UNIVERSITY OF TORONTO

Sentiment Analysis

Canadian Election 2019

Prepared by: Eraj Ahmed
Student ID: 999177037

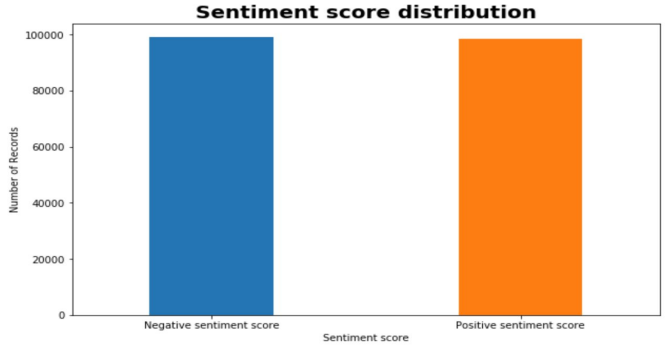
Course: Introduction to Data Science & Analytics
Course Code: MIE 1624

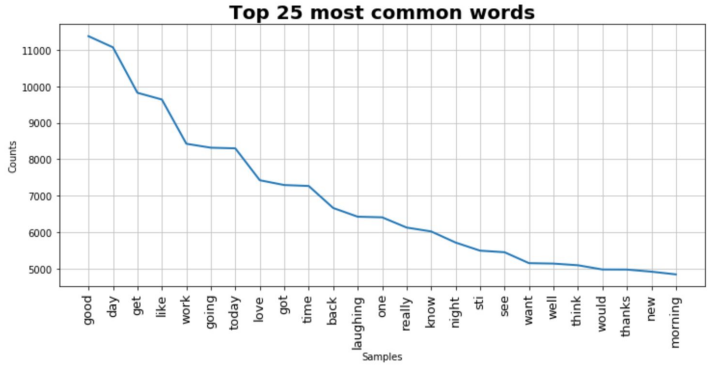
Exploratory Analysis (Generic Tweets)

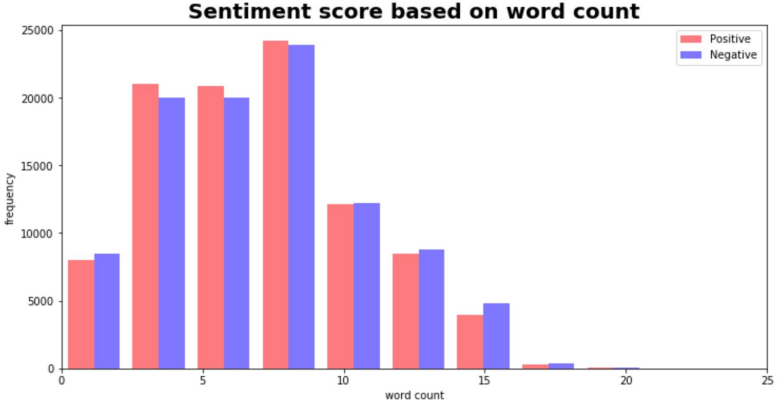
Initially I carried out exploratory analysis to get a better sense of the dataset and what are the common words appearing frequently.

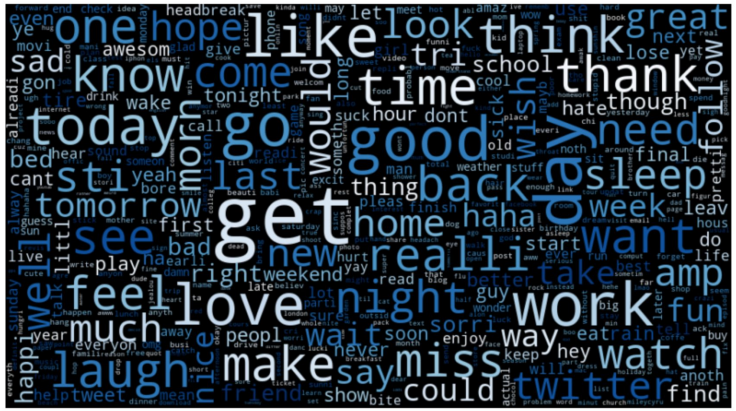
Bonus: Multiple Visuals

```
Number of positve tagged sentences is:  99020
Number of negative tagged sentences is: 98553
total length of the data is:            197573
```







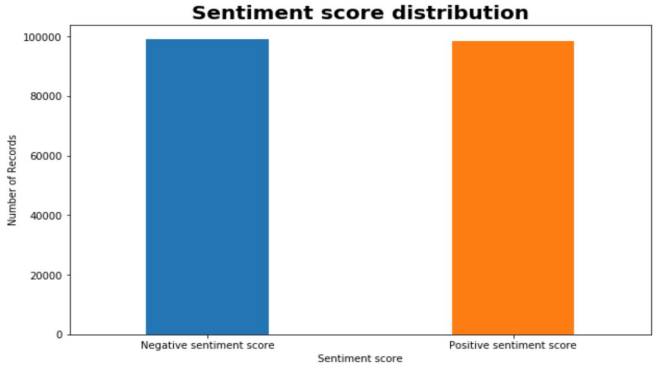


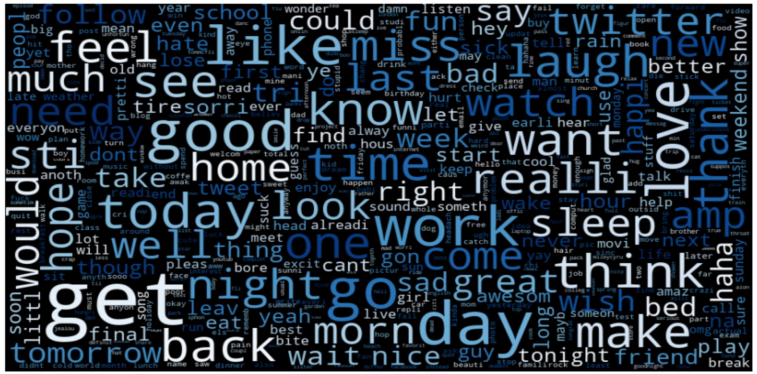
Exploratory Analysis (Election Tweets)

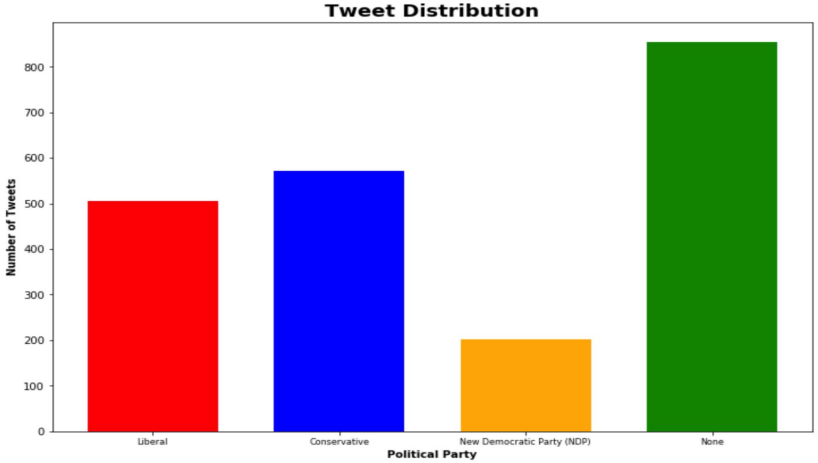
Then I carried out exploratory analysis on election dataset to understand the data.

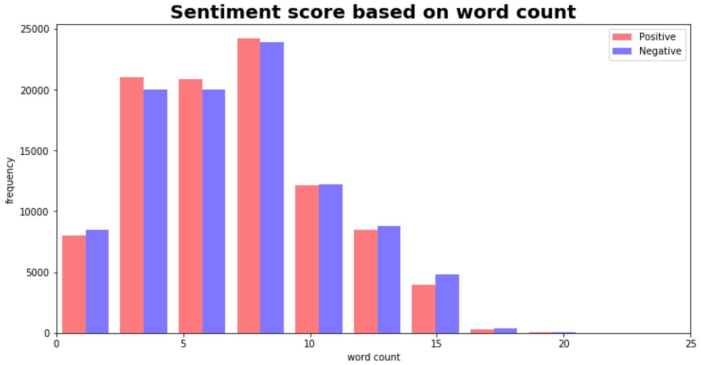
Bonus: Multiple Visuals

```
Number of positve tagged sentences is:  1127
Number of negative tagged sentences is: 1006
Total length of the data is:            2133
```



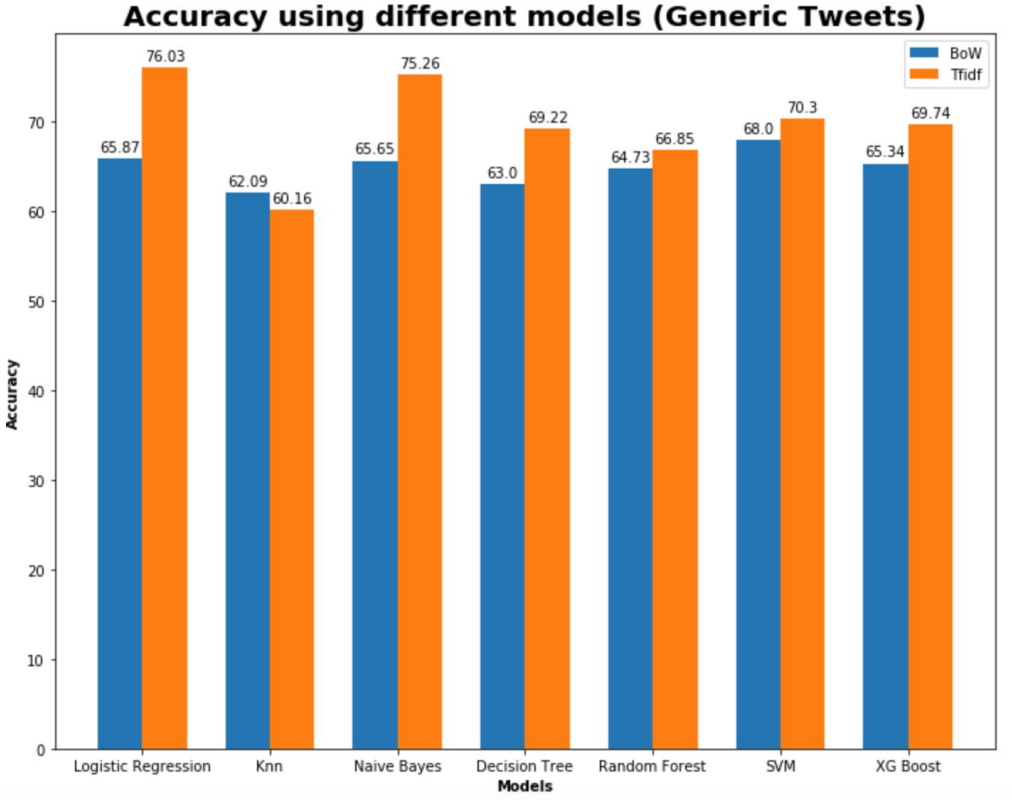






Model Implementation

In this step I all trained all the models on generic tweets dataset and plotted the prediction result.



From the plot on the left hand side I can see that logistic regression model is working best. Therefore, I use this model and train it on election tweets.

```
Accuracy of logistic regression model on election tweets: 71.12
```

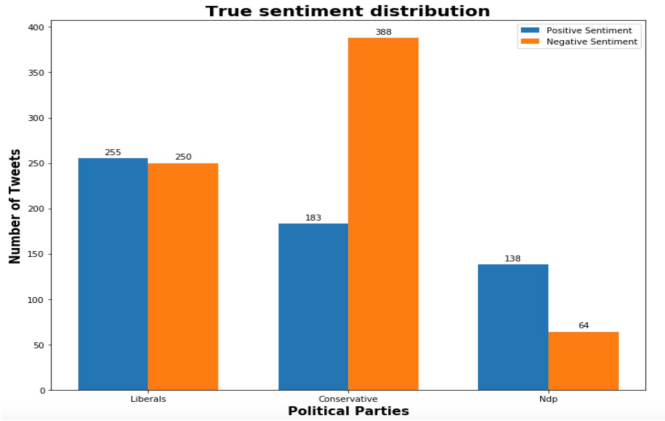
Bonus: Hyperparameter Tuning

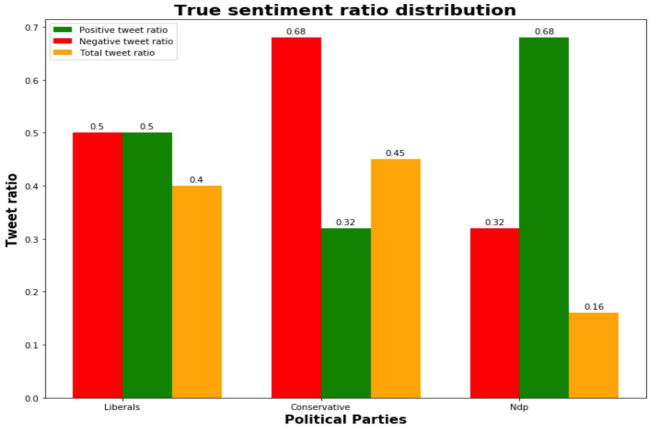
I got an accuracy of 71.12% which was then improved by tuning the hyperparameters. I used gridsearch and the performance of the model improved to 74%.

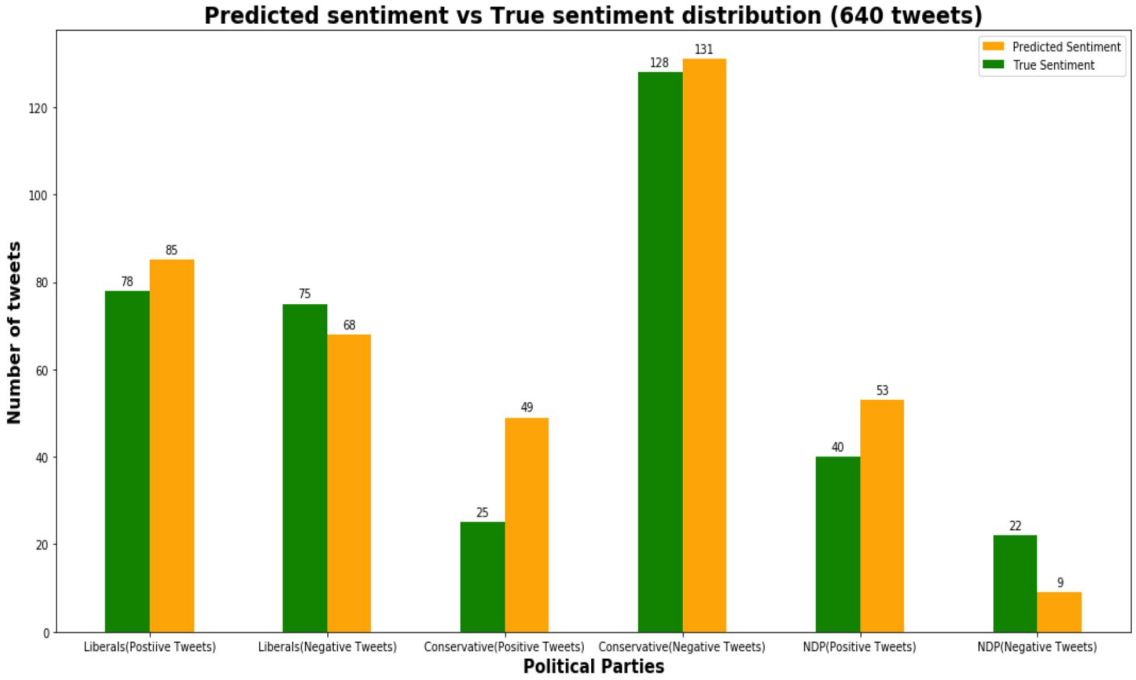
	precision	recall	f1-score	support
0	0.76	0.67	0.71	306
1	0.73	0.80	0.76	334
accuracy			0.74	640
macro avg	0.74	0.74	0.74	640
weighted avg	0.74	0.74	0.74	640

Sentiment Distribution

In this step I first found true sentiment distribution and then compared it with sentiment prediction results for three political parties.

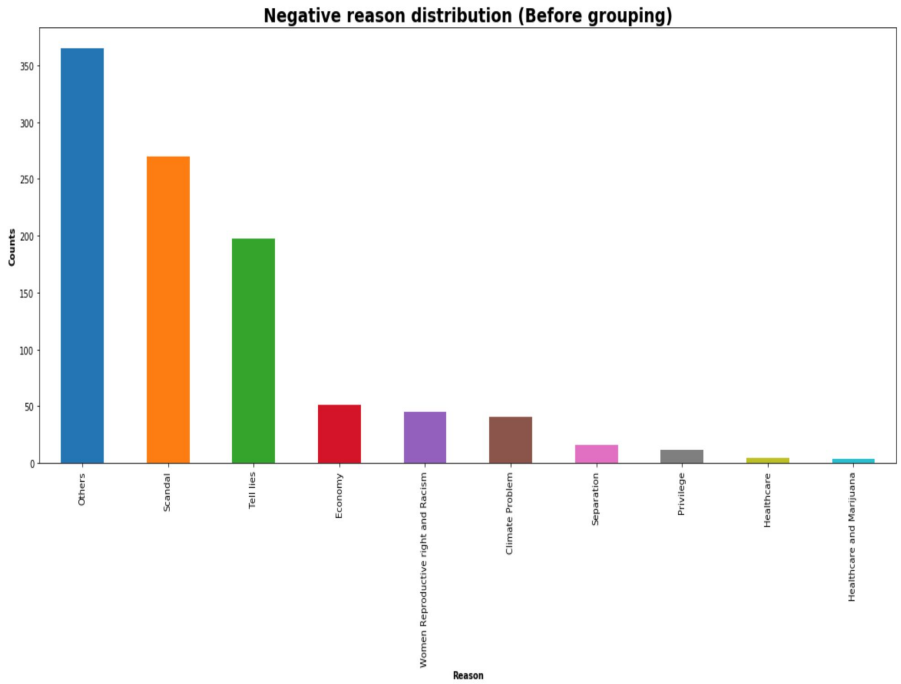




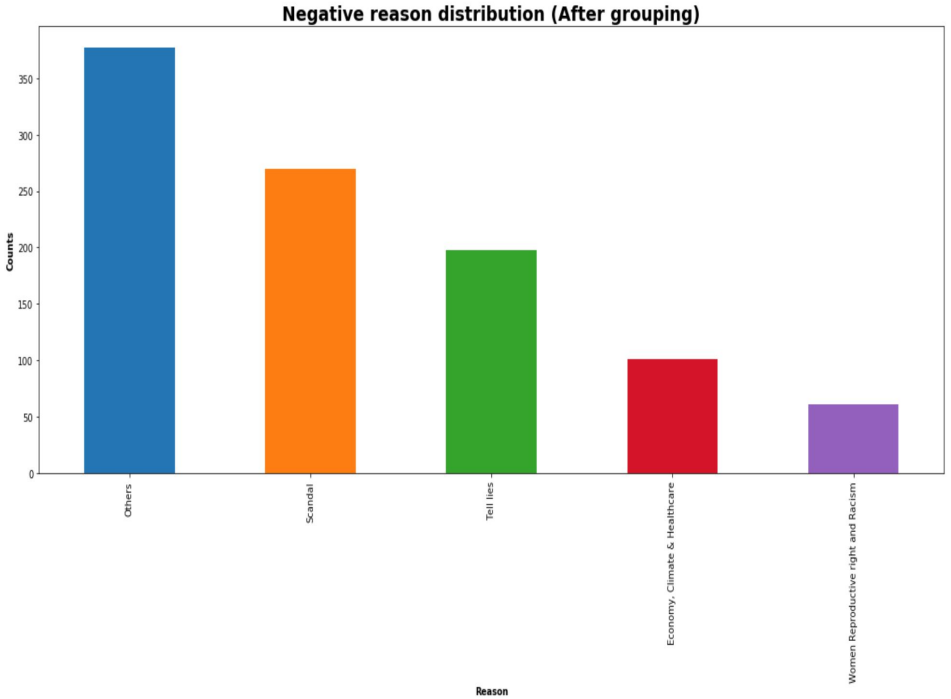


Multiclass classification (Negative Reason)

In this step I first found the distribution of labels and then regrouped some of the classes as it is a severely unbalanced dataset. The reasoning behind the regrouping process can be found in the Ipython notebook.



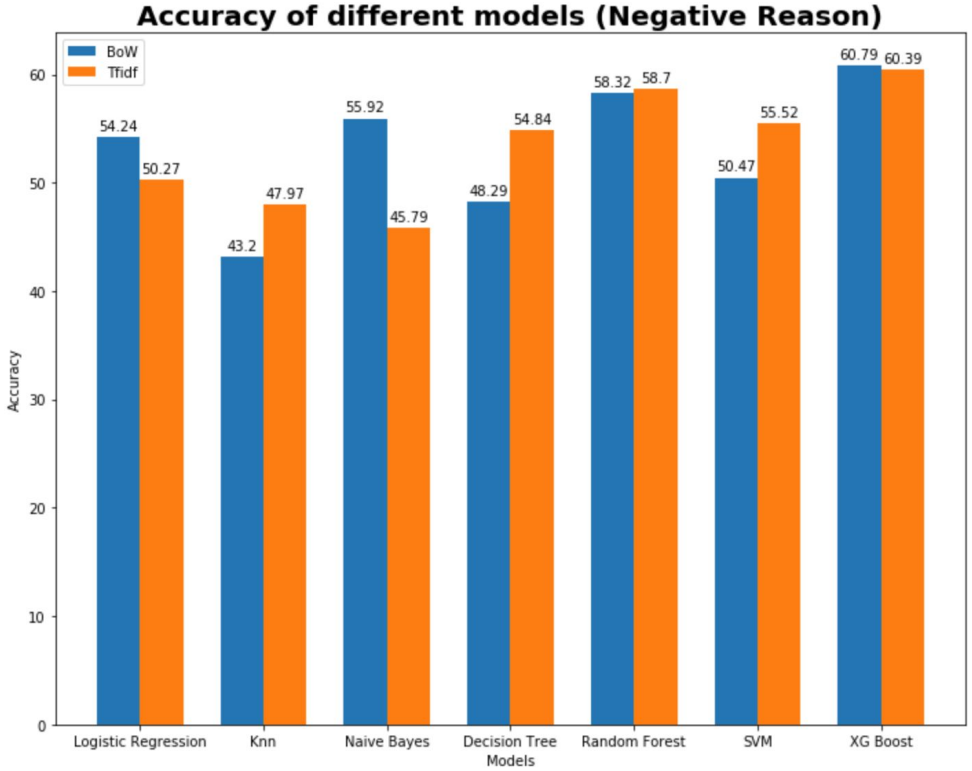
Before Grouping



After Grouping

Multiclass classification (Negative Reason)

In this step I trained all the models to identify negative reason in the elections dataset.



Bonus: Hyperparameter Tuning

From the plot on the left hand side I can see that XGBoost is giving highest accuracy (60.79%). Therefore, I do hyperparameter tuning on this model using gridsearch in order to improve the performance of the model. The new accuracy is 63%.

	precision	recall	f1-score	support
Economy, Climate & Healthcare	0.80	0.26	0.39	31
Others	0.54	0.86	0.66	125
Scandal	0.69	0.49	0.58	83
Tell lies	0.69	0.42	0.52	48
Women Reproductive right and Racism	0.40	0.12	0.19	16
accuracy			0.59	303
macro avg	0.62	0.43	0.47	303
weighted avg	0.63	0.59	0.56	303