Data Science for Sentiment Analysis

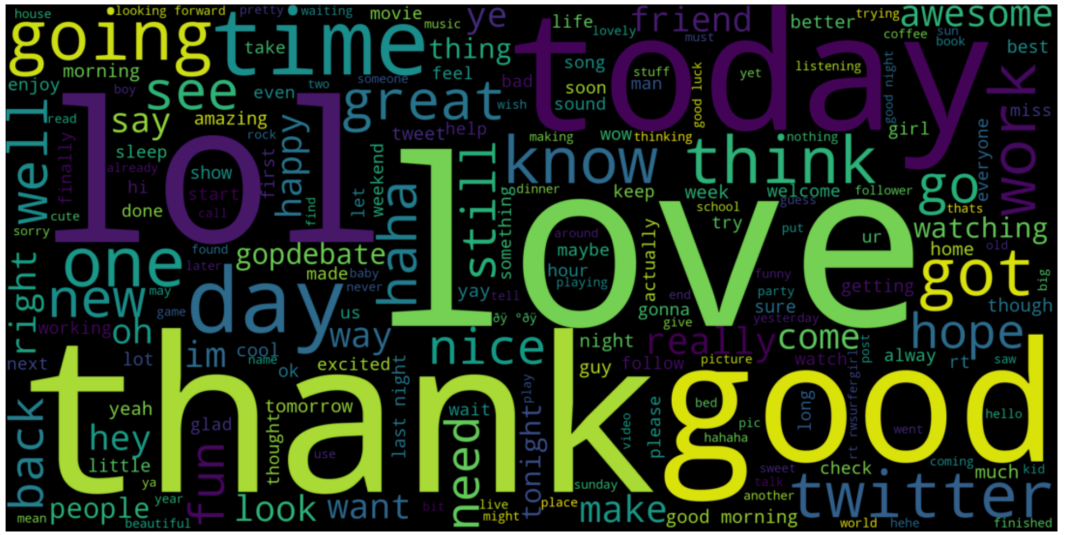
MIE 1624 Introduction to Data Science and Analytics

Name: Zichuan Wang
Student ID:1000474300

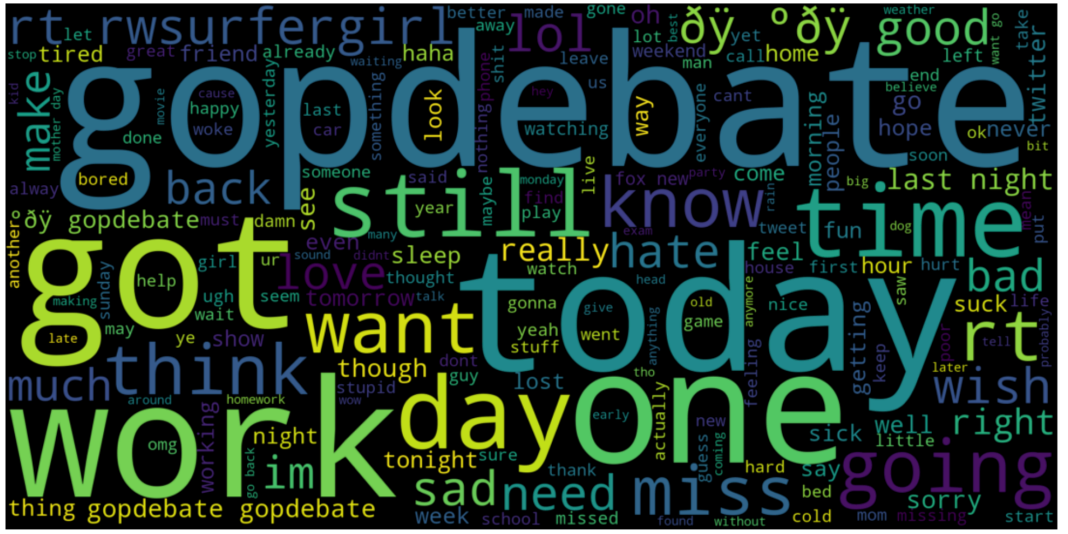
Exploratory Analysis: Generic Tweets

From the word cloud of population's positive voice, we can see that positive tweets have some key words like "good", "love", "great". And some other key words including "hope", "going", "will". It seems that positive tweets show people's hope and plan about future.

From the word cloud of population's negative voice, we can see that negative tweets have some key words like "work", "hate", "bad". It seems that negative tweets show people's reluctance to work.



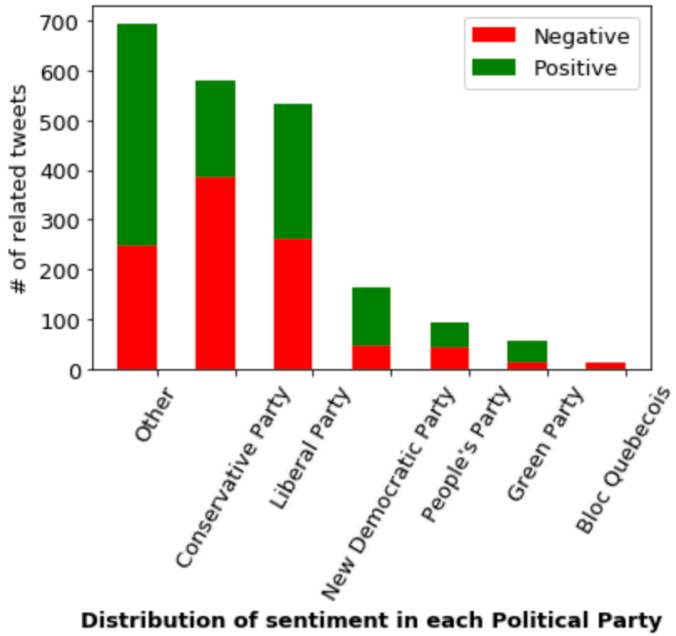
Word Cloud for Generic Tweets with Positive Sentiment

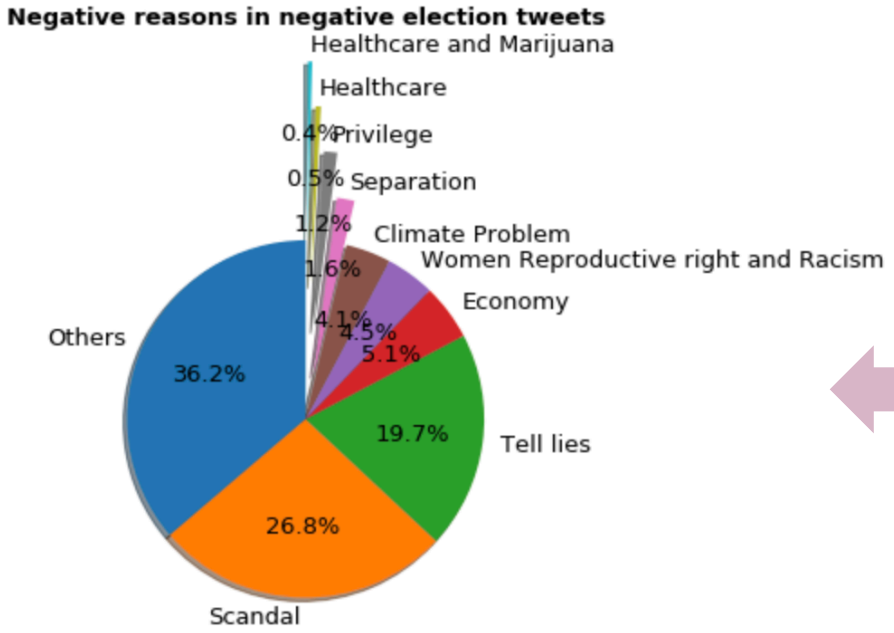


Word Cloud for Generic Tweets with Negative Sentiment

Exploratory Analysis: Election Tweets

Distribution of sentiment in each political party figure demonstrates that Conservative Party has the highest popularity, followed closely by Liberal Party. But the negative portion of the Conservative Party is the highest amongst all parties. This is a good indicator on why Conservative Party lost in the election.



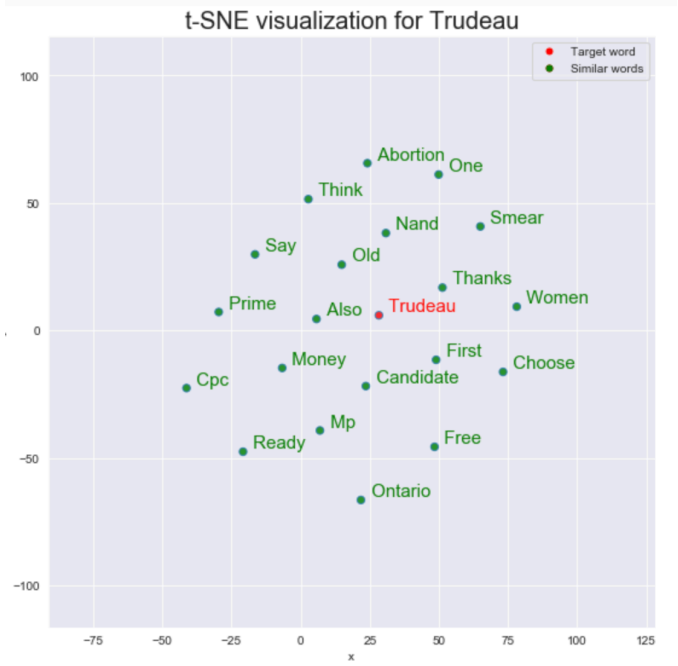


The negative reasons figure demonstrate the negative reasons distribution amongst all negative election related tweets. Scandal and Tell lies have the highest percentage besides others. This shows that people cares more about whether the elected government is honest or not, rather than things like economy.

Model Feature Importance

TF-IDF:

The tf–idf value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word, which helps to adjust for the fact that some words appear more frequently in general. tf–idf is one of the most popular term-weighting schemes today. That the reason why tf-idf is the main feature engineering technique used through out this assignment.



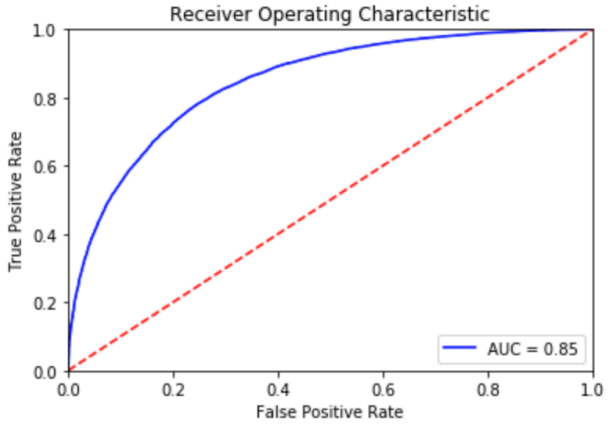
Word Embedding:

Word embedding represents words as vectors in multi-dimensional space. As an example, Trudeau is selected as the target word in the election tweets. Plot shows related words with reduced dimensionality. Since Trudeau is the candidate from Liberal Party, it makes sense that word “candidate” is closer than word “CPC”.

Model Results and Visualization

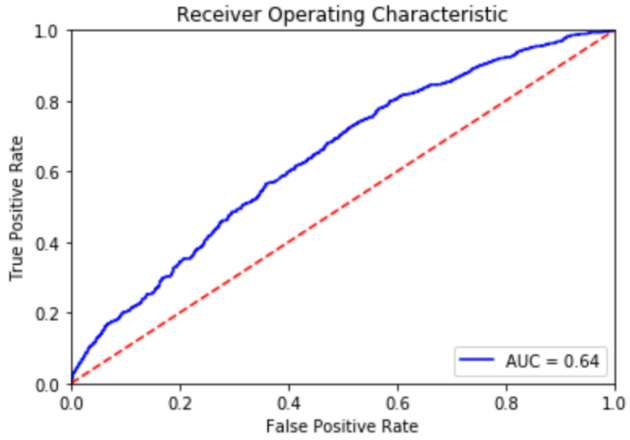
Classifier	train_acc	test_acc
Logistic Regression	86.3%	76.7%
Naive Bayes	85.9%	75.6%
K Nearest Neighbor	71.3%	57.5%
Linear SVC	84.4%	76.7%
Decision Tree	99.4%	70.0%
Random Forest	99.4%	74.6%
Gradient Boosting	62.3%	62.1%

7 different classifiers trained on generic tweets are compared using grid search and cross validation. From the test accuracy, we can conclude that Logistic Regression, despite its simplicity, is the best classifier among all algorithms.



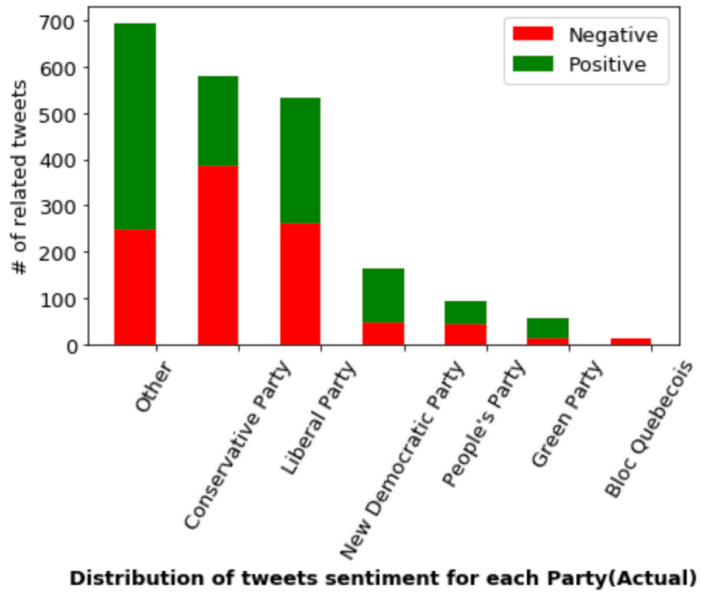
Model performance on generic tweets

Logistic Regression model trained on generic tweets performs worse on election tweets than on generic tweets. This model gets a test accuracy of **60.1%** compared to **76.7%**, and gets s AUC value of **0.64** compared to **0.85**. Potential reason why this happens is that different topics have different text features.

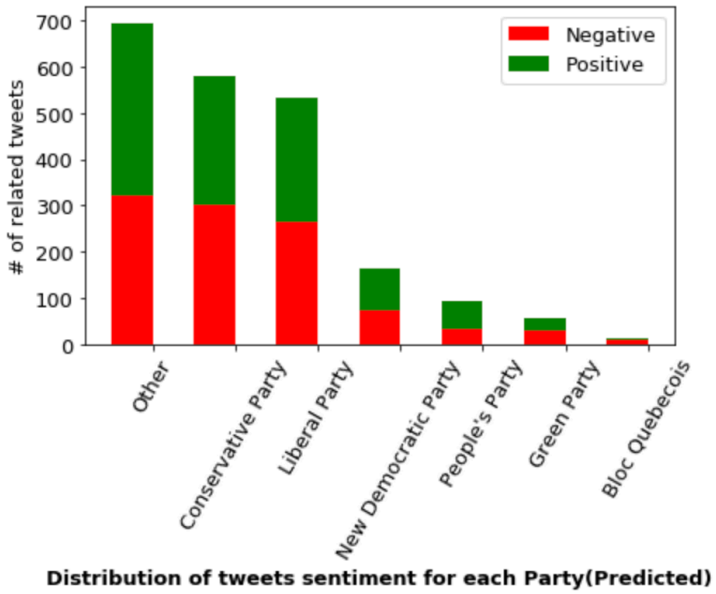


Model performance on election tweets

Model Results and Visualization



Distribution of tweets sentiment for each Party(Actual)



Distribution of tweets sentiment for each Party(Predicted)

Based on the Distribution of tweets sentiment for each Party(**actual**), Although total number of Conservative Party related tweets is slightly higher than that of Liberal Party related tweets, the negative portion of the Conservative Party related tweets is significantly higher than that of Liberal Party. This explains why Liberal Party wins the election.

Based on the Distribution of tweets sentiment for each Party(**predicted**), the ratios between negative and positive tweets for all parties are close to 50:50. This is consistent with the overall 60% test accuracy of the model. This indicates that model that based on generic tweets doesn't generalize well to election related tweets. And topics of tweets can be quite different in terms of input features.