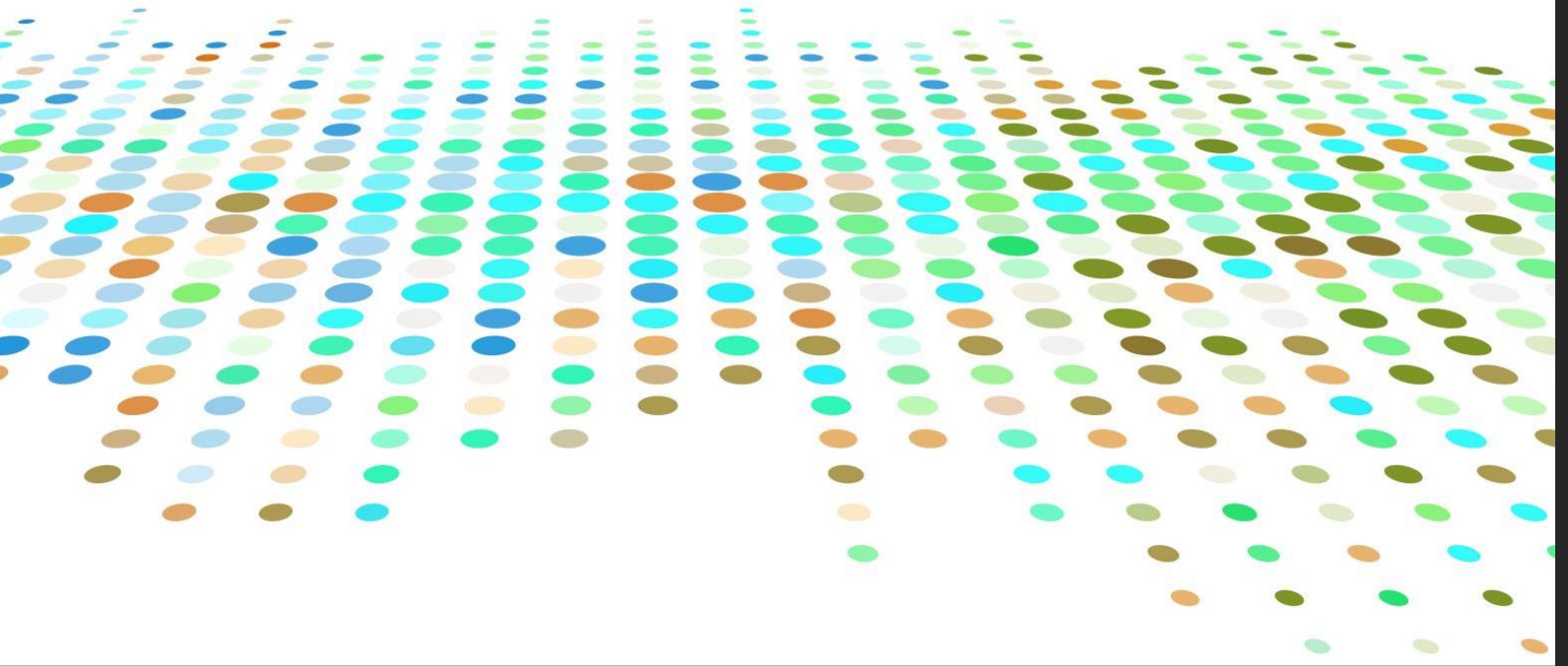


MIE 1624: INTRODUCTION TO DATA SCIENCE AND ANALYTICS

ASSIGNMENT 2

NAME – SHREYAS CHOUDHARY

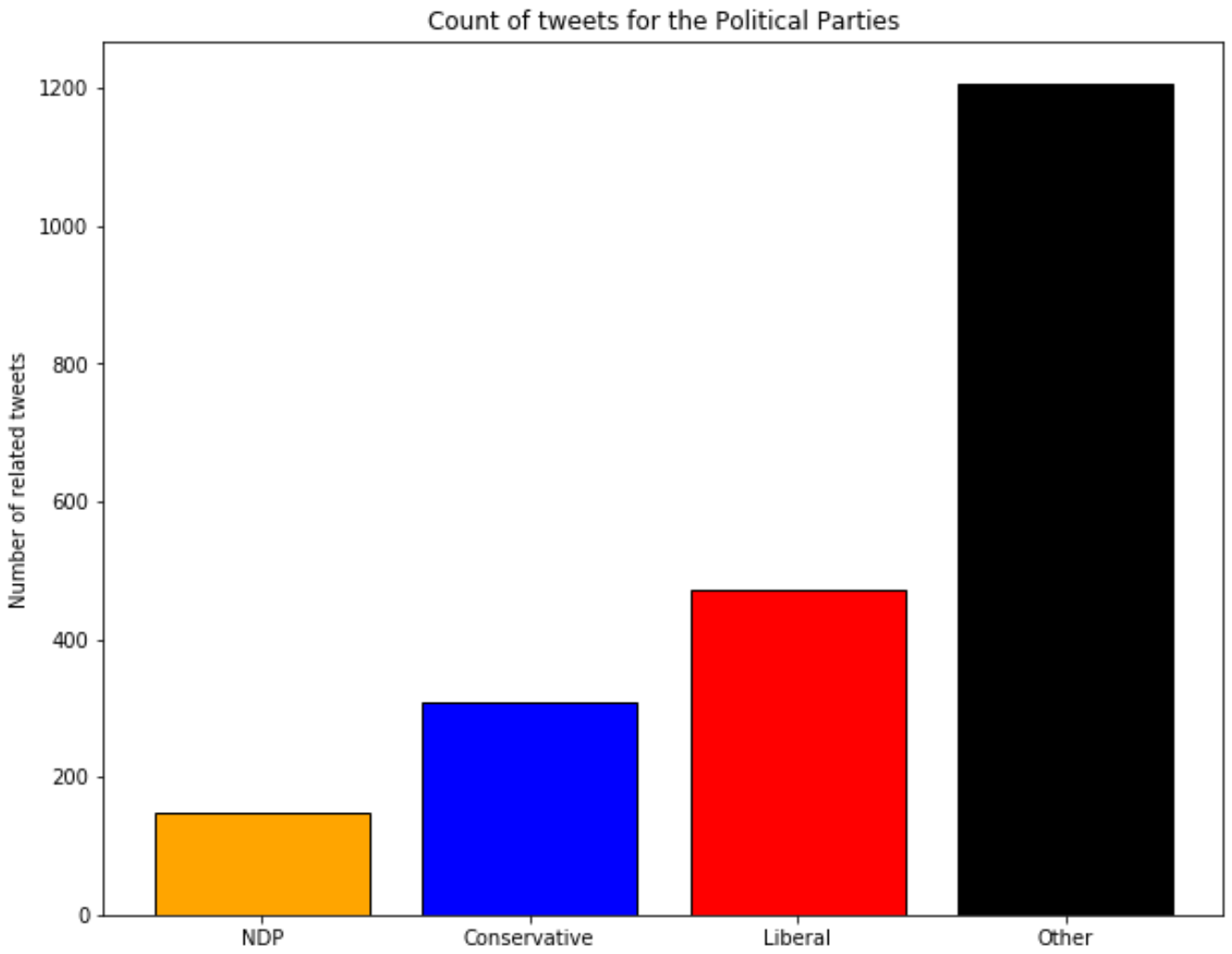
STUDENT NUMBER- 1006376217

DATA CLEANING

- Both the datasets contained many stop words, characters, people mentions, which needed to be gotten rid of as a part of cleaning the data for the analysis.
- Besides the stop words already existing in 'nltk' library, the stop words txt file provided for the assignment was also added to it in order to increase the list of stop words needed to be eliminated from the data.
- Hashtags were kept as it is and not eliminated and later a hashtag visualization was done which gave the results that corroborated with the sentiment analysis.

EXPLORATORY ANALYSIS

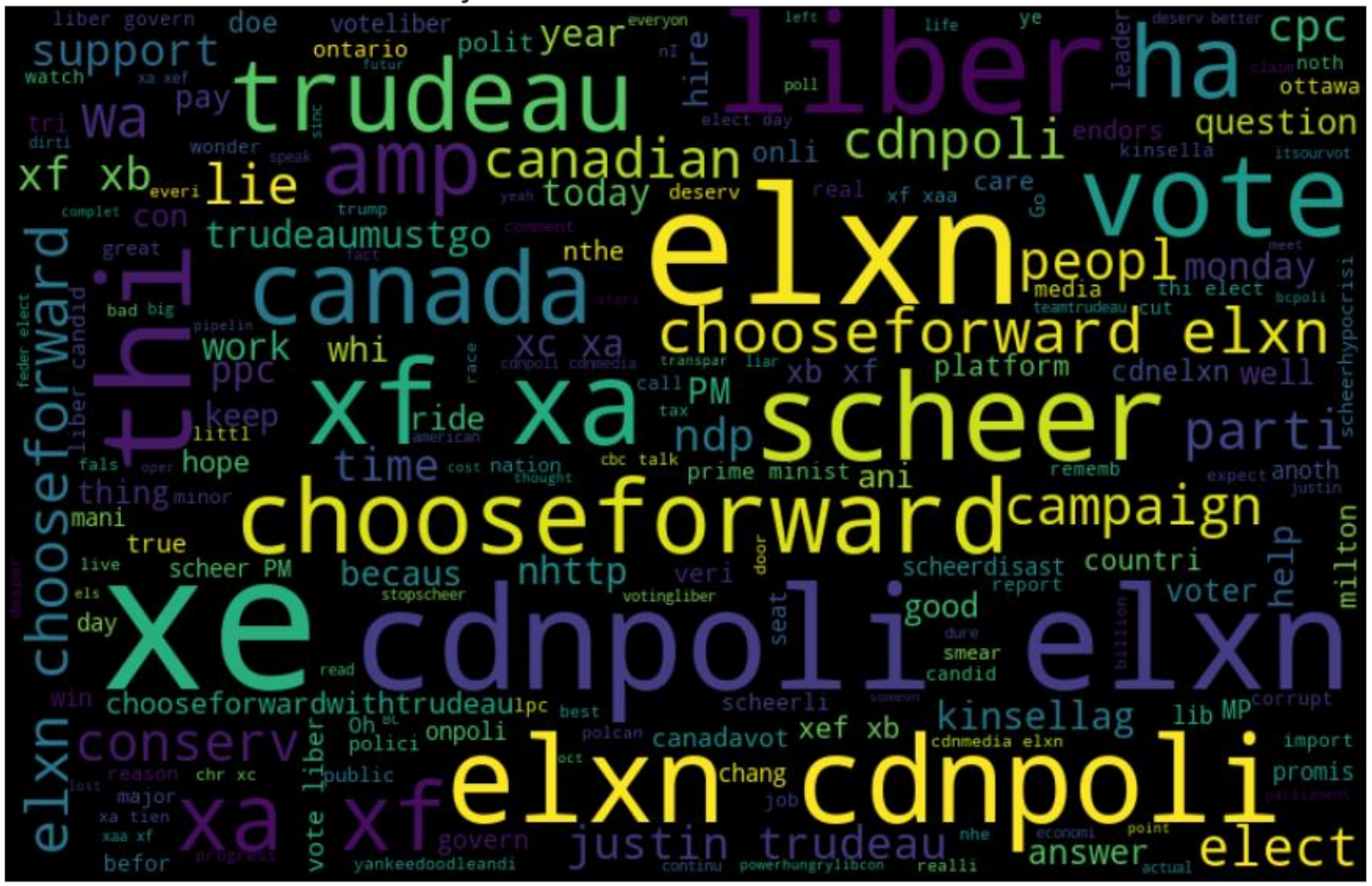
- First the tweets were categorized according to the major political parties contesting the elections. The main parties used were the Liberals, Conservatives, NDP and the rest was classified under the label of Others.



EXPLORATORY ANALYSIS

- Next, I constructed the word cloud in order to reflect on the keywords used more for the tweets according to each major parties contesting the elections. The figure to the right shows the word cloud obtained for the Liberals.

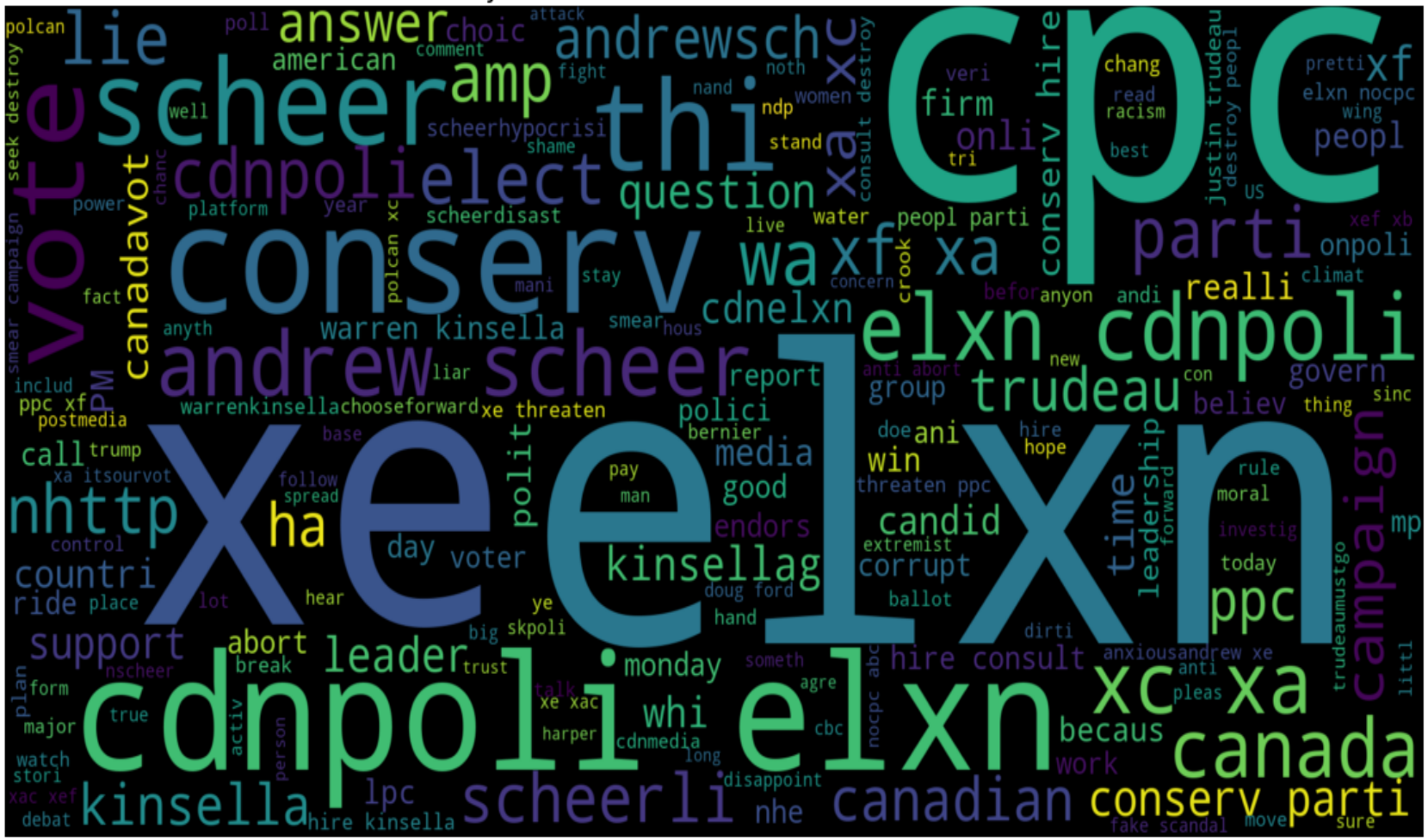
Keywords of Tweets about Liberal



EXPLORATORY ANALYSIS

- The figure to the right shows the word cloud obtained for the Conservatives. These keywords were also used to segregate the tweets about the parties initially.

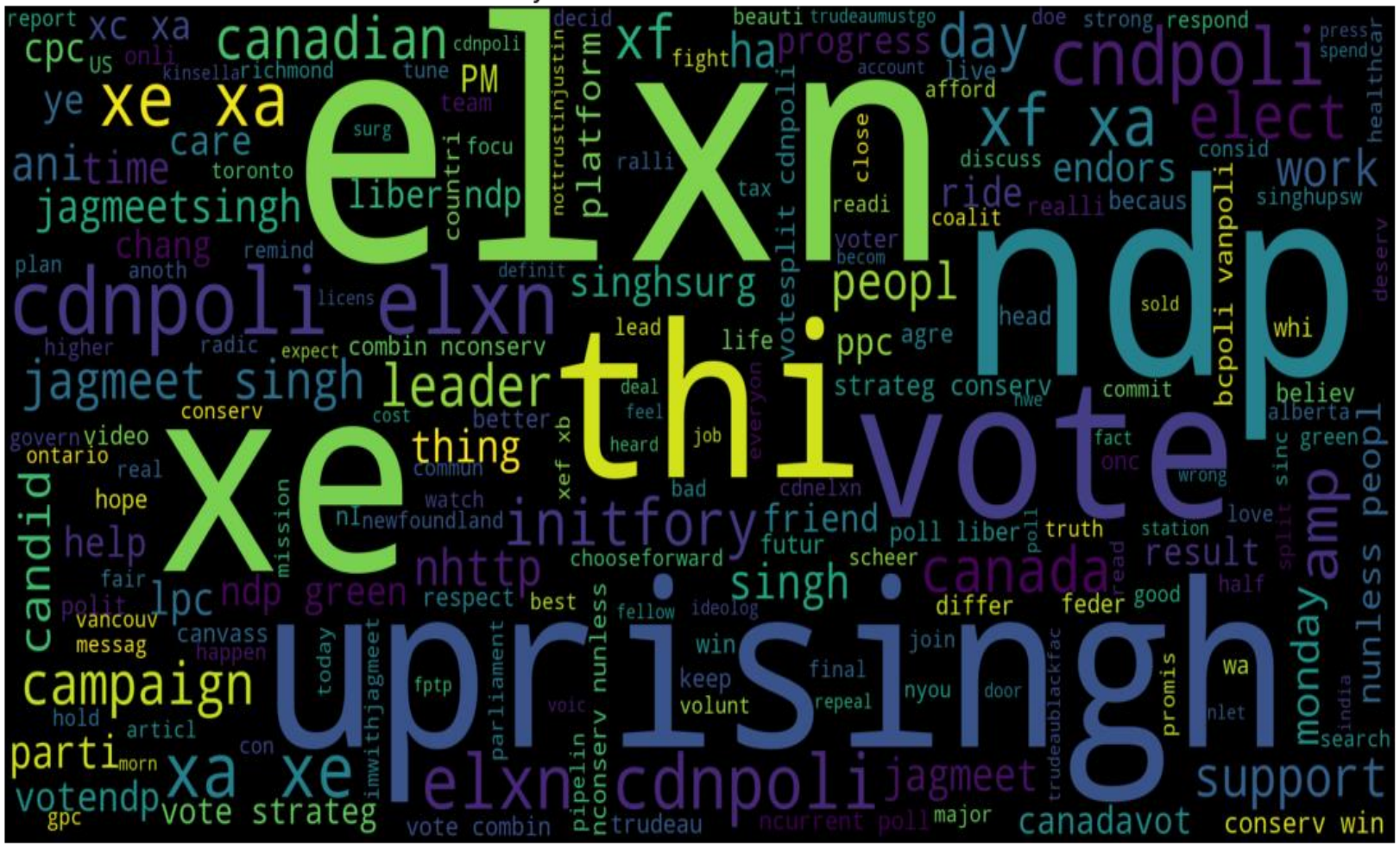
Keywords of Tweets about Conservaties



EXPLORATORY ANALYSIS

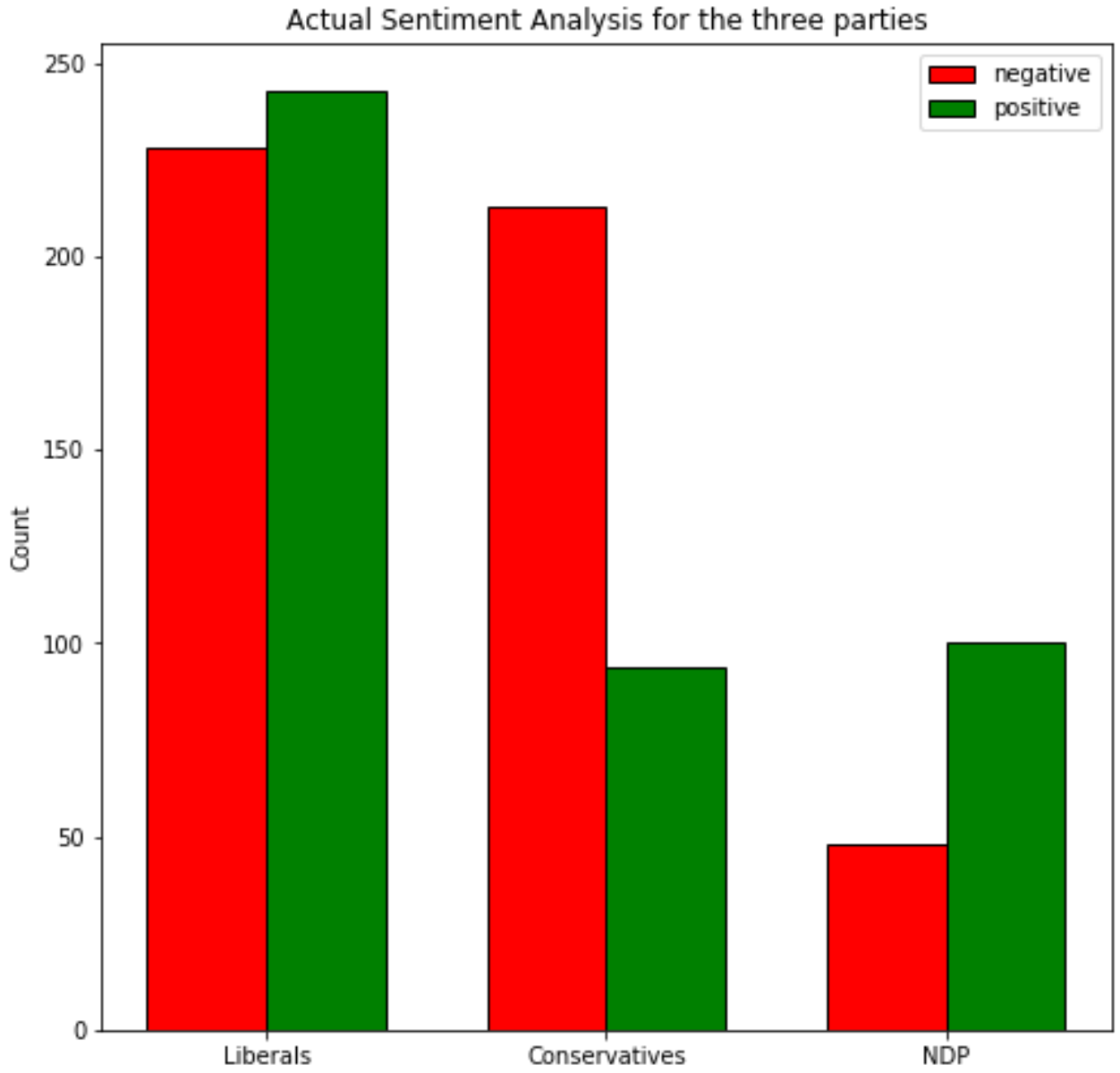
- The figure to the right shows the word cloud obtained for the NDP.

Keywords of Tweets about NDP



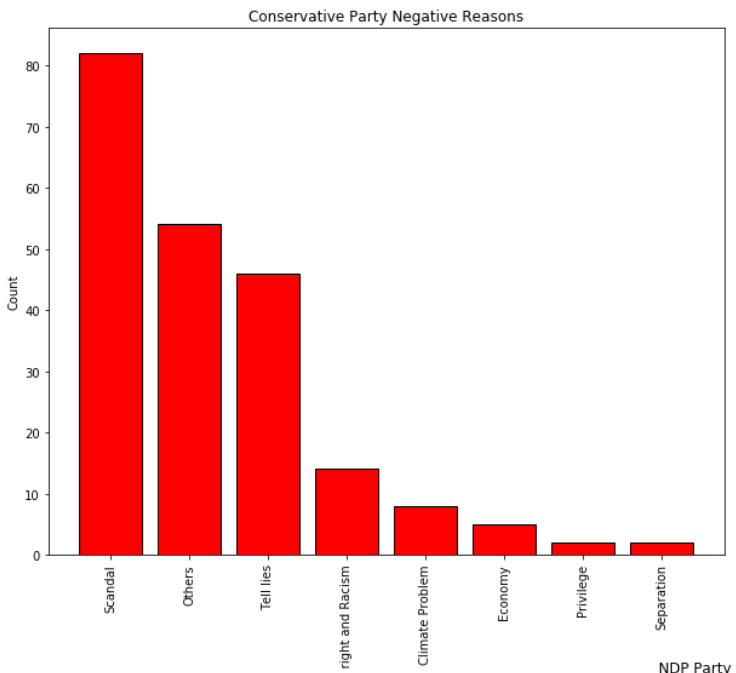
EXPLORATORY ANALYSIS

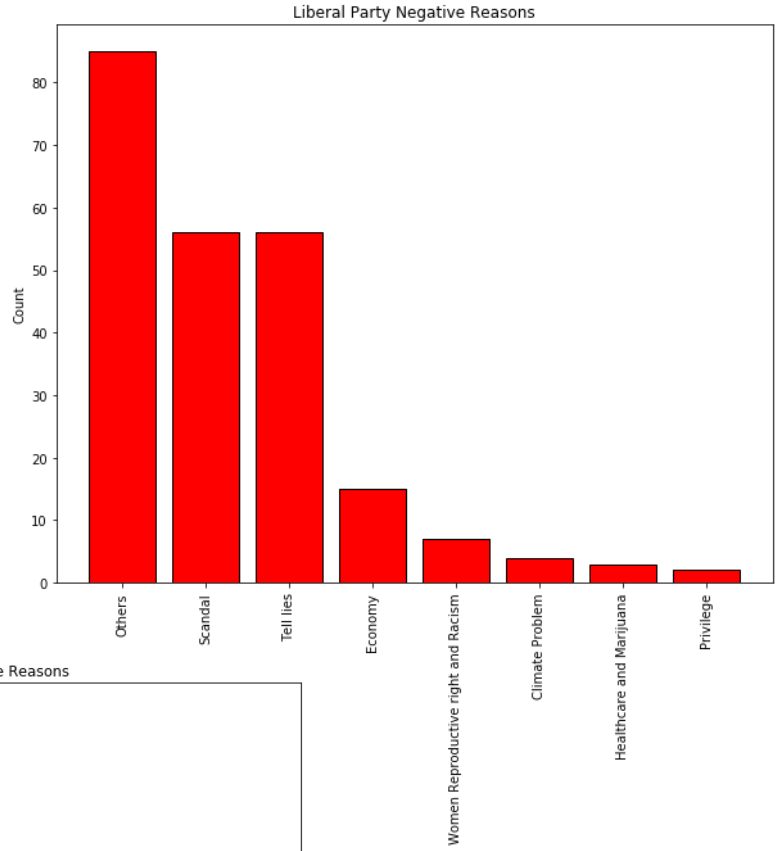
- Actual Sentiment model gave the segregated count of the positive and negative sentiment tweets according to the major parties which is shown in the figure to the right. This is further used to compare with the values predicted by our model.

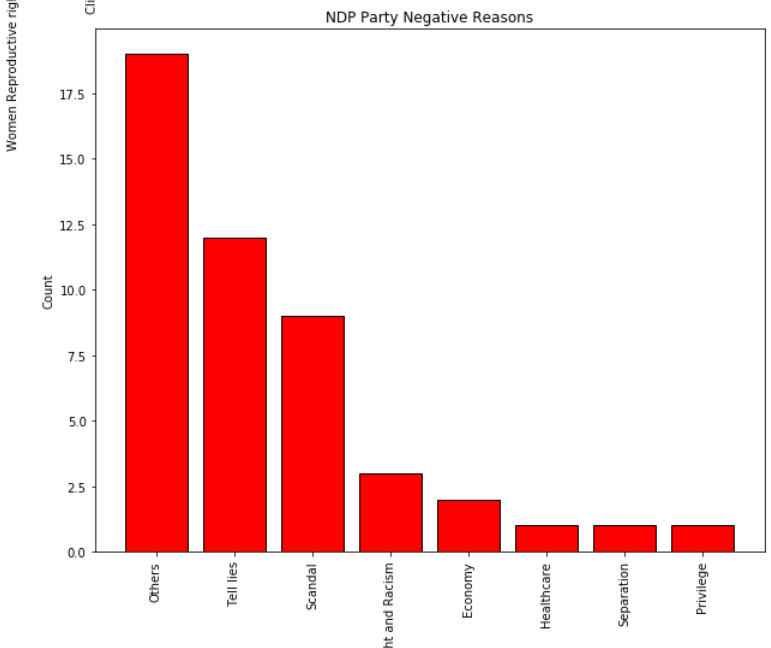


EXPLORATORY ANALYSIS

- Next, the negative reasons were assessed for each major party in the elections and the analysis gave some astonishing results. Most negative reasons for the conservatives were categorized as Scandals, that for the Liberals were Others and same for the NDP party. The reasons for NDP are significantly less than the other two parties and hence it really is not of much use in our analysis.

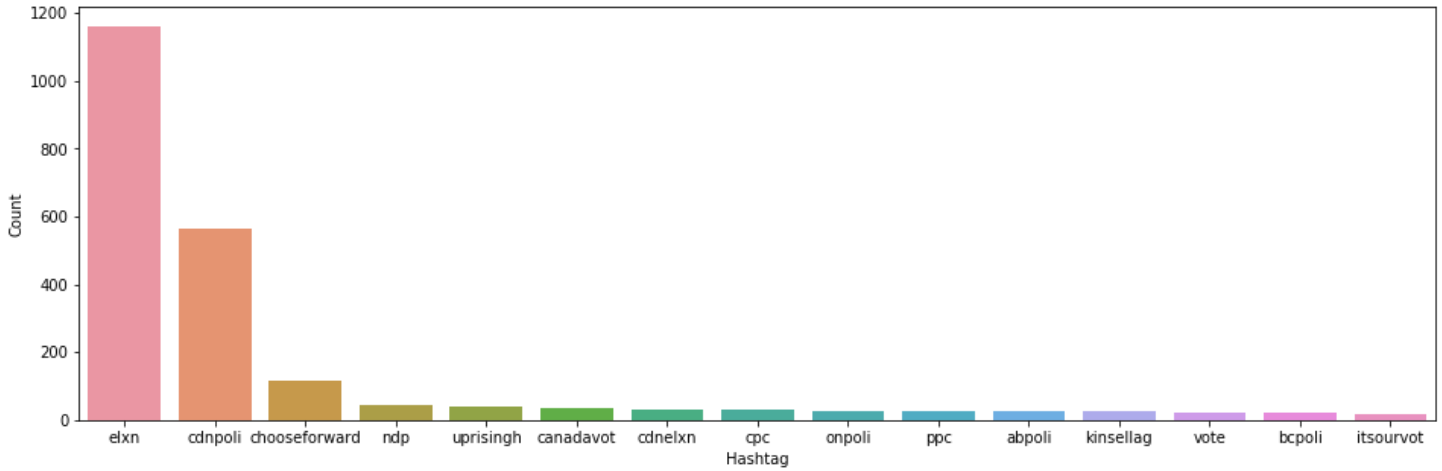




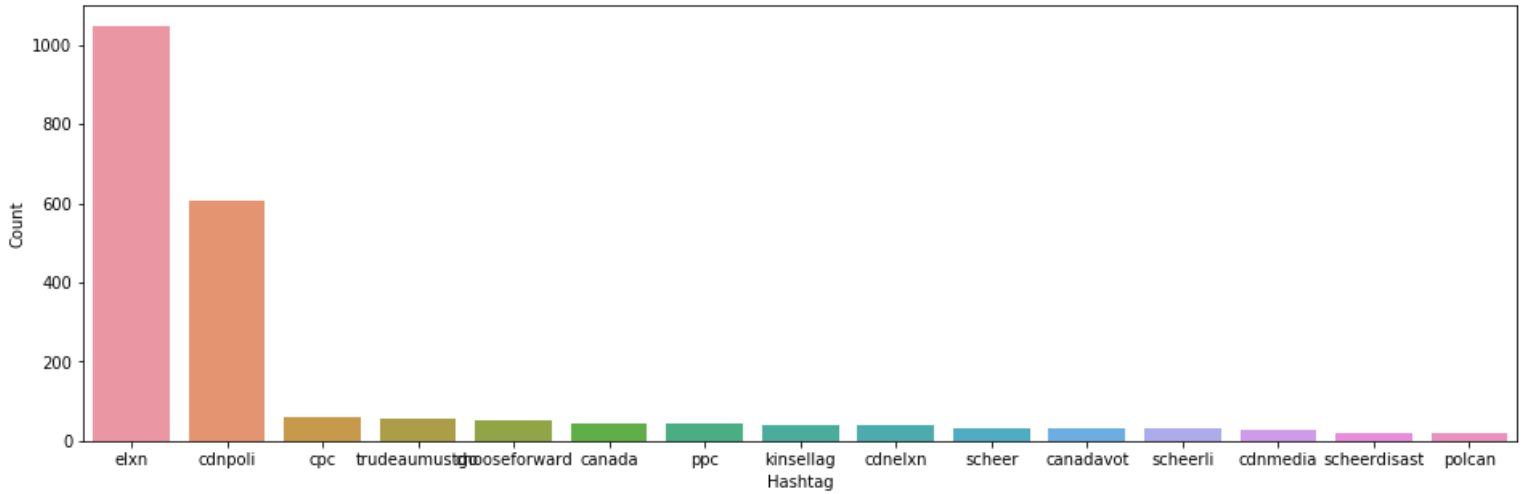


EXPLORATORY ANALYSIS

- As mentioned before, a hashtag visualization was done for the positive & negative sentiments in the dataset. The results corroborate with the all the results of our exploratory analysis. #elxn, #cdnpoli was common in both the graphs, but we can see that #chooseforward was the most used hashtag in the positive sentiment tweets which is related to the Liberals , whereas #cpc was the most used hashtag in the negative sentiment tweets, which is related to the Conservatives.



Positive sentiments Hashtags



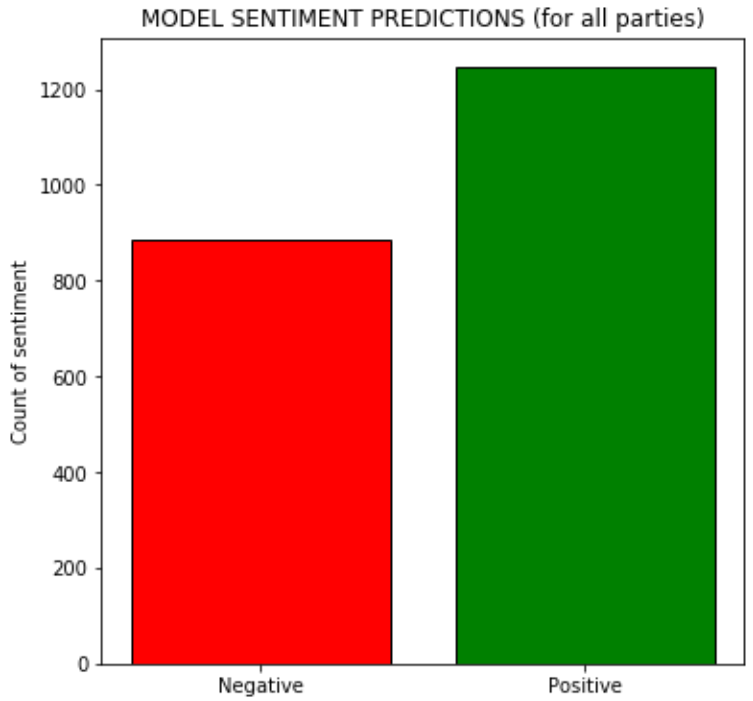
Negative sentiments Hashtags

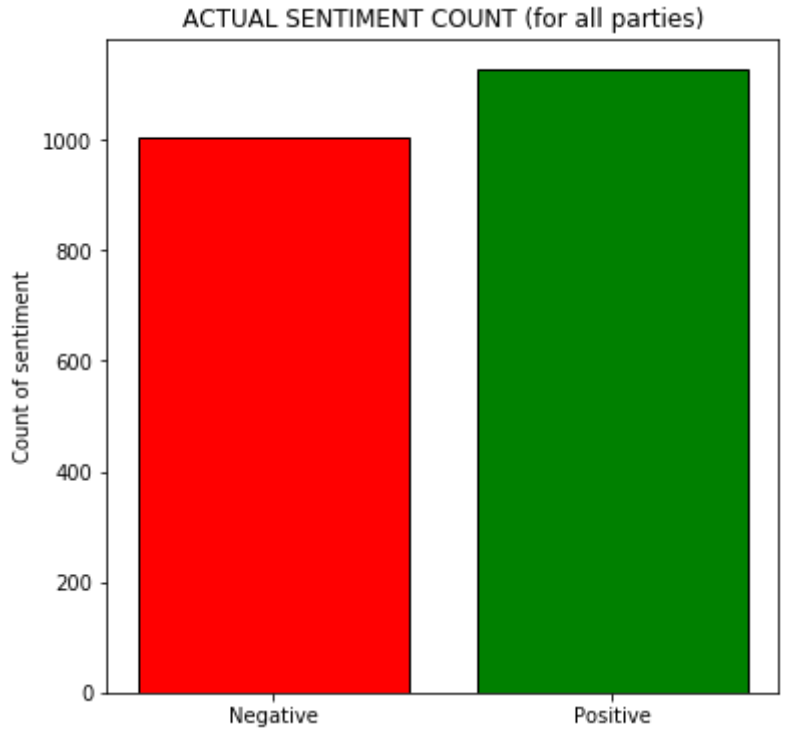
MODEL PREPARATION

- Worked on the generic sentiments dataset and applied various models (logistic regression, k-NN, Naive Bayes, SVM, decision trees, ensembles (Random Forest, XGBoost)). The Logistic regression model with CountVectorizer performed the best among all with an accuracy of 73.74%, which was further implemented on the Canadian elections 2019 elections dataset.

MODEL IMPLEMENTATION

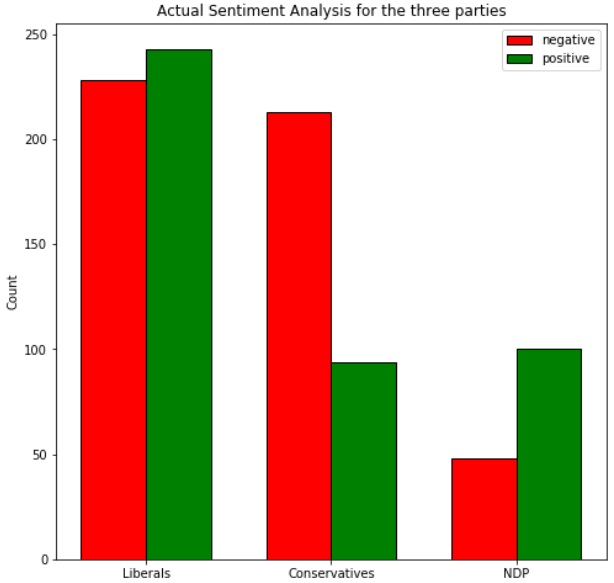
- The figures on the right show the actual sentiment count as well as the model predictions and we can say that the model fared well with respect to the actual one. The Logistic Regression model with CountVectorizer gave an accuracy of about 64% when implemented on the Canadian Elections 2019 tweets dataset.

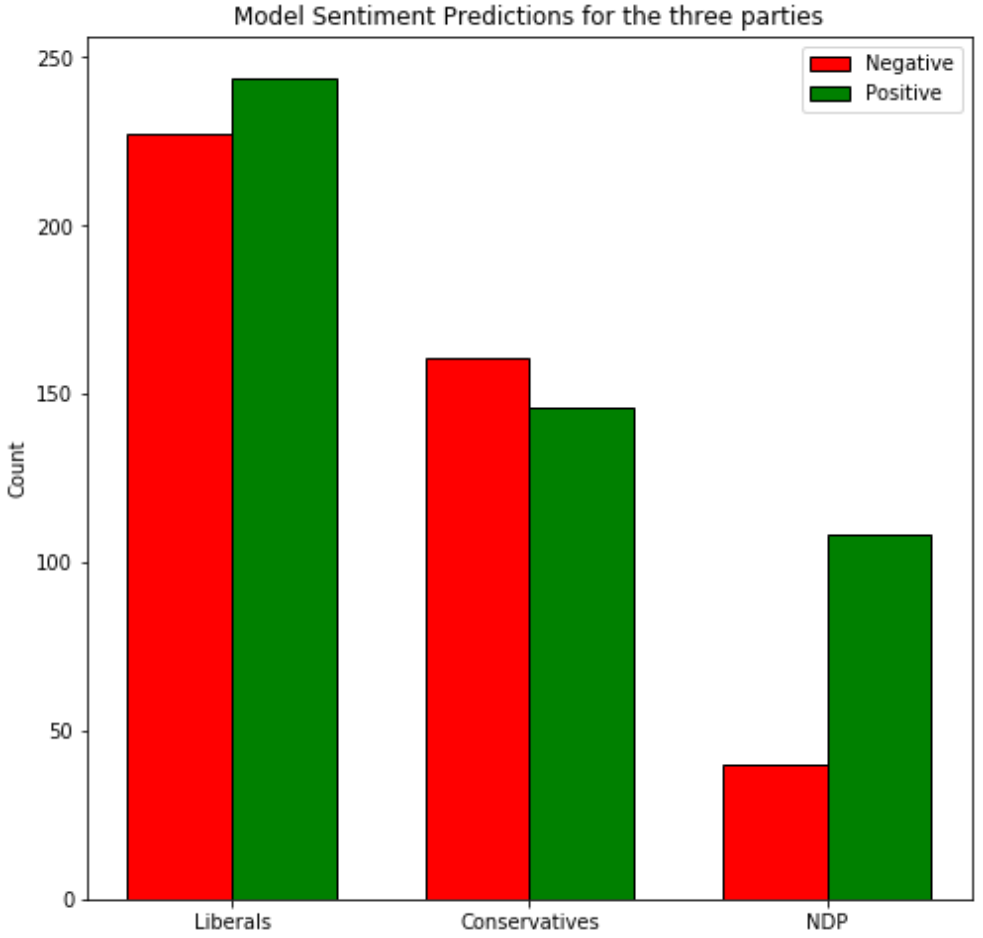




MODEL IMPLEMENTATION

- As stated before, the comparison of actual sentiments and the model predicted sentiments are pretty much the same. In the model predictions, some negative sentiment tweets have been classified under the positive sentiment tweets which can be because of the misunderstanding of some sarcastic language or emotion being used in the tweets, or failing to convert the emojis to the text or tokens for the analysis.





RESULTS

Explain how each party is viewed in the public eye based on the sentiment value

Liberals - The actual and the predicted sentiment predictions for the Liberals had some contrast. The actual sentiment model for the Liberals had more positive sentiment tweets than that in the predicted sentiment model. There were more tweets predicted negative in the predicted sentiment models. But altogether, the liberals seems to have been constantly in the limelight as the people were constantly talking about them. It is probably why the Liberals were able to nail the elections.

Conservatives - The same trend of the model failing to classify the negative tweets properly was observed. However, the actual positive tweets observed are still lower than that of the Liberals. They are seen as negative in the eyes of the people on social media, which the hashtag visualizations done above also suggest. It is probably why they were left behind by the Liberals in the recent elections.

NDP: Despite the overall count of the tweets being low, the model did a better job in classifying the tweets for this party. This means that NDP was not as much discussed on the social media as compared to the Liberals and the Conservatives, which gives an idea about their popularity in the country. However, they are the only party amongst the three who have more positive tweets than negative ones. NDP could use this information to focus on more reasons to concentrate on their shortcomings and could fare well in the next elections.

RESULTS

Discuss whether NLP analytics based on tweets is useful for political parties during election campaigns.

Sentiment analysis always has been used quite often in data science, mainly for retrieving the sentiments of the customers based on the reviews of the products or a service. This plays an essential role in providing the essential areas where the product or the service can be improved, thus giving a superior product/service. One way that NLP analytics can be used is in the sentiment analysis on the tweets before any big event, as Canadian Elections in our case. It can be used to detect the current mood of the people as many of the people nowadays resort to online platforms like twitter to express what they feel about a certain event. The sentiment analysis done for an election campaign can be used in many ways by the parties participating, they could analyze the negative sentiment , if it exists and can also extract reasons for the sentiment. They can use this information to prioritize the issues and concerns of the public sentiment and can influence elections to a large extent.

A good example of how sentiment analysis can be closely related to the actual outcome of the elections can be seen from our example. The negative reasons behind the negative tweets for the conservatives were mainly the leaders of the party involved in the scandals, telling lies and some other reasons. This could actually be one of the major reasons why Liberals got an upper hand in the recent elections. But at the same time same reasons propped out for the negative tweets on the Liberals which probably could be the reason they weren't able to gain a full majority in the parliament like the last time. Such information can play a crucial role in determining major things such as the agenda of the elections.

RESULTS

How can we improve the accuracy of the models we implemented?-

---> Sentiment Predictions tweets model accuracy could be increased by:

- Incorporating the emoticons and translating them into words so it would help assess sentiment more precisely.
- Updating the dataset used to prepare model(in our case Sentiments dataset) to match the current and recent trends in order to improve the accuracy of the model implementation on the target dataset(in our case Canadian Election 2019 dataset).
- Using the hyperparameter tuning for the models that are implemented (in order to identify optimal parameters which in turn can help in boosting the accuracy of the model).

---> Canadian Elections 2019 tweets model accuracy could be increased by:

- Using more advanced ML algorithms such as deep neural networks for more accurate results of the sentiment analysis.
- Creating more data in the form of more categories in order to create a balanced dataset. For instance having some more reasons in the others category of the negative reasons in our example.
- Using the hyperparameter tuning for the models that are implemented (in order to identify optimal parameters which in turn can help in boosting the accuracy of the model).
- In our case, accuracy can also be increased by grouping the more weighted options (the first three options in the negative reasons in our case), but this would fail to account for rest of the categories and even if we'll get the accuracy boosted, it would still be on the skewed dataset.

RESULTS

For the second model, based on the model that worked best, provide a few reasons why your model may fail to predict the correct negative reasons. Back up your reasoning with examples from the test sets.

As already mentioned above that the prediction of the correct negative reasons could be because of number of reasons getting concentrated in the top three reasons out of which, the 'others' reasons is non definitive. It would have helped if that reason could have been broken down further into some other options and then we could have been able to regroup the reasons more effectively, thus contributing in improving the accuracy of the model. Further, we can explain this by creating a confusion matrix of the reasons, so as to support the above mentioned reasoning further.

Since accuracy may not be a good measure when the dataset is skewed, confusion matrix is plotted which allows us to understand how many times were the predicted and the true labels the same.
As can be seen from the confusion matrix above, the data in majority is being classified in mainly three reasons (which account almost 85% in the dataset) and therefore is contributing in creating bias in the model. In particular, the 'others' reason listed in the negative reasons seems to go overboard in dominating, so it can be broken down further into individual reasons so that the split is balanced and that would also help in an efficient regrouping as we did above, which could further lead to an increase in the accuracy of the model.
Since, three reasons(other, scandals and telling lies) constituted almost 85% of the reasons column in the dataset, as discussed above it can lead to a skewed dataset, which further could've resulted in the accuracy being somewhat compromised.