Datacleaning

Design a procedure that prepares the Twitter data for analysis:

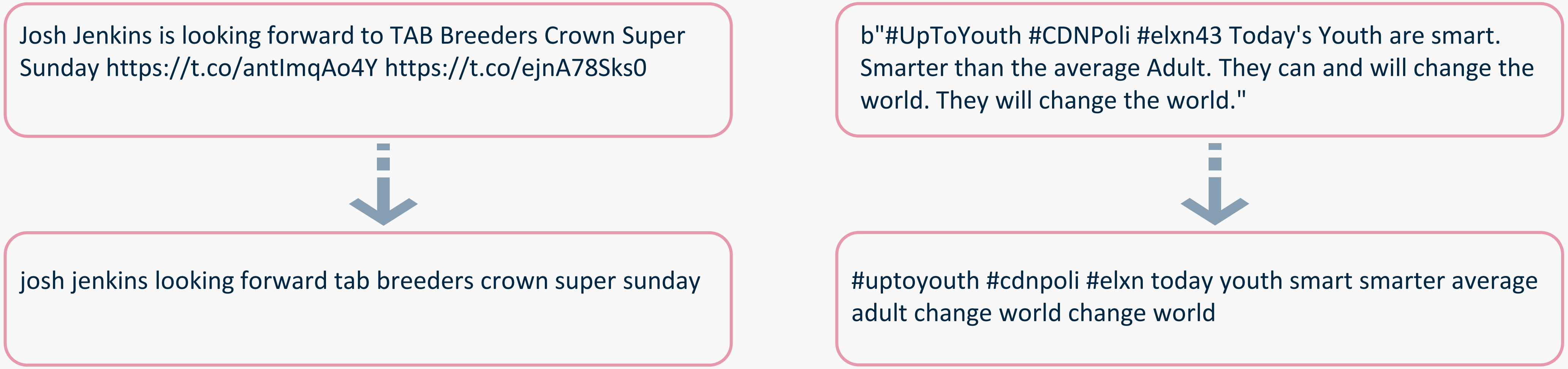
Genetic tweets:

- Remove all html tags, attributes, all stop words, and URLs;
- Html character codes are replaced with an ASCII equivalent;
- Remove the words after @, all punctuations, and all digital;
- Remove all non-ASII words, “rt”, and “\n”;
- Lower all text.

Canadian election tweets:

- Remove all html tags, attributes, all stop words, and URLs;
- Html character codes are replaced with an ASCII equivalent;
- Remove the words after @, digital, first letter “b”, and “\n”;
- Remove punctuations except “#”;
- Lower all text.

Cleaning results(sample tweet):



Josh Jenkins is looking forward to TAB Breeders Crown Super Sunday https://t.co/antImqAo4Y https://t.co/ejnA78Sks0

josh jenkins looking forward tab breeders crown super sunday

b"#UpToYouth #CDNPoli #elxn43 Today's Youth are smart. Smarter than the average Adult. They can and will change the world. They will change the world."

#uptoyouth #cdnpoli #elxn today youth smart smarter average adult change world change world

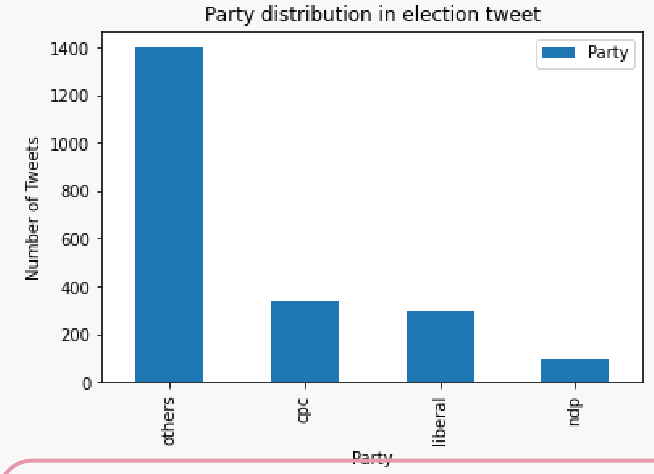
Exploratory analysis

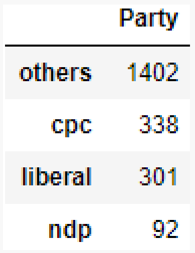
Define the political party for all tweets in Canadian election data

Selected the follow key words to identify three political parties:
For Conservatives(cpc): 'cpc', 'scheer', 'andrew', 'andrew scheer','conservative','conservatives';
For Liberal: 'lpc','trudeau','trudeaumustgo','liberal','justin','justintrudeau','kinsella';
For ndp: 'ndp','ndc','democratic','jagmeet','singh','jagmeet singh’.

All these keywords beased on my personal thinking and looking for total words frequency result from the previous part. The identification keywords are not perfect.

Party's frequency distribution daigram from the counts dataframe

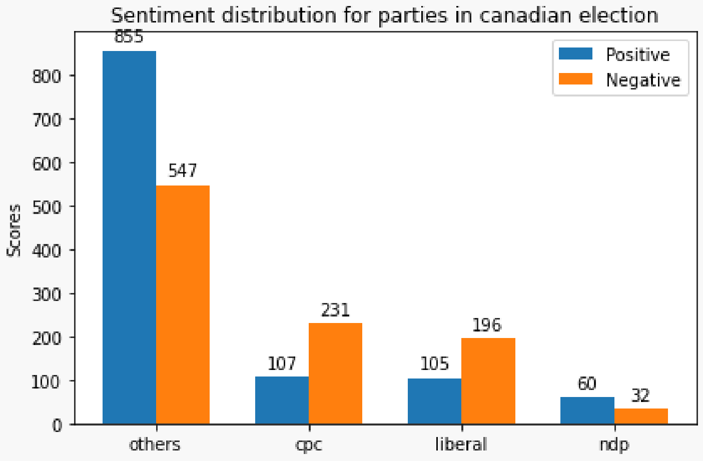




	Party
others	1402
cpc	338
liberal	301
ndp	92

"cpc" and "liberal" have similar score value and "ndp" is lower. party "cpc" and "liberal" have higher attention and more powerful than the party "ndp"

positive and negative sentiment distribution for different parties

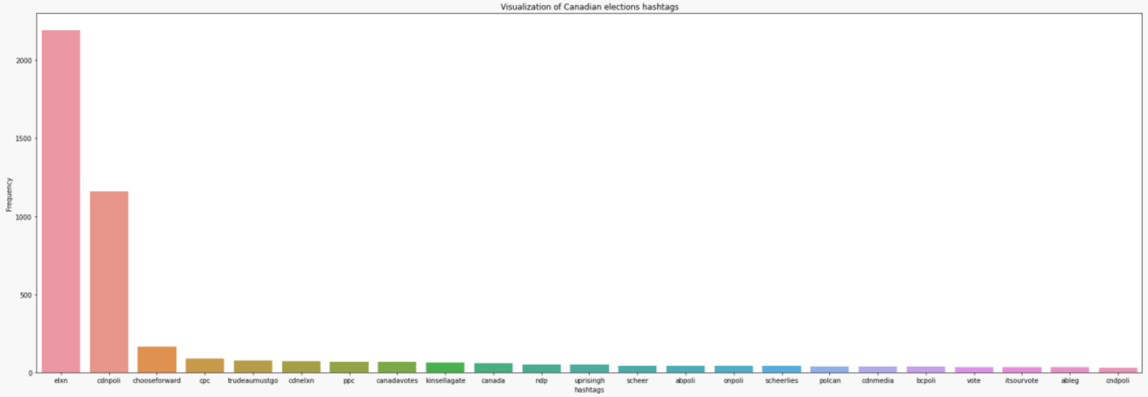


People has more negative sentiment for "cpc" and "liberal"the result shows party "others" and "ndp" have higher positive score than negative. However, part "cpc" and "liberal" have higher negative score than negative. People has more negative sentiment for "cpc" and "liberal".

Exploratory analysis

Hashtag analyzation about the Canadian election data

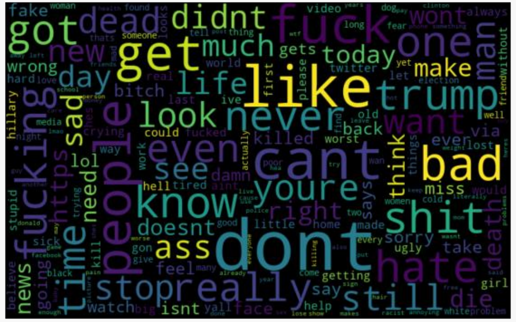
The plot shows the most popular and highest frequency hashtag words are elxn and cdnpoli which are far higher than others.



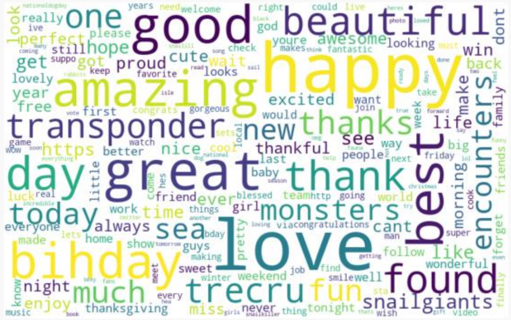
Negative and Positive Words Frequency Cloud in Canadian election tweet(right)and genetic tweet (left)

For these two genetic tweet figures, the highest score negative words are: like, dont, cant, and bad. The highest score positive words are: happy, love, great, and amazing. The results are close to the real life which can see the data cleaning process is reliable.

Genetic Words Frequency Cloud for Negative Sentiment

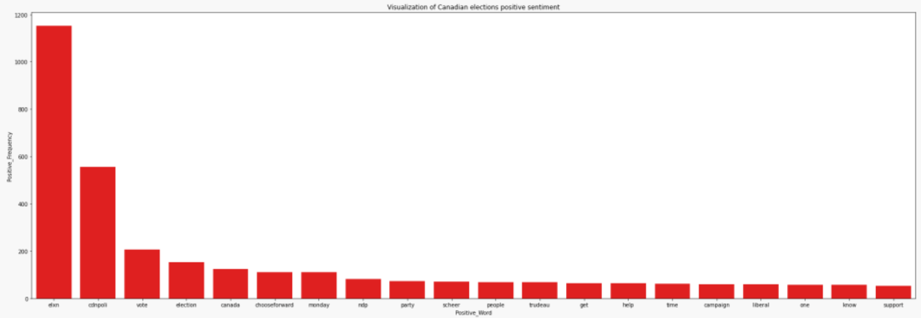


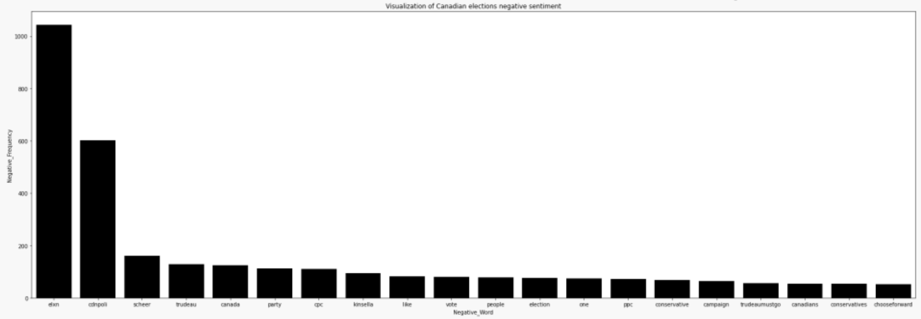
Genetic Words Frequency Cloud for Positive Sentiment



Negative and positive sentiment visualization of Canadian elections tweet keywords

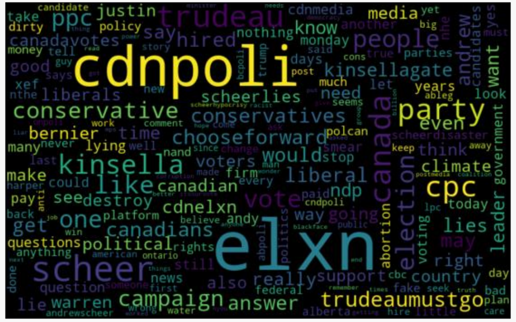
These two distribution bar plot are similar, but in negative words distribution there are more words about parties.



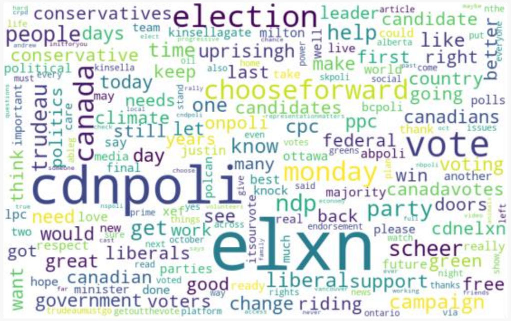


For these two election tweet cloud figures, we can clear see the highest score words for both negative and positive are elxn and cdnpoli. There are more election topic words in negative figure.

Election Words Frequency Cloud for Negative Sentiment



Election Words Frequency Cloud for Positive Sentiment



First model to predict genetic tweet sentiment

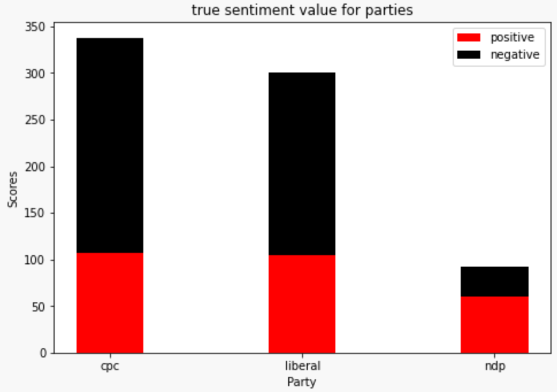
Split the genetic tweets dataset into train and test sets. Try seven different classification algorithms to predict genetic tweet sentiment with two feature extraction methods:

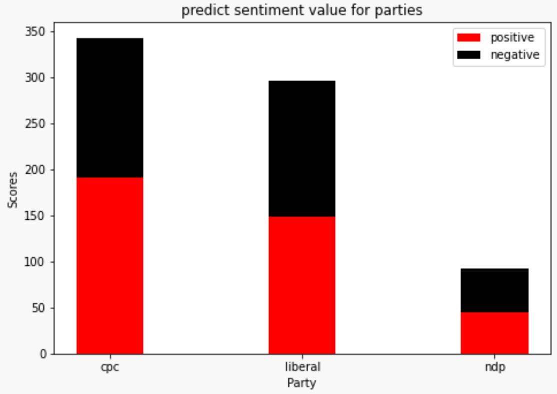
		Logistic regression	K-NN K=100	Naive Bayes	SVM	Decision Tree	Random Forest	XGBoost
Bag of Words	Train accuracy	0.918	0.892	0.872	0.687	0.962	0.962	0.849
	Test accuracy	0.916	0.891	0.872	0.689	0.912	0.921	0.848
TF-IDF	Train accuracy	0.917	0.839	0.866	0.707	0.962	0.962	0.852
	Test accuracy	0.916	0.828	0.866	0.708	0.913	0.923	0.850

Best prediction model:
Random Forest Classifier with TF-IDF

Hyperparameter tunning select best prediction model： Parameters are {'min_samples_split': 10, 'min_samples_leaf': 1, 'max_features': 'auto', 'max_depth': 200}

Visualize the sentiment prediction results and the true sentiment for each of the 3 parties.





First model to predict Canadian election tweet sentiment

Use genetic trained best model to predict the election sentiment, the prediction accuracy is 0.493 for both features. Try other genetic trained models which is logistic regression to predict the election sentiment, the prediction accuracy is 0.517 which is higher than the best genetic trained model.

When fit election data with four simple classification model to predict the election sentiment, the highest prediction accuracy is 0.71 with logistic regression and TF-IDF feature.

It's hard to use genetic tweet trained model in election prediction. However, when apply the classification model to predict the election sentiment can get higher accuracy then using genetic trained model, which shows possibility to use NLP in political party analyzation in during election campaigns

Second model to predict Canadian election negative reason

Split the negative Canadian elections tweets into training data (70%) and test data (30%).

	Others	Scandal	Tell lies	Economy	Women reproductive right & Racism	Climate problem	Separation	Privilege	Healthcare	Healthcare & Marijuana
score	364	270	198	51	45	41	16	12	5	4

Change negative reason Healthcare & Healthcare and Marijuana to 'Others' because of the low counts.

Three multi-class classification models to predict the reason for the negative tweets:

		Logistic regression	Decision Tree	Random Forest
Bag of Words	Train accuracy	0.939	0.997	0.997
	Test accuracy	0.533	0.497	0.589
TF-IDF	Train accuracy	0.737	0.997	0.997
	Test accuracy	0.487	0.497	0.556

Hyperparameter Tunning:

		Logistic regression	Decision Tree	Random Forest
Bag of Words	Train accuracy	0.859	0.712	0.932
	Test accuracy	0.556	0.500	0.579
TF-IDF	Train accuracy	0.672	0.746	0.980
	Test accuracy	0.566	0.447	0.583

Results

“What can public opinion on Twitter tell us about the Canadian political landscape in 2019?”

From the fist model NLP analyzation, there are three output plots can help us to understand the Canadian political landscape.

The hashtag words frequency distribution figure shows which topics have more attention.

The negative and positive sentiment words frequency distribution plot by bar figure and word cloud. Both showed the key words people feel negative or positive sentiment.

The party's sentiment frequency distribution shows how tweet's sentiment distribution for three different political parties.

Second model analyzation

The best hyperparameter tunned model in election negative reason prediction is Random Forest Classifier with TF-IDF feature, the accuracy is about 0.583.The second model accuracy is lower than the element sentiment prediction model. Which means my second model fail to predict the correct negative reasons. The reason can be following factors:

Small sample data size; The target negative reason features; Text feature extraction method

Both models suggestion

1.Increase the election tweet sample size. which is most importance factor in this assignment.

2.Increase the number of word features. Based on my computer limitation, I can support 500 features, more features will increase performance.