NLP: Sentiment Analysis and Negative Tweet Reasoning

Learning Objectives:

- *Text parsing and data cleaning in the context of NLP*
- *Preliminary data visualization and analysis*
- *Sentiment prediction for generic tweets and negative reason prediction for US election tweets*
    - *Logistic Regression, kNN, Naive Bayes, SVM, Decision Trees, RF, XGBoost*
- *Result visualization and making data driven inferences*

Samuel Atkins - November 2020

Data Cleaning + Exploratory Analysis

Original Dataset Size:

Sentiment: 550391, 3 *US Elections: 2552, 3*

Text Data Parsing Example:

Original Text:

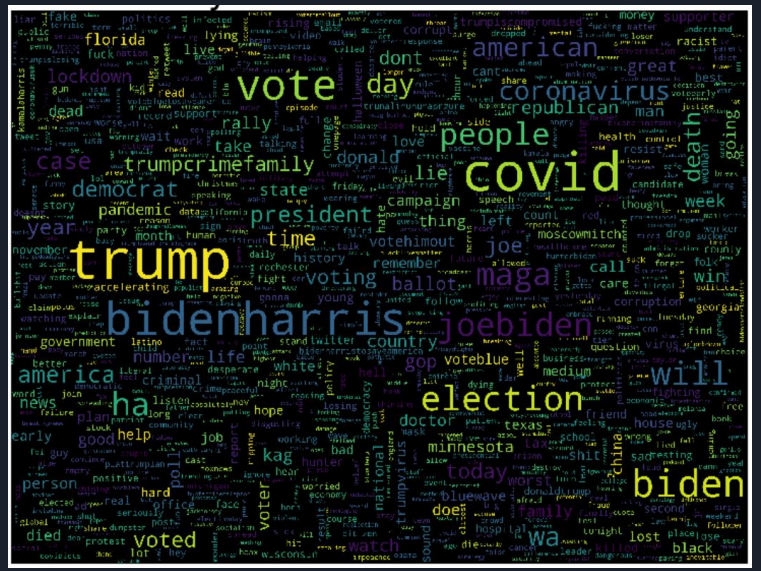
"b\"RT @MelissaTweets: I didn't think all the freak out was about control in the beginning of \#Covid.\n\nNow, it's clear that's ALL it's'"

Parsed Text:

['didnt', 'freak', 'wa', 'control', 'covid', 'clear']

Exploratory Data Analysis:

US Election Tweet WordCloud:



Model Preparation, Implementation, & Results: Sentiment Analysis

1. **Words Vectorized**
- Both BoW and TF-IDF were implemented

2. **Data Split 70/30**

3. **Models Implemented**
- Logistic Regression, kNN, Naïve Bayes, SVM, Random Forest, XGBoost implemented
- 2000 features used

Optimal Model:
Logistic Regression Model:
- C = 1
- penalty = “l2”
- class_weight = None

```
Model accuracy on general tweet validation set: 93.022%

Model accuracy on tweet test set: 93.002%

Model accuracy on US election tweet test set: 63.708%
```

Model Preparation, Implementation, & Results: Negative Reason Prediction

Three different models implemented:

- *Logistic Regression*
- *LinearSVC*
- *Random Forest*

Hyper-parameter tuning with 10-fold cross-validation was performed

Optimal Model:

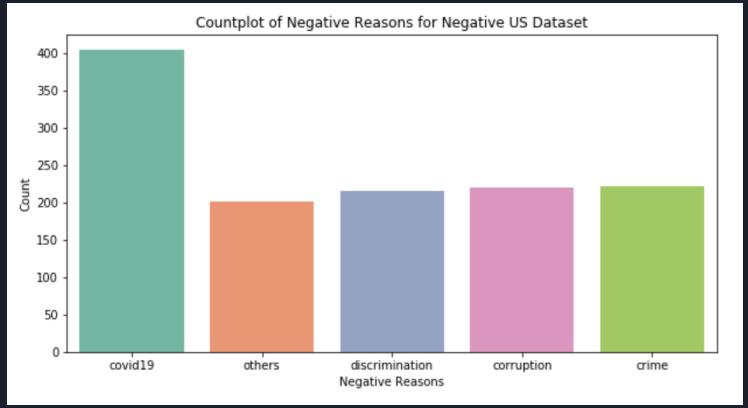
```
Random Forest Classifier:
-   n_estimators=1000
-   criterion="gini"
```

```
Model Accuracy on Test Set: 33.421%
```

Recall is very low for all classes except for the "covid19" class

This is due to the imbalanced nature of the dataset

If the model simply predicted the mode then it would obtain an accuracy of 33.5%




```
corruption
precision = 0.32
recall = 0.15

covid19
precision = 0.33
recall = 0.81

crime
precision = 0.35
recall = 0.11

discrimination
precision = 0.44
recall = 0.17

other
precision = 0.24
recall = 0.08
```

Possible Improvements + Rebalancing Bonus

Improving the accuracy of the sentiment analysis model:

- Implement a learned model that takes as input a sequence of word inputs represented by GloVe vectors
- Rebalance the dataset so that the positive and negative sentiments appeared approximately the same number of times
  - This was implemented as a bonus:

```
Negative sentiment recall increased from 0.88 to 0.91

Model accuracy on tweet test set increased from 93.002% to 93.088%
```

Improving the accuracy of the negative reason predictor:

- Increase the number of data points because the size of this dataset is very small compared to the generic tweet dataset
- Up-sample/down-sample the non-COVID/COVID negative reason tweets to obtain a more balanced dataset resulting in better recall for the other classes