Data Cleaning

1.Regular Expression is a useful tool to match text contents.

2.Nltk can be used to tokenize the raw text and perform other natural language processing techniques.

3.Lower case:Series str has api lower to convert characters to lowercase

4.remove stop words:First, use nltk to tokenize the string then remove all words in the stopword list.

Exploratory Analysis

1.Find the single word after @ and # symbol:these two symbols are always followed with topic name or party leader account, which can be used to identify the party that the tweet belongs to.
2.Select the words that each party is represented and combine them into three lists.
3.Tokenize the tweets into list of tokens and match the tokens with the party key word list to identify the party.
4.Count the number of each party's occurence in the same tweet and mark the tweet with the party name that occured the most, and if no party token appeared in the tweet, just mark it as none.

Model Feature Importance

To encode the tweets into numerical values, I used two methods, tf_idf and bag of words(term frequency).
To implement tf_idf , I used sklearn TfidfVectorizer which first vectorize the tweets using CounterVectorize then calculate the tf_idf score, to limit the dimension of feature space I limit the maximum number of features to 2000 then sum the tf_idf score of words after that rank the words and select the top 500 word as features.
For bag of words model, I simply count the number of each word and select the top 500 most common words as features, then judge if a word appears in the tweet to vectorize the feature.

Model Results

	tf_idf train acc	tf_idf test acc	bog train acc	bog test acc
logistic regression	0.70	0.69	0.699	0.6983
Naive Bayes	0.67742857142	0.675883333	0.6695714	0.67208333333
SVM	0.6996	0.69885	0.6992357142857	0.69785
Decision Tree	0.5752	0.56825	0.58194285	0.577633333
Random Forest	0.63340714	0.62995	0.6445214	0.6442833
XGBoost	0.5712214	0.56735	0.57030	0.56975
RNN	0.70199287	0.70022		

Model Visualizations

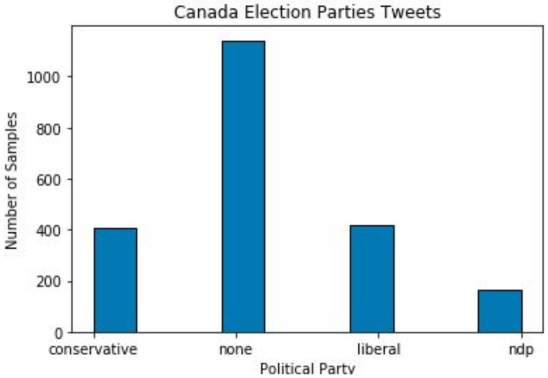

fig1. relation between party tweet number

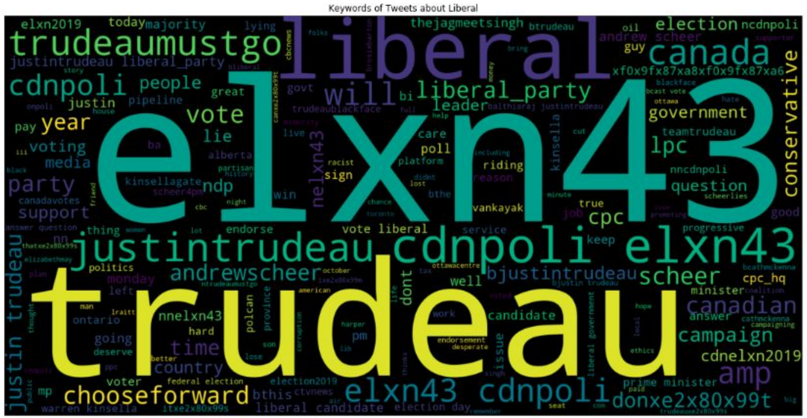

fig2. Liberal word cloud

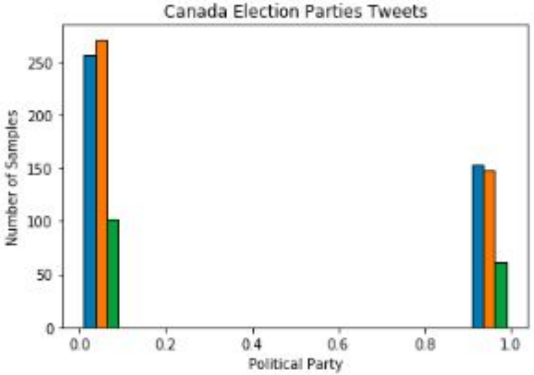

fig3. Election Sentiment Prediction

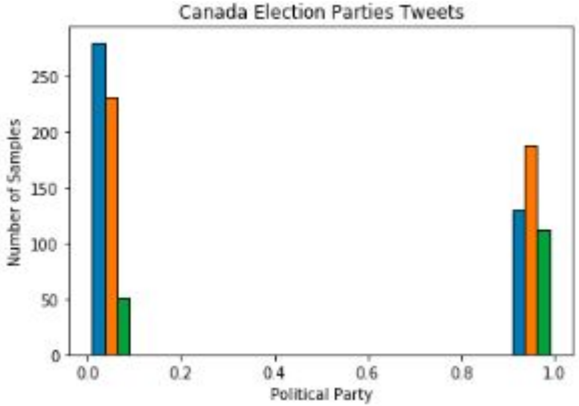

fig4. Election Sentiment Truth