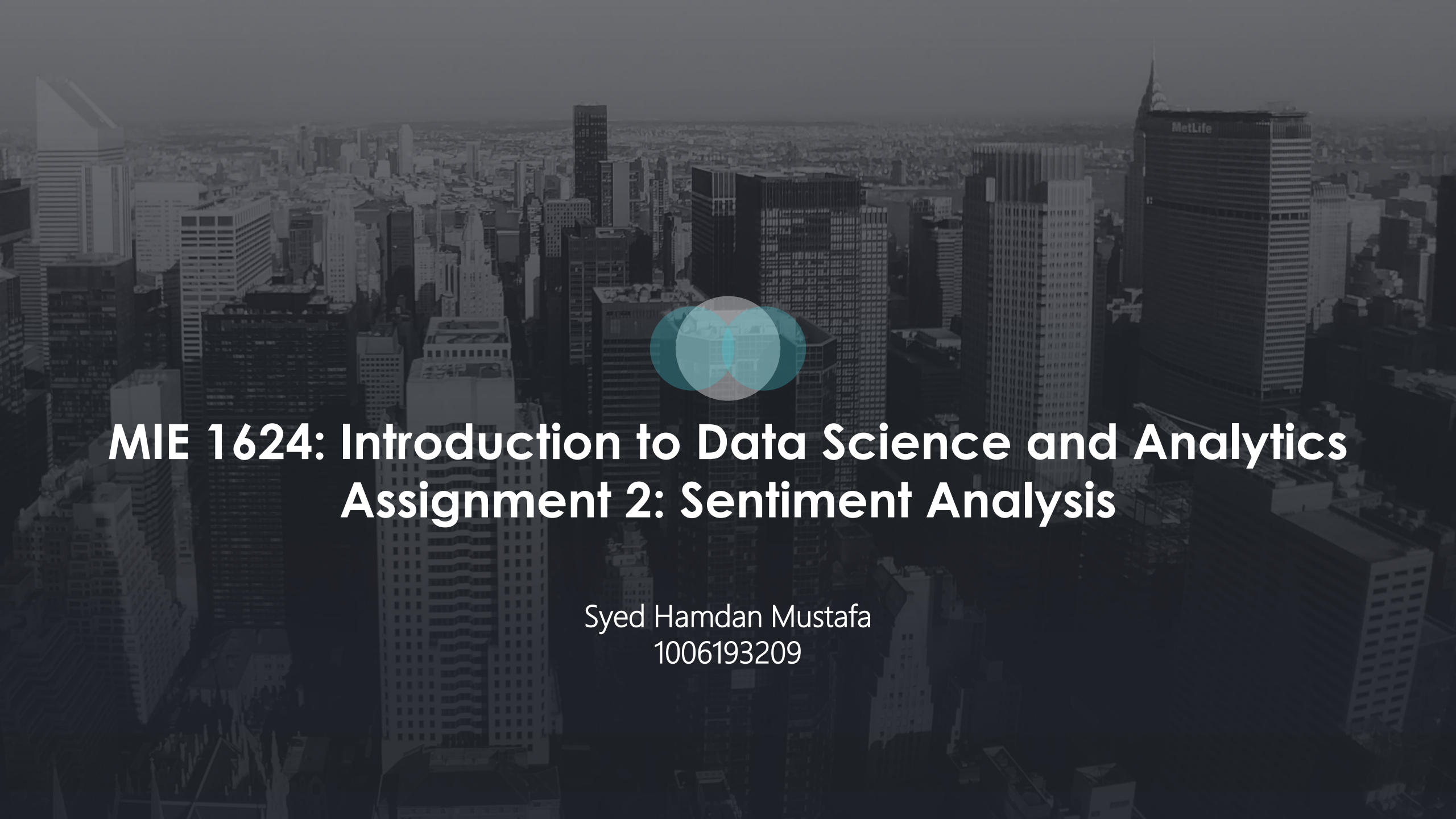
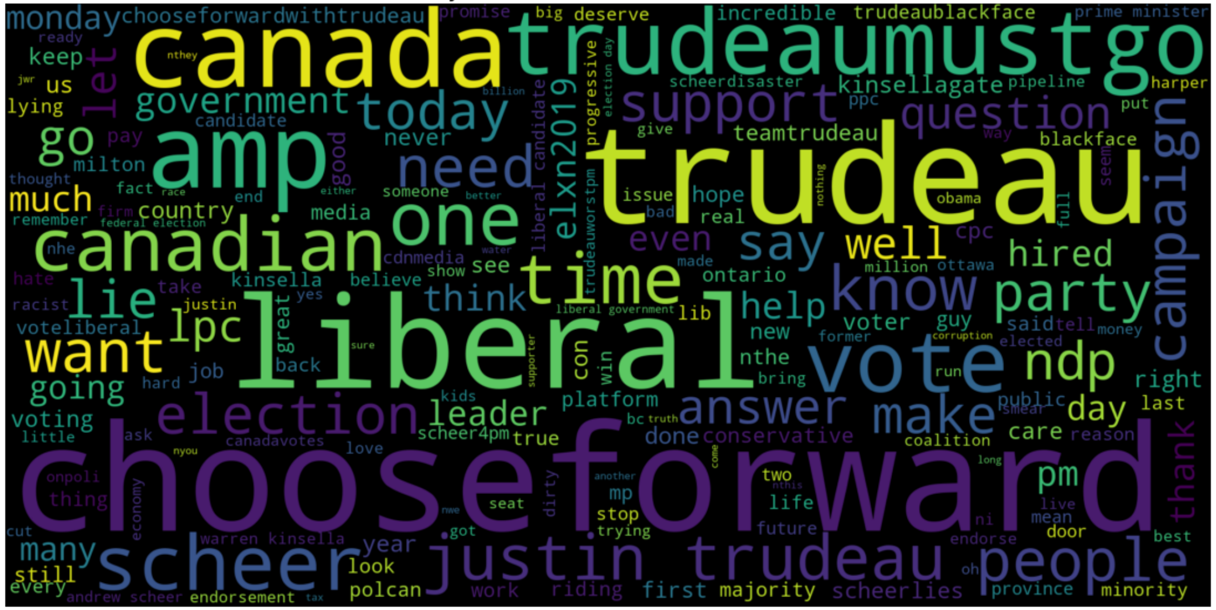


MIE 1624: Introduction to Data Science and Analytics
Assignment 2: Sentiment Analysis

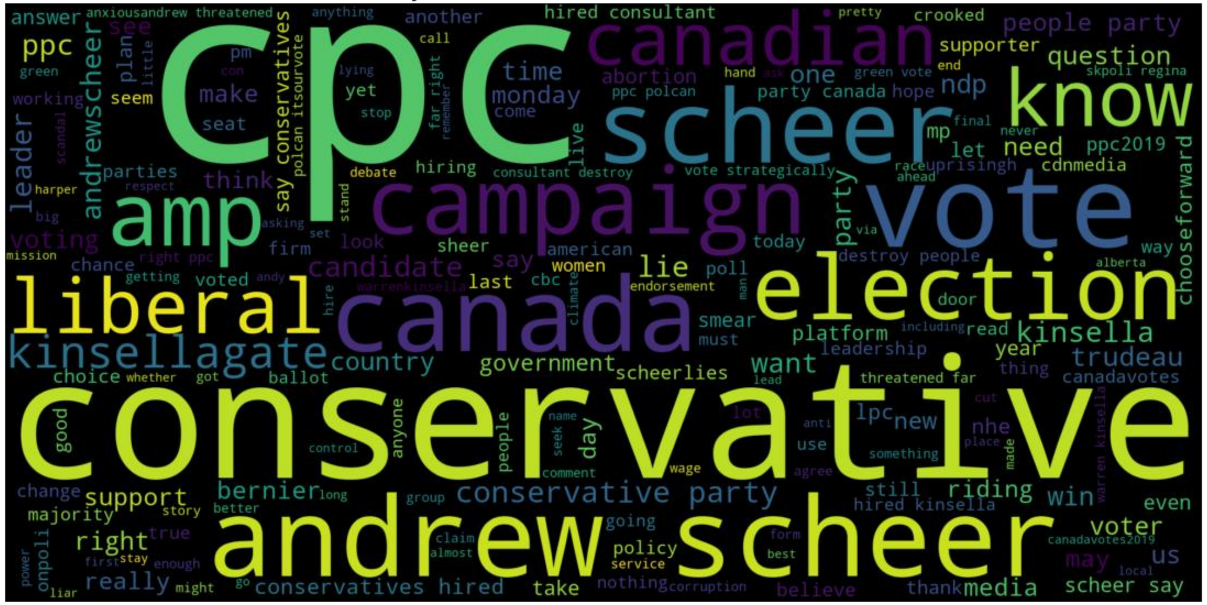
Syed Hamdan Mustafa
1006193209

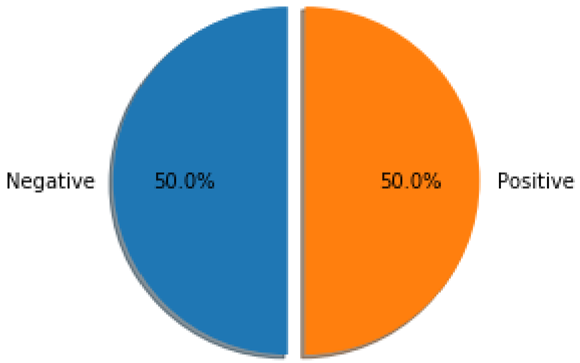
Exploratory Analysis

Keywords of Tweets about Liberal



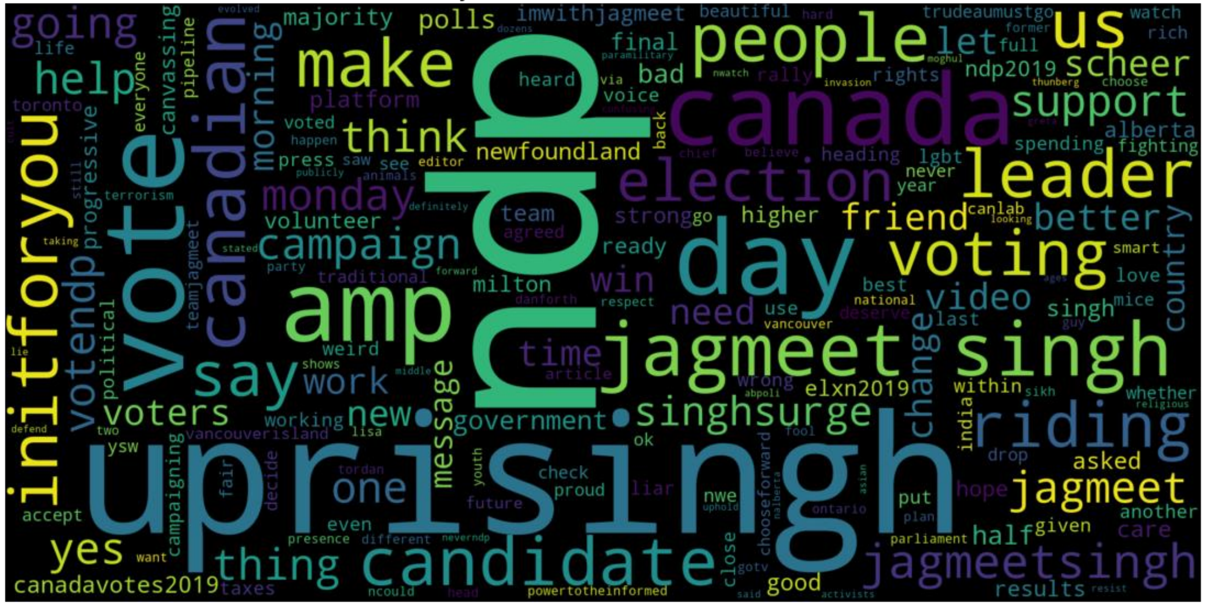
Keywords of Tweets about Conservaties





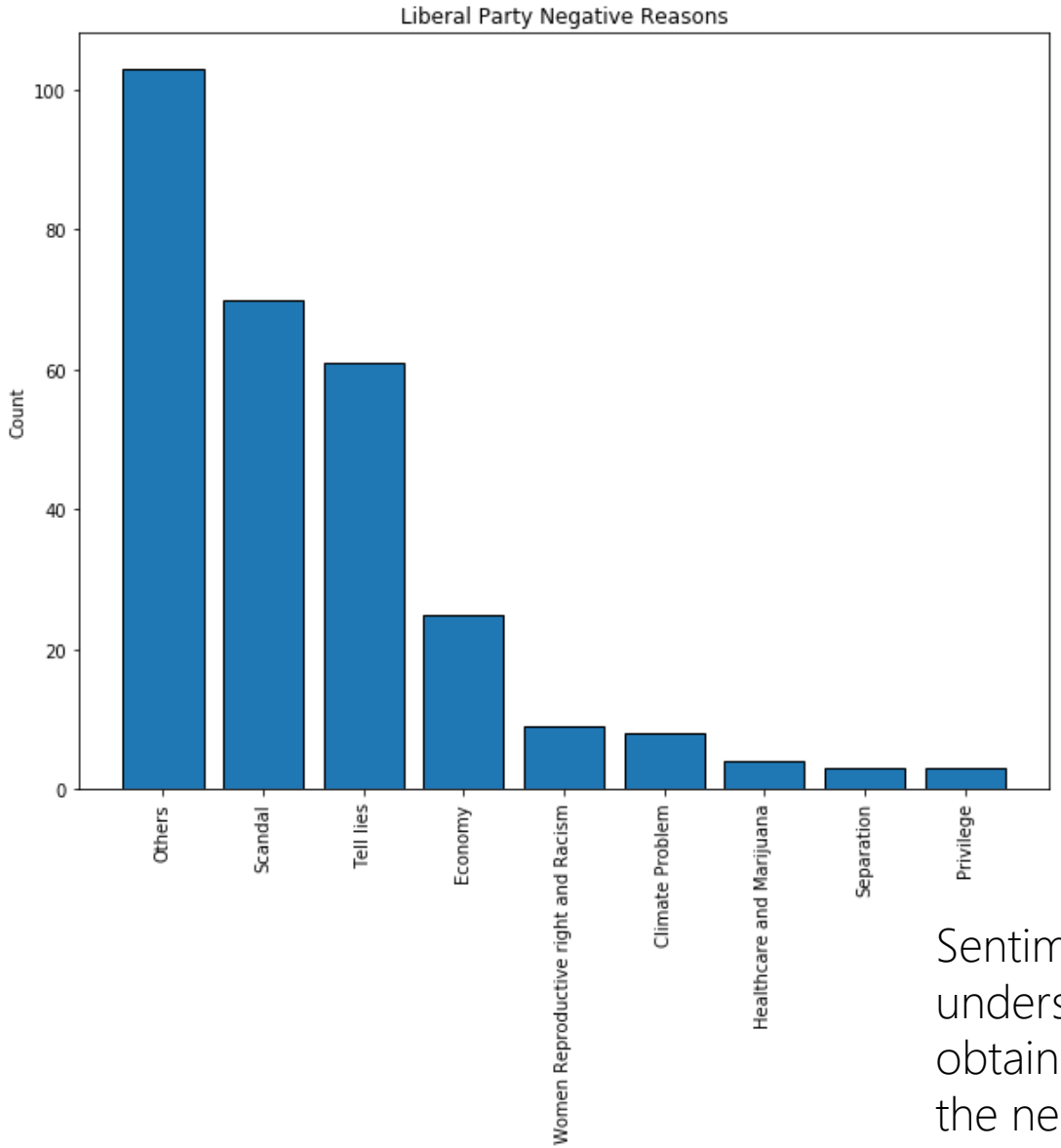
Generic tweets sentiments were separated 50-50

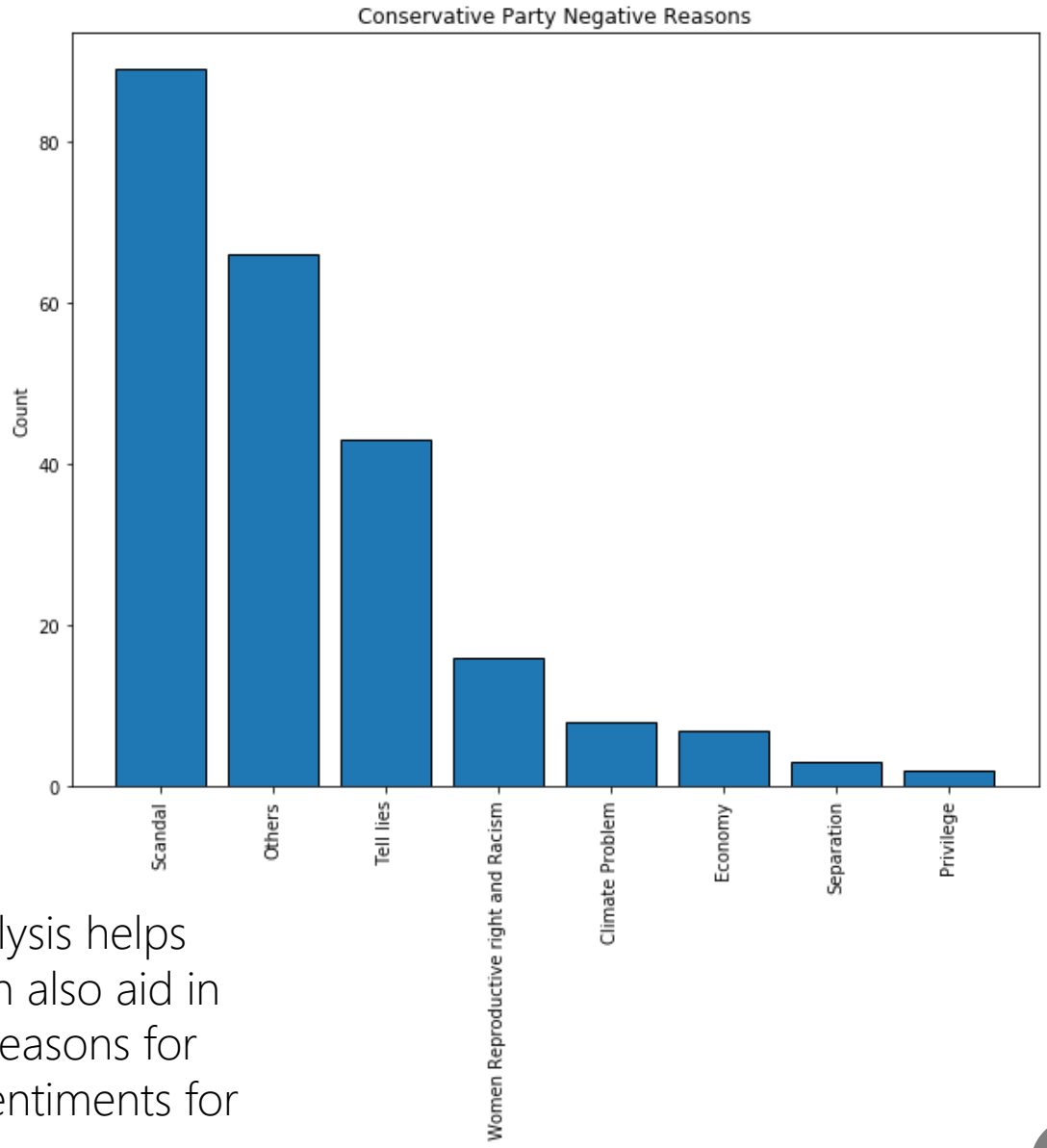
Keywords of Tweets about NDP



Word clouds generated for all the parties to see key words

Exploratory Analysis





Sentiment Analysis helps understand can also aid in obtaining the reasons for the negative sentiments for the parties.

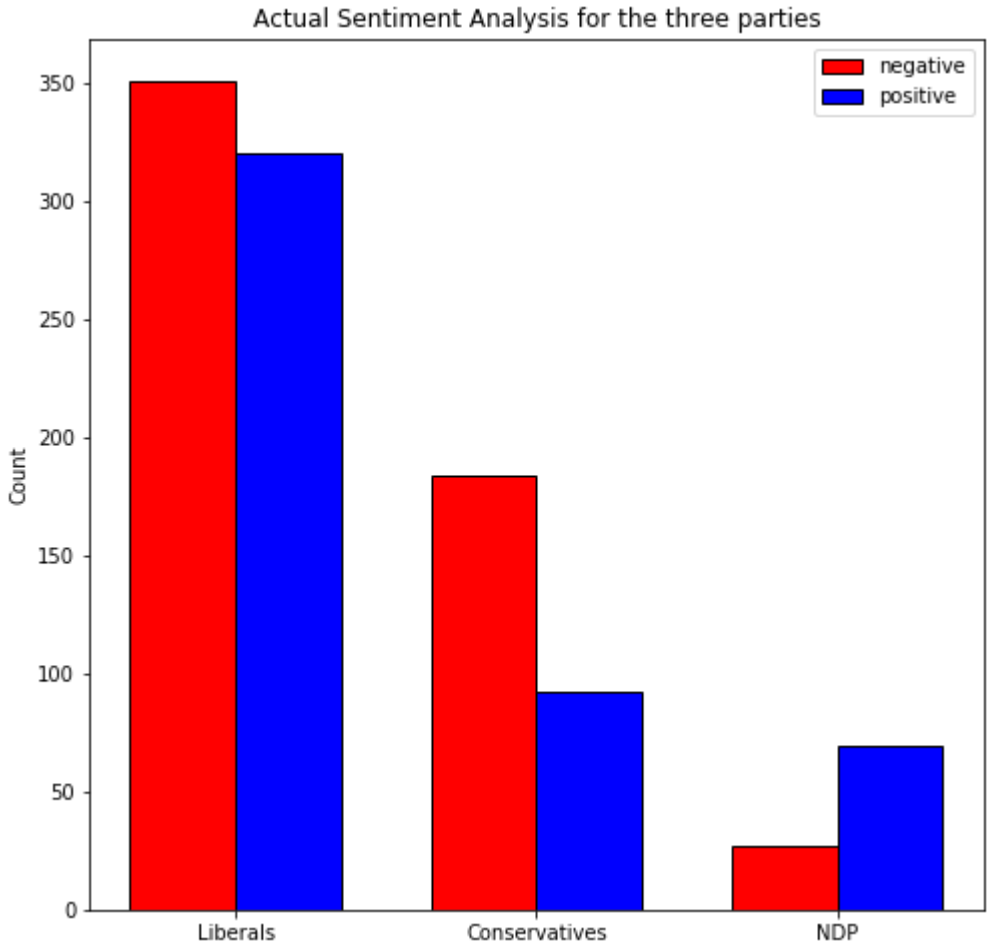
Feature Selection and Model Implementation

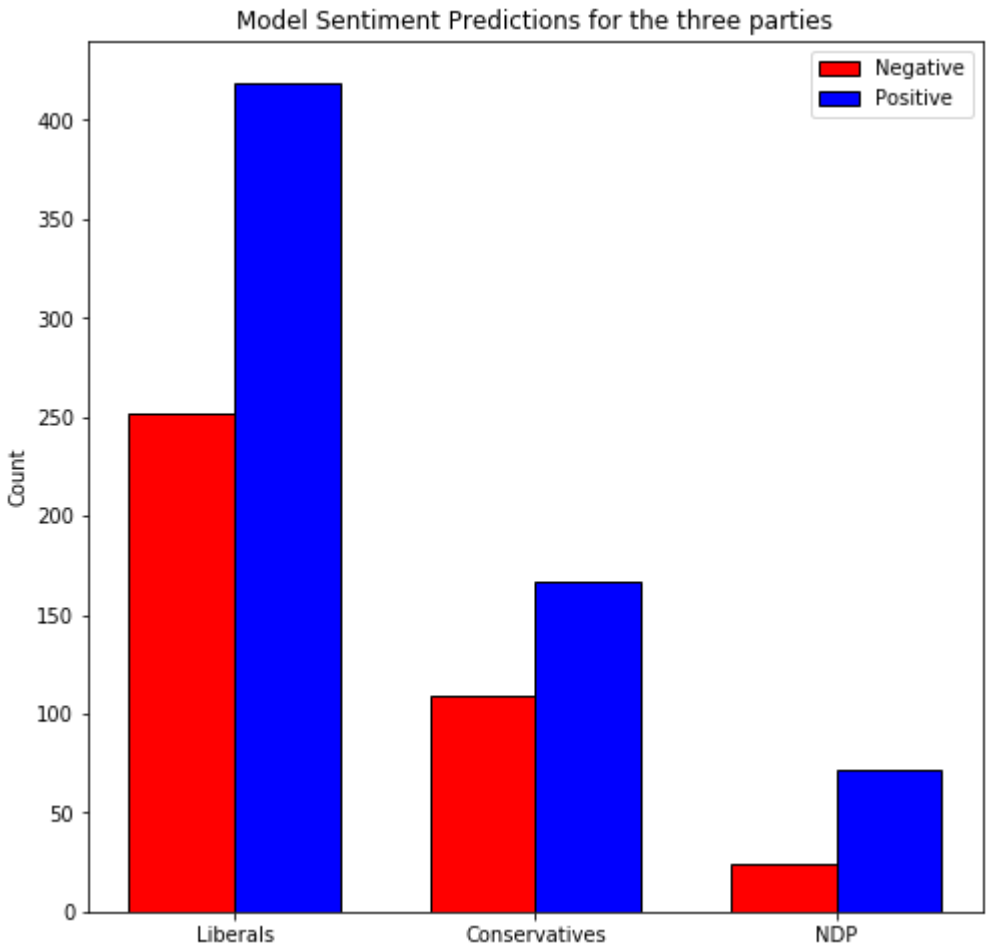
Two feature selection techniques used- Count Vectorizer and TF-IDF.
The models were trained on the Generic Tweets and tested on the same dataset after splitting the data.

Models	Count Vectorizer			TFIDF		
	Test Accuracy	Precision	Recall	Test Accuracy	Precision	Recall
Decision Tree	67.3%	65.7%	73.1%	67.8%	66.1%	73.5%
Random Forest	71.7%	72.5%	70.4%	72.5%	73.4%	71.1%
XG Boost	66.2%	61.3%	88.6%	66.2%	61.3%	88.7%
KNN	67.3%	65.5%	73.6%	63.7%	62.6%	68.8%
SVM	75.4%	73.9%	78.9%	75.4%	74.4%	78.1%
Naive Bayes	74.8%	75.5%	73.6%	74.5%	74.9%	74.1%
Logistic Regression	75.5%	74.3%	78.5%	75.7%	74.8%	78.0%

There were 7 models used- both using Count Vectorizer and TF-IDF.
XG Boost did a commendable job in terms of Recall.
The best results were through Logistic Regression (TFIDF) and this was used on the Elections tweets to predict the sentiments.

Model Implementation on Elections Dataset

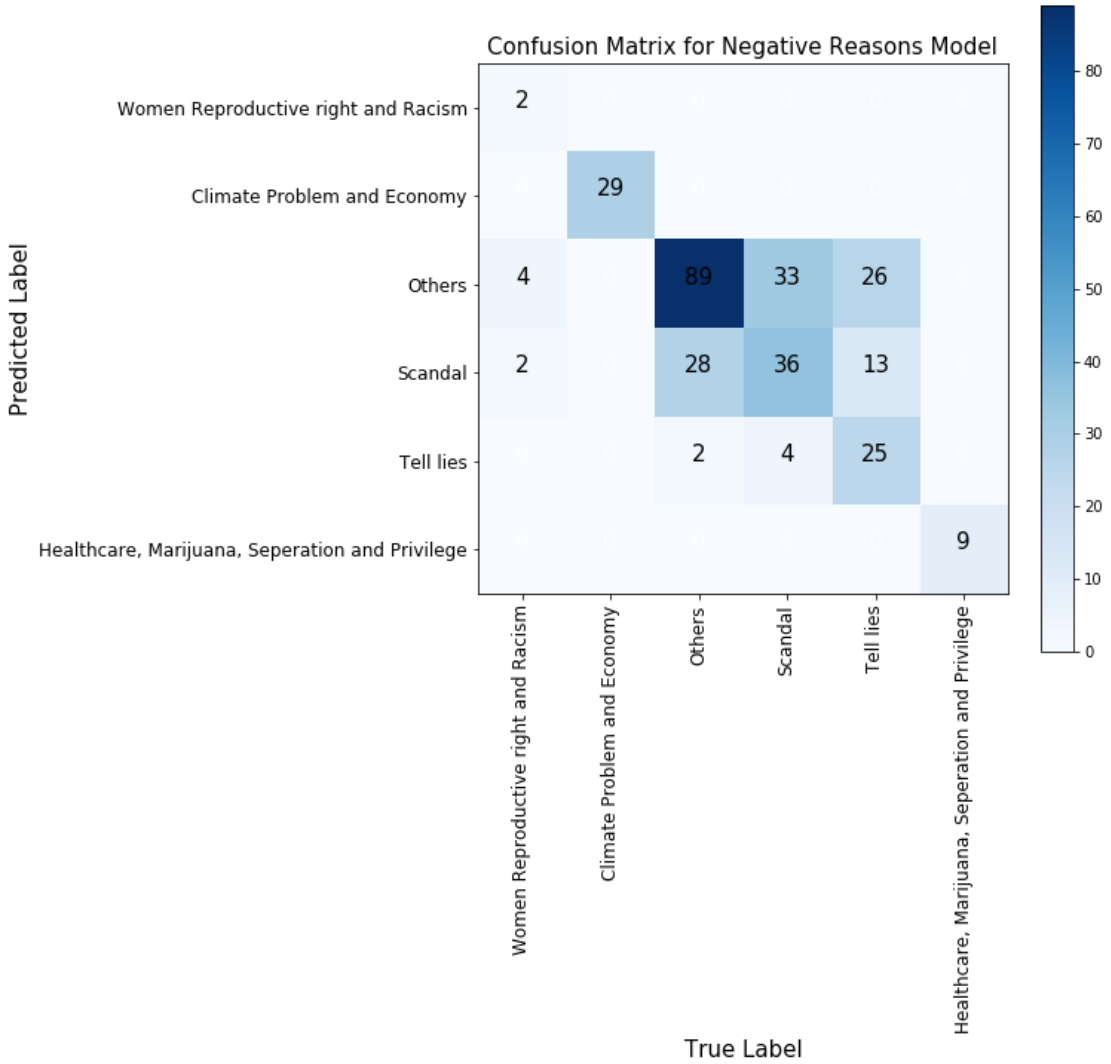




The model accuracy is 60%. The model has failed to classify the negative tweets properly and classified them as positive tweets instead.

Classifying Negative Reasons

There were 4 models used to classify the negative reasons. The initial results showed poor accuracy and overfitting of the model. This was rectified by changing the number of negative reasons category from 10 to 6 and using fewer TFIDF features (200 seemed to work best).



The best results were through Multinomial Logistic Regression with an accuracy of **63%**.

Implementation of Convolutional Neural Network did increase the accuracy up to **66%**.

The model failed to classify properly between Scandal and Tell Lies because they could be containing of similar words. There was not enough data available for the other groups.


THANK YOU!