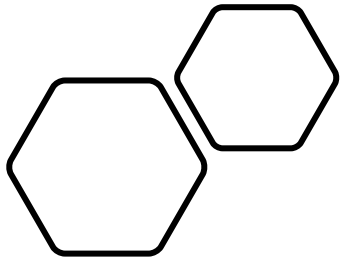
Sentiment Analysis of tweets of the Canadian Election 2020

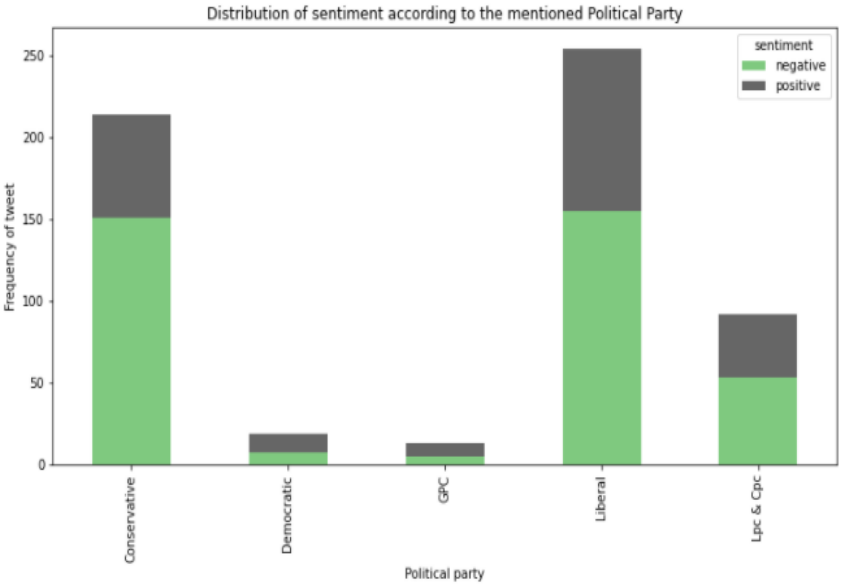


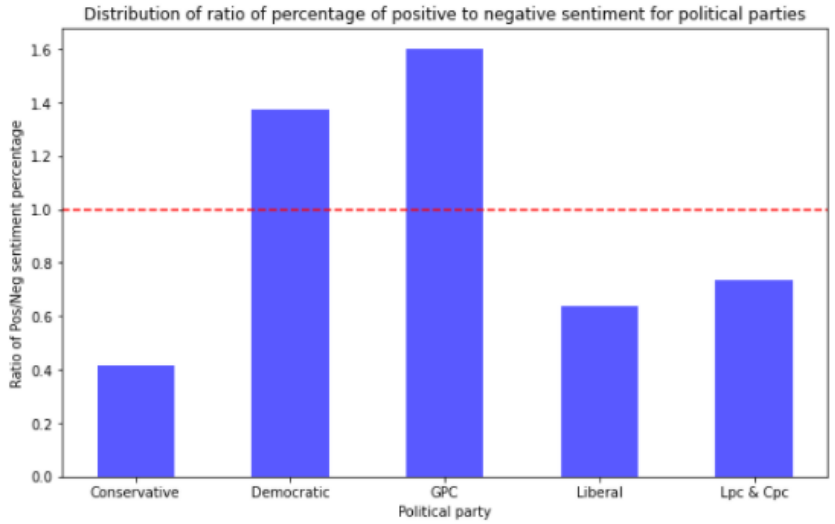
Assignment 3

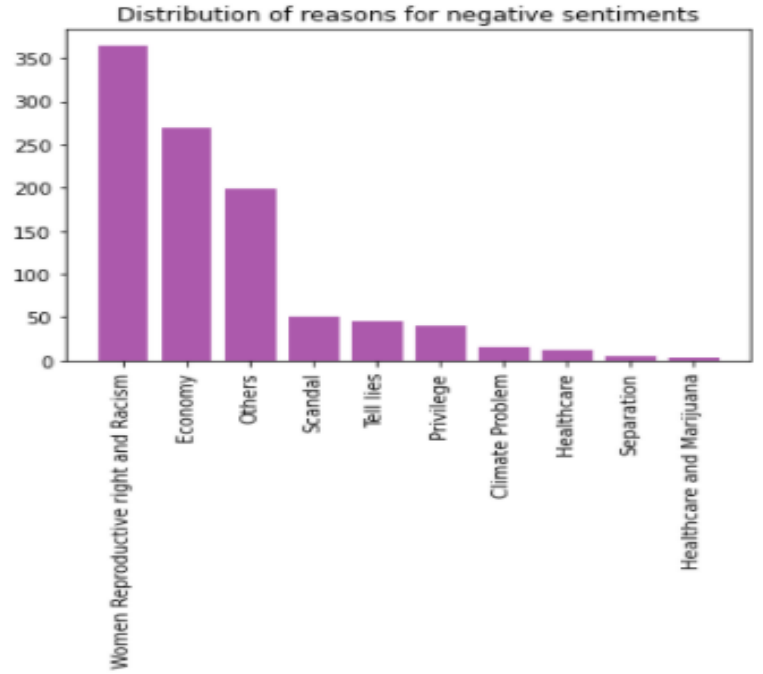
Name: Ruchit Shah

Student Id: 1005677830

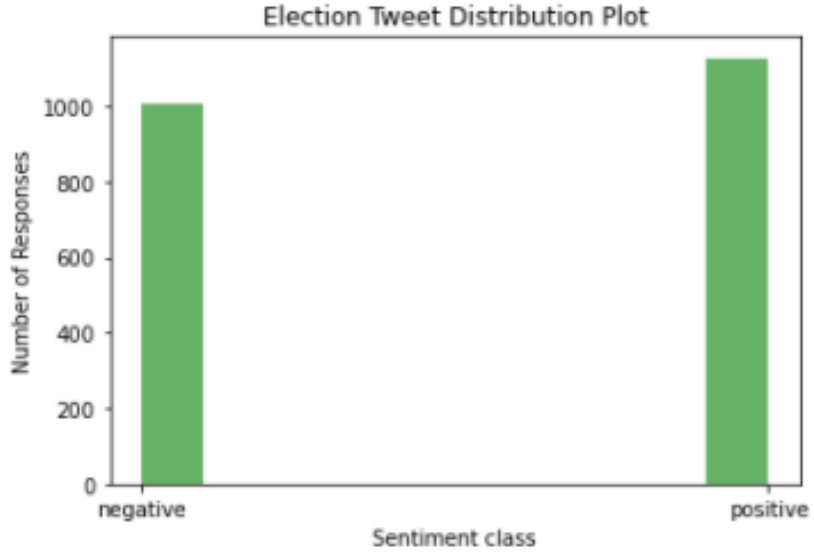
Exploratory Data Analysis



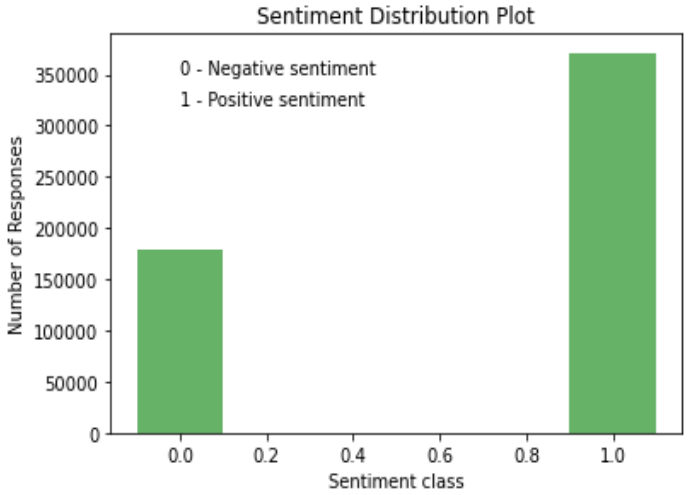




Election tweet sentiment analysis according to the political parties: The green bar shows the negative tweets for the particular party. The tweets are separated according to the party based on the hashtags words. It is observed that most of the tweets for conservative and liberal are negative. The frequency of tweets for the democratic and Green political party is quite less. From ratio of positive to negative plot, it is noted that the democratic and GPC has ratio more than 1 which implies more positive tweets than negative. While for the rest of the parties, there are more negative sentiments in the tweets.
Moreover, the histogram of negative reasons for the election tweets infers that the women reproductive and racism is the main reason which has frequency around 350. Whereas, the healthcare and marijuana is the least important reason. This shows Canadian people believes negatively towards political parties mostly because of women rights and economy.



Election Tweets Sentiment Distribution: It seems to have more positive sentiments in the election tweets which is around 1100. While there are around 900 negative sentiment tweets. Though positive tweets are more, the ratio is near about 1 only.



General sentiment tweets: The histogram plot for the distribution of sentiment dataset shows that most of the sentiment are positive. The positive responses are around 350k which is twice that of negative which is around 180k.

Word Cloud Analysis for the General Sentiment dataset: It is observed that the words like people, fuck, trump, hate, look, know, fucking, time, shit appear more frequently in the general negative sentiment tweet dataset. For the positive sentiment general tweet dataset, words like happy, love, thank, good, amazing, best, found, birthday, great, love appear more frequently which shows some positive words included.

Word Cloud Analysis for the Election Tweets:

- **Negative**: The words like conservative, Trudeau, Kinsella, Andrew, Canada, party, election, liberal, scheer are used most frequently. It indicates the people's negative sentiment is towards the political parties – liberal and conservative.
- **Positive**: The words like election, liberal, conservative, uprisingh, Canada, federal are used most frequently. This indicates the positive tweets are also there towards the political party – liberal, conservative and NDP.
- It is noted that NDP party has mostly positive sentiment which is proved from the previous negative to positive tweet ratio histogram. Moreover, the positive and negative words are used for both liberal and conservative party nearly in same ratio which proves from the previous sentiment distribution plot.

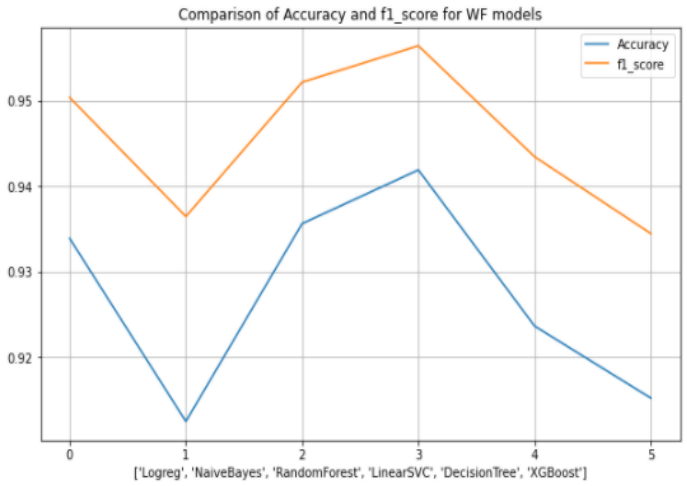
Model Feature Importance

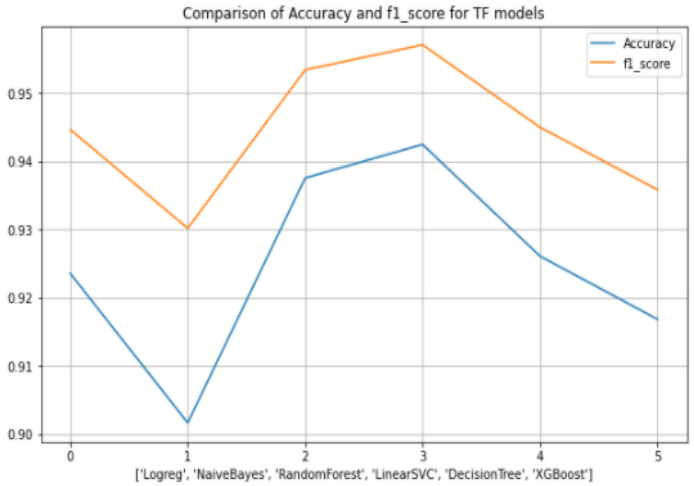
Bag of Words: This feature method represents the occurrence of words within a document. It involves a vocabulary of known words and a measure of the presence of known words. The order of words is not considered and only concern was made on the occurrences of the particular words. So, it is known as a 'bag' of words.

TF-IDF (TF – term frequency, IDF – inverse document frequency): It is a statistical measure that evaluates how relevant a word is to a document. It multiplies two metrics: how many times a word appears in a document, and the inverse document frequency of the word across a set of documents.

The two methods mentioned is used for the feature selection. The seven models are incorporated namely: logistic regression, Random Forest, Multinomial Naïve Bayes, Decision Tree, SVM, k-NN, and XGBoost.
The sentiment dataset is used to prepare a model. The dataset is split into the 30% test dataset. This step was taken first. So that the features are selected based on the test and train data respectively.
For the Bag of Words, CountVectorizer is used and for TF-IDF, TfidfVectorizer is used. The maximum feature is considered to be 2000 for the analysis in order to speed up the performance.

Results and Visualizations

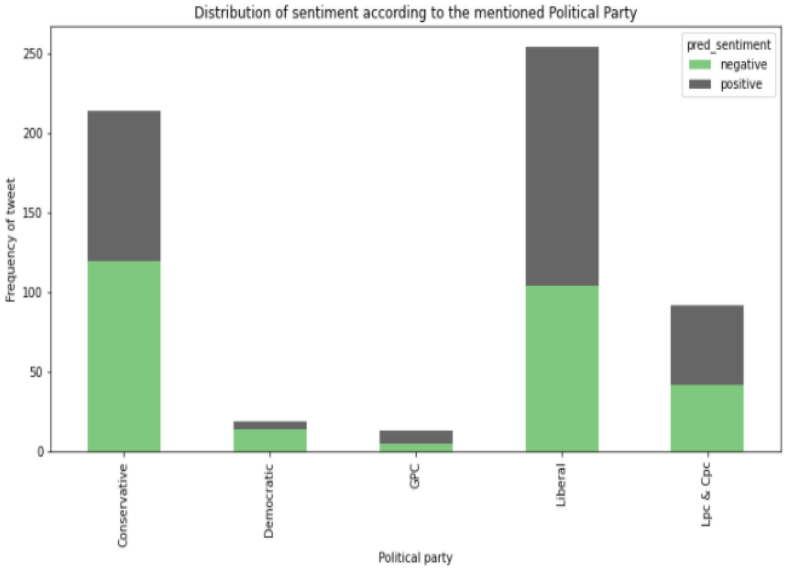


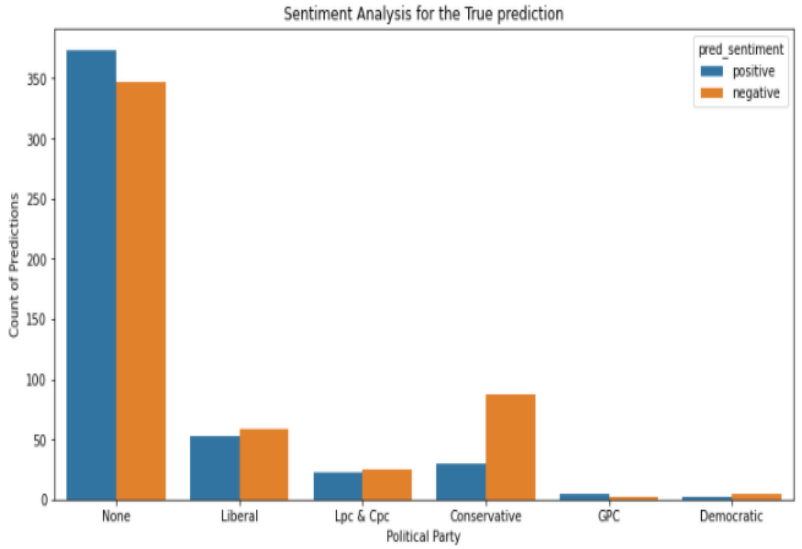


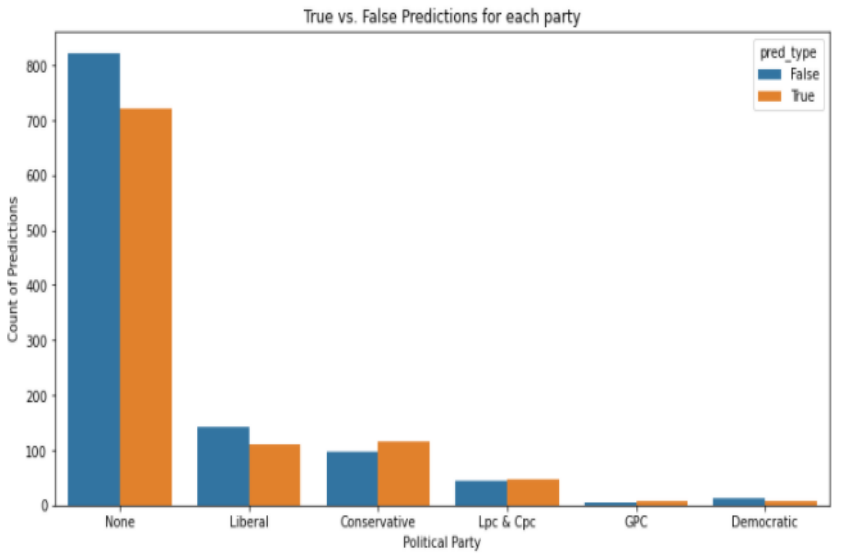
The comparison of the methods – logistic regression, random forest, naive bayes, decision tree, Linear SVC, and XGBoost is shown in the plot for both the Word Frequency (Bag of Words) and TF-IDF features methods for the sentiment dataset to prepare the best model for implementation. The "**accuracy**" as chosen as metric for comparison.

The **SVM model (Linear SVC)** gives the best result with the accuracy score of **94.18% for WF** and **94.24% for TF-IDF**.

Therefore, the SVM model is used for the implementation on the election tweet dataset. Also, the accuracy was compared for the WF and TF-IDF method using Linear SVC model.







Linear SVC model: **WF** has accuracy of **46.64%**. **TF-IDF** has accuracy of **47.30%**. Therefore, TF-IDF method with Linear SVC model is used for implementation.

The predicted sentiment for the different political parties shows the main impact can be observed for the conservative and liberal party. The liberal party has more positive predicted sentiment compared to negative sentiment. While for the conservative party, there are more negative sentiments predicted.

When the predicted sentiment is compared with the actual sentiment, it is observed that the liberal party has more false prediction (i.e., the predicted value does not match the actual value) and conservative party has more correct predictions.

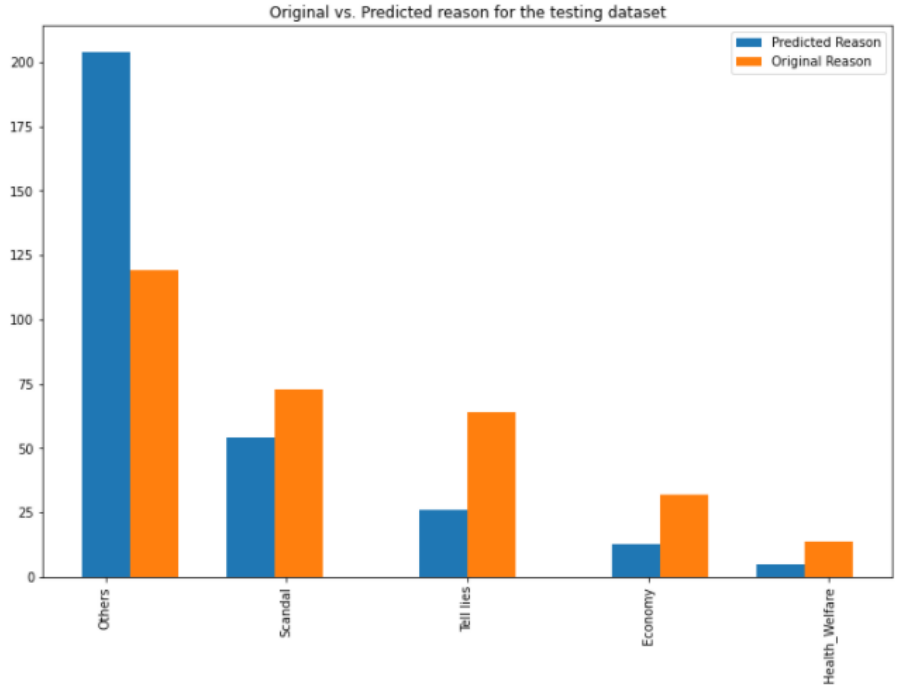
Therefore, the predicted sentiment plot was observed for the true predictions tweet only. From this analysis, it is found that the liberal party has more positive tweets and less negative tweets compared to the conservative party. This implies that the liberal party is on positive side compared to conservative party based only on the prediction made using the NLP machine learning algorithm. Considering the overall (true and false prediction) predictions also, it is noted that the liberal party has more positive sentiments than the conservative party.

Results: Based on the analysis, it is predicted that the **Liberal Party** will win the election. Also, the actual result is towards the Liberal Party only. Therefore, it is concluded that the NLP analytics is excellent tool to have a general opinion about the results.
The accuracy can be increased by tuning the hyperparameters and using the best optimized result.

Negative Reason Analysis

The negative reasons analysis is carried out using three methods: Logistic Regression, Random Forest, and Multinomial Naïve Bayes. The hyperparameter tuning is performed for all the three models and the best optimized parameters are used for all the three models. The comparison of the accuracy and f1 score for the best optimized three models is shown in the plot which implies that the **Random forest** classifier gives the best result among the three models with the accuracy of **57.28%**.

The predicted result for the negative reasons goes in the same flow as the actual reason with the only change that the reason 'Others' has more weightage. The model fails to have accurate prediction maybe because of the grouping of the negative reasons. Also, the WF and TF-IDF feature will account based on the polarity of the words. Now, there may be words which are in more than one reasons comparably and having same polarity. So, it fails to predict the most accurate results.



Bonus

A deep learning algorithm Keras TensorFlow model is used to analyse the sentiment. The WF feature is taken for the feature selection method in order to keep it simple. The 10 epoch were provided with the batch size of 200. The best model accuracy is found to be **96.73**%.
Though for this model, the predictions were not that accurate for the election tweet dataset.