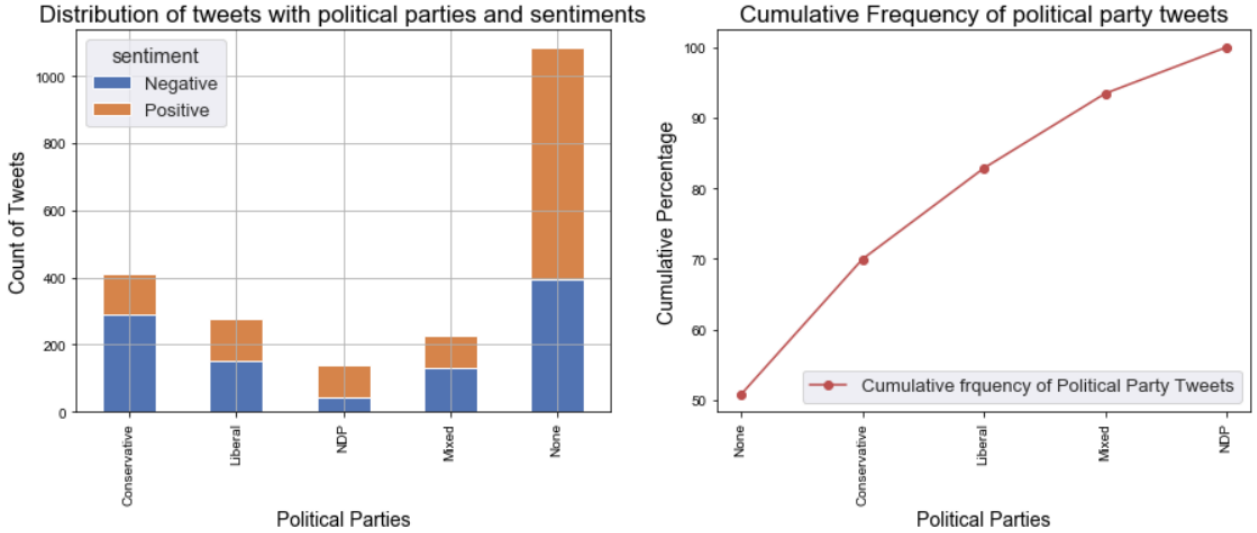
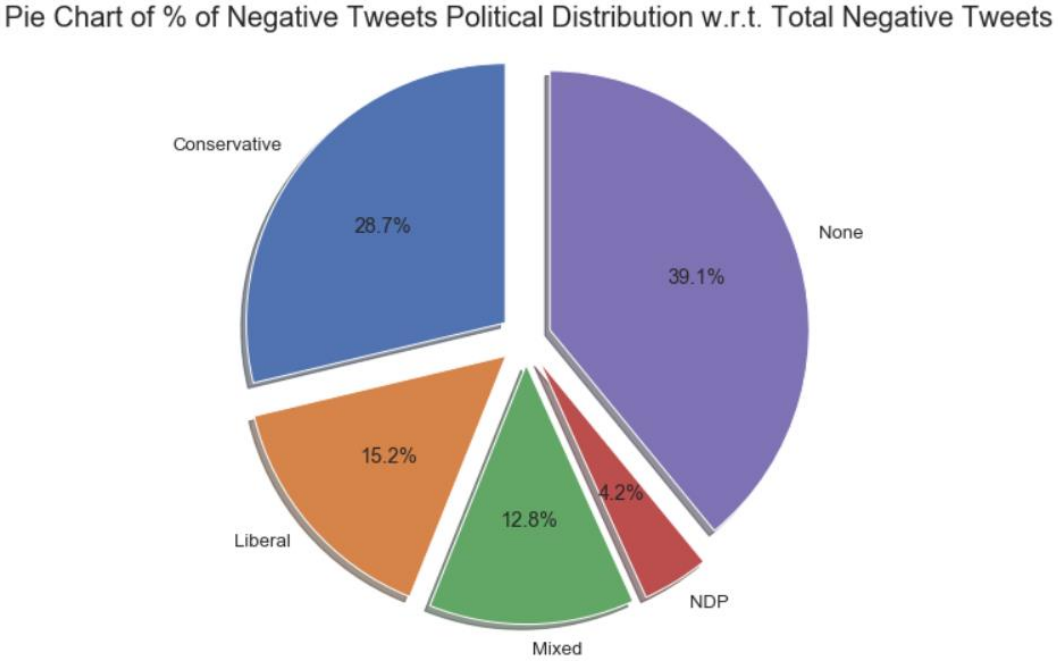
Twitter Sentiment Analysis of Canadian Political Landscape in 2019

By – Gautam Dawar, ID:

Section 1. Data Exploration



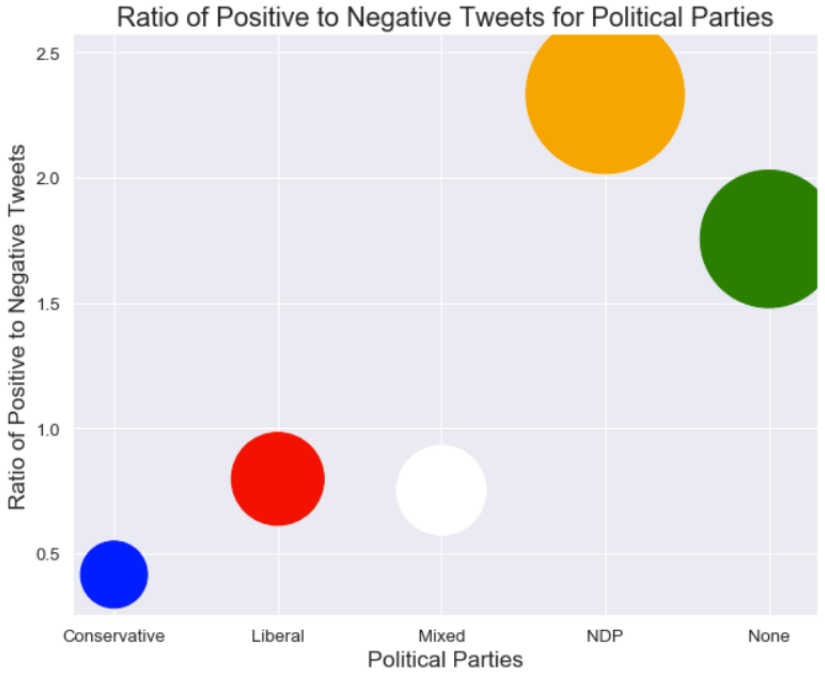


↑ Political spread of the tweets with respect to negative and positive tweets –

It is observed that the almost **50% of the tweets** belong **to neither** of conservative or liberals or NDP. Also, the number of **negative tweets** for **Conservatives and Liberals are significantly higher** than their positive tweets. Moreover, for **NDP the positive tweets are higher than their negative tweets.** Also, most of the tweets are affiliated to the Conservative Party according to this data.

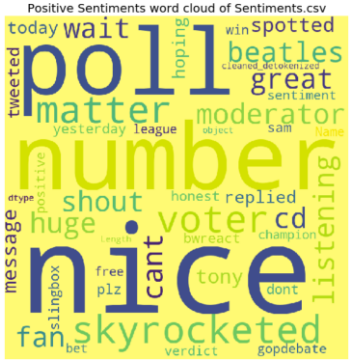
↑ Negative tweets political distribution–

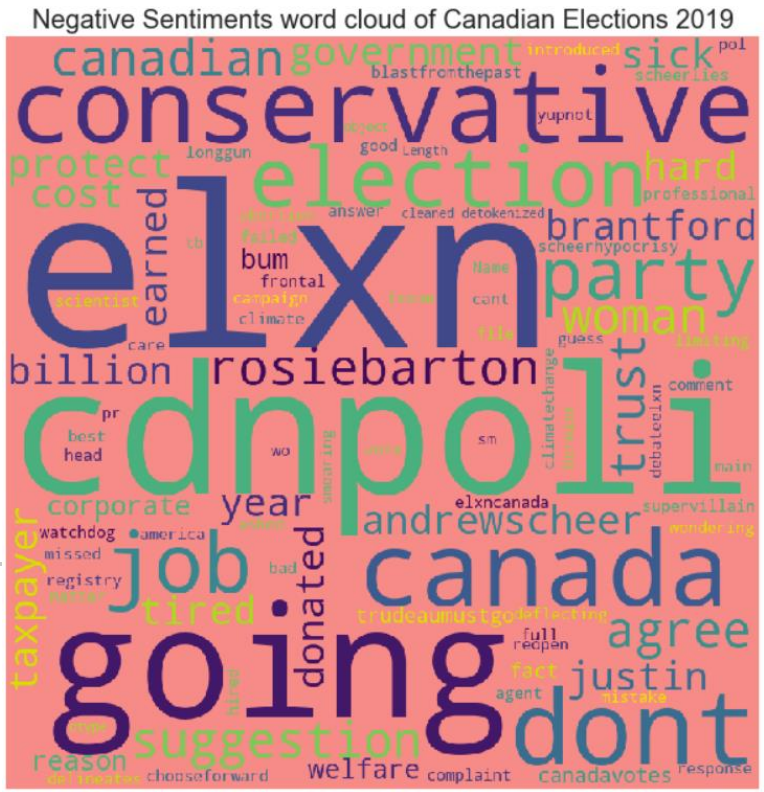
As we can see the negative tweets are most for the conservative party (28.7% of all the negative tweets) and NDP has the lowest number of negative tweets of all (4.2%)



↑ The positive to negative tweets ratio –

NDP has the maximum positive to negative tweets ratio (**2.33 positive to negative tweets**) and the lowest ratio is of **Conservatives (0.415 positive to negative tweets)**





↑ Word Cloud Analysis of Political Sentiments -

Negative Word Cloud of Elections = It seems Justin Trudue (Liberal Party) and Andrew Scheer (Conservative Party) appear most in the negative cloud. Hence this could highlight that people's hatred have been focused on the two major ruling parties in Canada i.e. Liberals and Conservatives

Positive Word Cloud of Elections = Besides 'elxn' and 'vote' appearing in the positive word cloud, words like 'thejagmeetsingh' and 'ppc' also appears in twitter. This suggested that public has a positive outlook for NDP and PPC political parties.

← Word Cloud Analysis of General Tweet -

Negative Word Cloud of Elections = It is evident that the words such as Night, Jealous, Watch, Love, Damn, Fight, UFC, Lost appear most of the times

Positive Word Cloud of Elections = It is evident that the words such as Voter, Moderator, Honest, Poll, Nice, Beatles appear most of the times

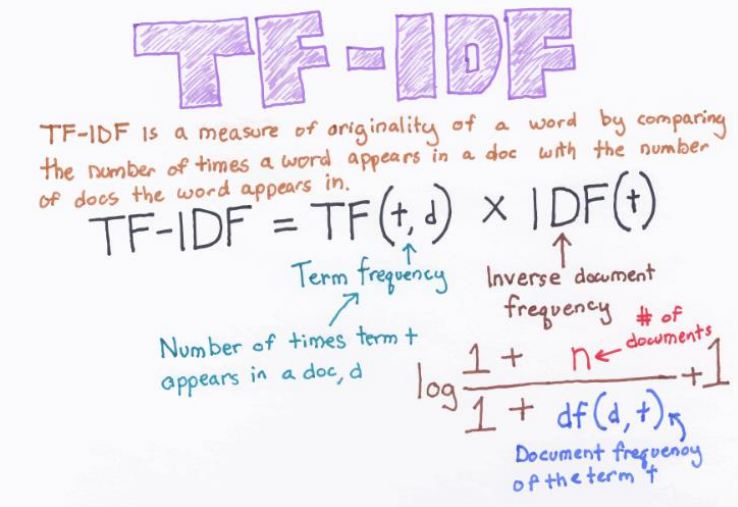
Section 2. Model Feature Importance

	abortion	absolutely	abt	access	accident
799	0	0	2	0	0
1167	0	0	1	0	0
1873	0	0	1	0	0
4323	0	0	1	0	0
6779	1	0	1	0	0

5 rows × 2000 columns

	abortion	absolutely	abt	access	accident
799	0.0000	0.0	16.6288	0.0	0.0
1167	0.0000	0.0	8.3144	0.0	0.0
1873	0.0000	0.0	8.3144	0.0	0.0
4323	0.0000	0.0	8.3144	0.0	0.0
6779	8.0982	0.0	8.3144	0.0	0.0

5 rows × 2000 columns



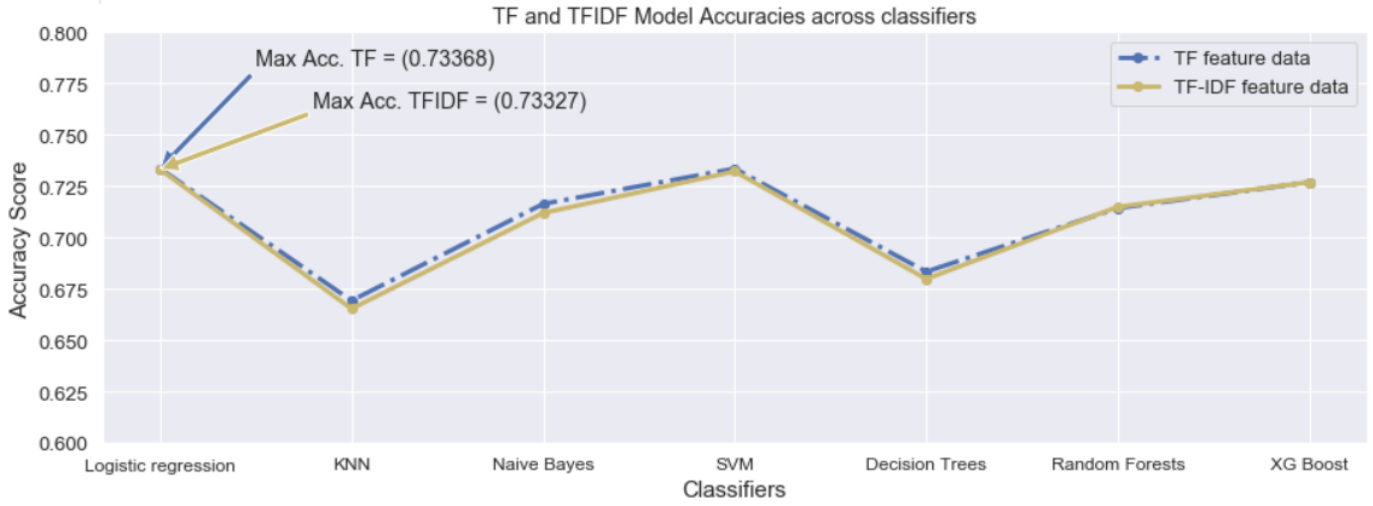
Bag of Words or Term Frequency (TF)

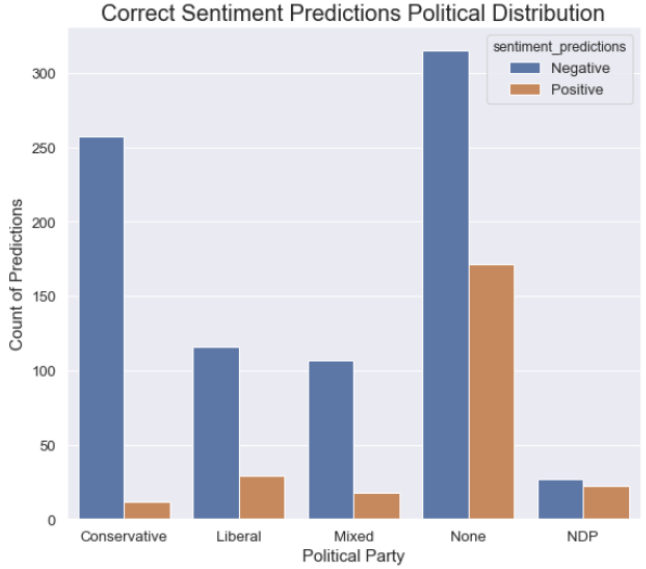
Bag of Words is a method to extract features from text documents. These features can be used for training machine learning algorithms. It creates a vocabulary of all the unique words occurring in all the documents in the training set.

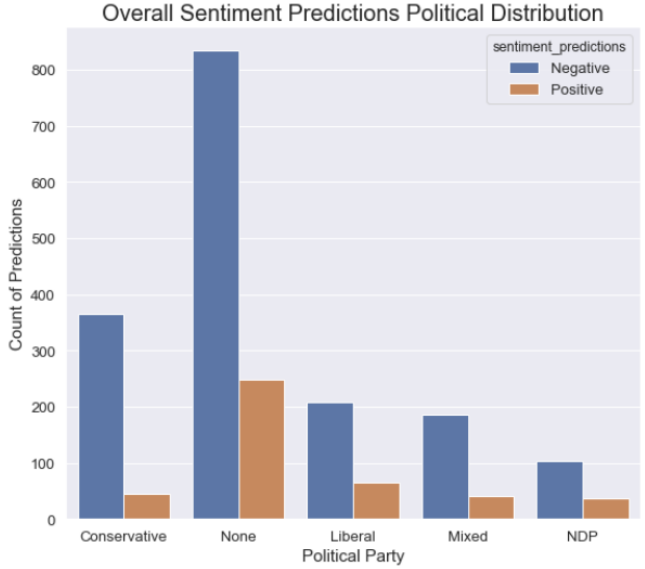
Inverse Document Frequency Data Preparation (TF-IDF) –

While computing TF, all terms are considered equally important. However it is known that certain terms, such as “is”, “of”, and “that”, may appear a lot of times but have little importance. Thus we need to weigh down the frequent terms while scale up the rare ones, by computing IDF, an inverse document frequency factor is incorporated which diminishes the weight of terms that occur very frequently in the document set and increases the weight of terms that occur rarely. IDF is the inverse of the document frequency which measures the informativeness of term t.

Section 3. Results and Visualizations







Analyzing the accuracy results of all the tuned models

TF = Best Accuracy was of Logistic Regression of 73.368%
TFIDF = Best Accuracy was of Logistic Regression of 73.327%

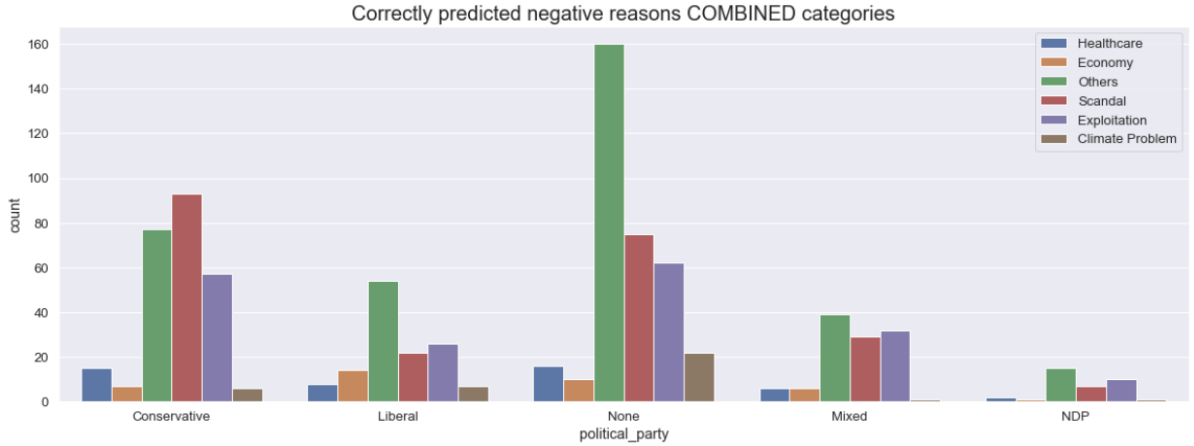
Achieved Tuned Accuracy Sentiments.csv = **73.368% (Logistic Regression on TF)**
Achieved Tuned Accuracy Election Dataset = **50.352% (Logistic Regression on TF)**

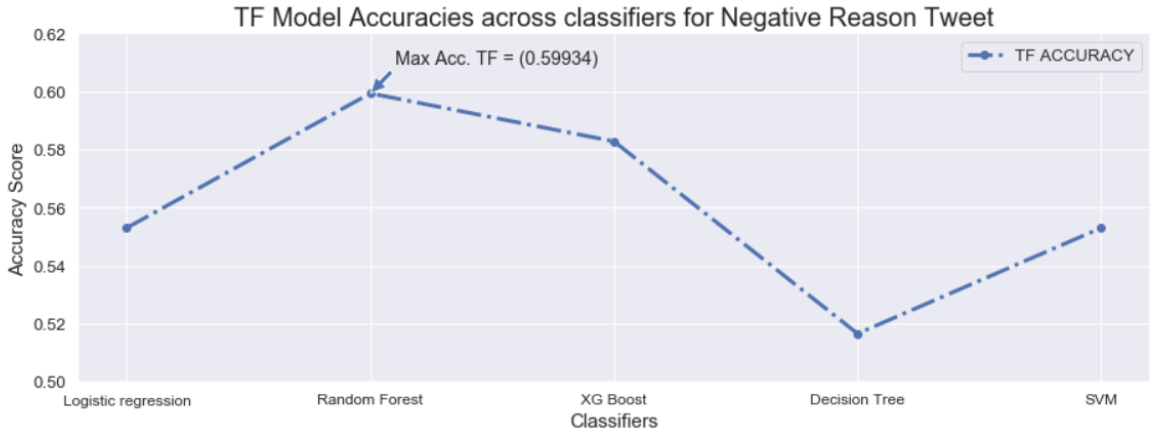
Exploring the predicted sentiments

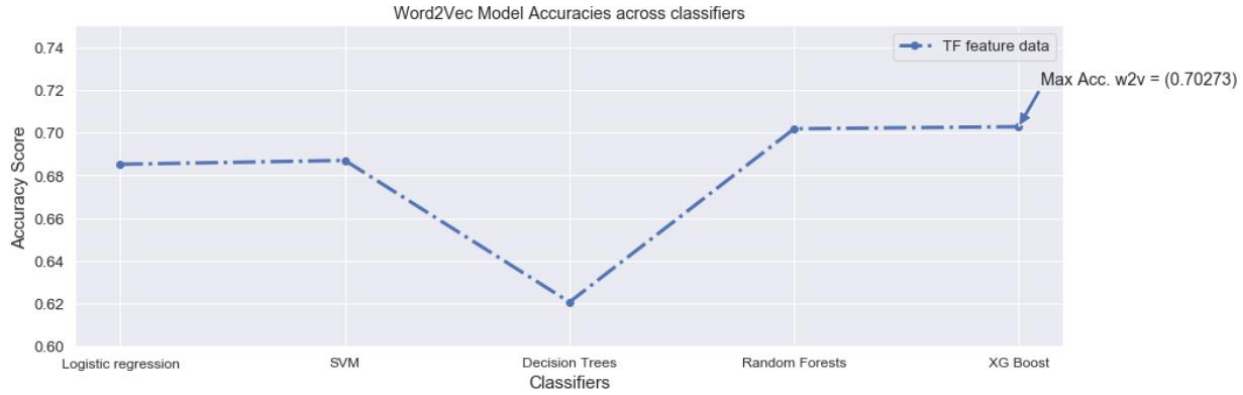
It is evident from the graphs that all three political parties have quite a lot negative sentiments expressed in tweets. However for the Conservative party, both the correct and overall predictions indicate the negative tweets are 5 times the positive tweet. This means that the conservative party has a lot of negative sentiment compared to positive sentiment. Moreover, the Liberal Party, although having many predicted negative sentiments, has a better predicted positive to predicted negative tweets ratio as compared to Conservative Party. Moreover, Liberal Party has the maximum number of predicted positive tweets amongst all three political parties. Additionally, NDP has the best (predicted) positive to negative tweets ratio amongst all 3 political parties. However, the overall tweets of NDP itself is very low in the dataset to produce any measurable impact on the precisions of election winner.

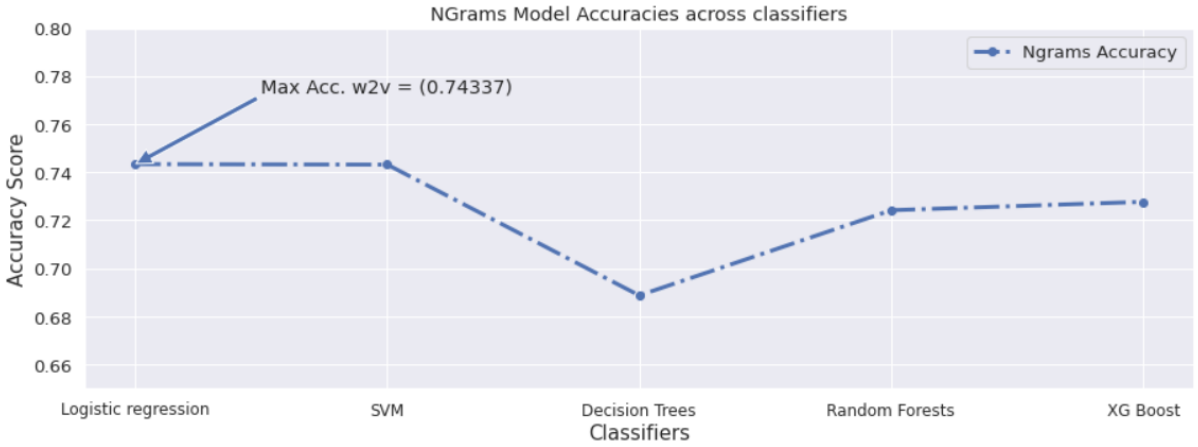
MODEL PREDICTED RESULT:
The Liberal Party should win according to the predicted model as discussed in the section above.

ACTUAL ELECTION RESULT:
The Liberal Party won the elections.
This shows that tweet NLP analytics can be a very useful tool to help us predict that the outcome of the elections and represent the general opinion.









↑ --------- RESEARCH QUESTION: --------- **Negative Reason Classification Max accuracy reached = **59.934% (Random Forest)**
Based on Model 1, What can public opinion on Twitter tell us about the Canadian political landscape in 2019?

--------- ANSWER --------- ** full answer in the notebook

Through NLP analysis of the tweets, we found that the Conservative party has a very high negative sentiment presence with respect to positive. Moreover, Liberals has a lower negative tweet count than the Conservatives, but they also had the maximum number of positive tweets out of all three parties. NDP had good negative to positive tweet ratio, but their overall presence over twitter was quite low suggesting low popularity amongst the public. Also, from the negative reason graph, it is evident that the major negativity around Conservatives is of 'Scandal' and 'Tell Lies', which aren't so prominent for the other parties.

↑ --------- BONUS: ---------
Based on Model 1, trying the classification models with word2vec and ngrams

--------- ANSWER --------- * full answer in the notebook

Max Accuracy **word2vec** 70.273% (XG Boost) < TF 73.365% (Logistic Reg)

Max Accuracy **Ngram** 74.337% (Logistic Reg.) > TF 73.365% (Logistic Reg)
The Accuracy obtained with ngram model with Logistic Regression is **49.742%** for election data set sentiment predictions