

President Trump Tweets

An initial content analysis

Ahmad Alshehri

May 11, 2020

GA - Part Time Data Science Program

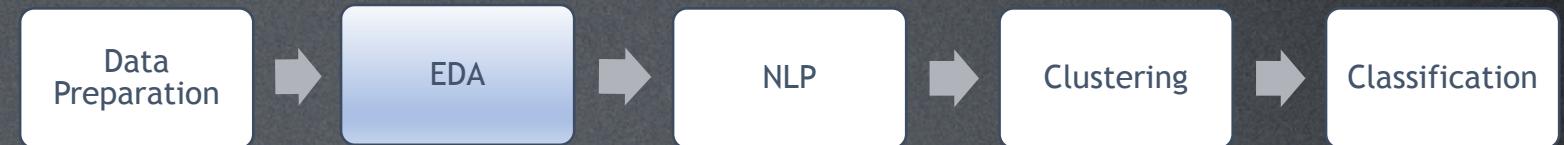


At times, Trump appears to tweet based on what has grabbed his attention at any given moment. What sort of topics grabs his attention and how people are reacting to it?

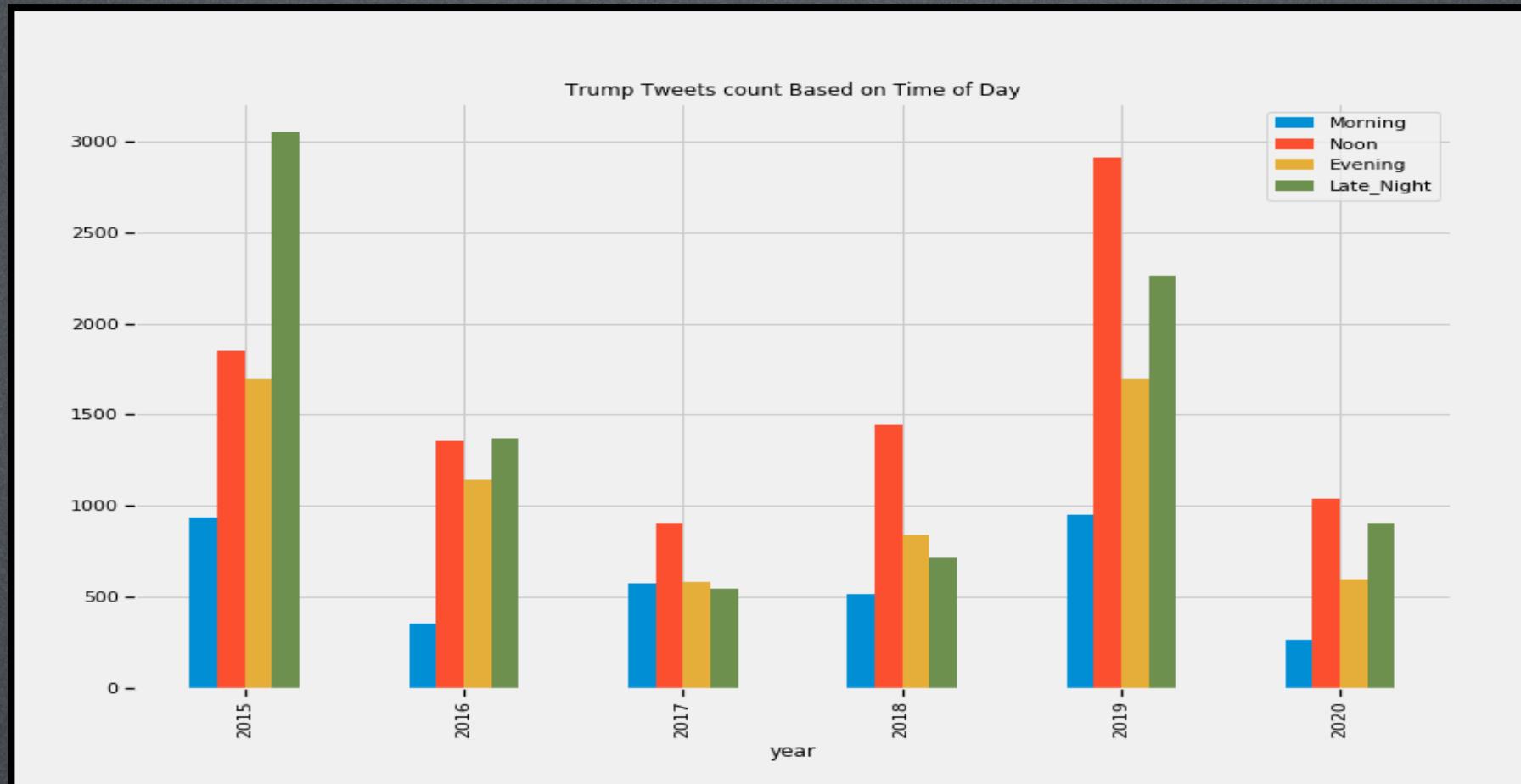
Objective is to build a classification model that can then be applied to unseen data in an automated tweet labeling system.

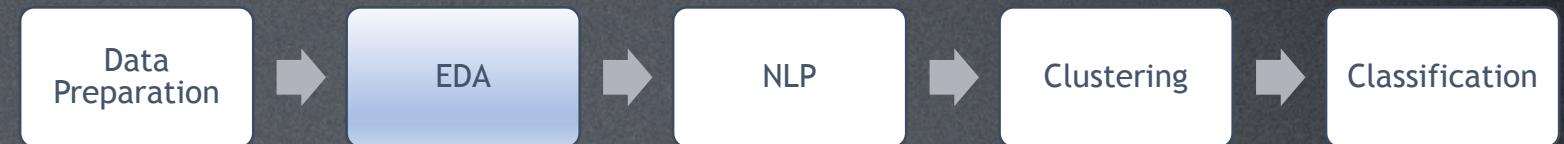
Project Workflow:



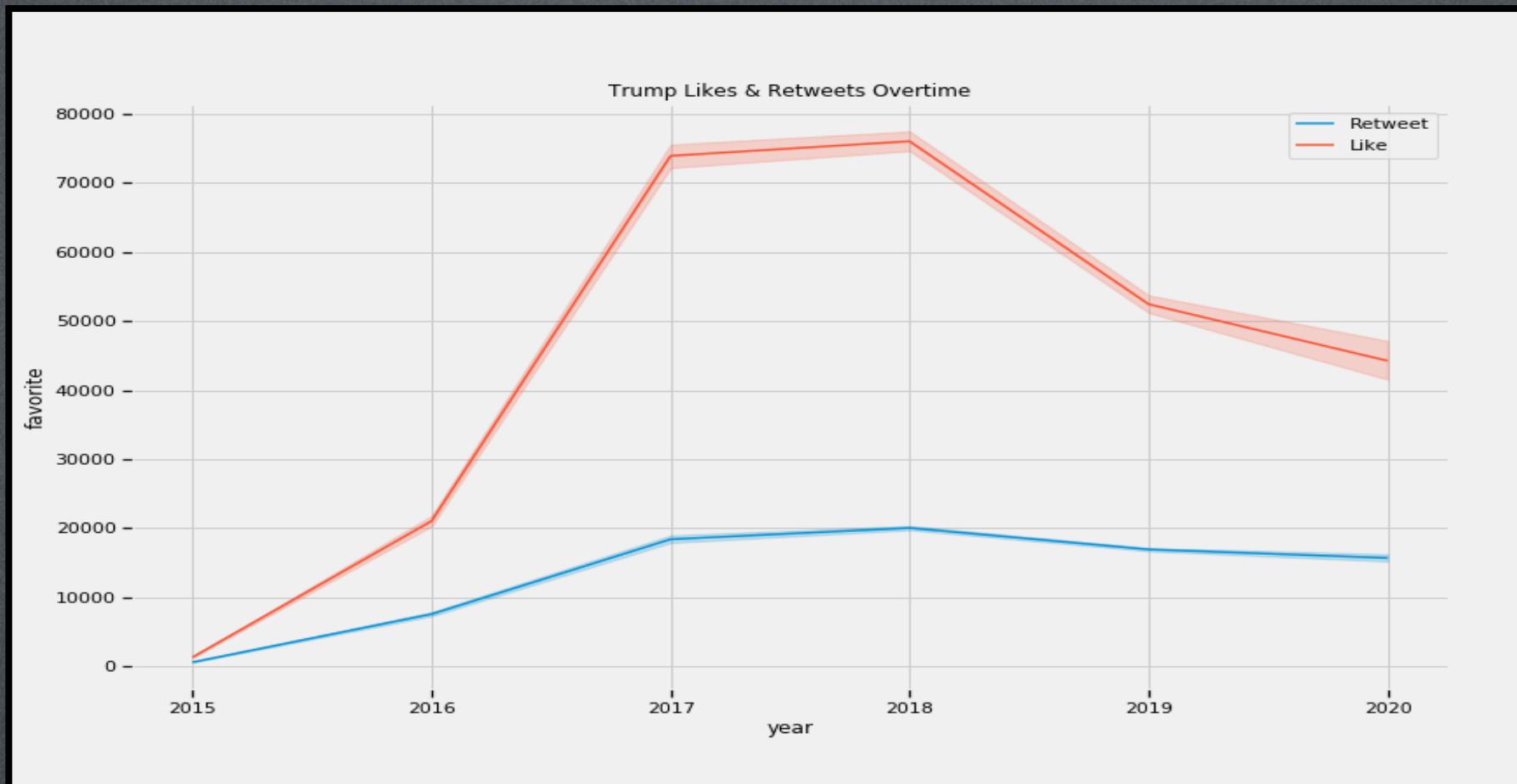


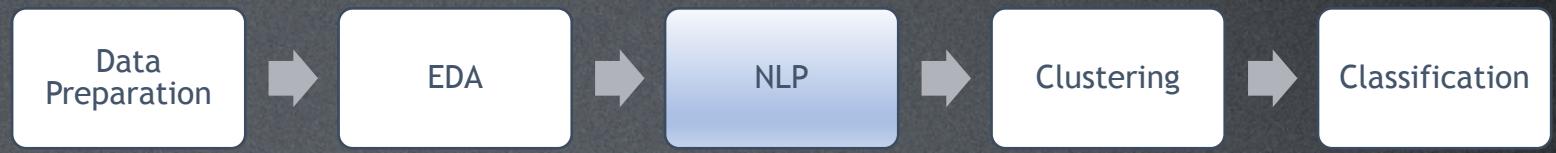
- from Jan 2015 - Mar 2020 total number of tweets: 28,498
- Average daily tweet rate: 16
- Retweets represents on 12%





- Popularity of tweets peaked in 2018 based





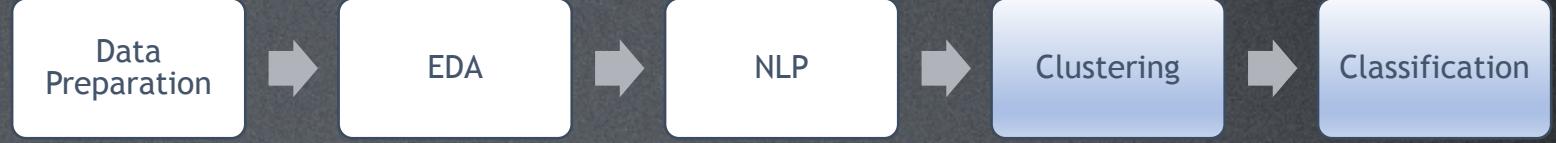
Steps:

- Removed punctuation
- Tokenization
- Remove stopwords
- Stemming
- lemmatization
- Vectorized using Td-Ifd

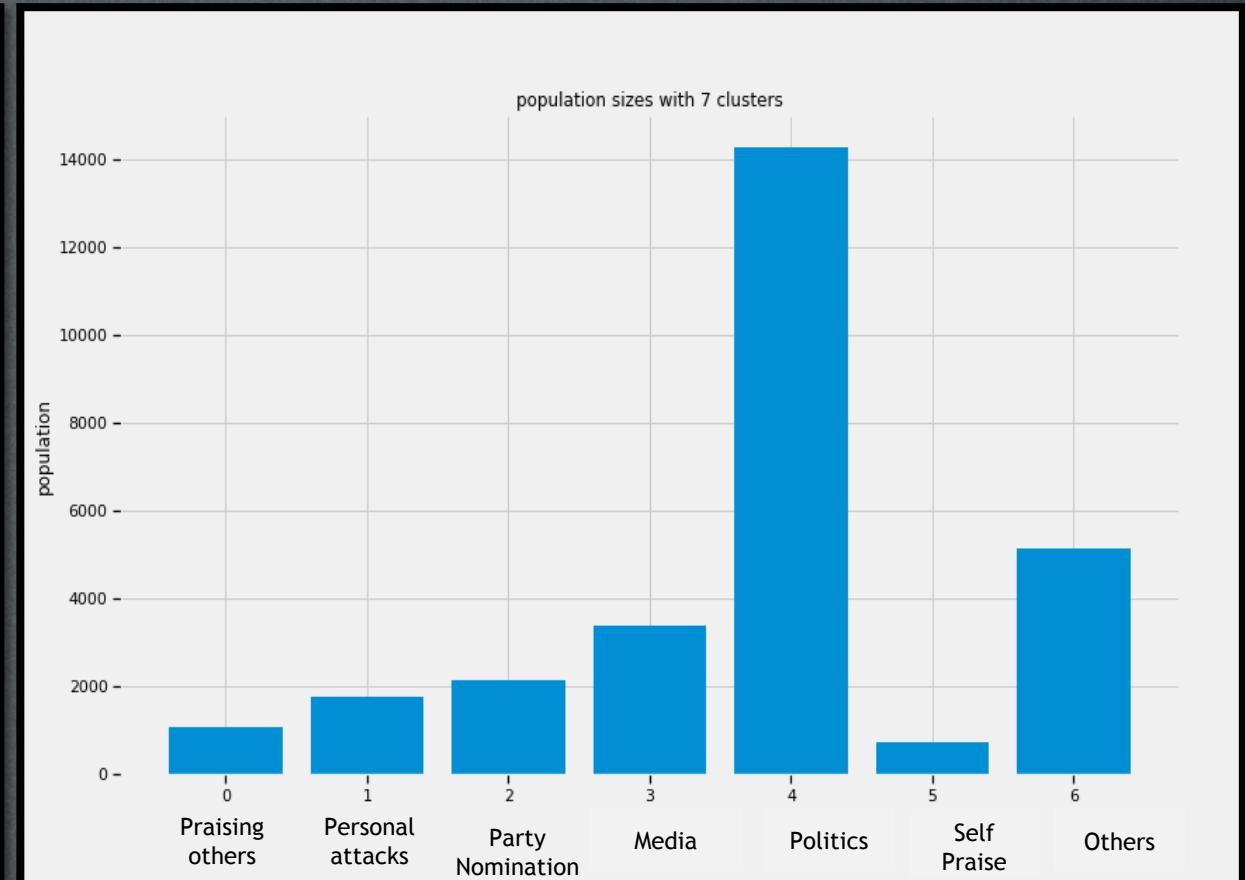
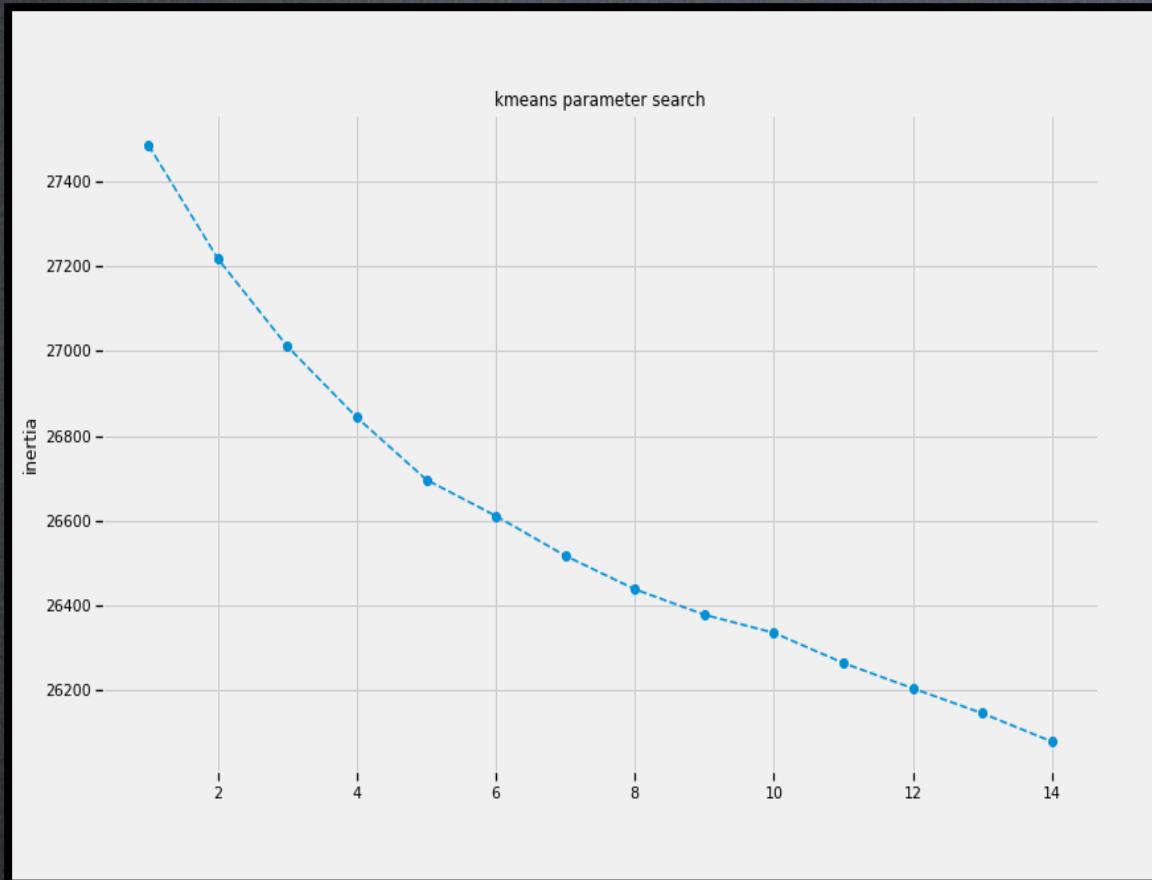
Resulted data frame had a total of 31720 words (features)

Specified a max feature of 2000 words due to limited computational capability





- Used Kmeans with K value of 7 after inspecting the inertia values for multible K values
- Used Logistic regression model with a prediction accuracy of 77%



Recommendation & Next Step

Moving forward there are a few approaches I feel would be very helpful to increase the model prediction accuracy:

- Work with other NLP libraries that might be able to analyze text . In this case, some of Trump tweets were in other languages and so having the right tool to analyze the text is important.
- Deep learning methodology such as neural network in order to analyze text data more efficiently. Traditional NLP often uses one-hot encoding to represent the word in a fixed vocabulary and uses a BoW to represent documents. While deep learning algorithms attempt to learn multiple levels of representation of increasing complexity/abstraction.
- As mentioned earlier, the total features extracted from the text were about 31498. Therefore, running the code using a cloud service such as AWS and Kaggle TPU might be helpful since they have a very high processing capability.

