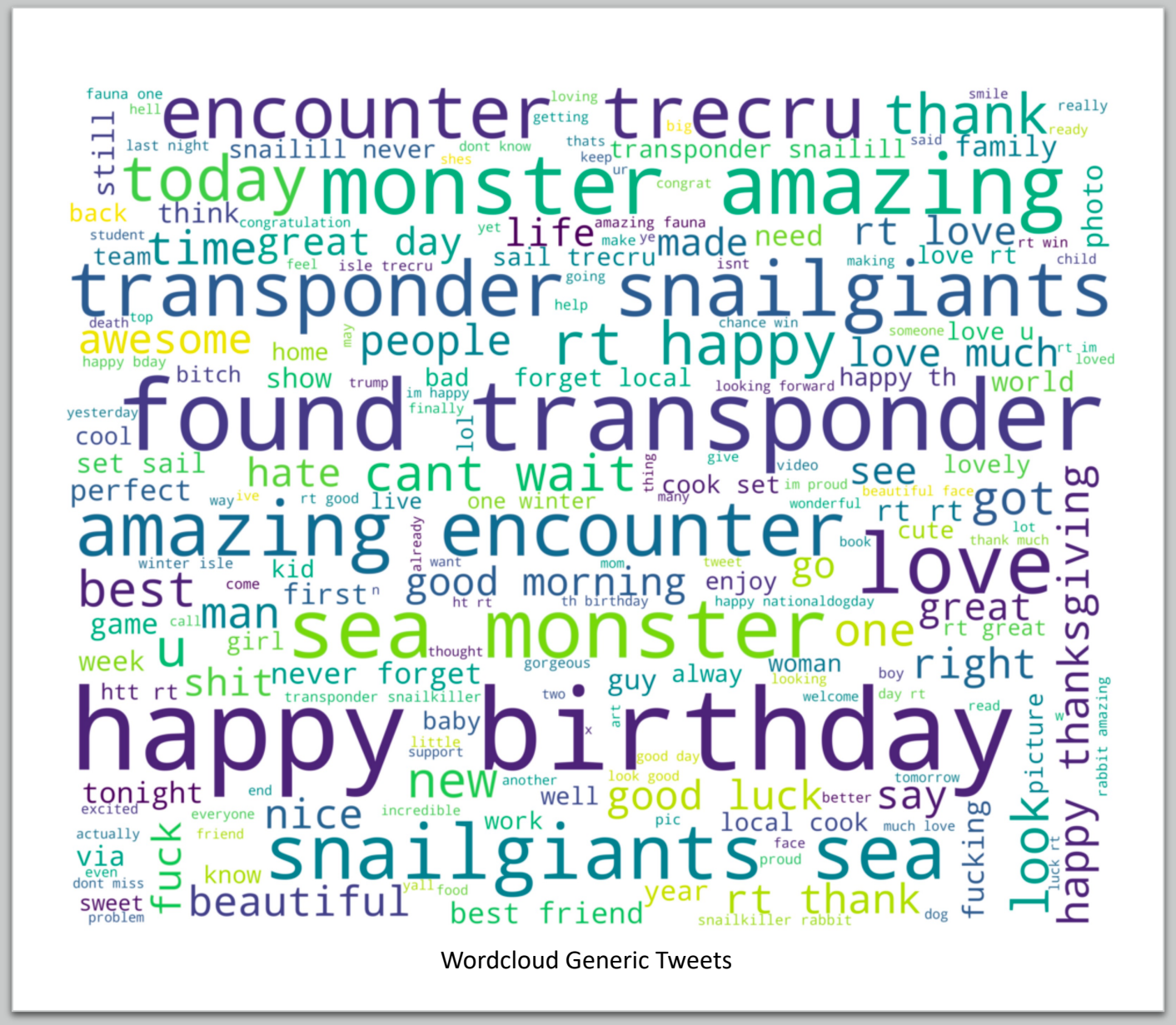
ASSIGNMENT 3

- **INTRODUCTION**: The goal of this project is to essentially use sentiment analysis on Twitter data to get insight into the 2019 Canadian Elections.
- **EXPLORATORY DATA ANALYSIS**: The wordcloud visualization for the sentiment analysis dataset showed that the common words in generic tweets are not related to elections. Common words are happy birthday, transponder, snailgiants sea, encounter, amazing encounter. This is an indication that public opinion on twitter has little to no connection with the 2019 Canadian election.



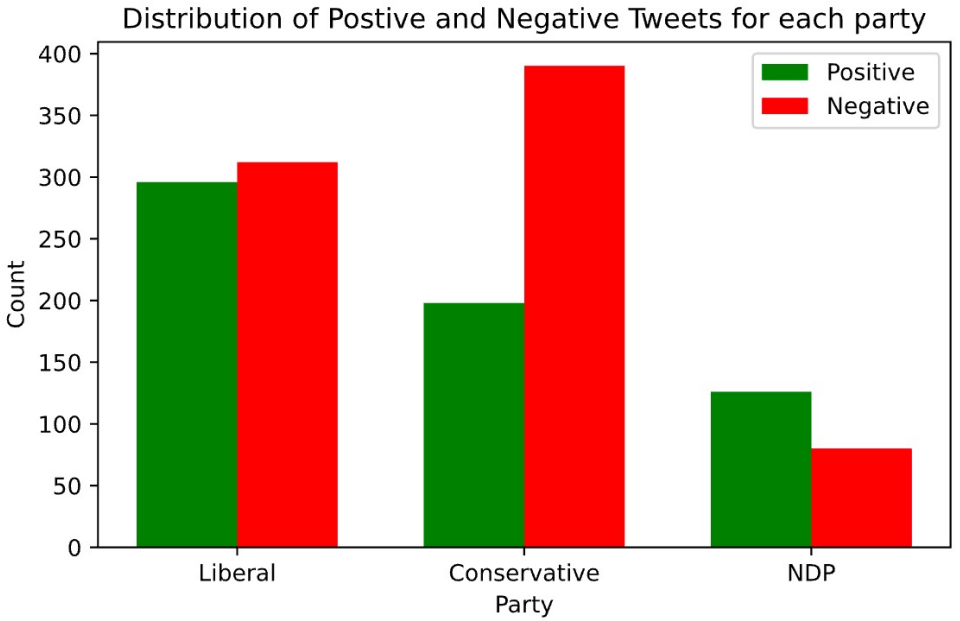
Wordcloud Generic Tweets

EXPLORATORY DATA ANALYSIS `CONTINUED

- The wordcloud of the Canadian election dataset shows common words like elxn43 (43rd election in Canada), cdnpoli (Canada Politics), slogans for different parties (trudeaumustgo, scheerlies, uprisingh).

Wordcloud Canadian Election

- The negative tweets of the Conservative Party double their positive tweets. There are few tweets about the New Democratic Party as compared to the other two parties but most of them are positive. The negative tweets about the Liberal Party is slightly more than its positive tweets.



MODEL RESULT: SENTIMENT ANALYSIS PREDICTION

- The RandomForestClassifier has the best performance among the other trained models. It has an accuracy of 94.4% on the TF-IDF feature of the generic tweets.
- Generally, the models performed better on the TF-IDF features than the Word Frequency feature.

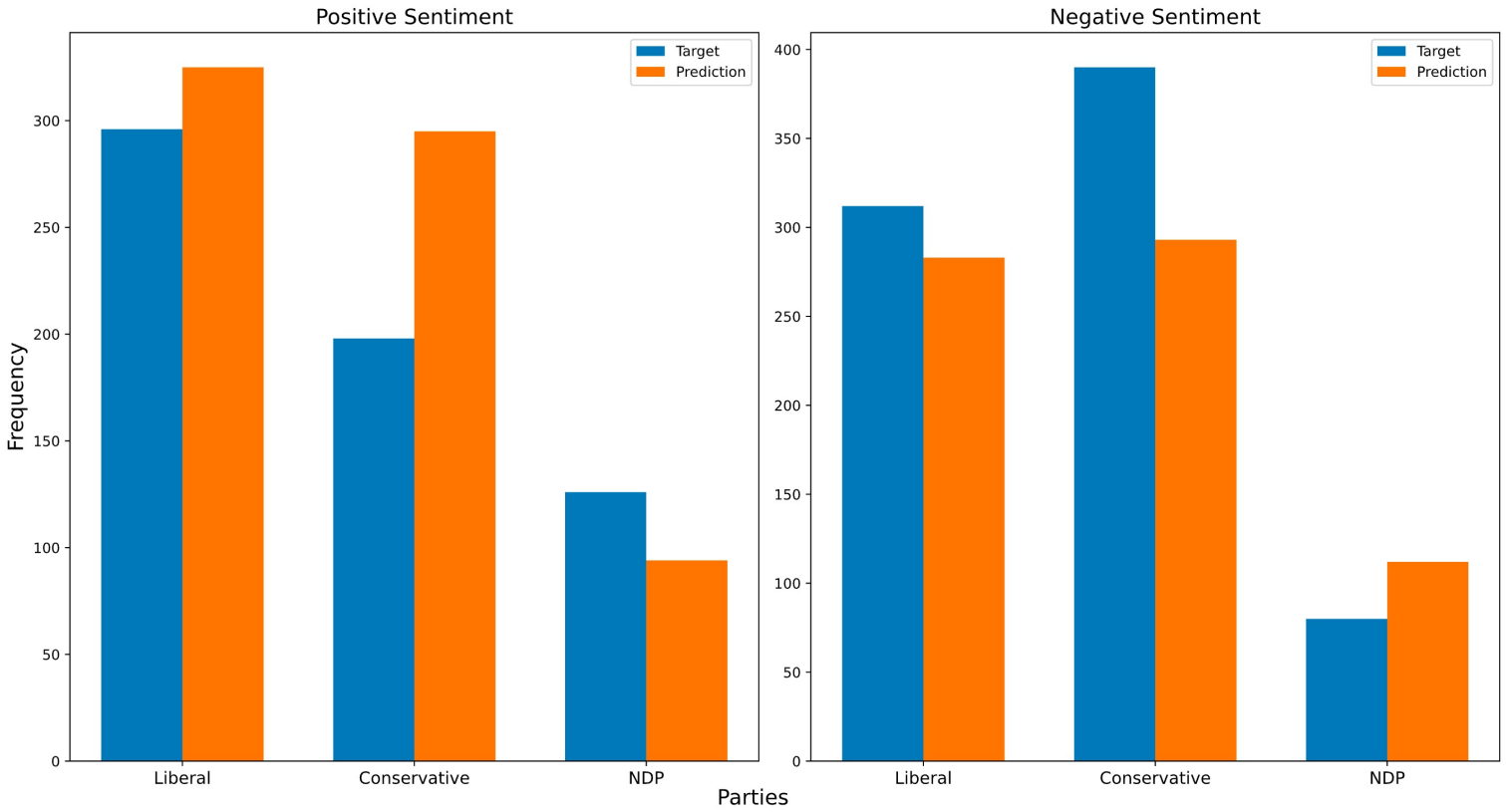
Model accuracy on BoW and TF-IDF feature fro generic tweets

	model	bow_score	tfidf_score
0	Logistic Regression	0.9432	0.9424
1	k-NN	0.9142	0.8722
2	Naive Bayes(MutinomialNB)	0.9075	0.8977
3	SVM(LinearSVC)	0.9424	0.9422
4	Decision Tree	0.9273	0.9291
5	Random Forest	0.9416	0.9438
6	XGBoost	0.9244	0.9248

MODEL RESULT: SENTIMENT ANALYSIS PREDICTION `CONTINUED

- The accuracy of the RandomForestClassifier model after hyperparameter tuning on the Canadian elections dataset was 52.4%.

- The model predicts more positive sentiments than there is for liberal and conservative and less positive sentiments for the NDP.



Model Sentiment Prediction VS True Sentiment for each Party

MODEL RESULT: NEGATIVE REASONS PREDICTION

- The best model was the SVM LinearSVC model with accuracy of 54.8%. The parameters of this model was further tuned to get the best performance out of the model. The accuracy of the model after hyperparameter tuning was 53.8%. The higher value of the default model is due to overfitting.

	model	score
0	Logistic Regression	0.5149
1	Naive Bayes(MutinomialNB)	0.5116
2	SVM(LinearSVC)	0.5479

- Among the three multi-class classification algorithm that was implemented, the SVM(LinearSVC) had the best performance with accuracy of 53.8%. From the confusion matrix plotted, it can be observed that the model performed well in predicting the Scandals and Others negative reasons but didn't perform well on the other categories.