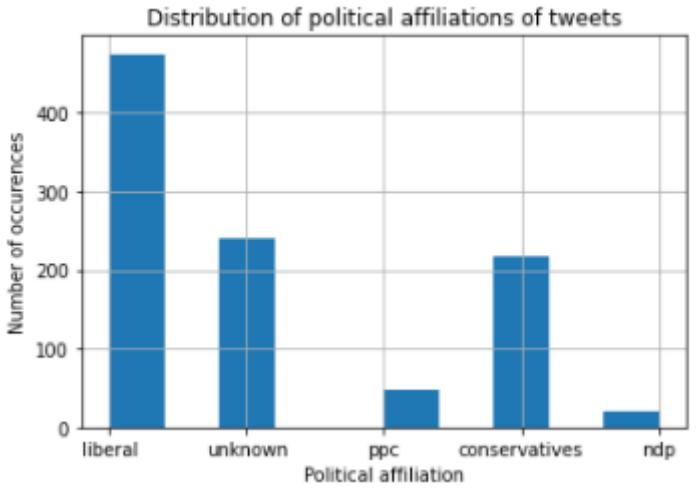
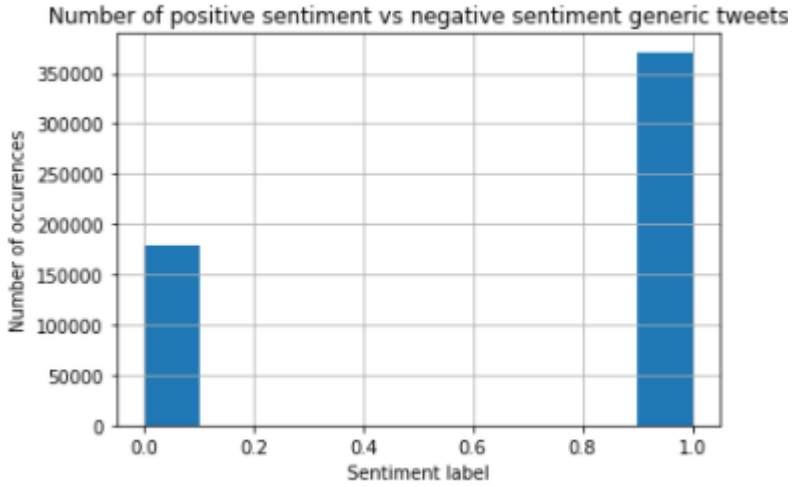
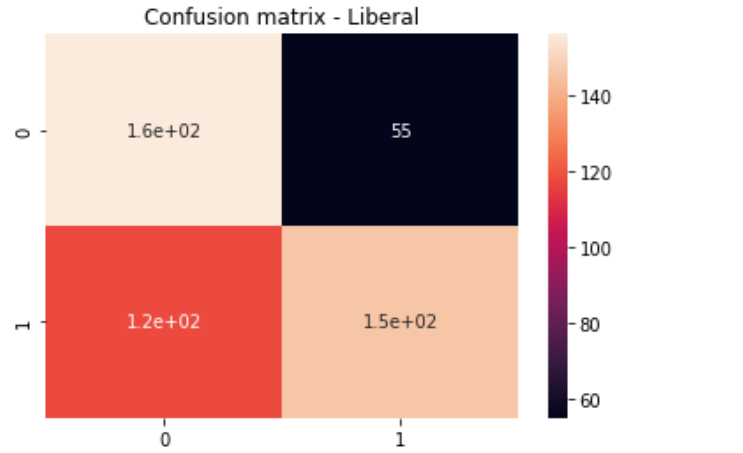
Exploratory Analysis



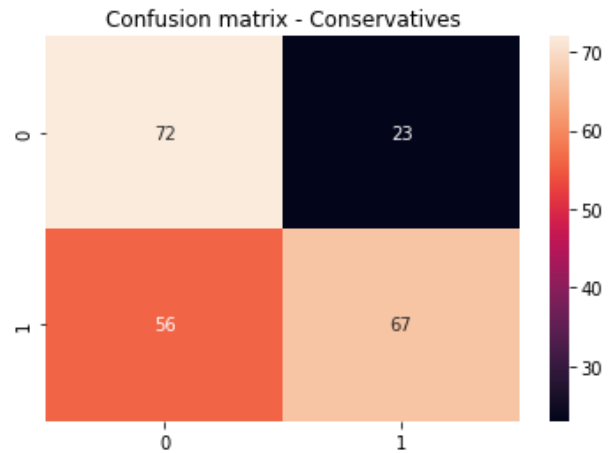
Most tweets are about the Liberal party and NDP has the least



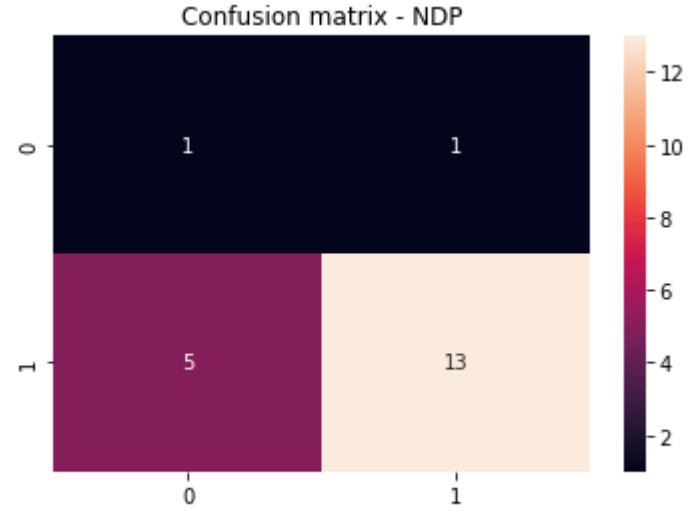
Most tweets (more than double) in the sentiment analysis are positive comments



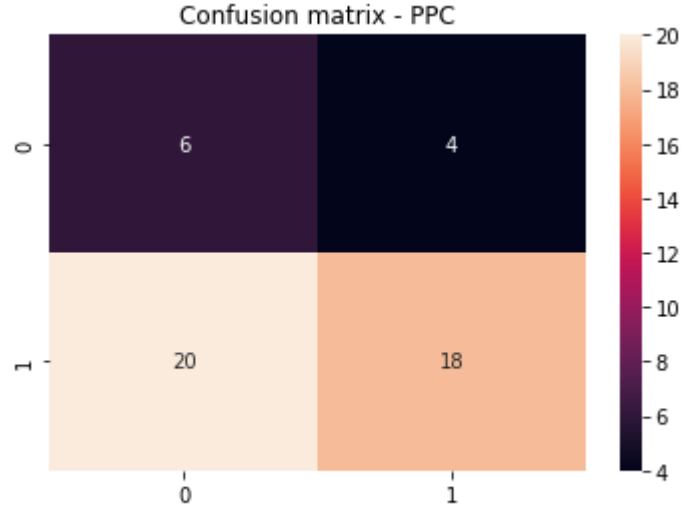
This model got an accuracy of 63.58% for liberal party



This model got an accuracy of 63.76% for conservative party



This model got an accuracy of 70.0% for ndp



This model got an accuracy of 50.0% for PPC

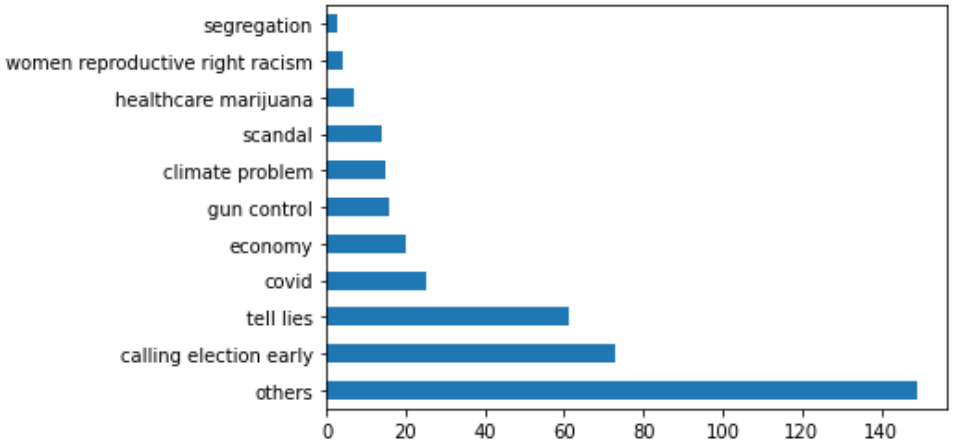
Most tweets are about the Liberal party and NDP has the least number of tweets. Liberal and Conservatives have many negative tweets about them, likely due to controversial nature of the elections.
Using Random Forest – TF-IDF model, we predicted the sentiment of the election tweets, our model predicts more ‘False Positives’.

Model Selection and Feature Importance

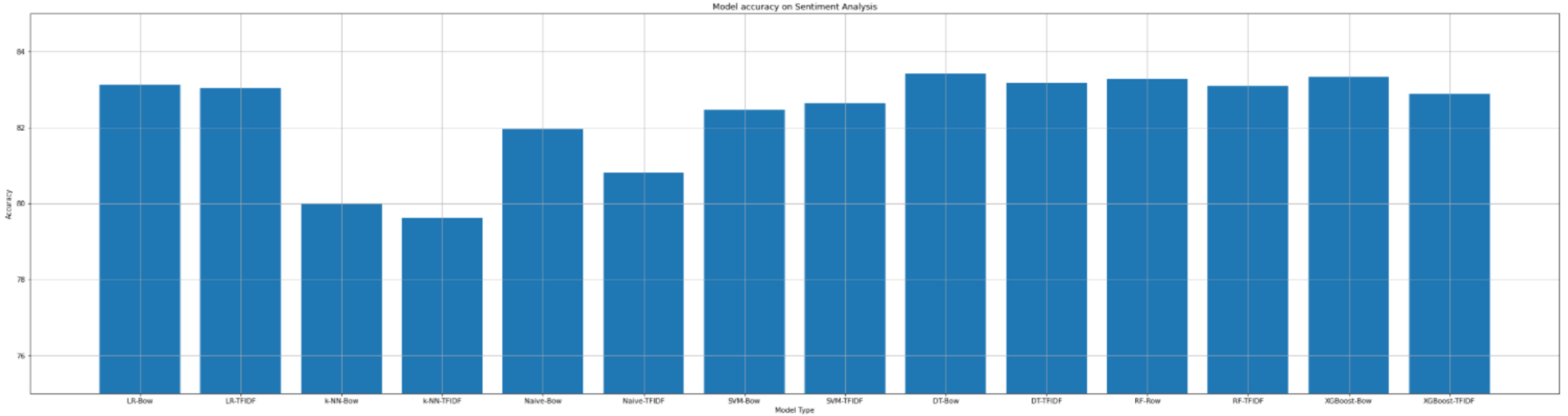
```
Index(['im', 'dont', 'like', 'cant', 'get', 'amp', 'people', 'bad', 'know',
       'fuck', 'one', 'trump', 'fucking', 'u', 'even', 'never', 'hate', 'got',
       'really', 'still', 'youre', 'shit', 'time', 'want', 'stop', 'look',
       'go', 'didnt', 'dead', 'see', 'ass', 'make', 'wont', 'man', 'doesnt',
       'new', 'us', 'much', 'day', 'need', 'killed', 'says', 'life', '2',
       'think', 'death', 'take', 'going', 'right', 'back'],
      dtype='object')
```

```
Index(['love', 'birthday', 'happy', 'great', 'amazing', 'good', 'best', 'amp',
       'day', 'thank', 'one', 'beautiful', 'im', 'thanks', 'new', 'much',
       'today', 'fun', 'hope', 'see', 'get', 'like', 'u', 'us', 'time', 'win',
       'make', 'wait', 'cant', 'look', 'night', 'awesome', 'ever', 'morning',
       'never', 'life', 'nice', 'always', 'proud', 'back', 'enjoy', 'dont',
       'way', 'perfect', 'miss', 'got', 'excited', 'real',
       'people'],
      dtype='object')
```

Here we can see the top 50 negative (left) and positive (right) words respectively. By adding more suitable stop words, we can change the decision and replace some words so the model can focus on words that have more sentimental meaning. We could also see some overlap between the positive and negative sentiments, which makes it very difficult for the model to differentiate.



We also differentiated and predicted the negative reasons based on the negative election tweets. The model had about a 53% accuracy on predicting the correct negative reason. The low accuracy may be due to the skewed amount of data, we can see that the majority of the negative tweet are due to three reasons.



Model Accuracy on Sentiment Analysis Dataset: The Decision Tree – Bag of Words provided the highest accuracy (83.42%). However, applying the model to the election dataset, it only had an accuracy of 62.28% which tells us that the model may be overfitting.