This project cleaned 550,390 labelled tweets as the training and validation dataset, transformed them into features by bag-of-words and TF-IDF transformers and trained ML classification models including logistic regression, SVM, decision tree and random forest. The best model was the random forest model with TF-IDF transformer, with a 0.949 validation accuracy.

The 1002 tweets regarding the election were firstly classified for their party affiliations, then transformed using the TF-IDF transformer applied to the training dataset, and predicted for their sentiments (negative or positive) using the trained random forest model. It was found that Liberal was the most popular, followed by CPC, leaving PPC and NDP far behind.

For those tweets with negative sentiment, by using 30% of the dataset with labelled reasons to train another random forest classifier with TF-IDF transformer, the negative reasons for the rest 70% tweets were predicted. The accuracy of this model could potentially be improved if the dataset were larger. The top reasons found were cancelling election early, lying and Covid.