SENTIMENT ANALYSIS (NLP)

RESEARCH QUESTION:

WHAT CAN PUBLIC OPINION ON TWITTER TELL US ABOUT THE CANADIAN POLITICAL LANDSCAPE IN 2019?

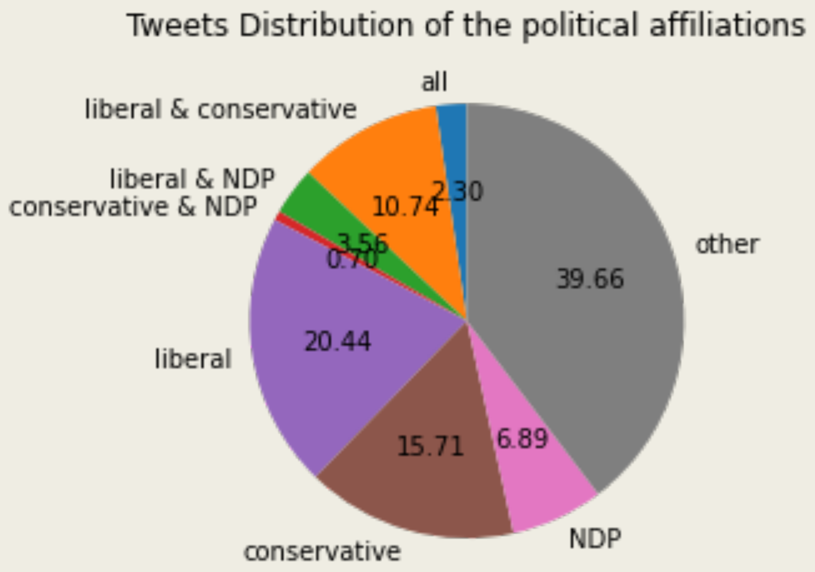
GOAL: USE SENTIMENT ANALYSIS ON TWITTER DATA TO GET INSIGHT INTO THE CANADIAN ELECTIONS

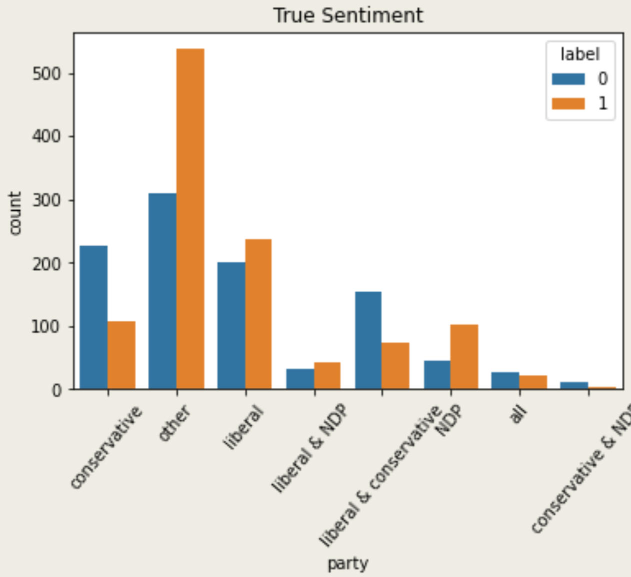
Data Exploration

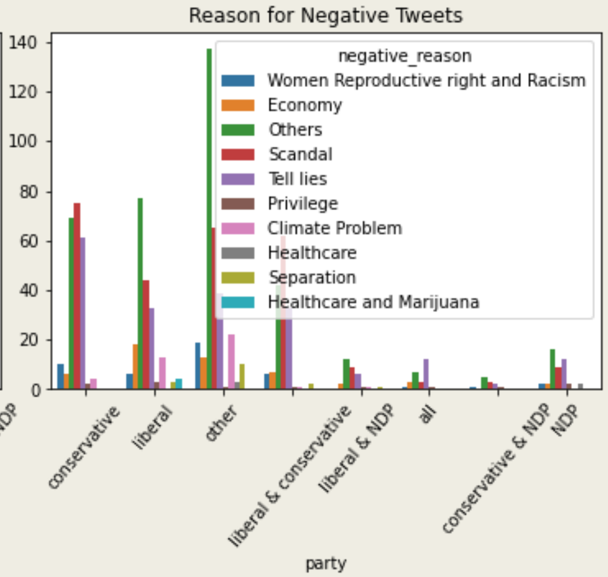
Political affiliation on Canadian Election tweets.

- Tweet related to single party: Liberal, Conservatives, NDP
- Tweet related to more than one party
- Tweet related to other party: Other

For tweets only relate to one party, liberal is the highly discussed topic on Twitter, around 20% tweets relate to this party. Using more explicit keywords related to the party would lead to more accurate results.


Tweets Distribution of the political affiliations


True Sentiment


Reason for Negative Tweets

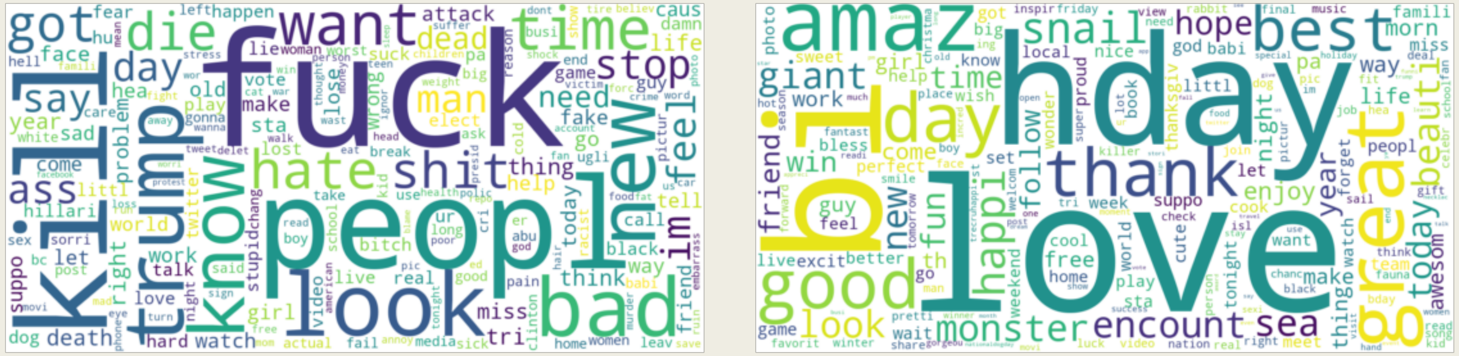
- Liberal and NDP: most tweets are positive.
- Liberal: more than 500 positive tweets, high popularity among the younger generation. Although there're also lots of opposition, the chance for liberal to win the election is still high, since the reason of most negative tweets is 'Others'.
- Conservative: most tweets are negative, and reasons for most of them are related to 'scandal' and 'tell lies, all indicates more negative public impression.

Text Exploration

Generic Tweet

Sentiment 0

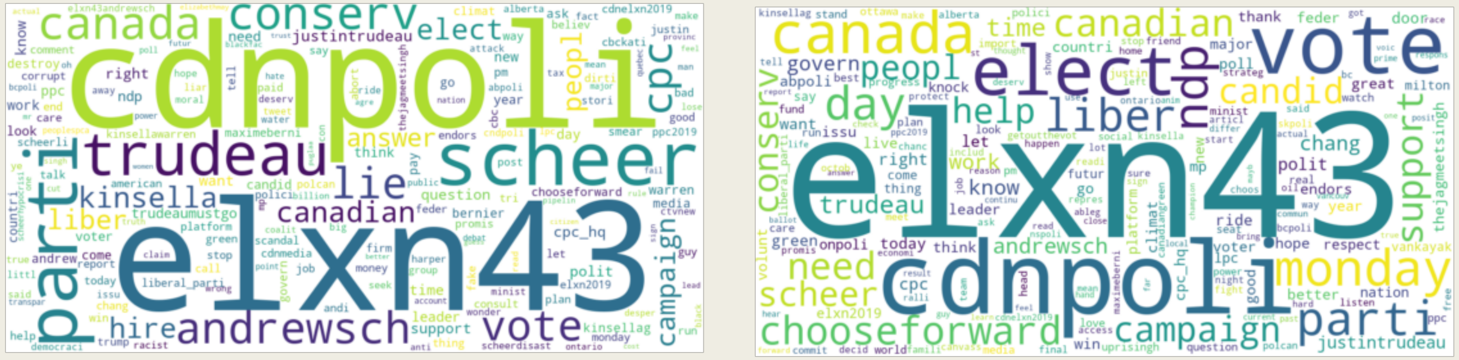
Sentiment 1



Canadian Election Tweet

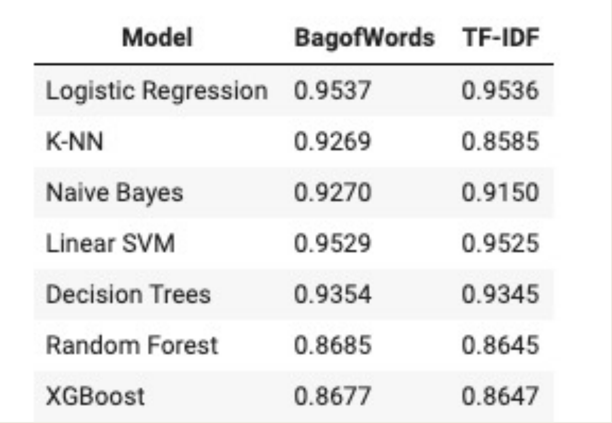
Sentiment 0

Sentiment 1

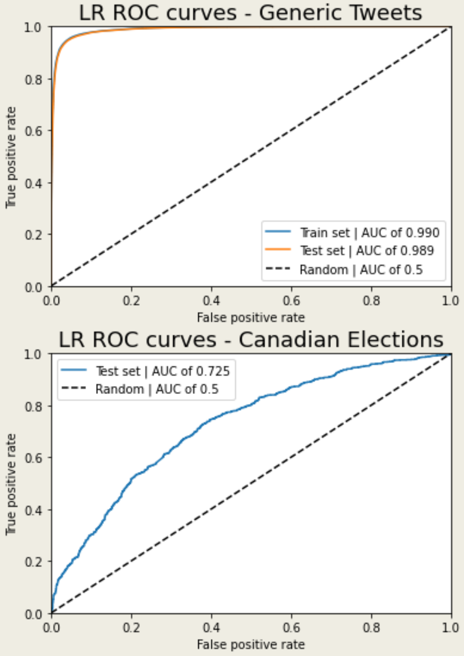


- Word cloud visualize the word frequency, larger word appears more often.
- Words like 'Trudeau' (in Canadian election tweet), word size in the negative sentiment plot looks larger than the word size in the positive sentiment plot, which imply the tweet contain this word is more likely to have negative sentiment.
- Words like 'elxn42', 'Canada', 'cdnpoli' appears in both positive and negative plots, they are neutral word related to the election but don't contribute useful information to classify the sentiment.
- Word cloud plots of 'generic tweet' and 'Canadian election tweet' vary a lot.

Sentiment Analysis Model (binary classification)



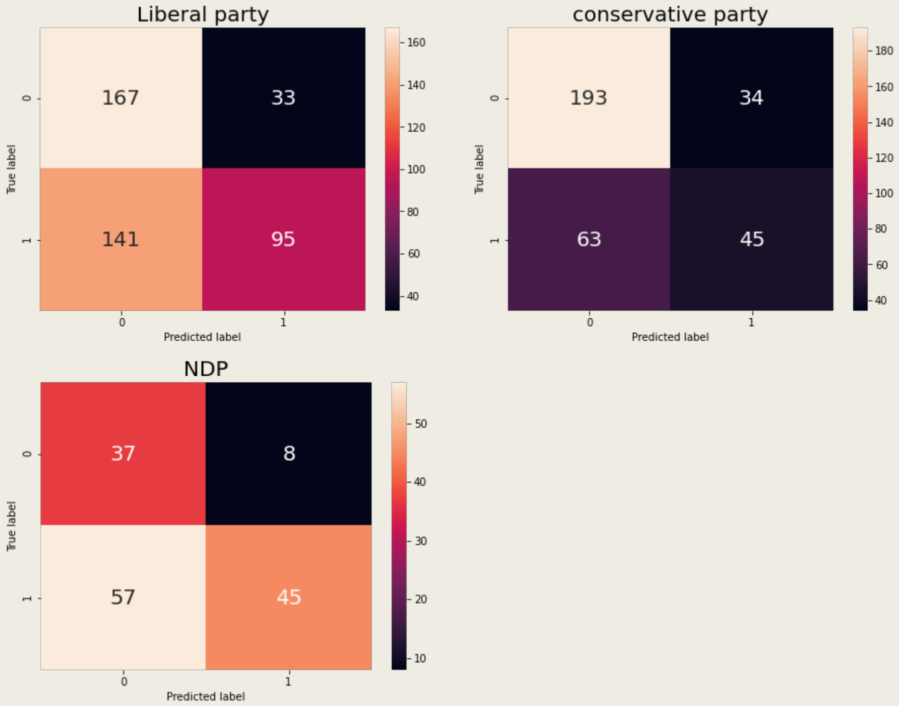
Model	BagofWords	TF-IDF
Logistic Regression	0.9537	0.9536
K-NN	0.9269	0.8585
Naive Bayes	0.9270	0.9150
Linear SVM	0.9529	0.9525
Decision Trees	0.9354	0.9345
Random Forest	0.8685	0.8645
XGBoost	0.8677	0.8647



- Train models on training data from generic tweets. Apply model on testing data to compute the testing accuracy.
- Best model: Logistic regression with Bagofword features

  test accuracy = 0.9537, AUC = 0.989
- Apply the model on Canadian election data

  accuracy = 0.6184, AUC =0.725

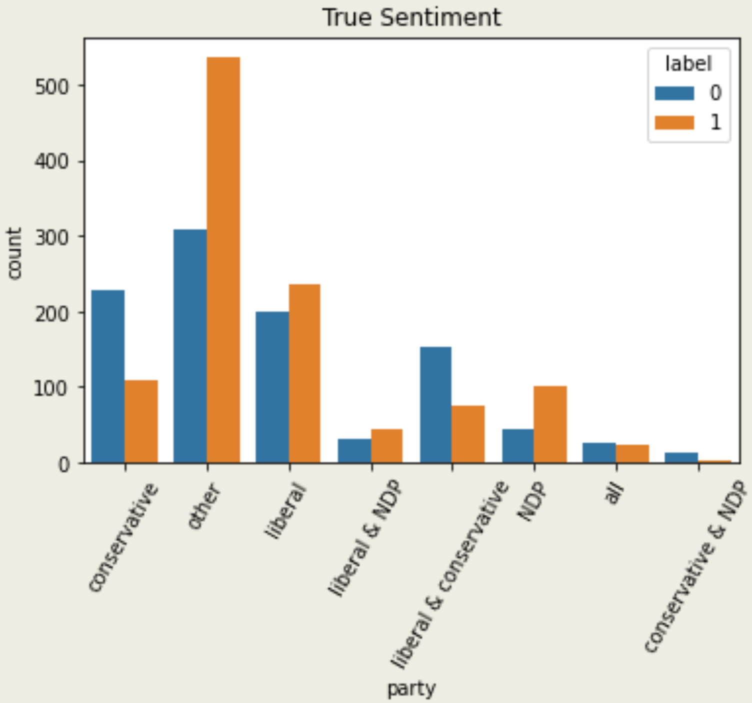


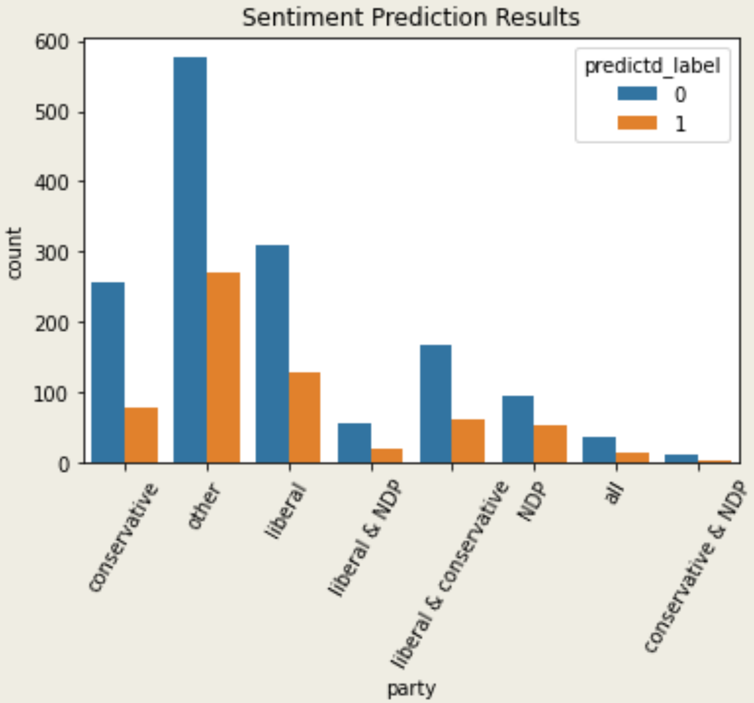
- Liberal: FN = 141
- Conservative: FN = 63
- NDP: FN = 57

For all three parties, large amount of positive tweets are predicted as negative tweets, large FN lower the testing accuracy and AUC score.

The model is trained on generic tweets, however the content between generic tweet and Canadian election are different. Some word indicate a positive sentiment in generic tweet, might contribute a negative sentiment in Canadian election data.

Sentiment Analysis Model (binary classification)



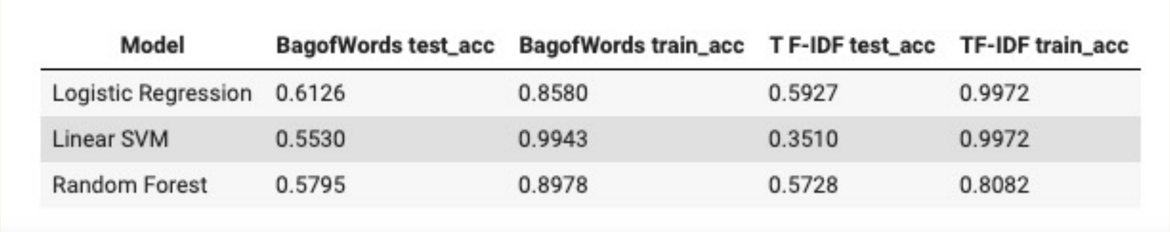


For all three parties, large amount of positive tweets are predicted as negative tweets,

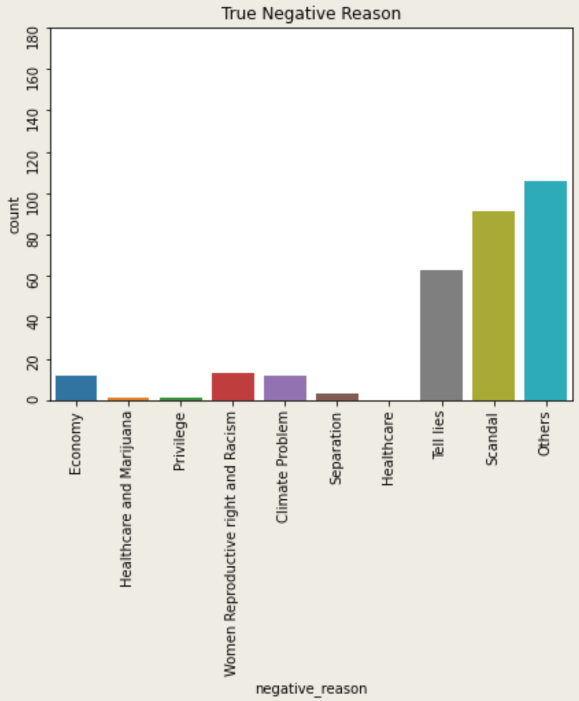
The sentiment prediction results shows, liberal have most positive tweets and negative tweets, NDP have least positive tweets and negative tweets, which is consistent with the true sentiment distribution.

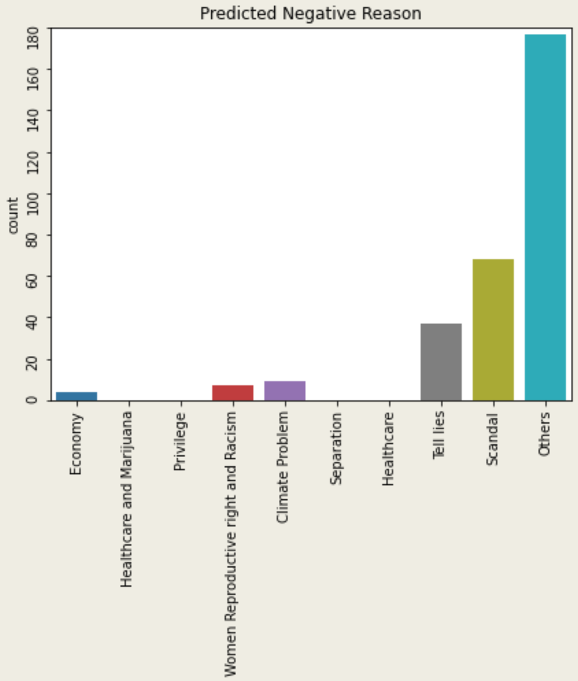
The Canadian election only count the number of vote, similar to positive tweets in this case. The liberal has high chance to win the election if twitter represent population well. And the predicted result is consistent as the result of Canadian election 2019, the liberal party won. So the NLP analytics based on tweets are still useful. Use tweets related to political topic might be better.

Negative Tweets Reason Analysis (multi-class classification)



Model	BagofWords test_acc	BagofWords train_acc	T F-IDF test_acc	TF-IDF train_acc
Logistic Regression	0.6126	0.8580	0.5927	0.9972
Linear SVM	0.5530	0.9943	0.3510	0.9972
Random Forest	0.5795	0.8978	0.5728	0.8082





- Model with highest testing accuracy

  Logistic regression with Bagofword features

- For all three models, the training accuracy is much higher than testing accuracy. The models are overfitting. The max_features are set to 5000, but there are only 1007 negative tweets. The data does not contain enough information to build complex model.

- The dataset is also imbalanced. Three models are all sanative to the imbalanced data. The small weighted class tend to be classified into larger weighted class. There are few negative tweets with reason ’Privilege’. But there’s no predicted negative tweet with this reason.

- Try increase the range of hyperparameter to tune or decrease the number of text features.