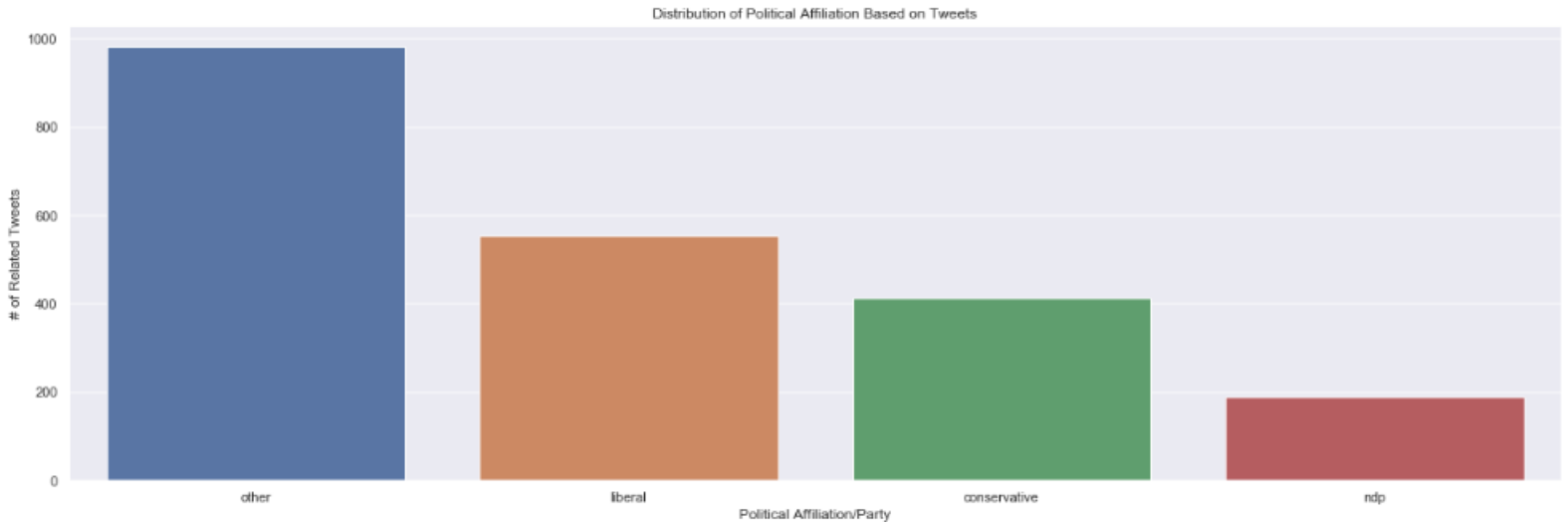
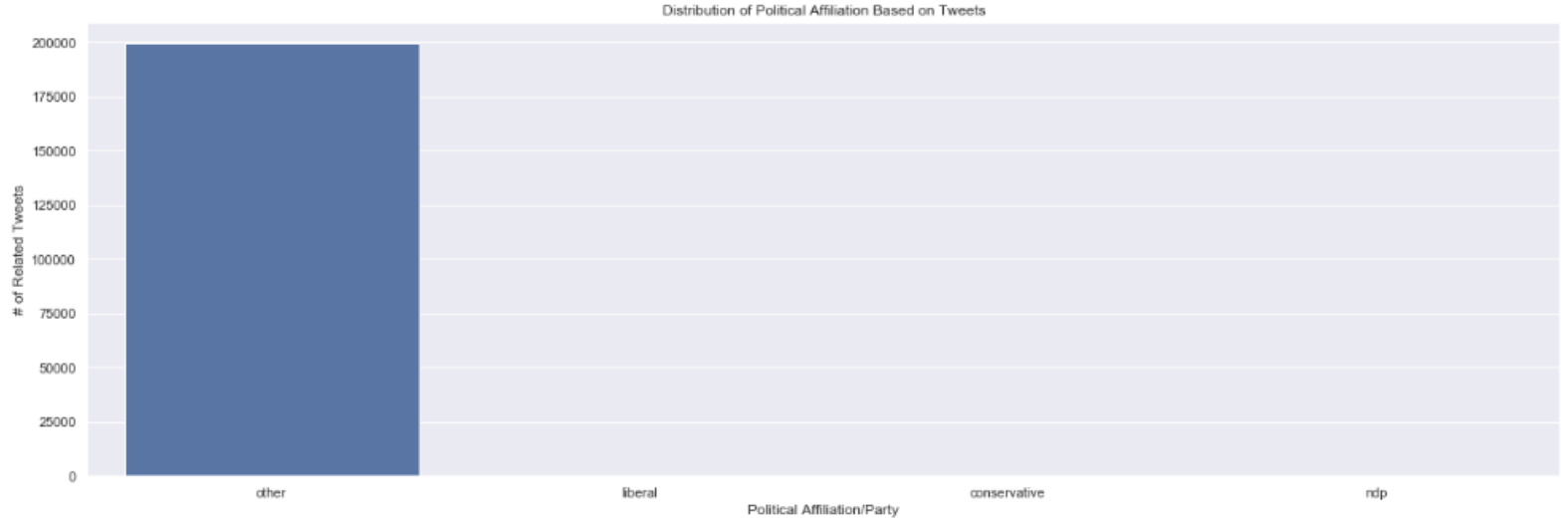
Exploratory Data Analysis – Sentiment Prediction

Political Affiliation/Party Distribution for Election 2019 Tweets



This figure represents the political affiliation/party distribution of election 2019 tweets. The class, ‘other’ has the largest distribution followed by Liberal, conservative and NDP. For election 2019, Liberals gained the most seats, followed by conservatives, then the NDP which aligns with the figure.

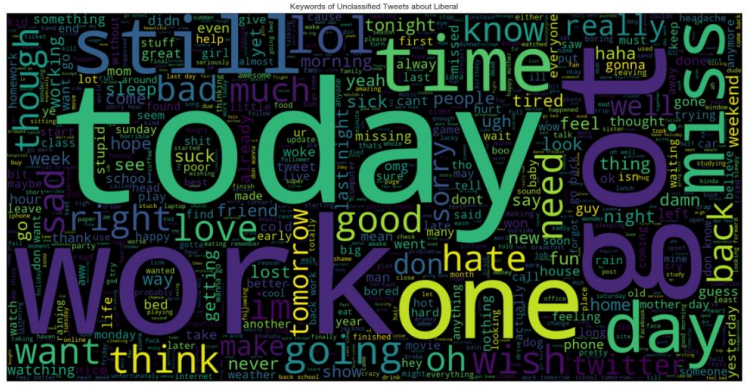
Political Affiliation/Party Distribution for Generic Tweets



This figure represents the political affiliation/party distribution of generic data. Almost all data were classified as the "Other", and very few were classified as either Liberal, Conservative and NDP. This is mainly because the tweets have been extracted in the 2009s which mean our keyword such as slogans and leader names aren't applicable as the classification methods.

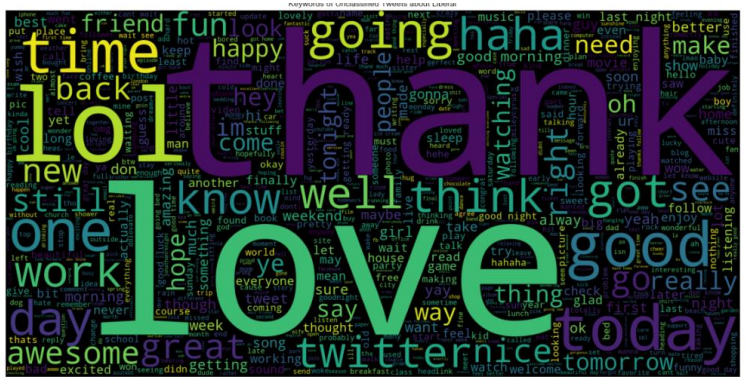
Exploratory Data Analysis – Sentiment Prediction

Word Cloud of Negative Tweets



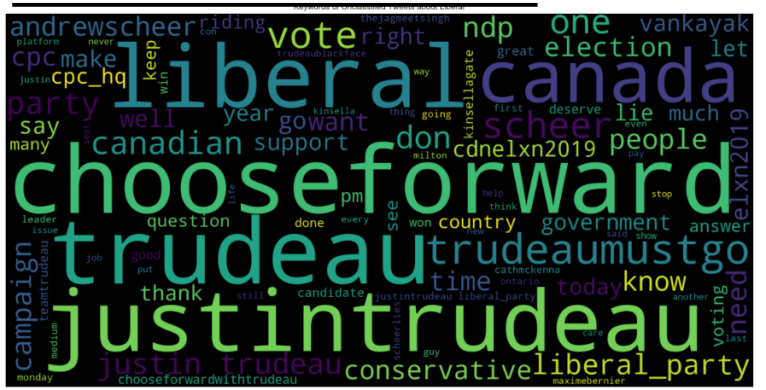
In the generic tweets that were classified as negative, words with high frequency seems to have neutral sentiment values. Words like hate, suck, and other negative words can be found in the word cloud.

Word Cloud of Positive Tweets



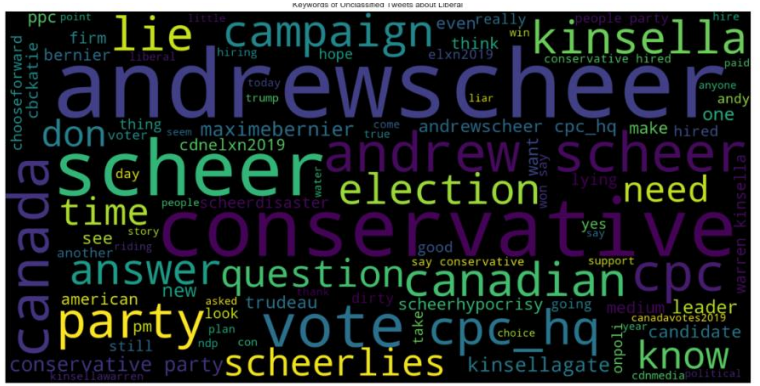
In the generic tweets that were classified as positive, words with high frequency seems to have positive sentiment values. Large words like "thank", "love", "lol", etc. are words with positive sentiments.

Word Cloud of Liberals Tweets



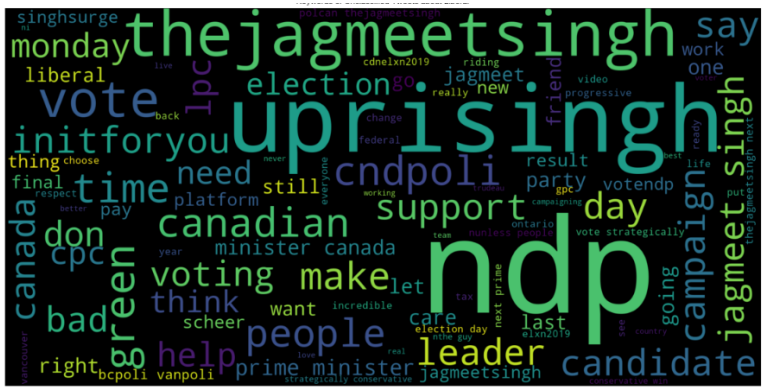
We can see keywords like chooseforward, trudeau, liberal, justintrudeau, treaumustgo, etc. which are all related to the liberal party in 2019 election

Word Cloud of Conservatives Tweets



We can see keywords like andrew scheer, conservative party, conservative, etc. which are all related to the conservative party in 2019 election

Word Cloud of NDP Tweets



We can see keywords like ndp, initforyou, thejagmeetsingh, uprising, etc which are all related to ndp party in 2019 election

Model Implementation & Results – Sentiment Prediction

Model Implementation Steps

Data → Feature Engineering → Data Modeling

Feature Engineering
- Bag of Words
- TF-IDF
- TF-IDF + 2-Gram

Data Modeling
- Logistic Regression
- Naïve Bayes
- SVM
- XGBoost
- KNN
- Decision Tree

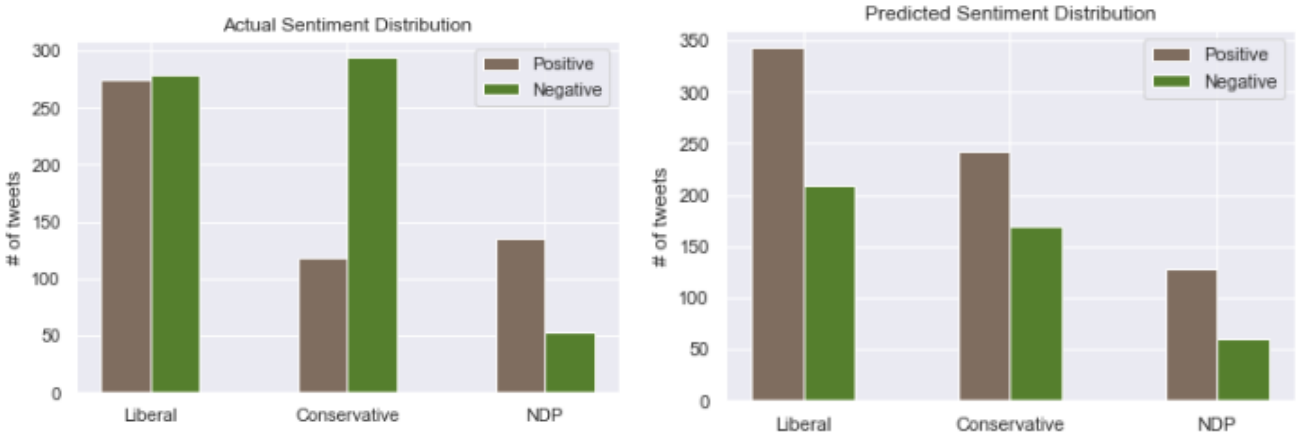
To achieve the highest accuracy, different combinations of classifiers and the feature extractions are performed.

Results

After training the generic tweet data using different feature extraction methods and classifiers, the best test accuracy was computed to be 76.07% when trained using Logistic Regression on dataset featurized using TF-IDF + 2-Gram.

When predicting on election data using the same model, the accuracy was computed to be 63.1%

Discussion



Findings

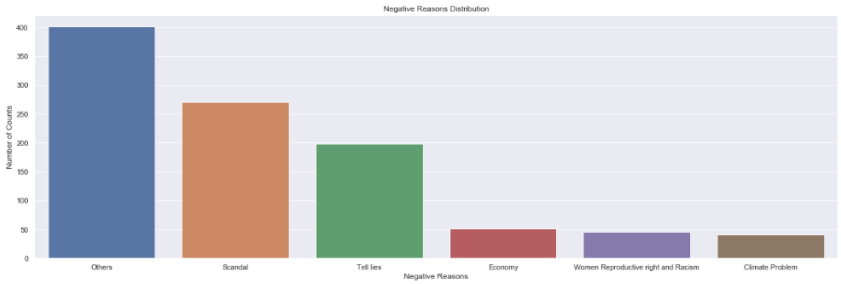
- The model does not perform well on election data compared to the generic tweets data.
- The model is biased towards “Positive” and does not do well on predicting “negative” sentiments. This can be seen in the distribution graph above.
- Even if our model predicted well, sentiment does not show strong relationship between the result of the election. For example, despite NDP having more positive tweets compared to the conservative, the number of seats they have won are significantly less. Sentiment values can give us useful insight about the election, but cannot help us make conclusions about which parties are going to win.
- The size of the network/tweets are proportional to the number of seats won.

Next Steps:

- Apply complex feature extraction methods like word embedding and use complex models like BERT to increase the accuracy.

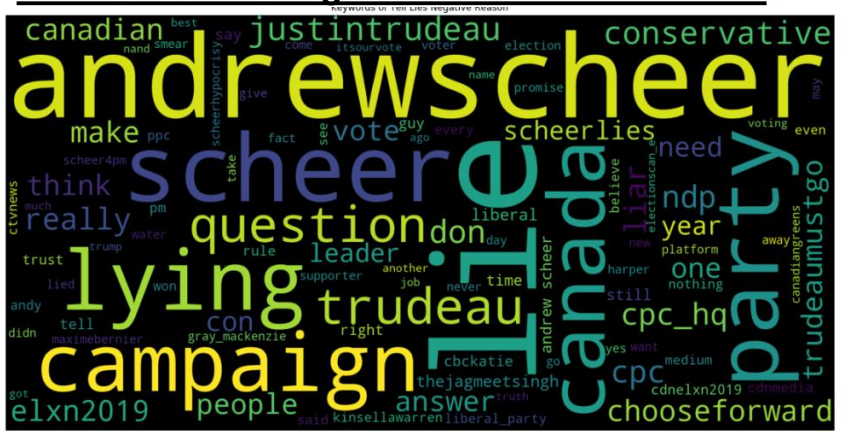
Exploratory Data Analysis – Negative Reasons

Distribution of Negative Reasons



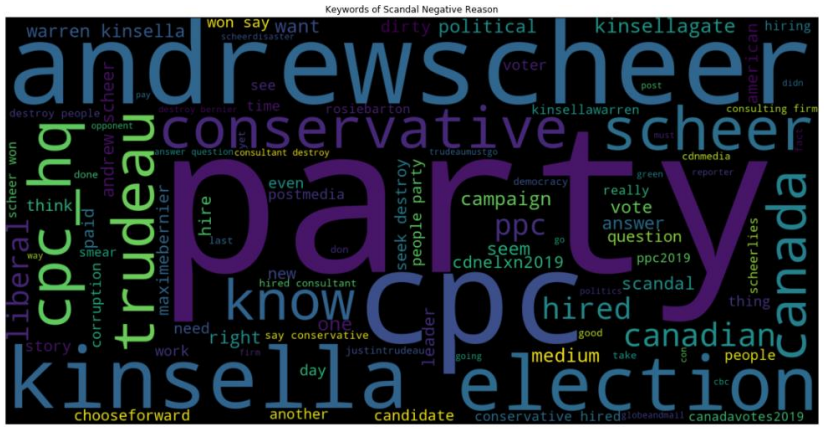
Despite eliminating few classes, the dataset is skewed meaning it is imbalanced. Also, the dataset is significantly smaller compared to the dataset used for sentiment prediction.

Word Cloud of Negative Reasons – Tell Lies



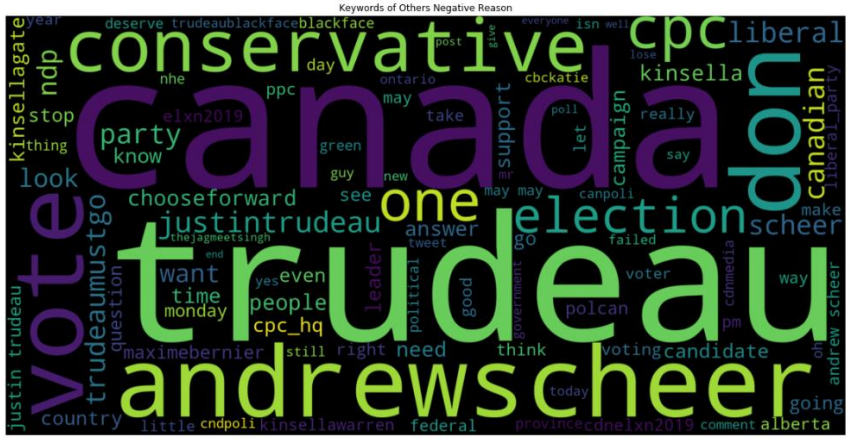
Again, the keywords in the word cloud are mostly associated with the conservative parties. Also, we can see words like lie which will help us with data modeling.

Word Cloud of Negative Reasons - Scandal



The major keywords in the word cloud are related to conservative party. Also, we can see words like 'scandal' which should help in our data modelling.

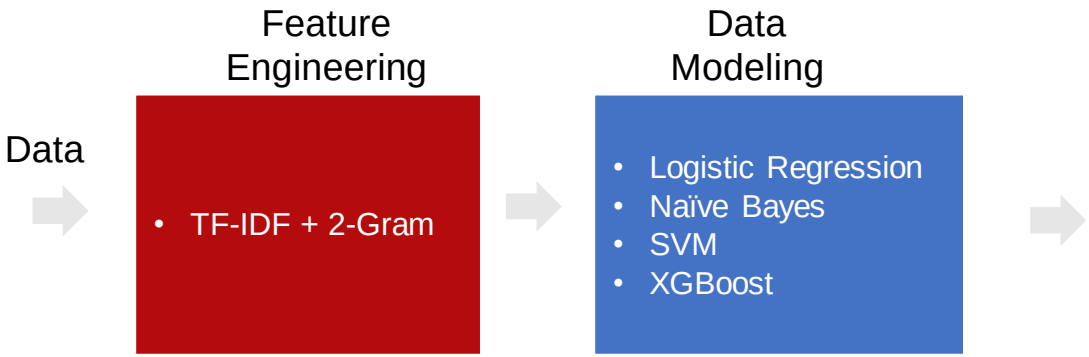
Word Cloud of Negative Reasons – Others



The keywords in the word cloud are associated with both liberals and conservative. The class 'other' doesn't seem to have words that can represent the class.

Model Implementation & Results – Negative Reasons

Model Implementation Steps

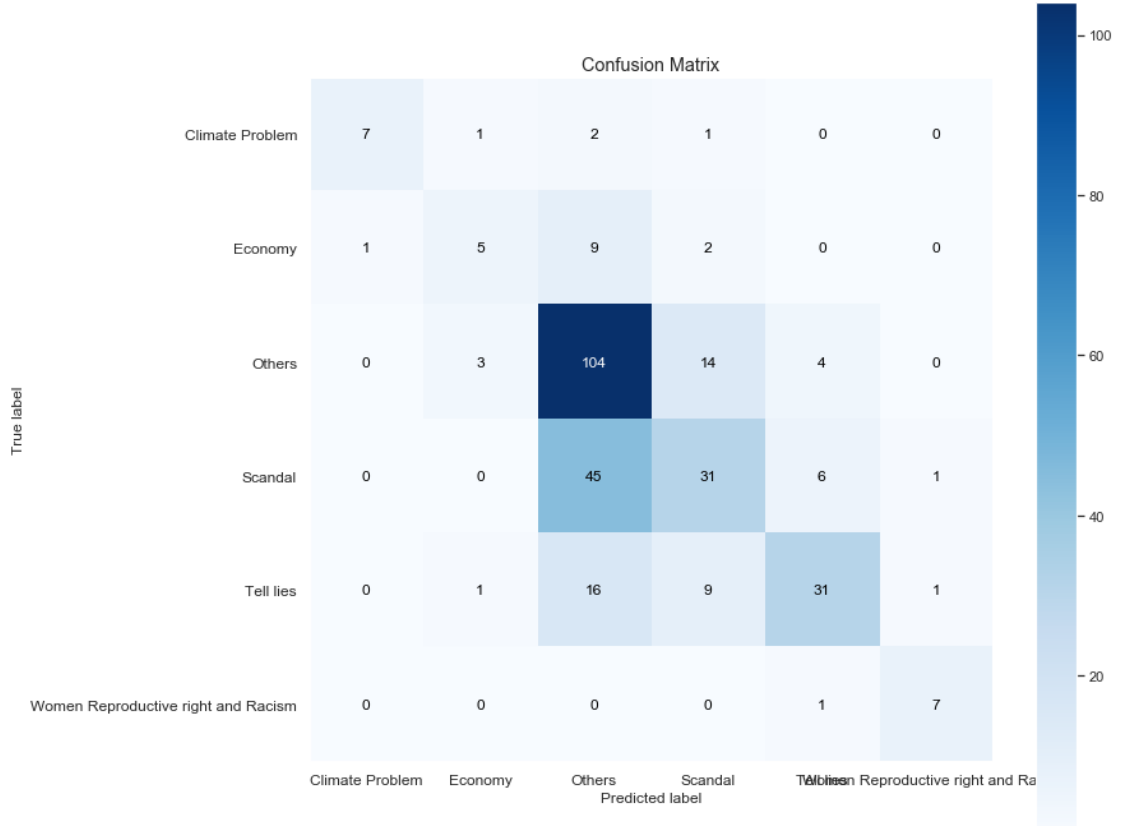


To achieve the highest accuracy, different classifiers are used for modeling the data.

Results

After training the generic tweet data using different classifiers, the best test accuracy was computed to be 64.24% when trained using XGBoost. Considering that this was a multiclass classification with 5 classes, 64.24% is pretty good.

Discussion



Findings

- Due to imbalance in the dataset, the model seems to be overfitted towards the class ‘other’ as seen in the confusion matrix above.
- Does not have enough data for the model to correctly distinguish different classes which causes the model to classify to the class with highest count.

Improvements:

- Train with a larger dataset with uniformly distributed class