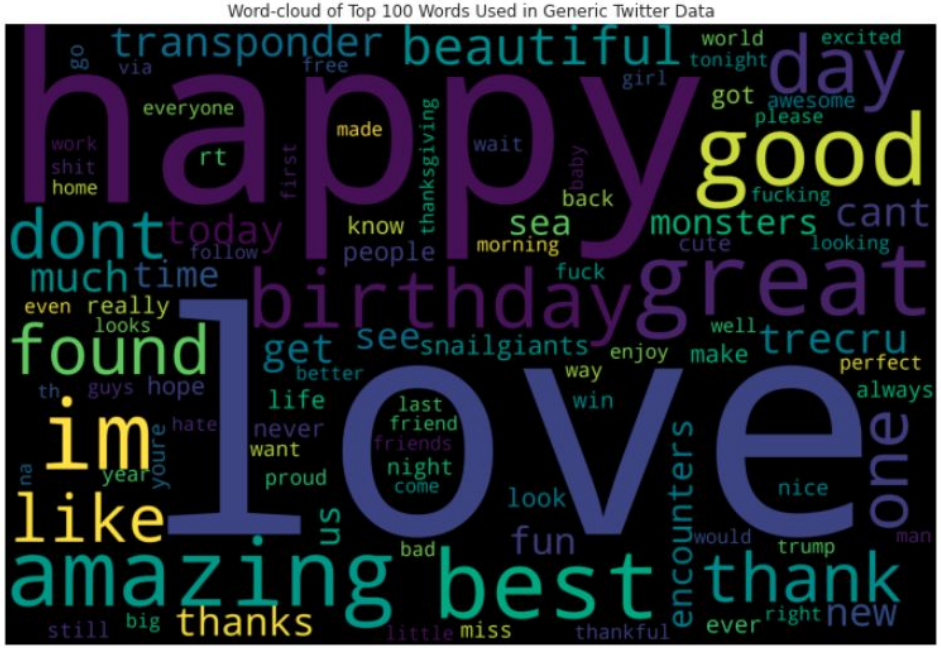
# Assignment 3 - Sentiment Analysis of Canadian Election 2019 data

## MIE1624 - Introduction to Data Science & Analytics

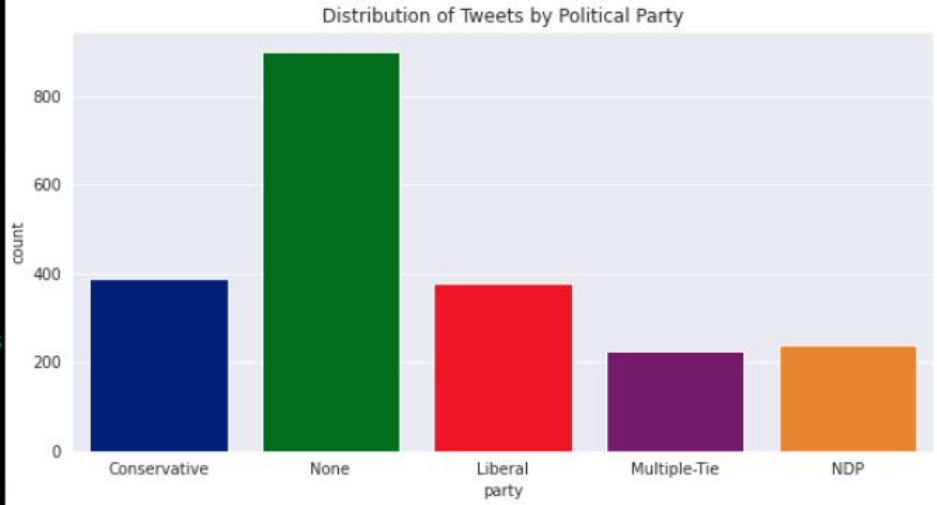
Sugumar Prabhakaran (id: 994126815)

29 Mar 2021

# Exploratory Data Analysis

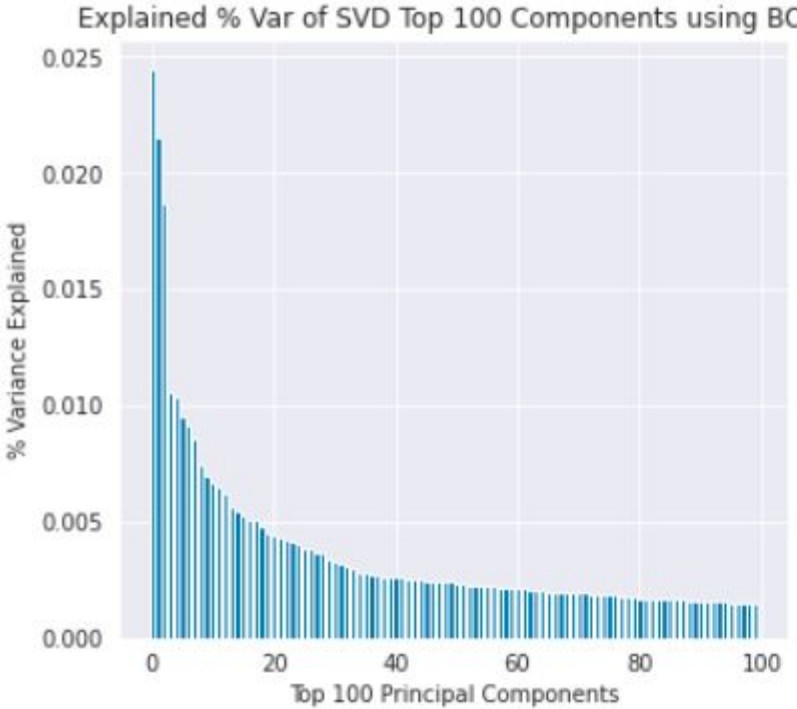


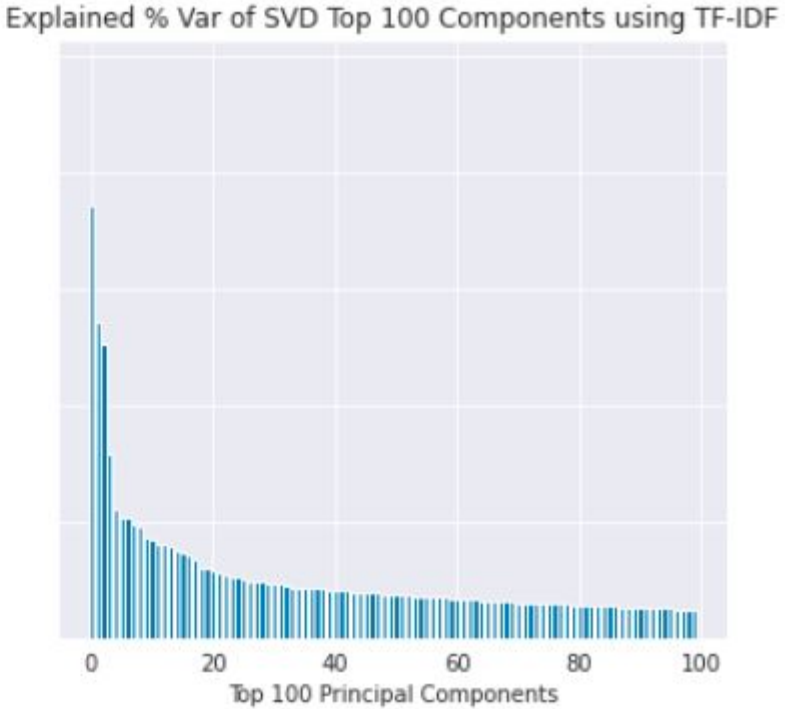

**Word-Cloud**: This image depicts the top 100 words with highest word count in the sentiment analysis data.




**Tweet Party Affiliation**: My procedure to predict party works by counting # of words in a tweet that belong to each party. The tweet belongs to the party with the highest count. The above graph shows the resulting distribution of this procedure.

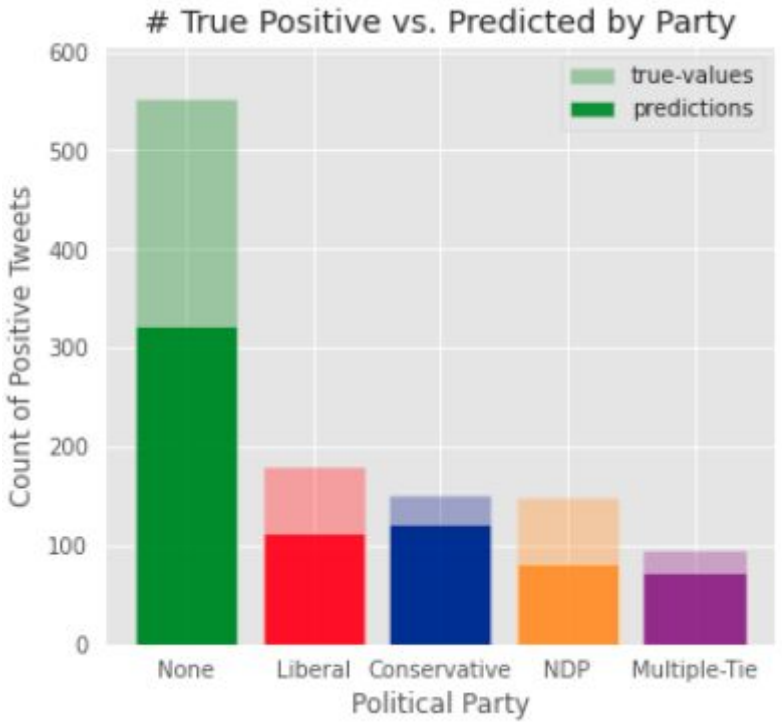
# First Model Feature Importance

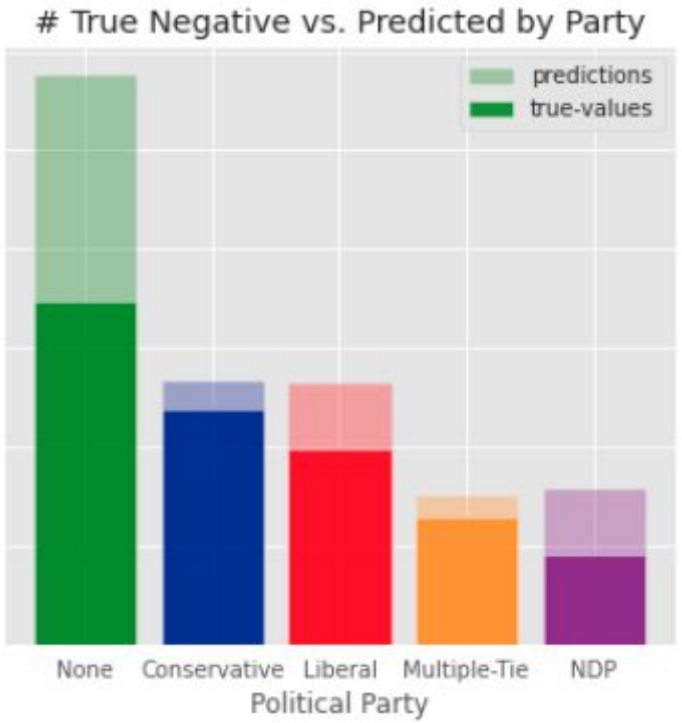



**Feature Importance:** After vectorizing with bag-of-words (BOW) or TF-IDF, we obtain 4700+ features, which made computation challenging. Using sklearn TruncatedSVD(), I reduced the # of principal components to 100, while retaining a significant % of the overall variance: **BOW: 37%, TF-IDF: 26%**. The plot shows the % of variance from each of the top 100 components in order of importance.

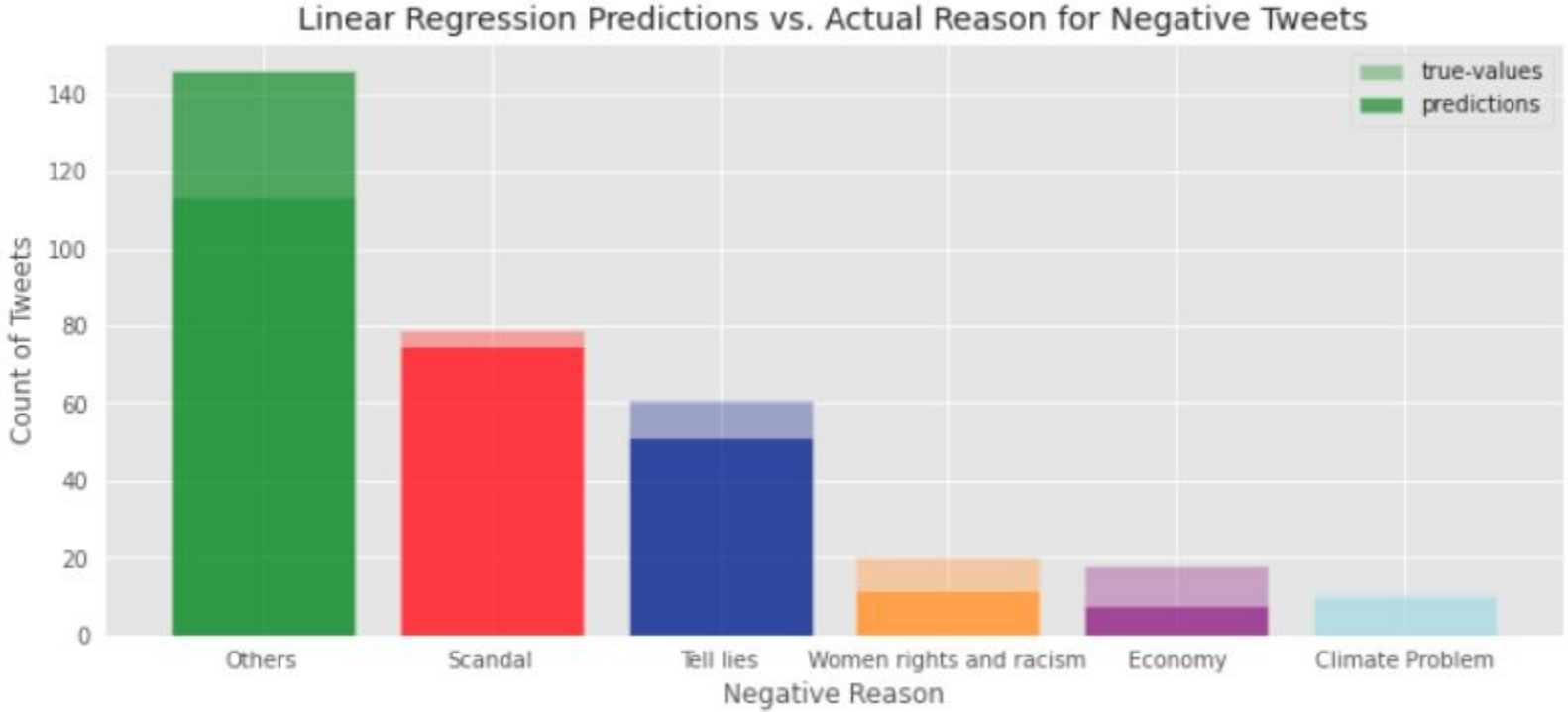
# First Model Results: Predict Canadian Election Sentiment




**Best Model Results (training: 94%, test: 53%)**: These results were obtained using Random Forest.

# 2nd Model Results: Predict Negative Reasons




**Best Model Results (training: 57%, test: 58%)**: These results were obtained using Logistics Regression.