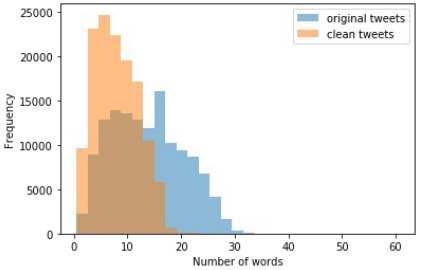
MIE 1624 Assignment 2:
Canadian Election Sentiment Analysis

Submitted by: Arun Shanmugam
Student No: 1005009884
UTORid: shanm159

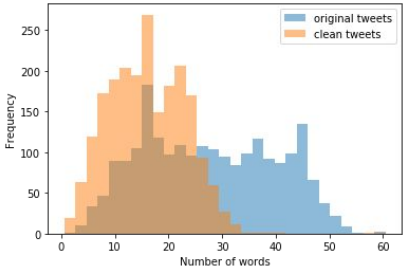
Exploratory Data Analysis

Tweets from Sentiment.csv & Canadian Elections were cleaned up by removing special characters, punctuations, stop words, tokenizing & then lemmatizing.Cleaned up tweets were subject to various Exploratory Data Analysis for visualization as shown.

Raw Data vs Clean-up

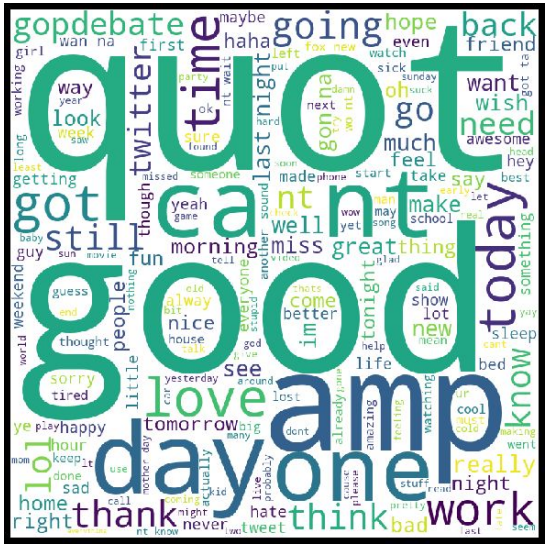


Sentiment Data

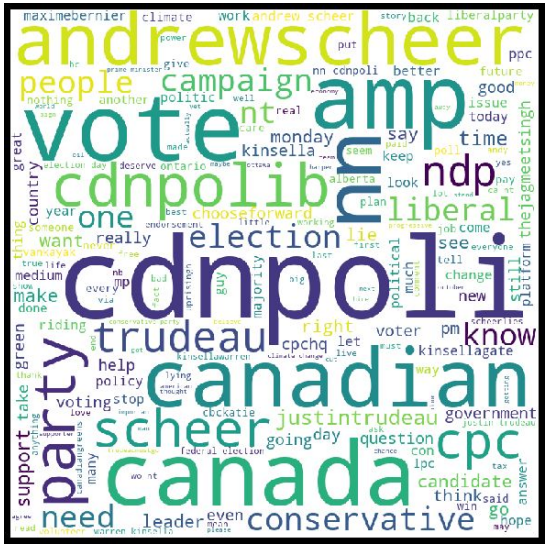


Canadian Election Data

Word clouds generated after data cleaning

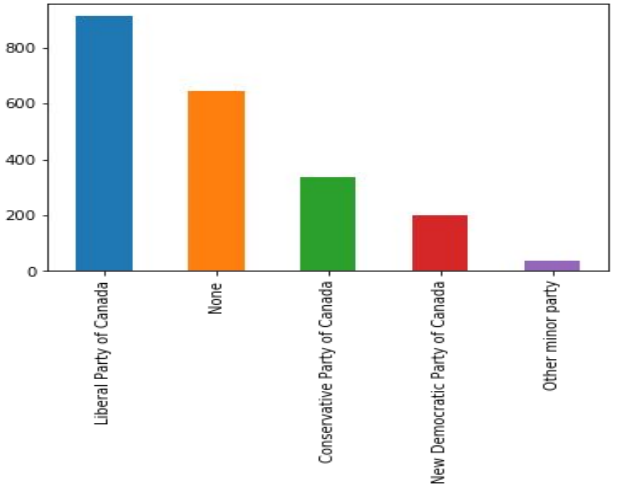


Sentiment Data



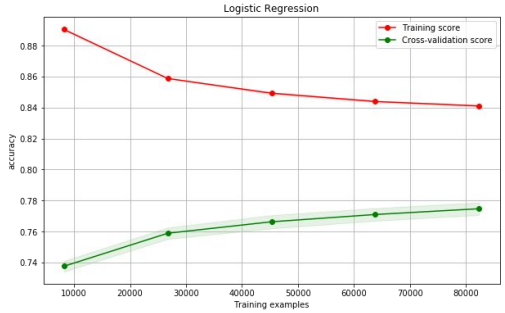
Canadian Election Data

A quick analysis of party affiliation of tweets

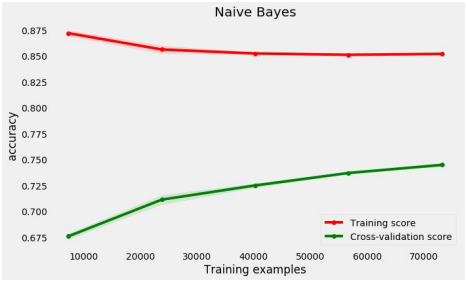


Data Preparation & Modelling

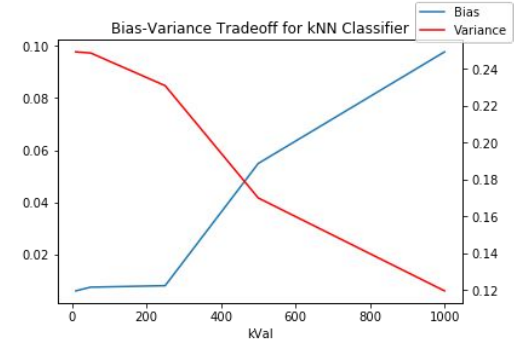
Learning curve for Logistic Regression



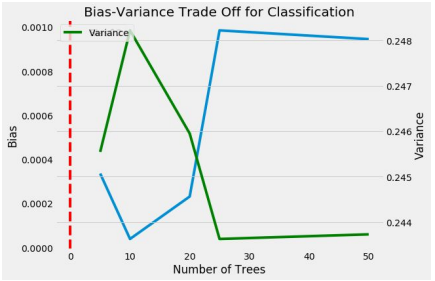
Learning curve for Naive Bayes



Bias Variance Tradeoff for kNN Classifier



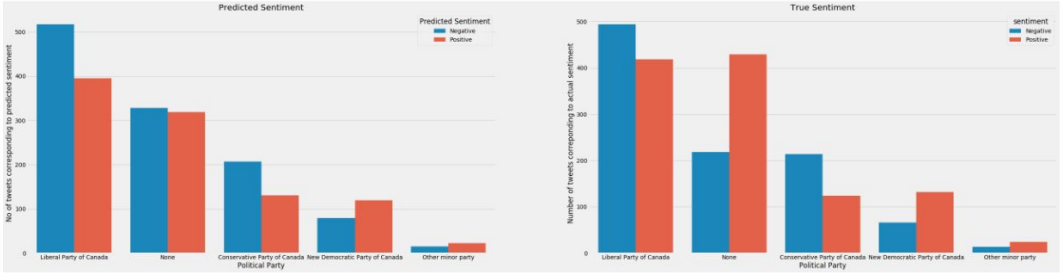
Bias Variance Tradeoff for Random Forest Classifier



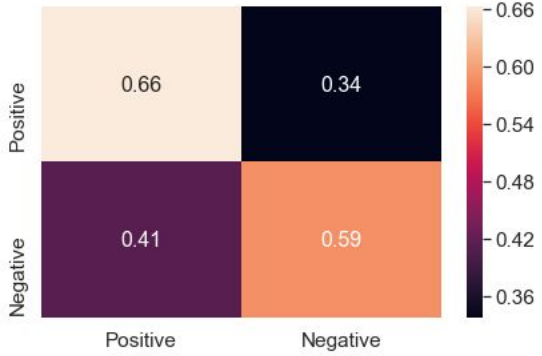
- Cleaned Tweets were prepared for ML modelling using Word Frequency & TF-IDF vectorizers
- The following models were tried out on the Sentiment.csv dataset:
  - Logistic Regression
  - k-Nearest Neighbours
  - Naive Bayes
  - SVM Classifier (long runtime)
  - Decision Trees
  - Random Forests
  - xGB Classifier
- Logistic Regression performed best with an accuracy of 77.793% on the test set of Sentiment.csv after Cross Validation & Hyper Parameter Tuning using grid search
- Bias Variance tradeoff curves were plotted for kNN, Decision Trees & Random Forest to understand relationship & better tune models

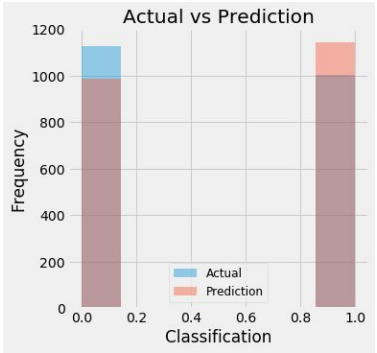
Model Implementation on Canadian Elections Data

- **Model choice:** Logistic Classifier model with tuned hyperparameters along with TF-IDF vectorization for Feature Engineering
- **Training:** Logistic classifier was trained on Sentiment data
- **Implementation:** Model was implemented on the Canadian Elections data and an accuracy of 62.21%



We can see that the model does a good job of predicting the sentiment for the four categories with misclassification occuring in the None category.



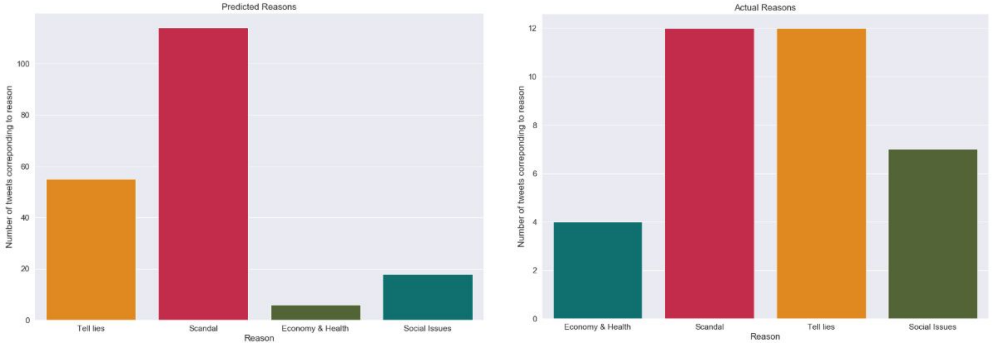


Discussion on Predictions:

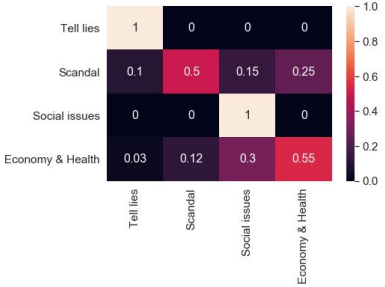
- Misclassification can be attributed to the word frequency feature engineering that eliminates context, which greatly dictates sentiment of tweets
- NLP analytics of elections is useful because voters vote with hearts than brains
- Since liberals have garnered the most number of tweets, it can be inferred that they would then win the election

Predicting reason behind negative tweets

Prediction vs. Actual



Logistic Regression

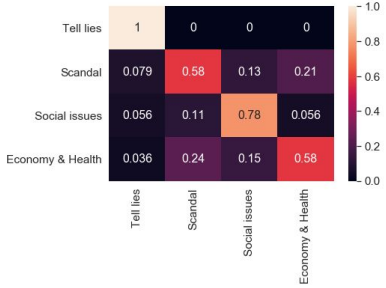


Accuracy: 0.5182

kNN Classifier

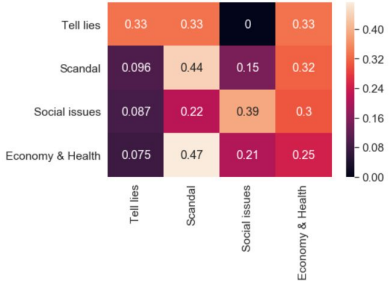
Accuracy: 0.5337

Random Forest



Accuracy: 0.6114

Random Forest (with Word Embeddings)



Accuracy: 0.378248

Discussion on reason predictions:

- The Random Forest model performs best but still has a fair degree of misclassification as evidenced by the above figure
- Misclassification can be attributed to:
  - Deletion of ‘Other’ tweets
  - Skewness to reasons that are more in number
  - Elimination of context
- As a bonus, Word Embeddings was carried out with a relatively small corpus due to memory and time considerations, and was passed to the Random Forest model