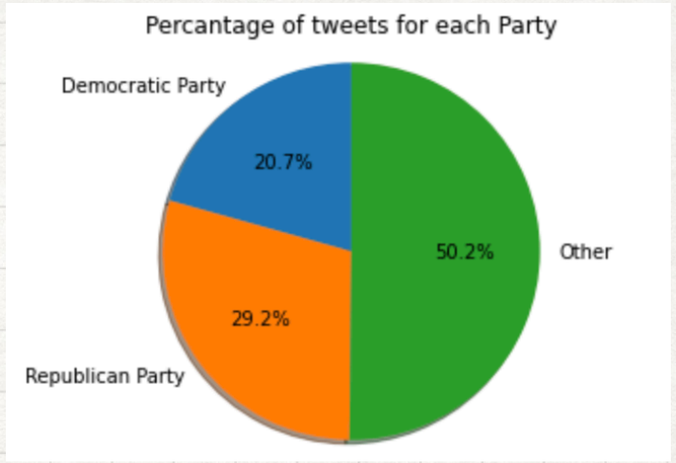
ASSIGNMENT 3
SENTIMENT ANALYSIS

MIE 1624 Introduction to Data Science and Analytics

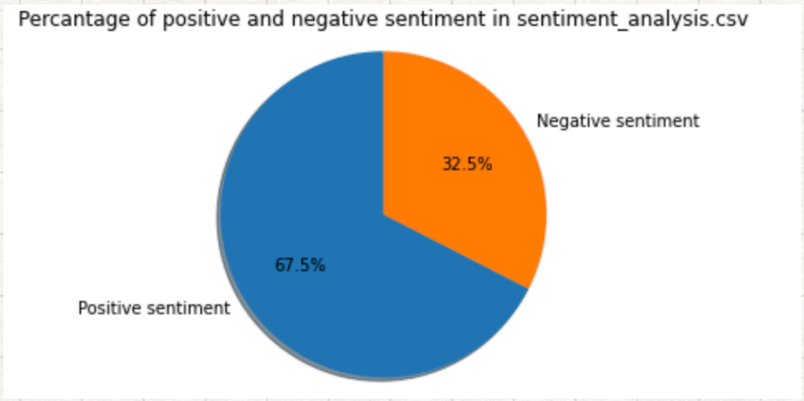
Name: Jiani Jia
Student ID: 1002226245

Exploratory analysis

1. Democratic Party Keywords: 'biden', 'democrats', 'diversityandinclusion', 'voteblue', 'chinajoe'
2. Republican Party Keywords: 'trump', 'buildthewall', 'votered', 'redwave', 'republican', ‘maga'
3. Among 2552 tweets, 20.65% tweets are about Democratic Party, 29.19% tweets are about Republican Party, and other contains 50.16% tweets. Half of the tweets are associated with the US election, in the election related tweet, over a half of the tweets are related to Trump.



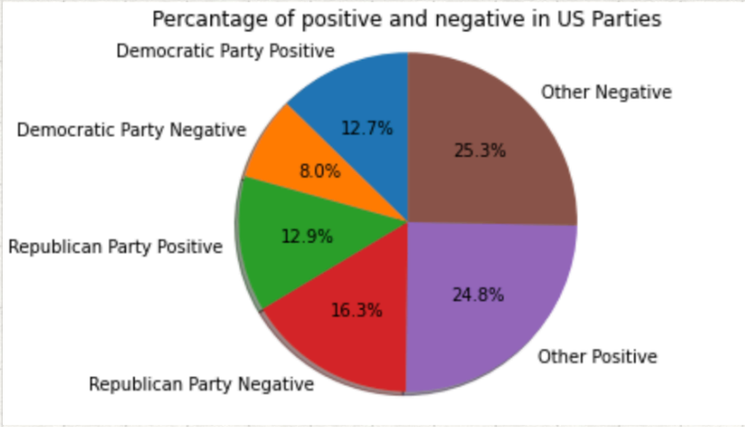
sentiment_analysis.csv

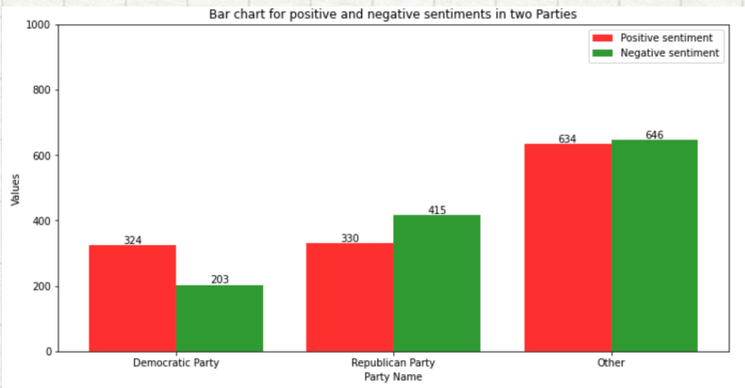


Pie chart shows that in sentiment_analysis.csv , among 550391 tweets, there are 67.5% tweets are positive sentiment and 32.5% tweets are negative sentiment.

Key word in the sentiment_analysis.csv can be visualized through wordcloud.

US_Elections_2020.csv





For Democratic party, the related positive sentiment tweets are more than negative sentiment tweets. And for Republican party, the related negative sentiment tweets are more than positive sentiment tweets.

Model Feature

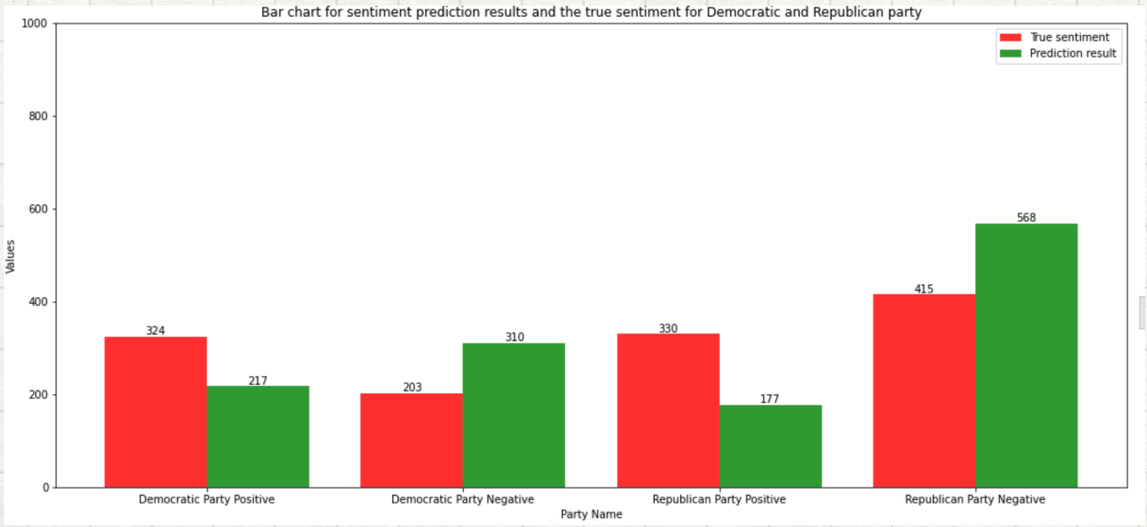
1. Text Preprocessing: it is an important step in NLP, it transform text into a more digestible from so that the machine learning can preform well.
2. Bag of words: is a way of extracting features from text for use in modelling, it is a representation of text that describes the occurrence of words within a document and only concerned with whether known words occur in the document, not where in the document.
3. TF-IDF: is a way to rescale the frequency of words by how often they appear in all documents, so that the scores of frequent word across all documents are penalized.
4. After Bag of words and TF-IDF, the sentiment_analysis.csv text become a 550391x317251 sparse matrix.

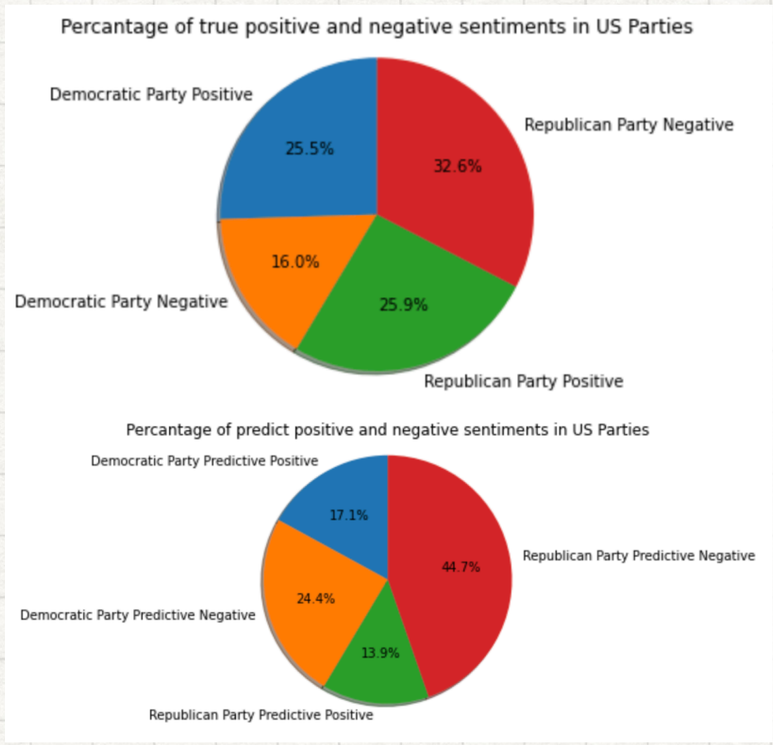
Model Implementation on 2020 US Election Tweets

- Training result on generic tweet:

Feature	Logistic Regression	Decision trees	k-NN	Naive Bayes	SVM	Random Forest	XGBoost
Bag of Words	0.95807	0.931933	0.89986	0.93191	0.955607	0.93999	0.84196
TF-IDF	0.95888	0.92929	0.62151	0.91117	0.95841	0.94217	0.84263

- Best model: Logistic regression for TF-IDF feature
- Training accuracy at generic tweet: 0.9588
- Test accuracy result on 2020 US election: 0.7128





Bar chart for sentiment prediction results and the true sentiment for Democratic and Republican party

Model Implementation On Negative 2020 US Elections Tweets

- Cross validation accuracy on three models:

Feature	Accuracy	Logistic Regression	SVM	Naive Bayes
Bag of Words	Training Accuracy	0.338	0.298	0.350
TF-IDF	Training Accuracy	0.353	0.339	0.3631

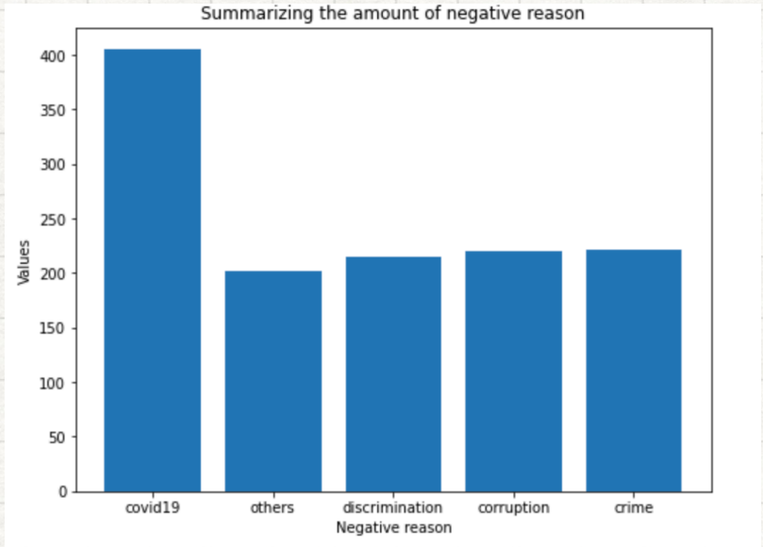
- Best model: Naive Bayes TF-IDF feature
- Test accuracy result on negative 2020 US elections tweets test set: 0.3579

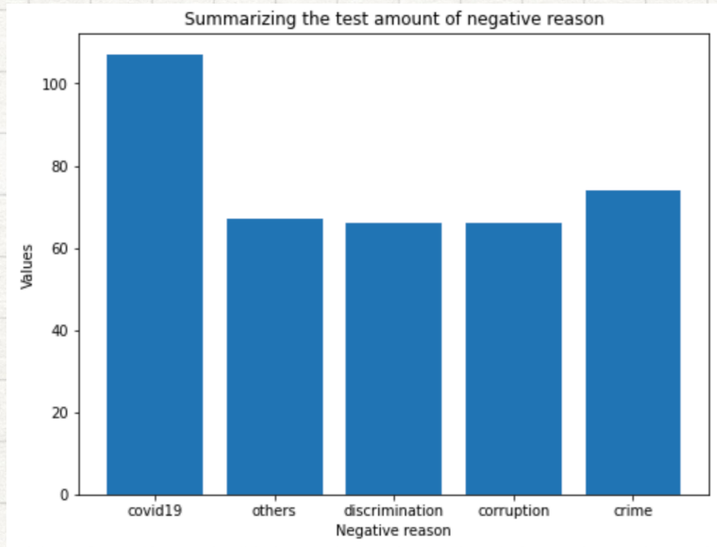
Model 1 Analysis

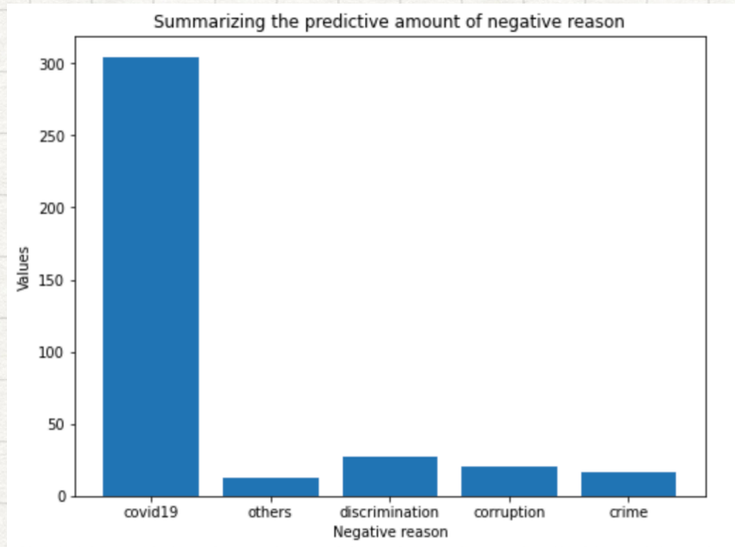
- After training on the generic tweets, logistic regression perform best. Test at the 2020 US election data and predict for the sentiment. The predict result shows 44.7% are Republican Party negative, 13.9% are Republican Party positive; 24.4% are Democratic Party negative and 17.1% are Democratic Party positive.
- Low test accuracy is because the generic tweet has 317251 word features due to the large amount of tweet in the dataframe, but 2020 US election data does not have that much word features.
- From the 2020 US election result, Joe Biden got 306 electoral votes and Donald Trump got 232 electoral votes. From the NLP analytics based on tweet, it shows that more public get Trump negative sentiment on tweet and more positive to Biden, so that NLP analytics is useful for political parties during election campaigns.
- Improve Model 1 accuracy method 1: The word features in the training set may not contain the keywords in 2020 US elections, the corpus in training set should consist of data from news sources e.g. recent tweets. This is because the vocabulary of a corpus varies with domains.
- Improve Model 1 accuracy method 2: eliminating features with extremely low frequency, because the keywords in low occurrence frequency in the corpus usually does not play a role in text classification. Reducing the feature can help to improve accuracy.

Model 2 Analysis

- Test accuracy on the US negative data: 0.358
- It is because the imbalance data, above bar charts count the number of the negative reasons in the whole dataset, test set and prediction. In the whole dataset, negative reason 'Covid19' contains 30% among the five reasons. The negative reason in test set also has around 30% of 'Covid19'. So that during the test, more of the 'Covid19' will be predicted, which is proofed by the last plot. Over 80% of the negative reason are predicted as 'Covid19', it leads test accuracy becomes very low.







- Improve Model 2 accuracy method 1: In order to fix the data imbalance, we can combine some closer reasons together to reduce the percentage of the main reason among the reason pool.
- Improve Model 2 accuracy method 2: classes reweight, it takes into account the asymmetry of cost errors directly during the classifier training, therefore the output probabilities for each class will already embed the cost error information.