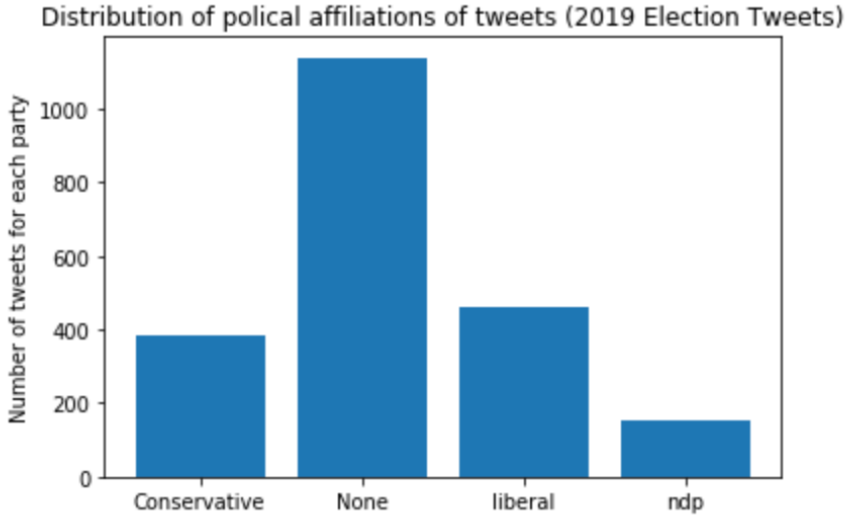
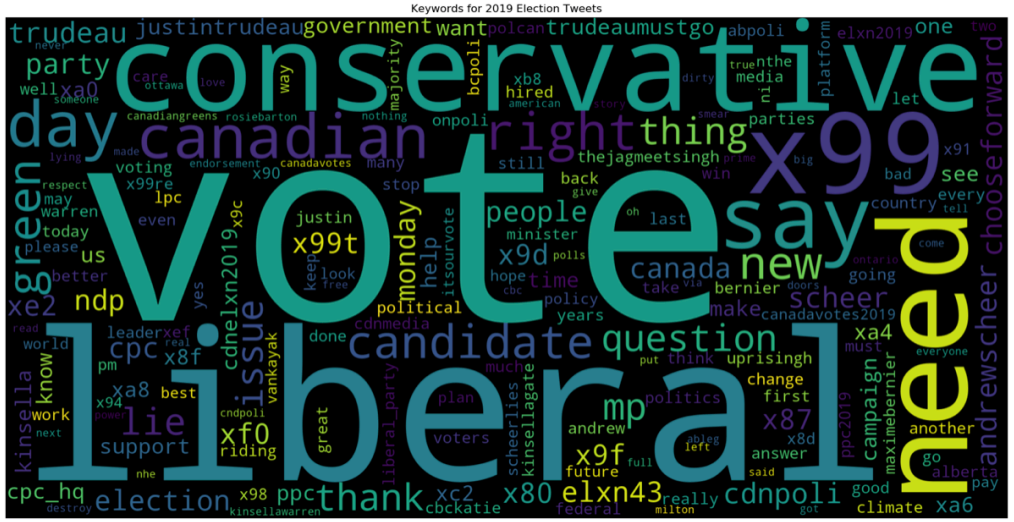
Student name: Yuchao Wu 1000651984

EXPLORATORY ANALYSIS – ELECTION TWEETS

- Extract the words followed by # and @ because these words are often related to the political party names and help to identify the party.
- Used Counter to count the number of occurrences for each mentioned and hashtag terms.
- Most of the tweets are from the "None" parties. These tweets could be from small parties, such as Green party and New Democratic party.
- The WordCloud plot shows that more tweets are associated with liberal party than any other parties.

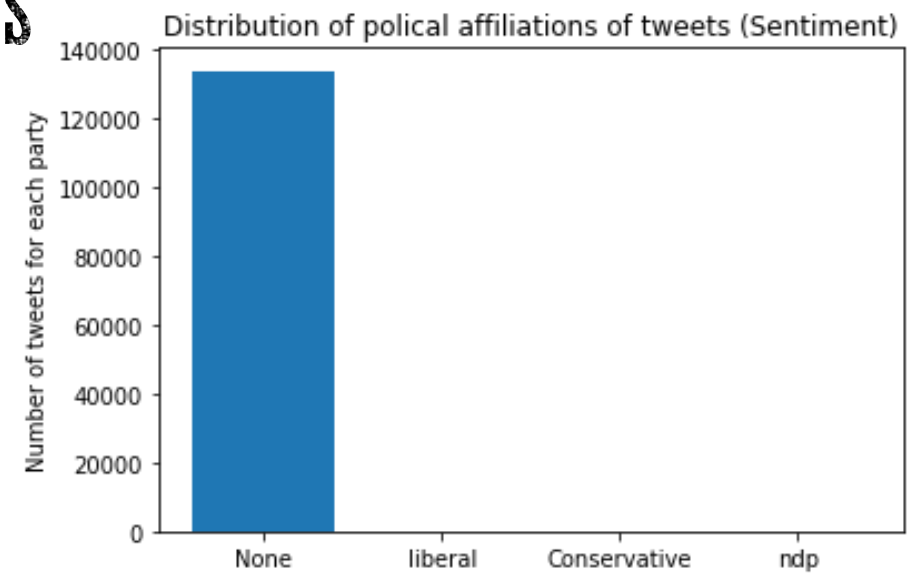


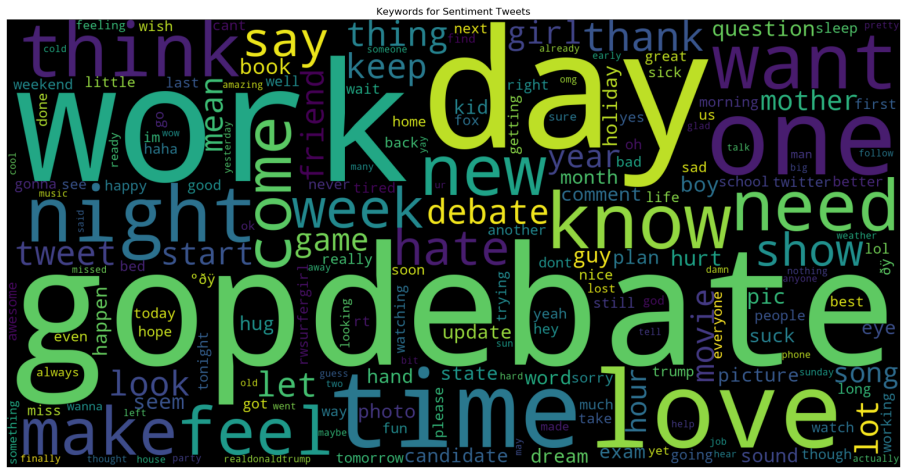


2019 Canadian Election Tweets

EXPLORATORY ANALYSIS – SENTIMENT TWEETS

- Most of tweets are classified as None type and less meaningful for giving more insights.
- Top 1000 most common words are used to plot WordCloud graph.
- The WordCloud plot shows that most common words in sentiment tweets, which are not very relevant to the key words stated in each party.
- This is the reason why most of tweets are classified as None type.





Sentiment Tweets

MODEL FEATURE IMPORTANCE

- Two ways to encode the tweets: TFIDF (TfidfVectorizer is used to calculate the tfidf score) and WF(CountVectorizer is used to vectorize the tweets).
- For both WF and TFIDF, 200 features are used to limit the feature space dimensions.
- The top 200 most common words from sentiment tweets are selected for the model preparation.
- The model trained with TFIDF tends to perform better than WF as TFIDF considers both words frequency and rareness of a term.

```
X_sentiment_WF_train.head(30)
```

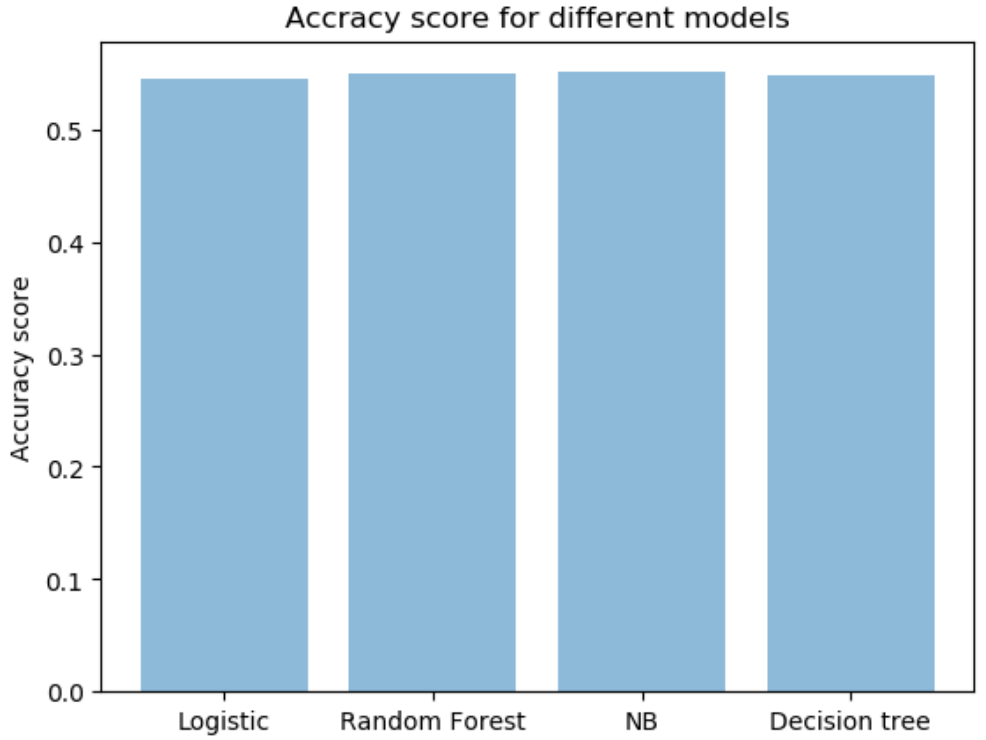
	actually	already	also	always	amazing	another	anyone	away	awesome
0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0

```
X_sentiment_tfidf_train.head(30)
```

	actually	already	also	always	amazing	another	anyone	away	awesome
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

MODEL RESULTS

- Hyperparameters for Four classifiers(logistic, random forest, NB and Decision Tree) are tuned using grid search and random search methods for model optimization.
- The test accuracy scores for 4 models are very closed to each other (around 55%).
- Finally, random forest is selected as the best model for the rest part of the assignment as it works better with high dimensional data.
- The test accuracy is around 47% when predicting the sentiments on 2019 Canadian election tweets with the best model.
- The performance is not very well since the feature space dimensions of election tweets are very different from that of sentiment tweets.



MODEL RESULTS VISUALIZATIONS

- More tweets are classified as negative sentiment for all three parties for predicted sentiments.(a similar trend is also observed in true sentiments)
- NPL analysis on tweets is, to some extent, useful for political parties, helping to predict the sentiment of election tweets.
- It is worth noting that there is actually a huge difference between the sizes of sentiment and election tweet, which may introduce some errors due to bias.
- For negative 2019 Canadian elections tweets, the random forest tends to perform better when working on multi-class classification than logistic, and decision tree.