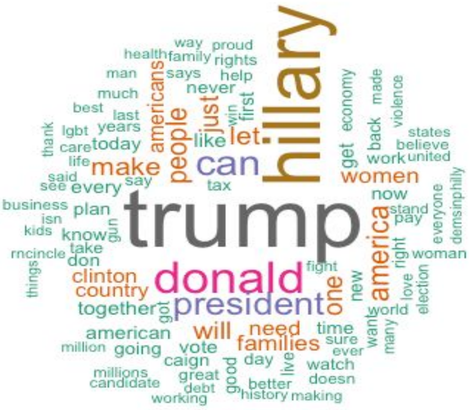


Figure 1: Most Frequent Words Found in Clinton’s Tweets *Figure 2: Most Frequent Words Found in Trump’s Tweets*

(Answers are not included here due to the designated limit; however, they are on the R script)

b. Word Cloud (wordcloud.R)

Word cloud visualizes what the tweets consist more intuitively. The dataset was preprocessed and transformed into a datatype of Corpus. Numbers were removed and capital letters were lowered for better visualization. The word count assigned a proportional size to each word based on their frequency. By setting the lower limit of the frequency and the upper limit of total words in the word cloud, it can be represented more intuitively.



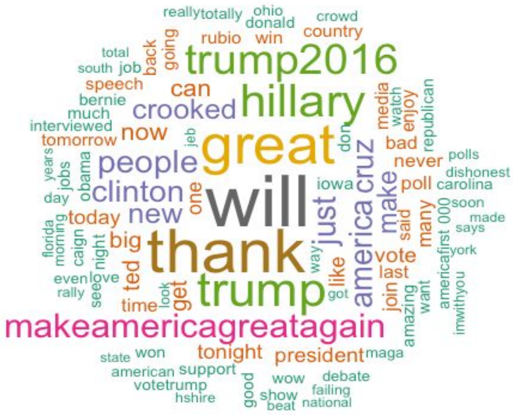


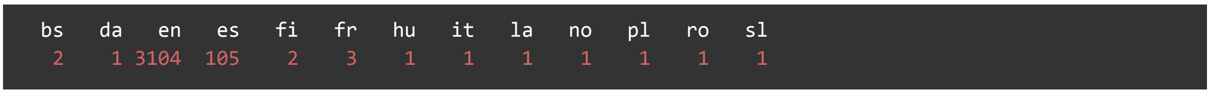
Figure 3: Word Cloud of Clinton’s Tweets *Figure 4: Word Cloud of Trump’s Tweets*

The word cloud above indicates that they both talked about each other frequently; however Clinton paid more attention to Donald Trump. Trump made some strong intention with words like 'great', 'will', or his slogan 'makeamericagreatagain'

c. Language Detection (language_detction.R)

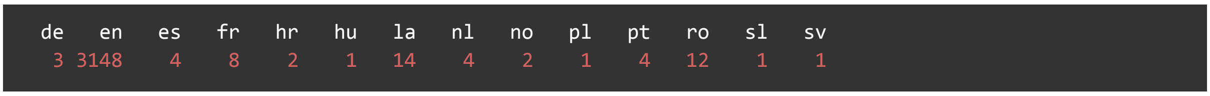
Though most of the tweets were made in English, they both made some tweets in other languages. To detect which language the tweet was written in, a library called 'textcat' was used. There are several profiles as an option to detect a language, and 'ECIMCI_profiles' was chosen due to its high accuracy in spite of its slow speed.

Clinton's tweets are in such languages:



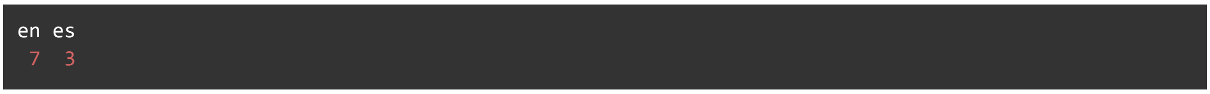
```
  bs   da   en   es   fi   fr   hu   it   la   no   pl   ro   sl 
   2    1 3104  105    2    3    1    1    1    1    1    1    1 
```

Trump's tweets are in such languages:



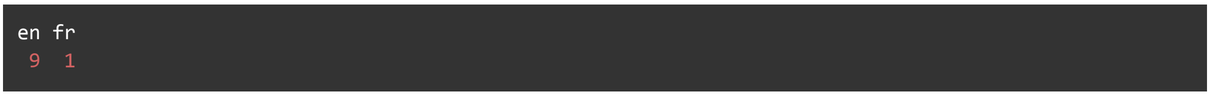
```
  de   en   es   fr   hr   hu   la   nl   no   pl   pt   ro   sl   sv 
   3 3148    4    8    2    1   14    4    2    1    4   12    1    1 
```

Clinton's top 10 most retweeted tweets are in such languages:



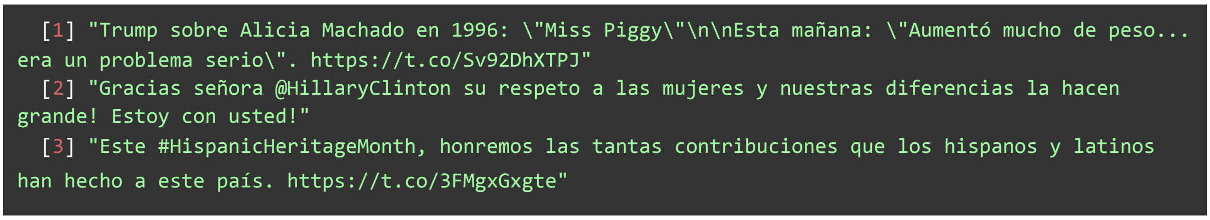
```
en es 
 7  3 
```

Trump's top 10 most retweeted tweets are in such languages:



```
en fr 
 9  1 
```

Clinton tweeted in Spanish more than Trump and her Spanish tweets were as below:



```
  [1] "Trump sobre Alicia Machado en 1996: \"Miss Piggy\"\n\nEsta mañana: \"Aumentó mucho de peso... era un problema serio\". https://t.co/Sv92DhXTPJ"
  [2] "Gracias señora @HillaryClinton su respeto a las mujeres y nuestras diferencias la hacen grande! Estoy con usted!"
  [3] "Este #HispanicHeritageMonth, honremos las tantas contribuciones que los hispanos y latinos han hecho a este país. https://t.co/3FMgxGxgte"
```

d. Sentiment Analysis (sentiment_analysis.R)

Sentiments and tones in a tweet differentiate how their message is delivered. The two candidates from two contrary parties showed different opinions throughout the election. In that, sentiment analysis on their tweets were made to see how positive or negative they were.

For sentimental analysis, library 'textdata' and 'tidytext' were utilized to tokenize the tweets and analyze their sentiments. Among various profiles, 'affin' fit this project the most because they were simple enough to distinguish the positive or negative sentiment but complex enough to show the strength of the sentiments in a scale between -5 and 5.

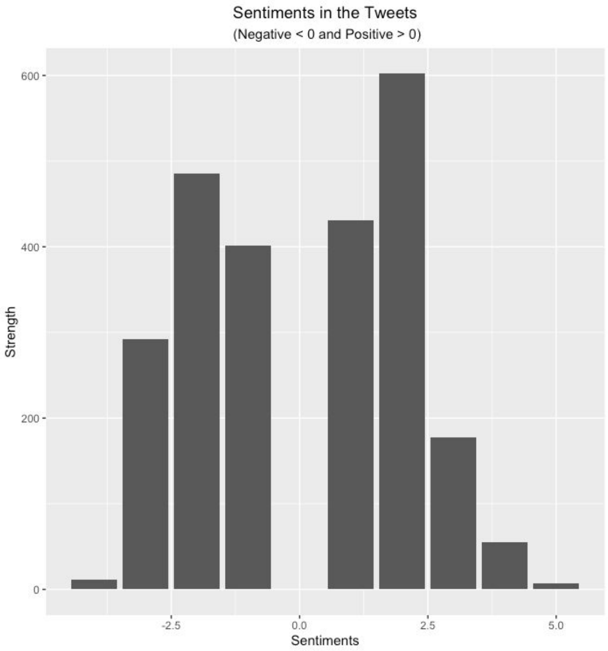


Figure 5: Sentiments in Clinton's Tweets

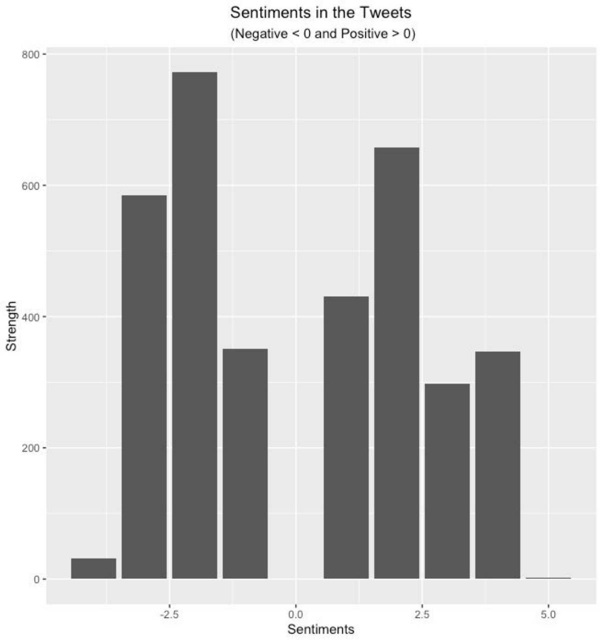
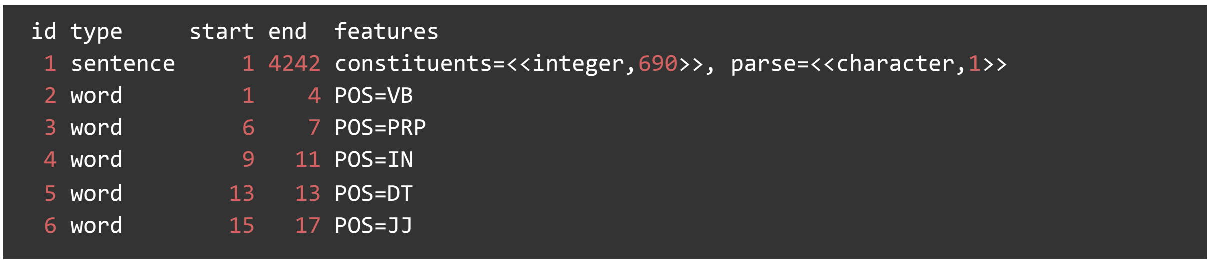


Figure 6: Sentiments in Trump's Tweets

e. Annotation Analysis (annotation_analysis.R)

Text annotation analysis delivers insights by tagging keywords, phrases, or sentences. By decomposing words by type, it can also compare texts how similarly they were structured. Unfortunately, due to the 140 characters limit in a tweet, the text holds similar structure and the dataset has low distribution in their diversity of types.
After decomposing the words in tweets, the annotation analysis was implemented. However, the result shows that they are mostly words not a sentence.



```
 id type     start end  features
  1 sentence     1 4242 constituents=<<integer,690>>, parse=<<character,1>>
  2 word         1    4 POS=VB
  3 word         6    7 POS=PRP
  4 word         9   11 POS=IN
  5 word        13   13 POS=DT
  6 word        15   17 POS=JJ
```

5. Conclusion

As much as their parties and political opinions are different, the way they present their opinions on twitter were distinguished. Clinton tweeted in a more positive tone where Trump tweeted in a more negative and strong tone. Clinton tweeted more frequently in another language, mostly in Spanish among them. Moreover, 3 out of 10 most retweeted tweets by Clinton were in Spanish. Above all the differences, they mentioned each other the most out of any other word.
Though it may not have been so long since twitter played a role in politics, politicians are communicating by tweets showing their sentiments, speaking another language, and posting immensely. Twitter would be a great source to see how politicians' messages spread and evolve.