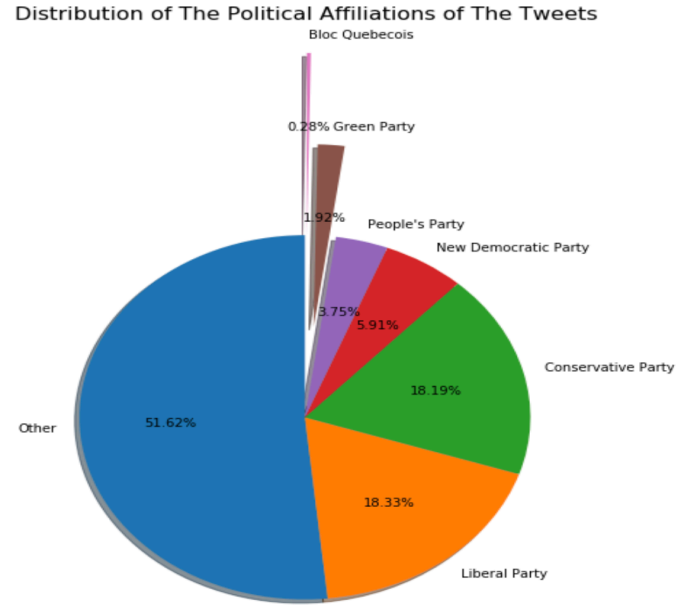
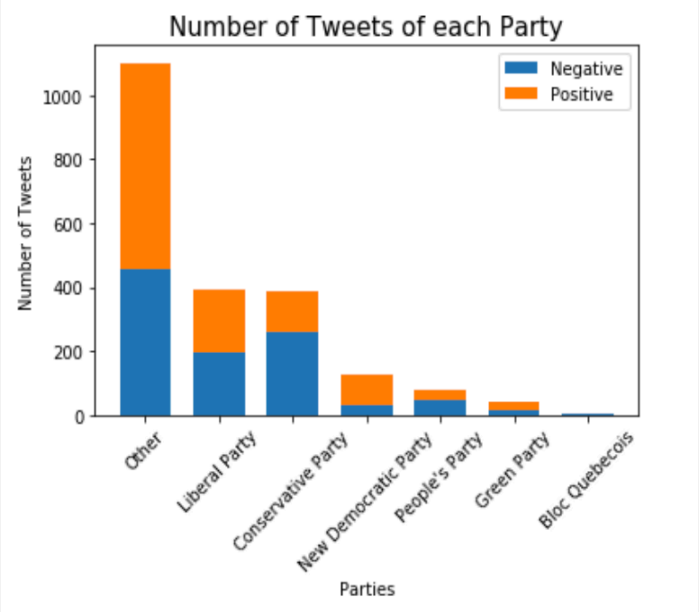
DATA SCIENCE FOR NLP

Zhaohui qu 1005783127

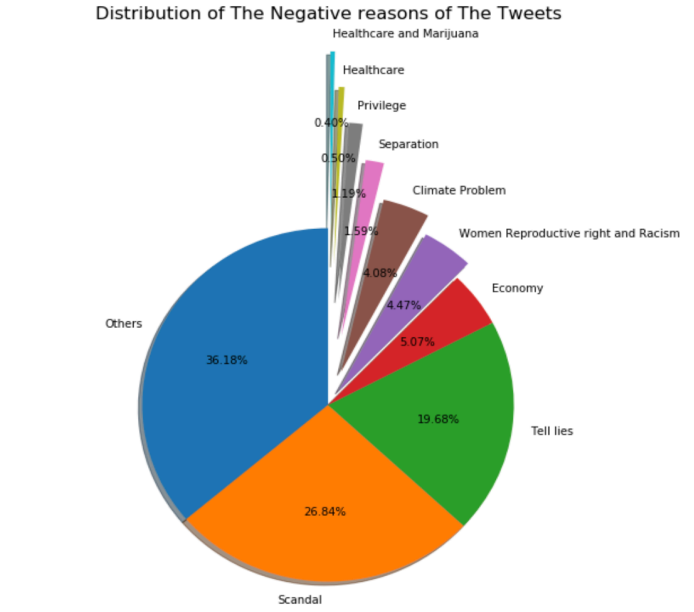
EXPLORATORY DATA ANALYSIS



From the distribution, we can see that the percentage of liberal party and conservative party are similar because they are main competitive parties in 2019 Canadian Federal election.



The above figure demonstrates that the percentage of negative tweets is higher than the percentage of positive tweets for conservative party, people's party and Bloc Quebecois, but the situation is opposite for new democratic party and green party, and the tweets related with liberal party have a similar ratio about negative and positive sentiment.



The above figure shows that except for 'Other' reason, the public can not tolerate scandal of party and telling lies to the public. So every party should be honest to voters.

MODEL IMPLEMENTATION

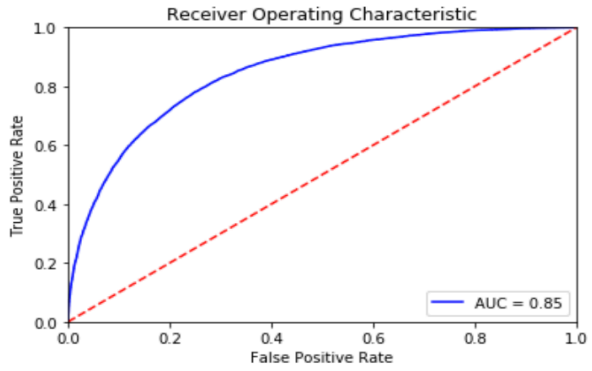
The following table demonstrates the accuracy values on Sentiment.csv file applied different models. We can see that when we use TF-IDF as feature and apply logistic regression model, we can obtain the highest accuracy value.

Feature	Logistic Regression	k-NN	Naive Bayes	SVM	Decision Trees	Random Forest	XGBoost
TF-IDF	76.63%	59.00%	74.05%	76.51%	67.38%	55.45%	67.66%
WF	76.41%	65.85%	75.55%	76.03%	67.94%	55.44%	67.86%

Logistic Regression Model

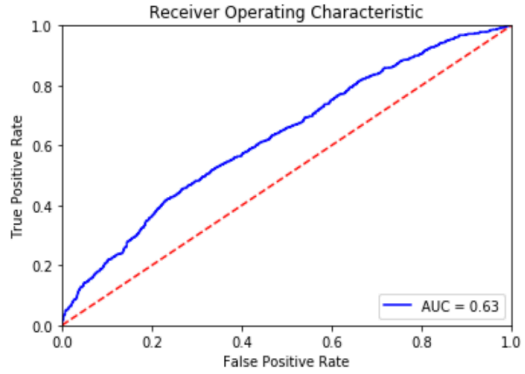
Prediction on generic sentiment

```
The accuracy of the model on train dataset is 76.69297%
The accuracy of the model on test dataset is 76.46722%
Test result report:

              precision    recall  f1-score   support

           0       0.78      0.79      0.79     22257
           1       0.74      0.73      0.73     17887

    accuracy                           0.76     40144
   macro avg       0.76      0.76      0.76     40144
weighted avg       0.76      0.76      0.76     40144

ROC plot and corresponding AUC:
```



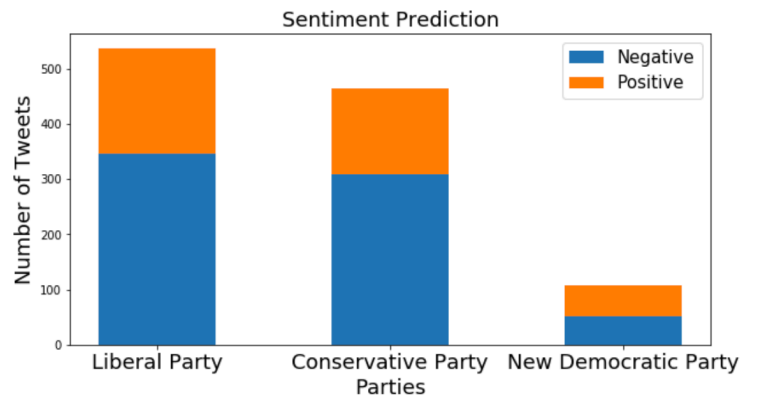
Prediction on Canadian elections

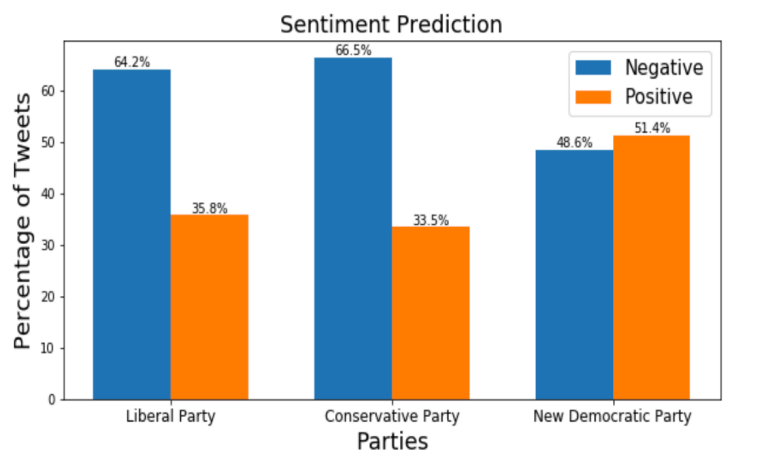
```
The accuracy of the model on the 2019 Canadian Elections Dataset is 58.46226%
Test result report:

              precision    recall  f1-score   support

           0       0.55      0.71      0.62      1006
           1       0.65      0.47      0.55      1127

    accuracy                           0.58      2133
   macro avg       0.60      0.59      0.58      2133
weighted avg       0.60      0.58      0.58      2133

ROC plot and corresponding AUC:
```



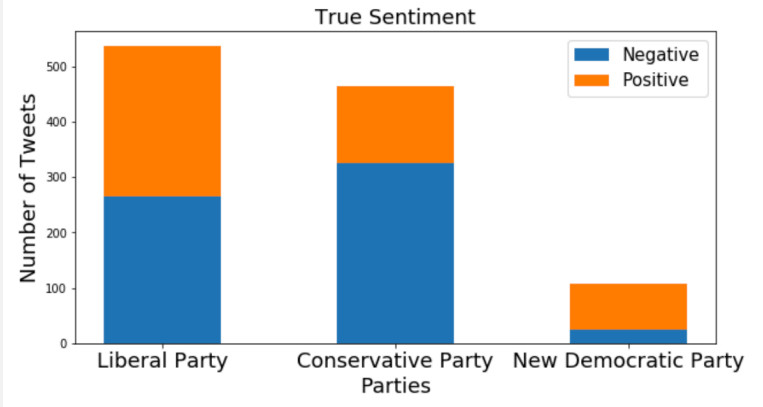
MODEL IMPLEMENTATION

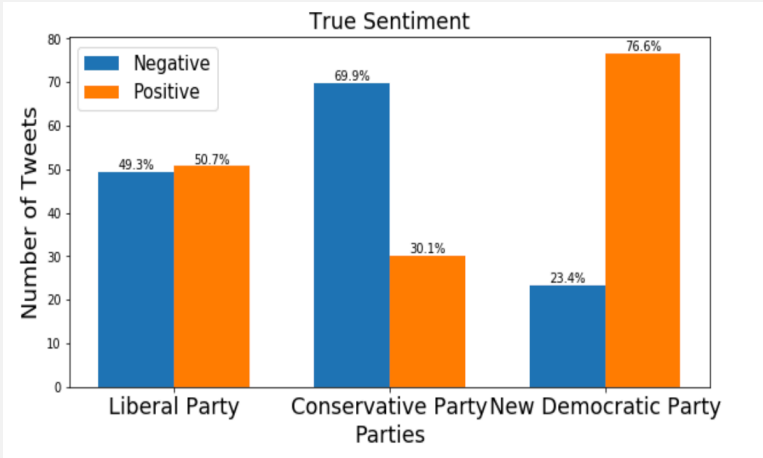
Predicted Sentiment





True Sentiment



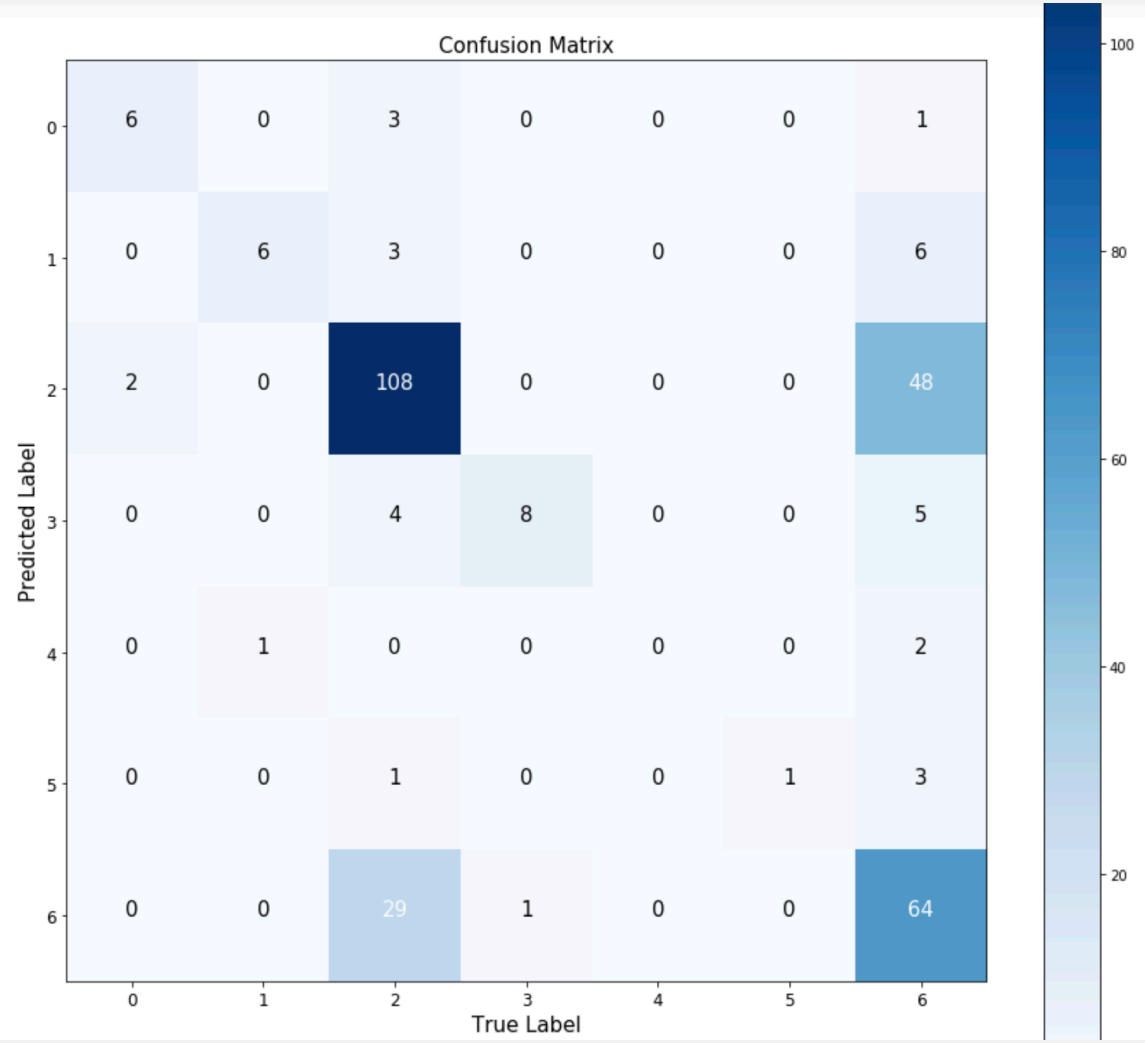


Although the accuracy of predicted sentiment is 58.46%, we can get the same conclusion from predicted sentiment and true sentiment analysis that conservative party and iberal party are two most prevailing parties in 2019 Canadian Elections. But it seems likely that more people support liberal party, because there are more people expressed negative sentiments to consercative party than liberal party.

From the results of 2019 Canadian elections, the percentage of seat total for liberal party, conservative party and new democratic party were 46.45%, 35.80% and 7.10%. The election results are similar with our sentiment analysis. Therefore, we can conclude that NLP analytics based on tweets is useful for political parties during election campaigns

MODEL IMPLEMENTATION

Predict the Reason for Negative Tweets



The accuracy of the logistic regression model on test dataset is 63.58%

From test result report in Logistic Regression Model, we can see that the logistic regression model does not have strong ability to classify the following reasons, including 'Healthcare', 'Separation' and 'Others'.

The accuracy of classifying negative reason of 'Healthcare' is 0. We print all tweets which negative reason is about Healthcare and Separation. We find that there are only 9 sample about healthcare and 16 sample about separation and the content of each tweet contains many meaningless characters or words. So lack of sufficient data and incomplete text cleaning are the main reasons.