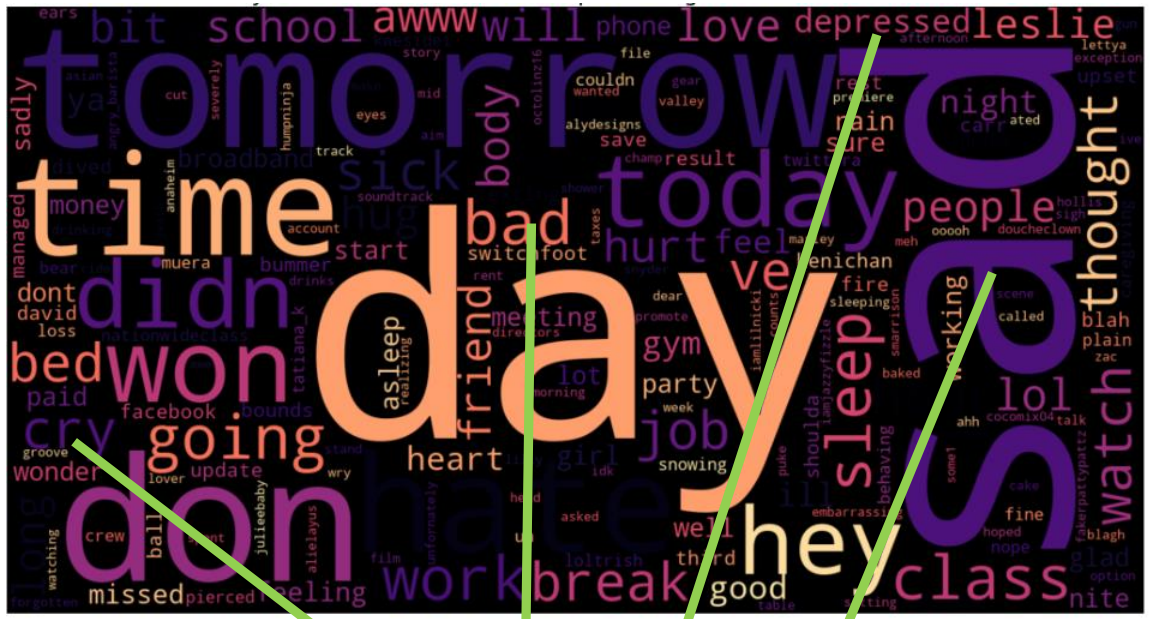
Project 2
Natural Language Processing of Tweets

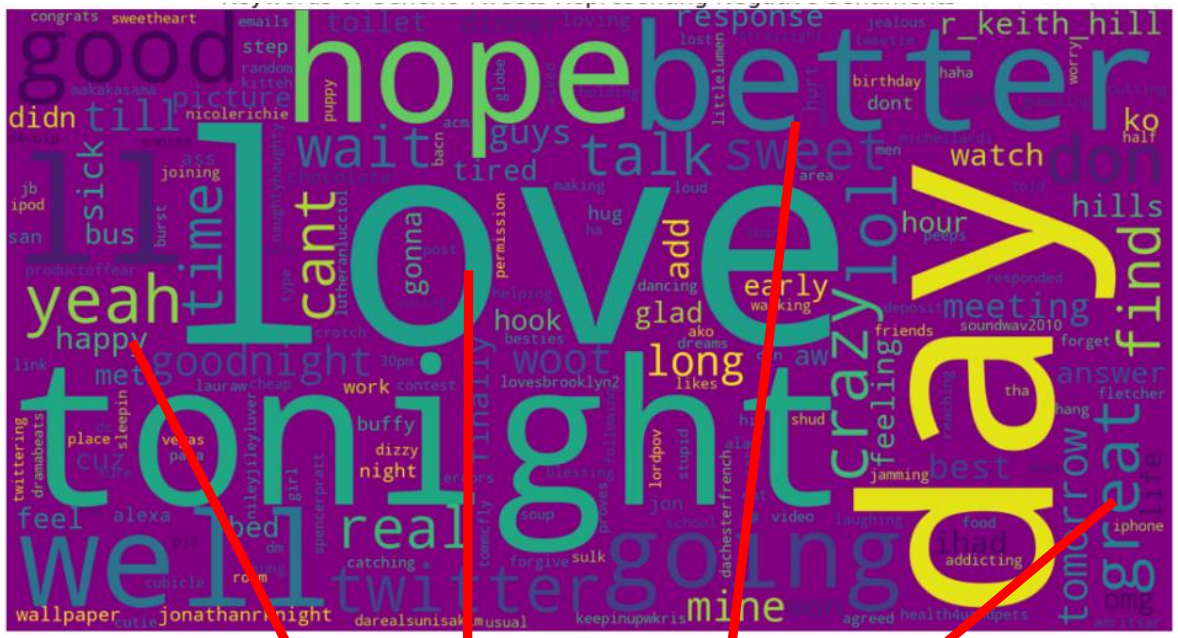
Saket Thavanani

Student Number-1005643145

Exploratory data analysis (Generic Tweets)

Word Cloud for Generic Tweets





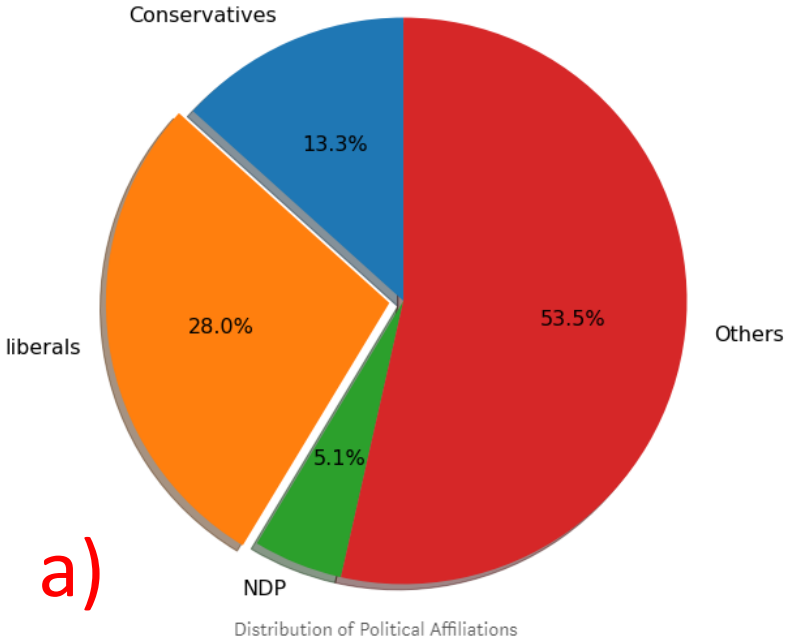
Words Representing Negative Sentiments

Words Representing Positive Sentiments

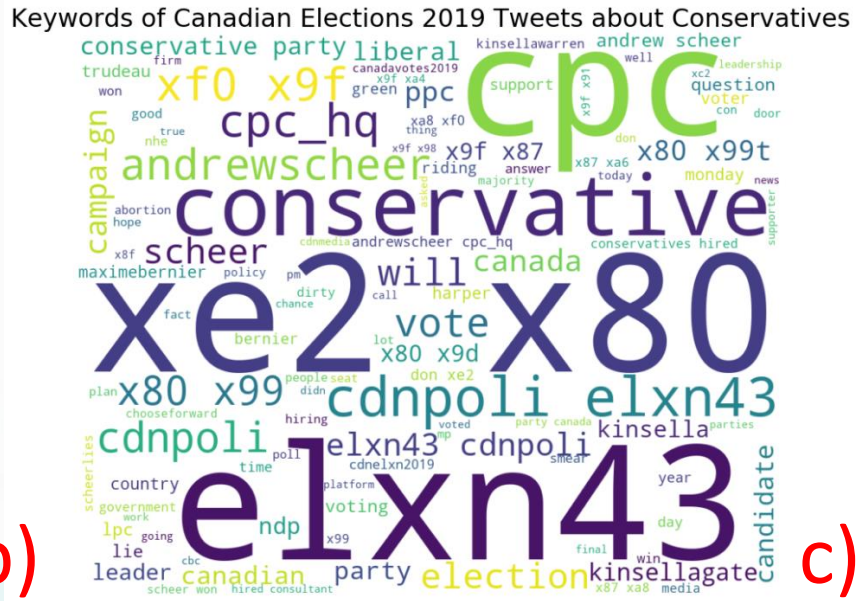
Inference-

- World cloud was generated for the generic tweets dataset.
- We can see that words which represent positive sentiments like love, hope, happy are on the right side.
- Whereas, words which represent the negative sentiments are on the left side such as sad, cry, bad,etc.

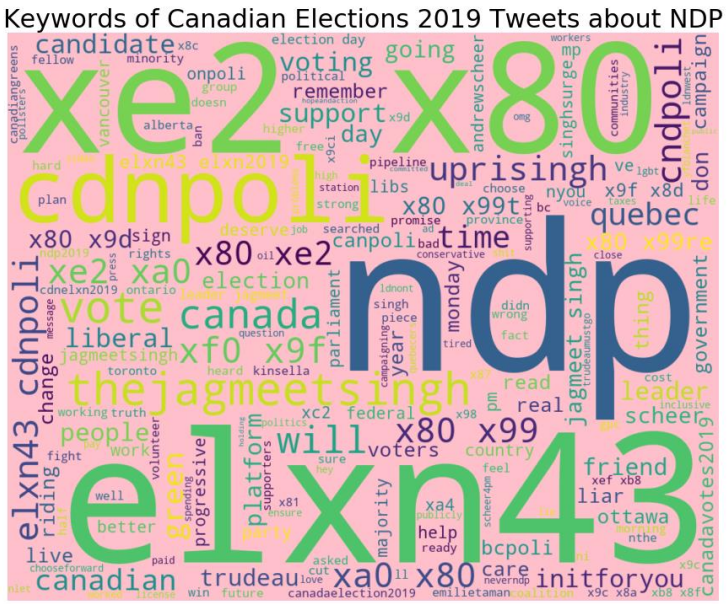
Exploratory data analysis(Canadian Election Tweets)



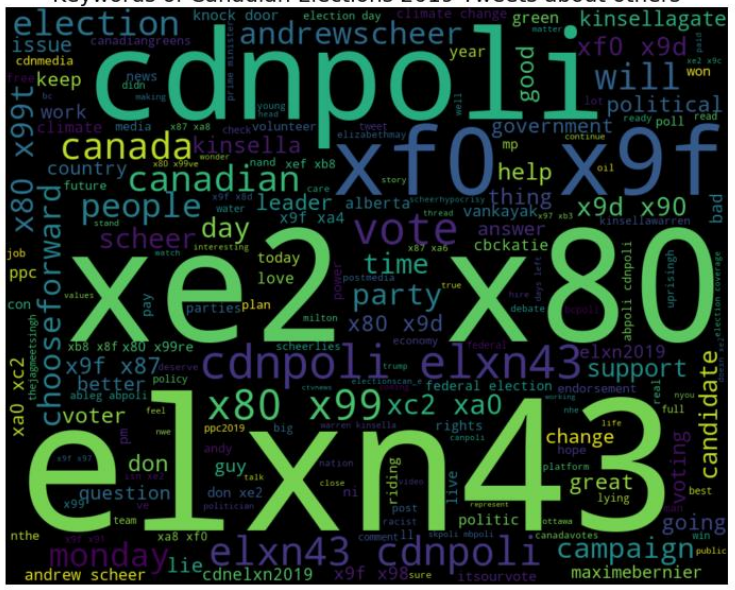
a)



b)



c)



d)

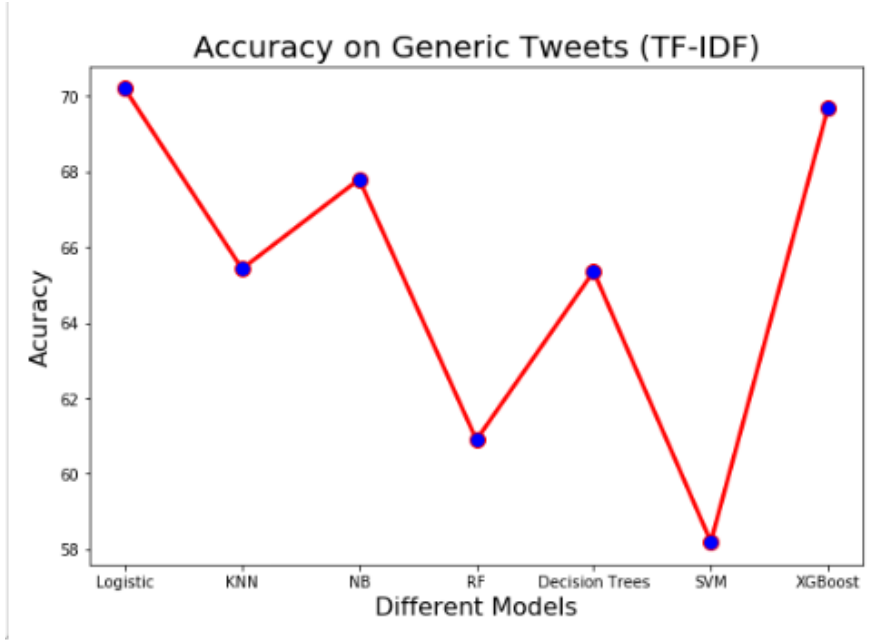
Inference-

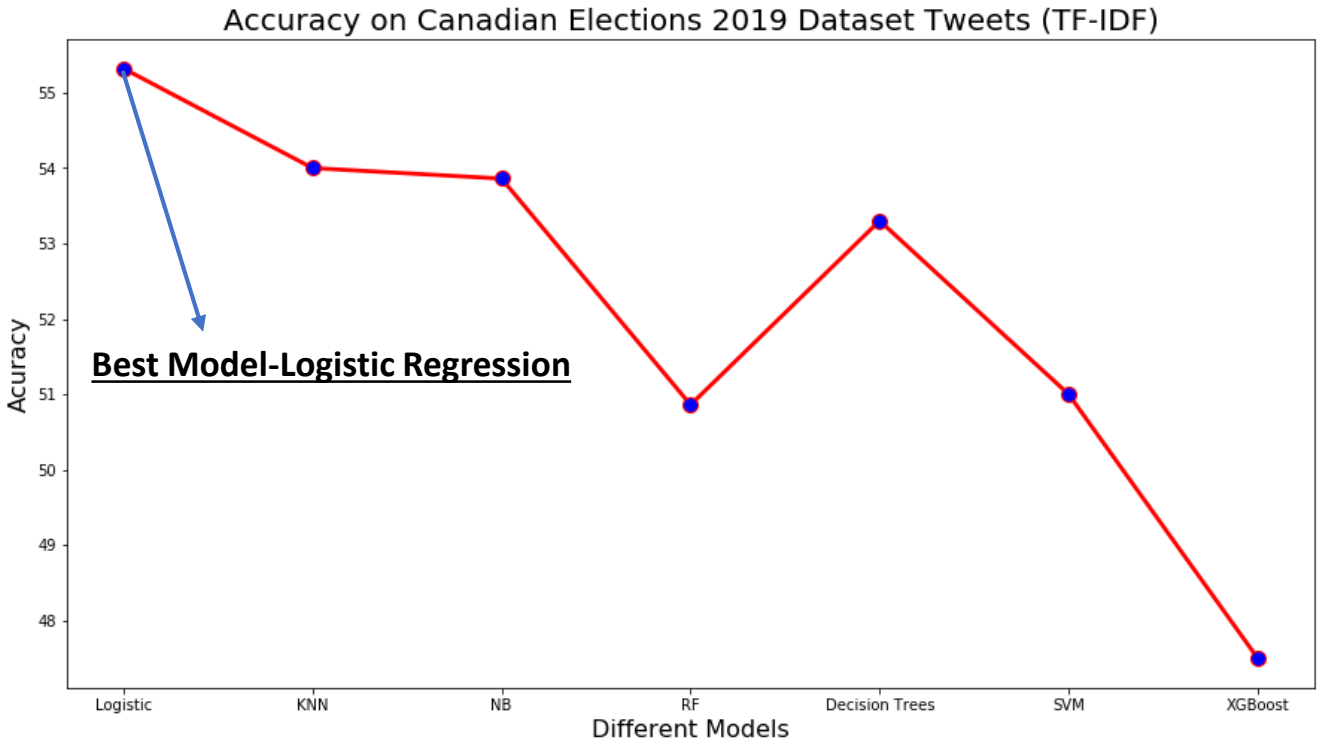
- Based upon the major keywords associated with various parties the tweets were classified into various political affiliations.

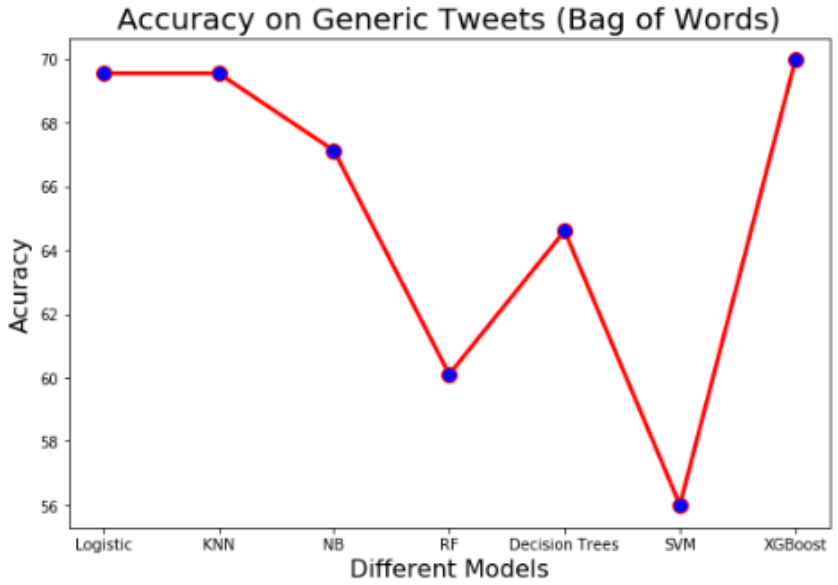
- We can see from the above plot that most of the tweets were classified as none (1142) . But among the political affiliated we had the most number of tweets of liberal party (511) followed by conservative party (384) further followed by NDP(96) at the lowest.

- Apart from that world cloud was plotted for all the political affiliations, we can notice the keywords such as Conservative for conservative party, ndp keyword for the NDP party

Model Results on Based on Features







Inference-

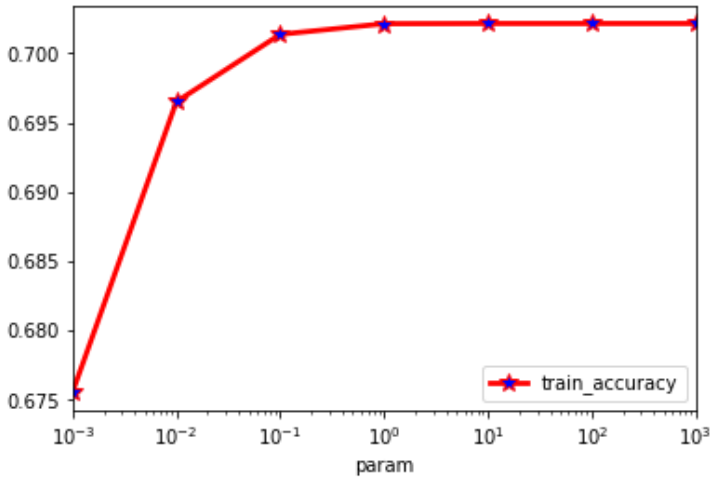
- All the models were implemented with using two sets of feature techniques- Tf-Idf vectorization and bag of words. All the scores of all the models are displayed above in form of the graphical plot.
- Based upon the above test results in form of graphical plot , the Logistic Regresion gave the best results for prediction of sentiments. Hence, for the model 1, I choose logistic regression to be the best model.
- We can also infer that XG-Boost model is performing pretty well on the generic tweets data set but it is not performing well on Canadian elections data set.

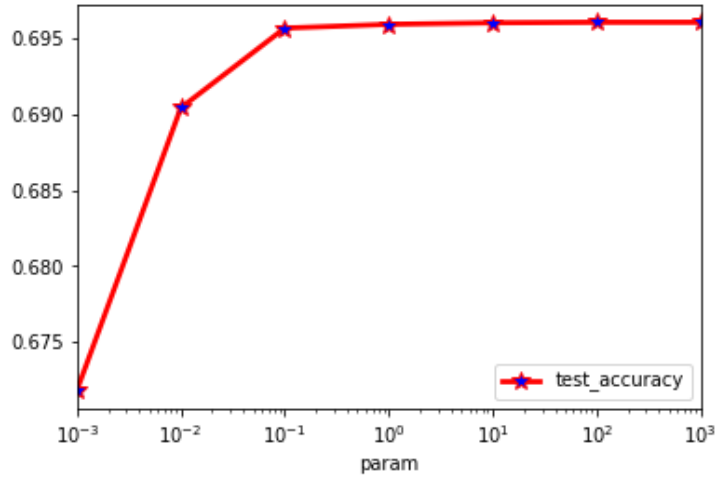
Best Model 1- Results (Displayed)

- Logistic Regression- Hyper-parameter tuning was also performed on the best model to obtain best parameters.
- These results below are obtained with the best Hyper-parameter C=1000 on Canadian Elections 2019 Dataset

- **Inference**- Here, param is the C hyperparameter using in logistic regression function. The best values of this C is obtained by tuning the hyperparameters first on generic tweets.
- Below we can see that the confusion matrix and ROC curve has been displayed to represent the results of my model visually.

Hyper-parameter Tuning on generic Tweets

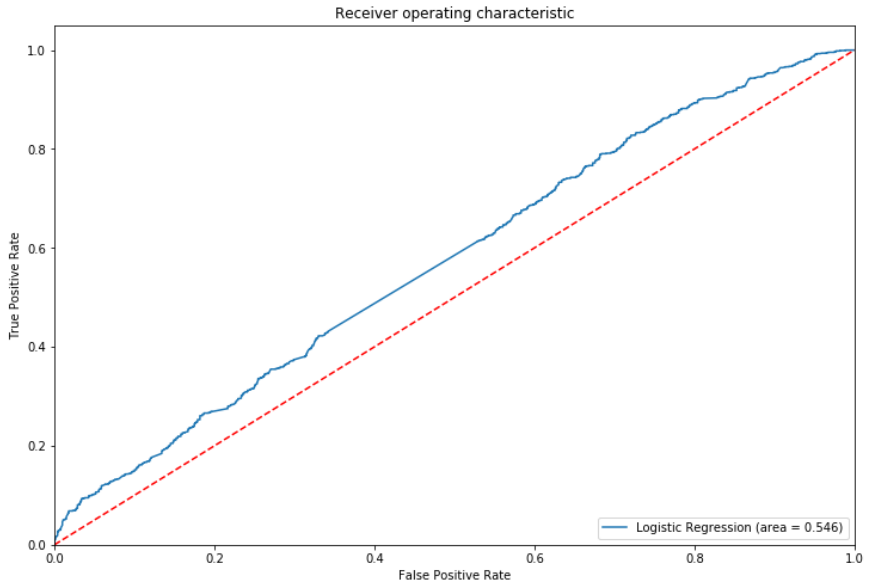




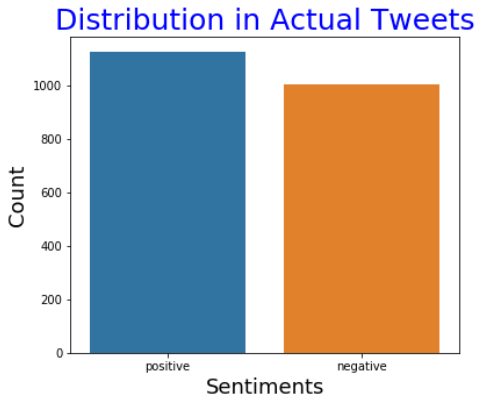
Accuracy= 56%

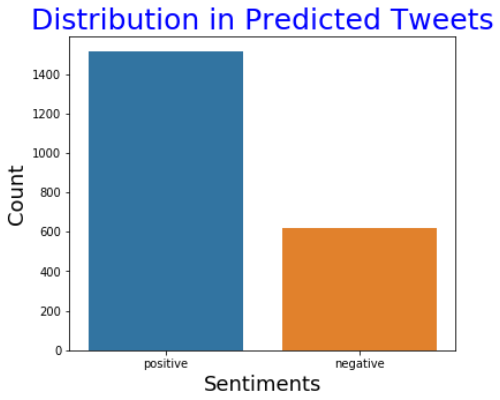
	Negative	Positve
Precision	**0.53**	**0.56**
Recall	**0.41**	**0.68**
F1-score	**0.47**	**0.62**

ROC Curve

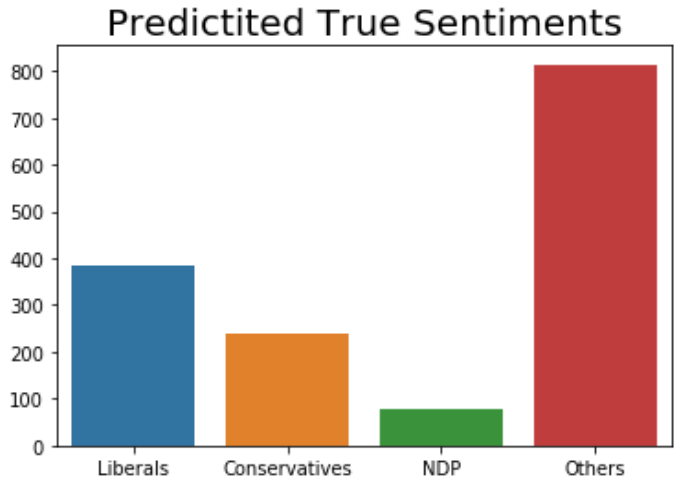


Visualization of Predicted Sentiments

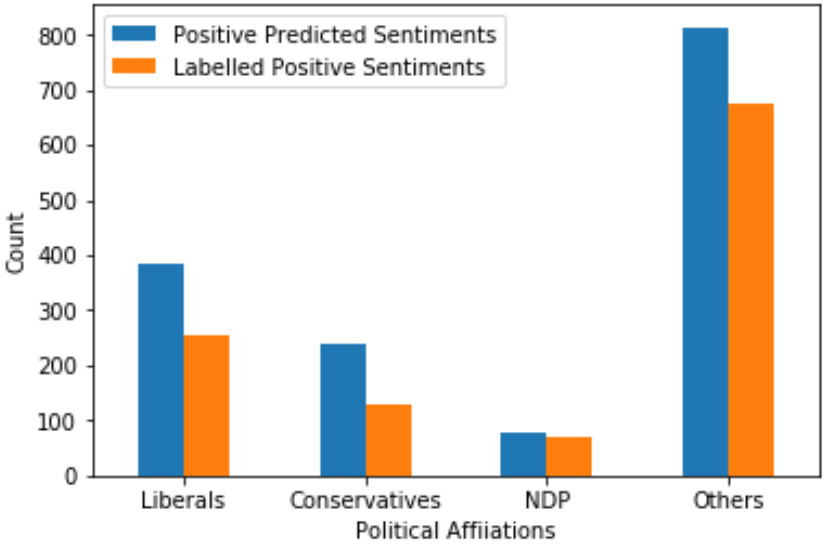




- **Inference**-There were more positive sentiment tweets in the predicted sentiments as compared to the actual labels in the Canadian election tweets

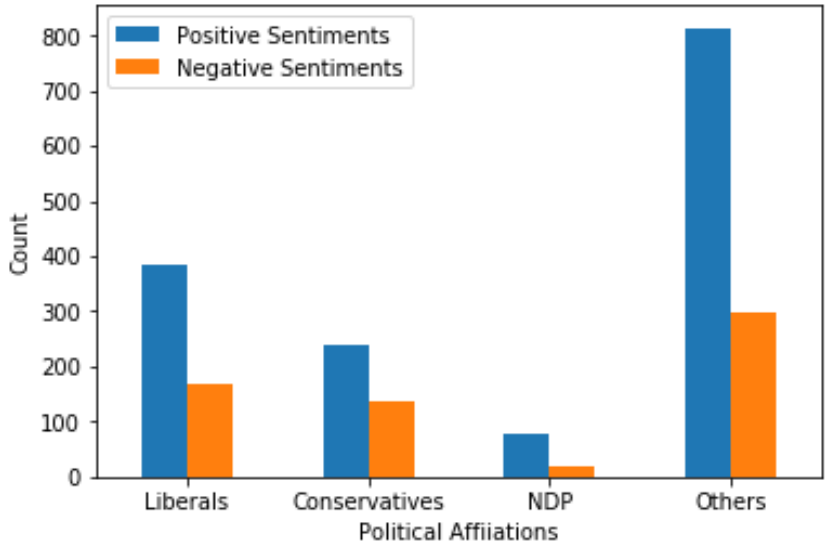


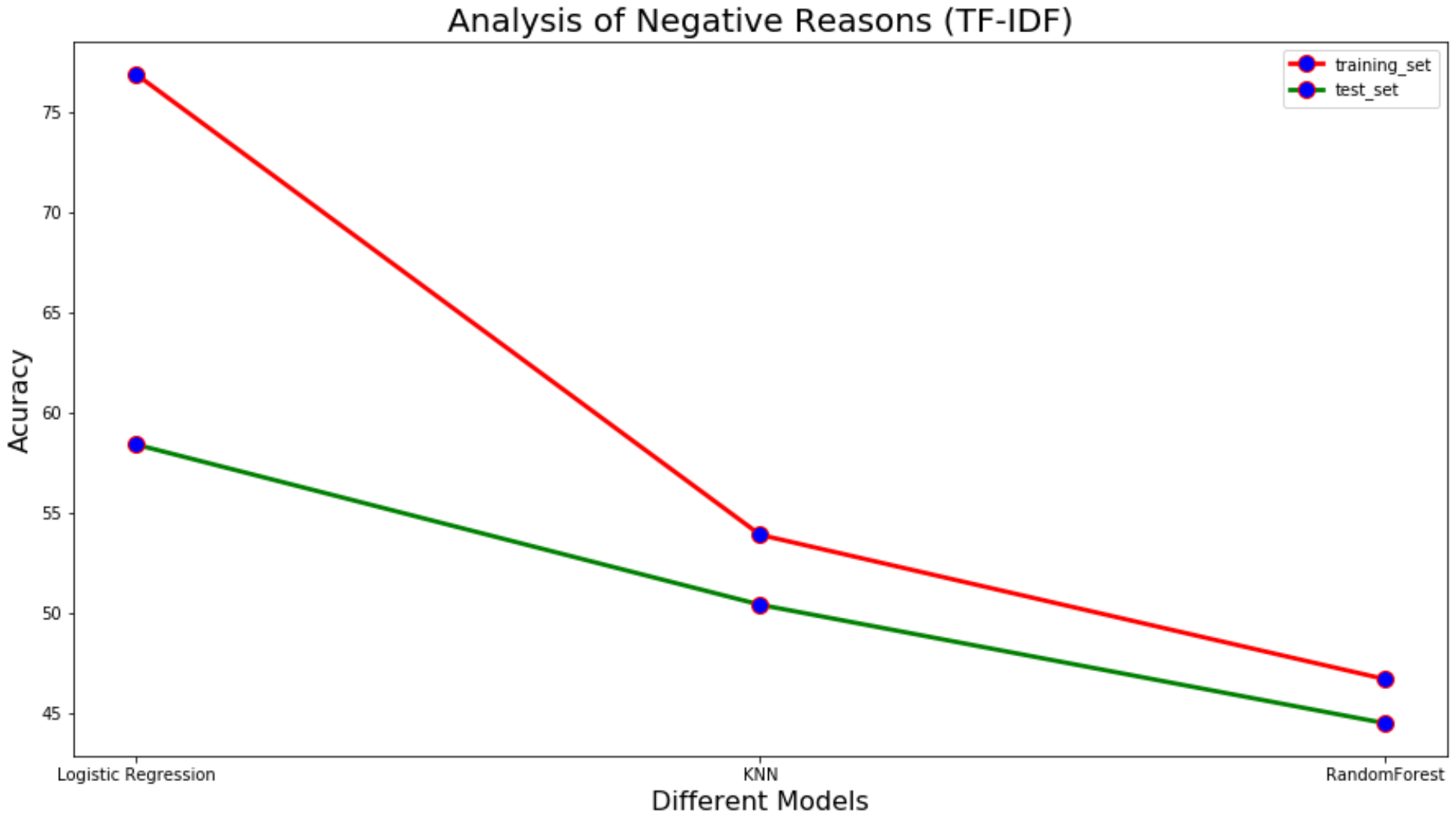
- **Inference**-When we see the visualization of predicted true sentiments based upon the political affiliation.
- We can see that the liberals had the most number of predicted true sentiments.
- This indicates that NLP analytics is pretty useful for predicting the elections outcome.



- **Inference**-When we see the visualization of predicted true sentiments compared to the actual true sentiments in labels based upon the political affiliation.
- We can see that the still liberals had the most number of predicted true sentiments.
- This indicates that NLP analytics is pretty useful for predicting the elections outcome.

- When we compare on the basis of predicted true and negative sentiments over all the liberals had the highest ratio vs true and negative sentiments indicating that liberals should winthe elections.





Inference-

- For analysis of negative reasons I used three machine learning models- logistic regression, KNN and Random forest.
- Out of them Logistic regression performed best of training and test set giving a score of 58% on test set which is reasonably good.