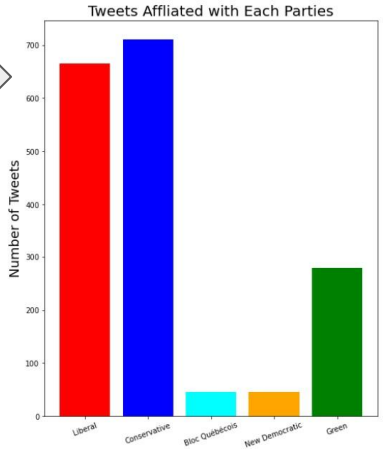
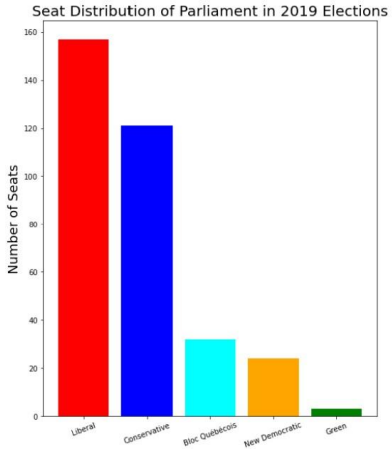
Assignment 3:
Natural Language Processing On Tweets

Zhi Xin Zhu
#1002117112

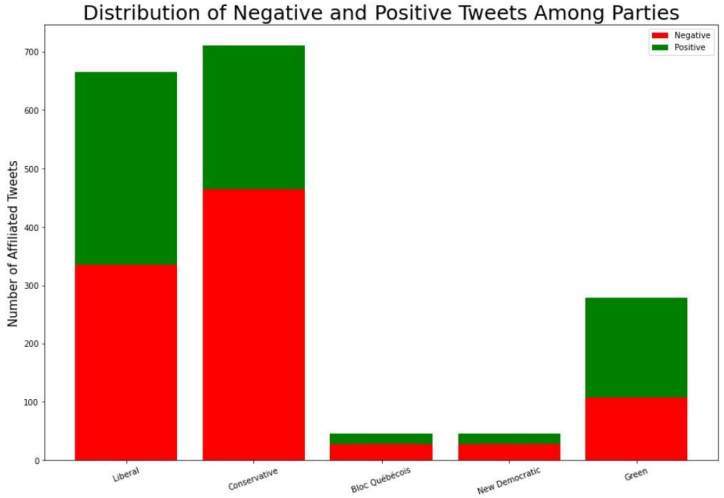
Exploratory Analysis

- The distribution of the political affiliation of tweets (left) generally reflects the result (right) of the election
- Tweets counts are closed between Liberal (red) and Conservative (blue); might be one of the reason why Liberal fails to form a majority government





- - The ratio of positive sentiment and negative sentiment tweets for Liberal party is around 50:50.
- - Majority of tweets affiliated with Conservative Party has a negative sentiment
- - Majority of the tweets related with NDP and Bloc Québécois has negative sentiment.
- - Green has more positive tweets than negative ones



Model Feature

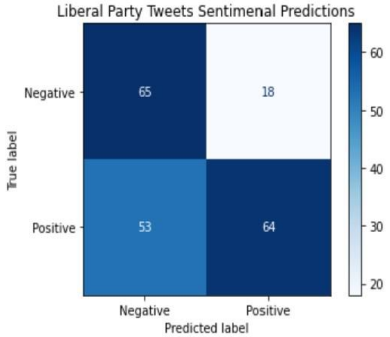
Text preprocessing is an important step in natural language processing. Text data needs to be transformed to numerical data in order to perform machine learning algorithm. The following text transform method are used:

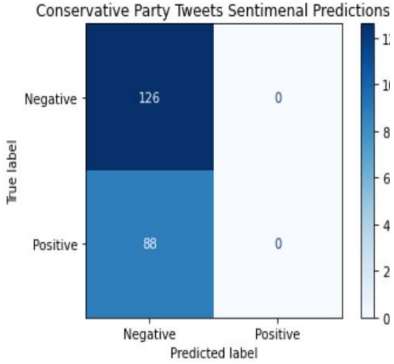
1. Bag of Word (Word Frequency)
    - A way of representing text data that describes the occurrence of words with a document; do not concern where the occurrence happened
    - Created 1000 features
2. TF-IDF
    - A numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. Rescale the frequency of words by how often they appear in ALL document
    - Created 2000 features

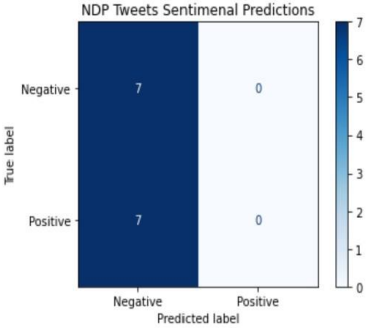
Model 1 Results: Sentiment Analysis for Each Parties

Based on the training and testing accuracy of generic tweets, the SVM model with TF-IDF features is selected to be implemented on the Canadian Election data for the three parties: Liberal, Conservative and NDP. Confusion matrices are constructed to provide insight on what the classification model is getting right or wrong.

Feature	Logistic Regression	k-NN	Multinomial Naive Bayes	SVM	Decision trees	Random Forest	XGBoost
TFIDF	94.68%	83.21%	88.44%	**94.77%**	92.57%	93.93%	94.33%
WF	93.7%	91.59%	87.63%	93.87%	92.6%	93.33%	93.48%







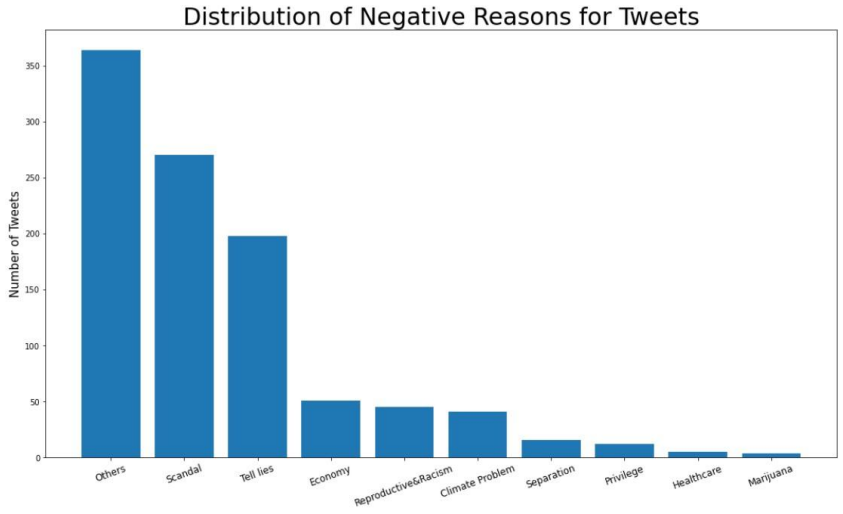
- For all three parties, the main prediction error is false negative, in which the prediction model label the tweet with negative sentiment while its true sentiment is positive.
- For Conservative and NDP, the model simply guess all tweets to have negative sentiment. these two models are simply making random guesses. This issue might be due to unbalance data.
- The predictions made by the NLP analytics provide a general picture of the sentiments among the voters. However, the estimation of party's support rate should not be solely based the predictions made by the NLP analytics.

Model 2 Results: Negative Reason Classification

Implemented 3 algorithms; results shown on the right.

Model	Parameter(s)	Train Accuracy	Test Accuracy
SVM	C = 1.0	48.71%	52.32%
Logistic Regreesion	C = 10, solver = liblinear	48.71%	51.32%
XGBoost	max_depth = 2	49.2%	56.62%

- The XGBoost has the highest test accuracy, which is 56.62%.
- Probable poor performance reason:
  - Lack of training samples. Unbalanced data for each categories
  - Noise in the original dataset. For example, the hashtags such as #elxn43 and #cndpoli provide no meaningful information to the algorithm.
  - The category 'Others' can be more specified. The model implemented mainly make the prediction to be 0 or 1. A more detailed classification that converts 'Others' into small sub-categories may improve the performance.



Improvement on Models

Model 1:

- Ensembling multiple SVM
- Feature selection is important to the application of SVM. Possible modification on the feature selection and feature extraction (tfidf) algorithm.
- Increase sample size for each party.

Model 2:

- Use a balanced dataset with larger number (such as at least 300) of training samples for each category.
- Improve feature selection algorithm to better eliminate features that are not useful to the prediction.