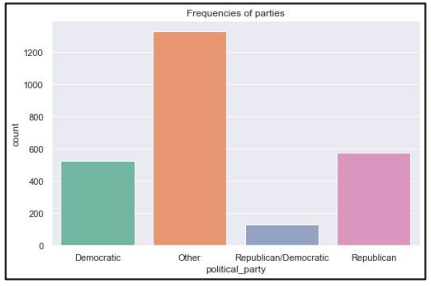
MIE1624 assignment 3
Jia Lei Qian (1002675544)

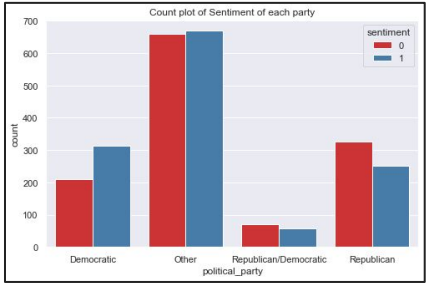
Sentiment Analysis

Exploratory Data Analysis

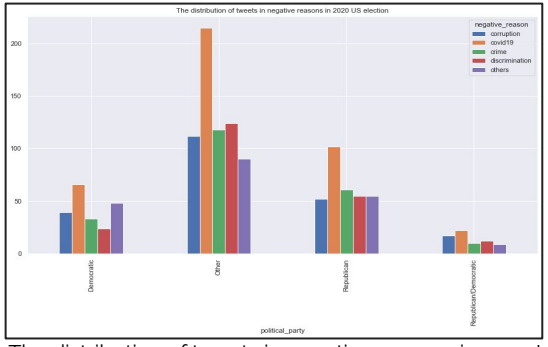
Distribution plots of political parties

- After classifying the dataset into four categories, I found that:
  - The majority of tweets are classified as "other,"
  - Only a small number of people vote for both parties.
  - The number of people who vote for republican and democratic are very close.
  - Two classes of sentiment are imbalanced for democratic and republican parties.
  - For the republican party, the number of negative tweets is more than positive tweets; in contrast, the positive tweets for democratic parties are more than negative ones, which is reasonable as Trump always posts offensive twitter speech.

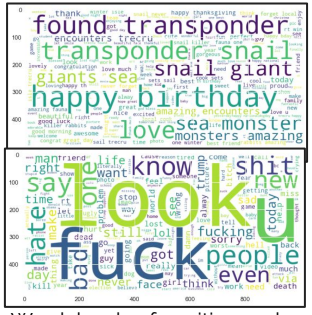




Distribution and count plots of sentiment political affiliations of the tweets



The distribution of tweets in negative reasons in 2020 US election



Wordclouds of positive and negative tweets of sentiment analysis data

Word Clouds of positive and negative tweets of sentiment analysis data

- The neural words occur in both positive and negative tweets. The positive tweets contain positive words such as "love" and "happy", while the negative ones have many negative words.
- It is interesting to find "trump" in the word cloud of negative tweets. Those words would help us to decide the sentiment of the text.

Distribution of tweets in negative reasons in 2020 US election

- The covid-19 is the main reason for negative tweets.
- Other than covid-19, people wrote negative tweets to attack the republican party due to crime and discrimination issues, while people wrote to attack the democratic party due to others and corruption reasons.

Feature Selection & Model Implementation on Sentiment Analysis data

Bag of Words (Word Frequency)

- Word frequency is a word frequency counting technique in which a sorted list of words and their frequencies are generated, where the frequency is the frequency of occurrence in a given composition.
- In this project, the CountVectorizer function is used. It converts a collection of text documents to a matrix of token counts.
- After training and splitting the dataset, there are 77054 samples and 2477 features in the training set

TF-IDF

- Tf-idf is a statistical method used to assess the importance of a word to one of the documents in a document set or corpus.
- The importance of a word increases proportionally with the frequency of its occurrence in the document, but decreases inversely with the frequency of its occurrence in the corpus.
- In this part, I use TF-IDF to convert a collection of raw documents to a matrix of TF-IDF features with maximum 300 features
- After training and splitting the dataset, there are 77054 samples and 300 features in the training set

Bag of Words (Word Frequency)

	Train Accuracy	Test Accuracy	Mean cross validation score after hyperparameter tuning
Logistic Regression	92.194%	92.023%	91.941%
KNN	91.704%	90.961%	90.490%
Naive Bayes	89.437%	89.198%	89.250%
SVM	92.432%	91.972%	91.878%
Decision Trees	93.425%	91.151%	91.022%
Random Forest	93.425%	91.591%	91.463%
XGBoost	89.875%	90.107%	89.776%

TF-IDF

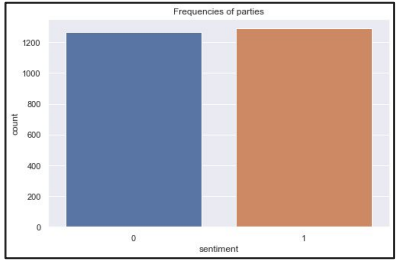
	Train Accuracy	Test Accuracy	Mean cross validation score after hyperparameter tuning
Logistic Regression	89.824%	89.359%	89.605%
KNN	87.797%	83.736%	84.643%
Naive Bayes	83.070%	82.882%	82.535%
SVM	89.724%	89.281%	89.527%
Decision Trees	94.512%	88.139%	88.332%
Random Forest	94.666%	89.114%	89.356%
XGBoost	89.500%	88.823%	89.065%

According to the tables shown above, the **logistic regression using the word frequency (Bag of words) features with hyperparameter {penalty:'l2', C: 10, solver:'liblinear}** performs the best.

Feature Selection & Model Implementation on U.S. Election data

Feature selection

- Since the class of target variable "sentiment" is balanced, there is no need to resample the model.



Model Implementation

- After apply the logistic regression model with tuned hyperparameters, the model has obtained an overall accuracy of 74%, f1-score of 71% for negative tweets, and fi-score of 77% for positive tweets.

	precision	recall	f1-score	support
0	0.81	0.62	0.70	383
1	0.69	0.86	0.77	383
accuracy			0.74	766
macro avg	0.75	0.74	0.74	766
weighted avg	0.75	0.74	0.74	766

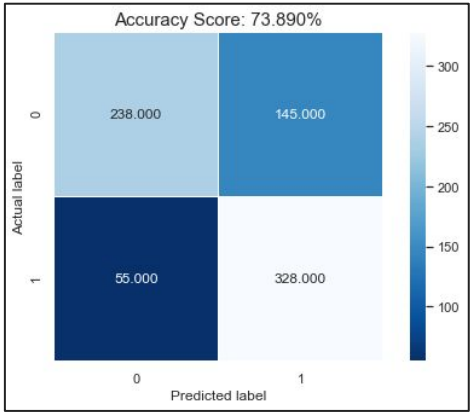
Classification report of the US election model

Accuracy for the Republican and the Democratic

- The accuracy for republican is slightly higher than the accuracy for democratic party, which are 78.443 and 77.273% respectively.

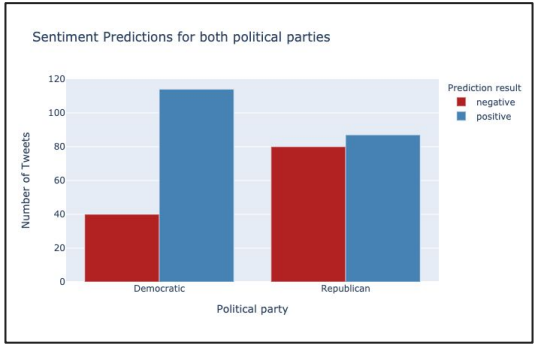
Party	Accuracy
Republican	78.443%
Democratic	77.273%
Overall	73.890%

Visualization of confusion matrix

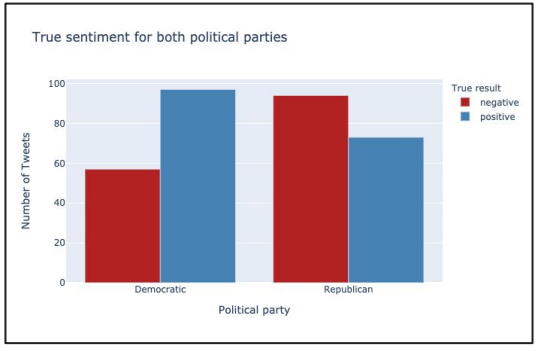


Visualization of Sentiment Prediction model on U.S. Election data

Visualization of sentiment prediction for different parties



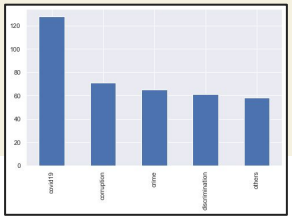
Visualization of true sentiment for two parties



Discussions

- The overall trend and number of tweets involving democratic and republican parties are similar.
- The democratic party has a higher positive sentiment result, while the Republican has a higher negative sentiment result.
- The democratic party has more positive sentiments than negative sentiments under sentiment prediction and true emotion. However, the republican party shows a reversal trend of forecast compared to the true result.
- Even though both parties' overall accuracy scores are similar, the model predicts the result for the democratic party more accurately.

In general, the logistic regression model captures the overall trend of the actual data for the democratic party but not for the republican party, which implies that the model could not predict the negative sentiment correctly. In reality, the democratic party wins the 2020 US election. The sentiment prediction result could represent the election result, and the NLP analysis can be used to observe the election trend. However, people still should not rely too much on the model to predict the results.



Negative reason classification on U.S. Election data

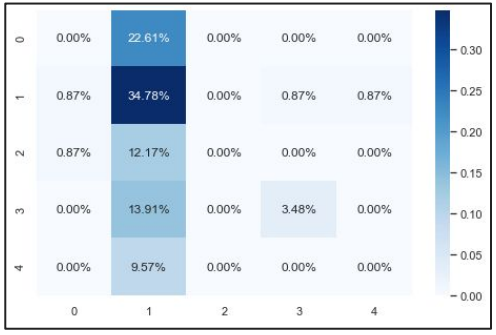
Feature selection

- Since all of those five reasons are different, there is no need to group those reasons. However, label encoding is required as the target values are a list of string

Model Implementation

- The logistic regression performs the best with the highest accuracy of 38%.
- According to the visualization of a confusion matrix for the multiclass logistic regression model, class 1 (covid-19) has the highest correct prediction rate, while class 0(corruption), 2(crime), 4(other) perform the worst.

	Accuracy
Logistic regression	38%
Random Forest	25%
XGBoost	30%



Visualization of confusion matrix

Discussion:

- According to the result of negative reasons prediction on the US election dataset, logistic regression still performs the best with an accuracy of 38%. However, the classification report implies overfitting as the accuracy for the training set is higher than the accuracy for the test set. Also, the size of the training set is really small, which is only 268. Also, refer to the visualization of negative reasons for each party, the "covid-19" counts for the majority, the classes of target values for the negative reason prediction are imbalanced.
- Possible improved methods:
  - Scraping more data from Twitter or combine similar negative reasons
  - Increasing 5-fold cross-validation to 10-fold cross-validation
  - Using word embedding (Word2Vec) or N-gram for feature selection instead of Bag of Words and TF-IDF
  - Using resampling method (oversampling/undersampling/SMOTE) to distribute the amount of data in each class uniformly