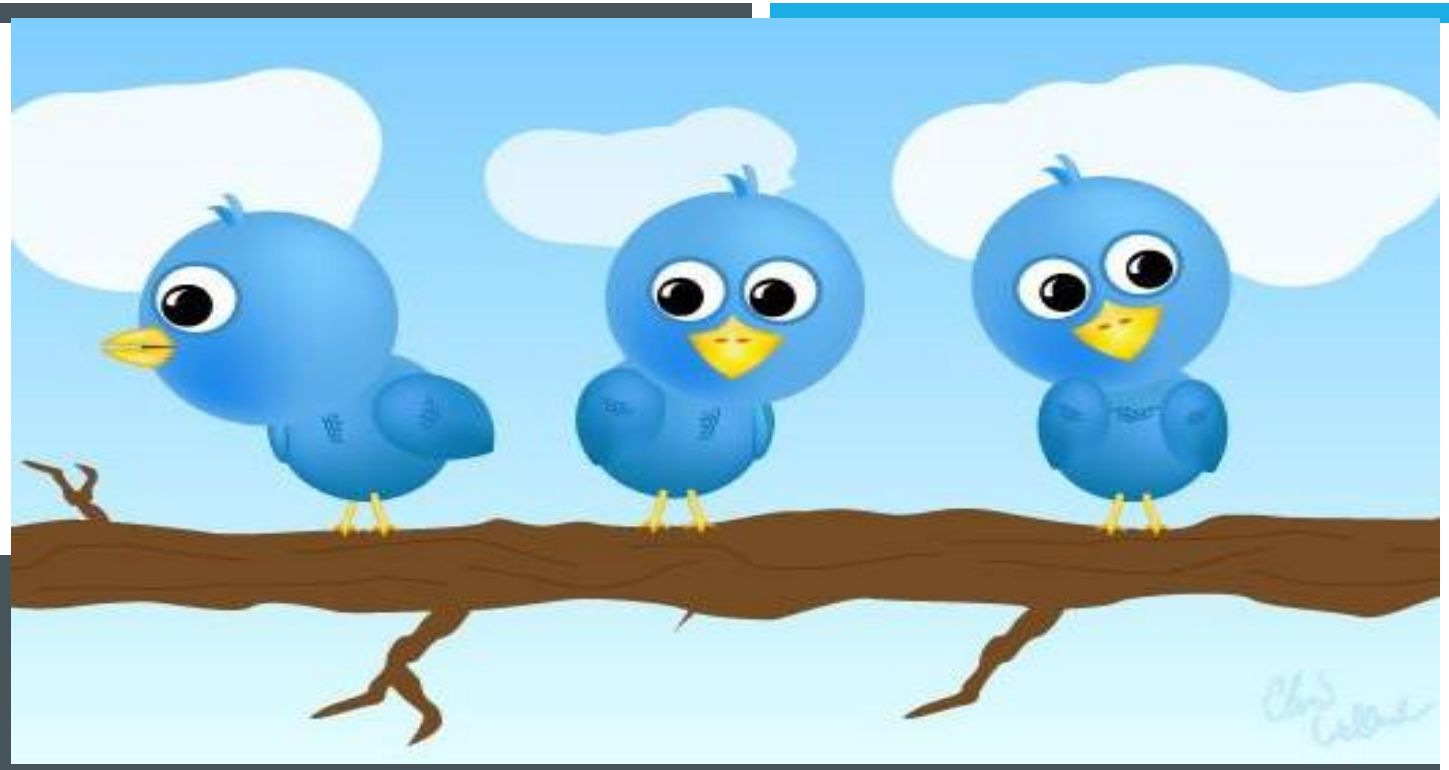
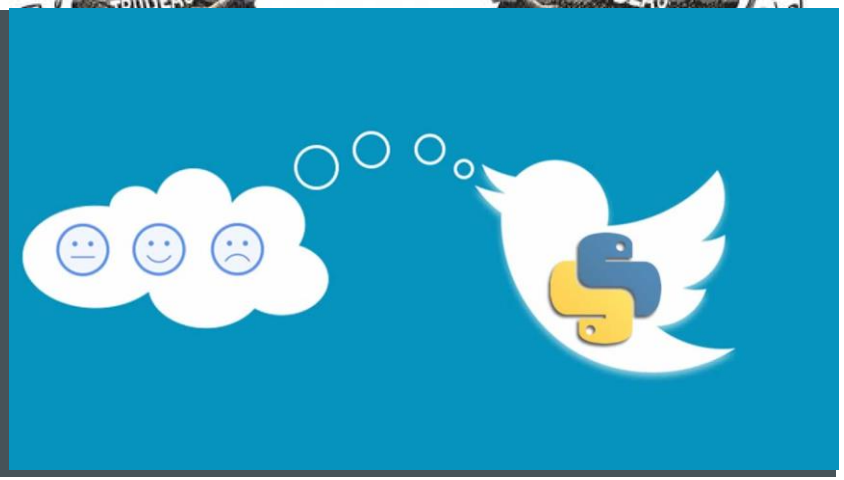


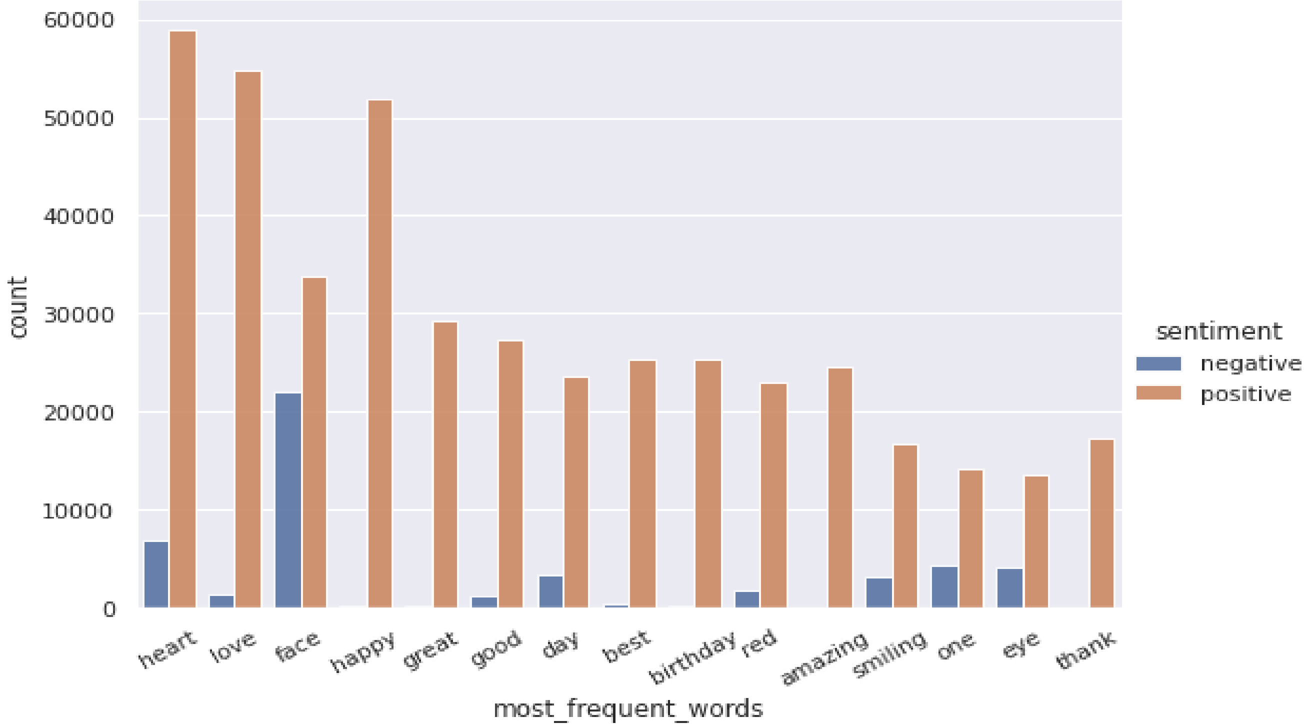
TWITTER SENTIMENT ANALYSIS

MARYAM MOADELI

1004824283



EXPLORATORY ANALYSIS OF GENERIC TWEETS



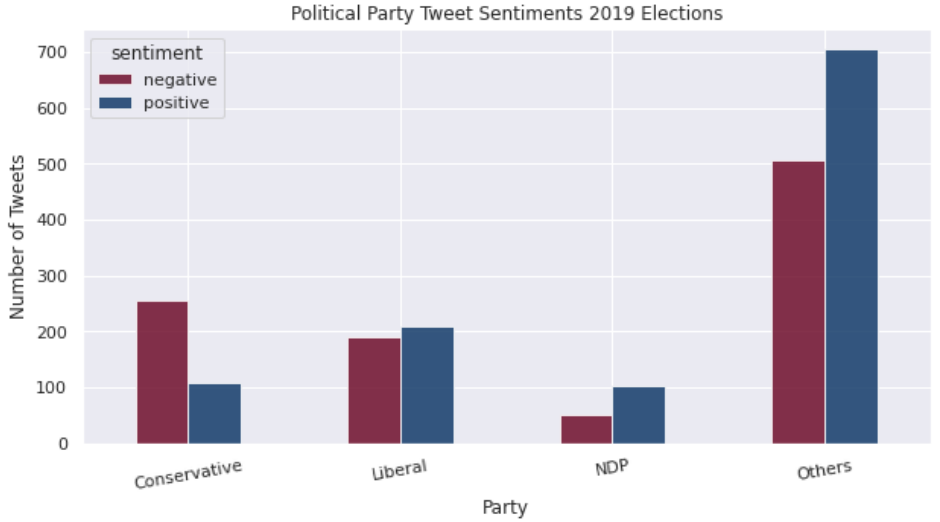
- As we see, words with positive sentiments such as happy, love, heart, best have higher positive counts in most frequent words plot that are extracted from tweets
- words in word cloud are totally representative of positive and negative sentiments in tweets, especially as we don't delete emojis and we substitute them with words, For example:

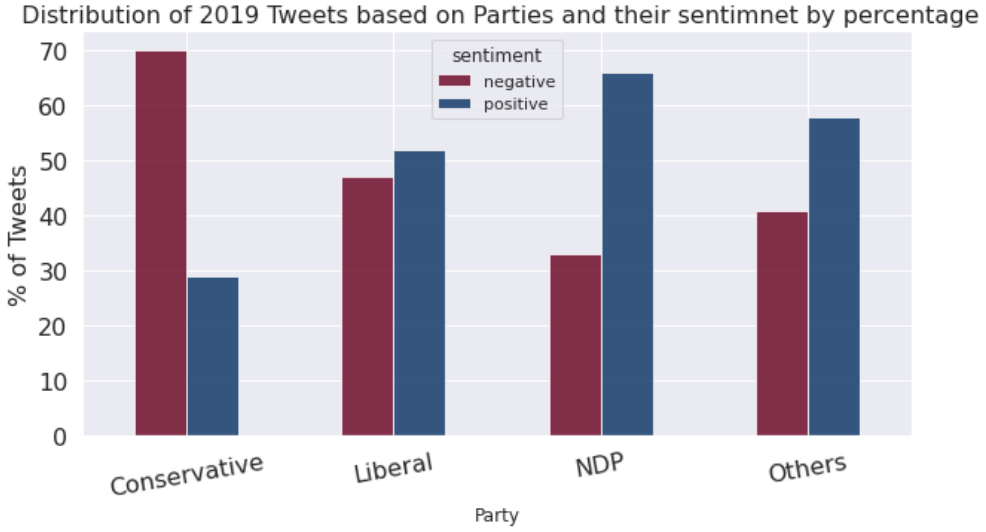
  Positive emotion words: read heart, smiling face, love, nice

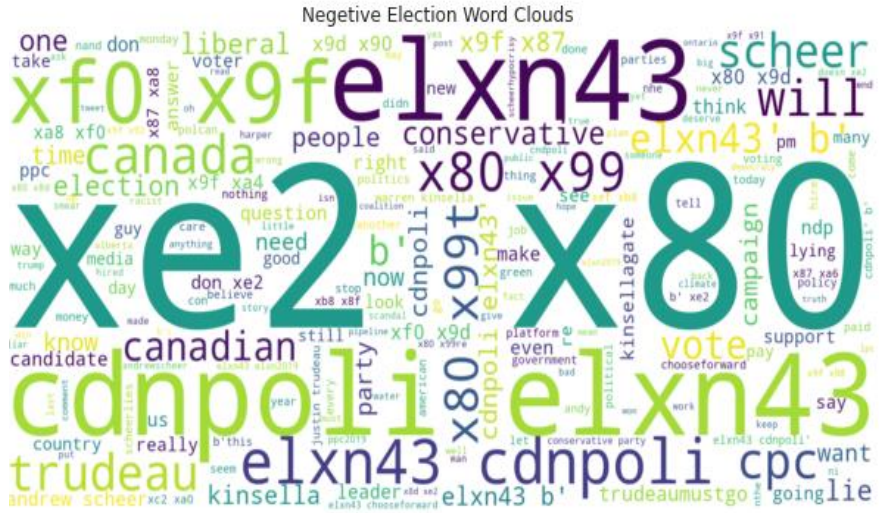
  Negative emotion words: tear, cry face, hate, hate

EXPLORATORY ANALYSIS OF ELECTION DATASET

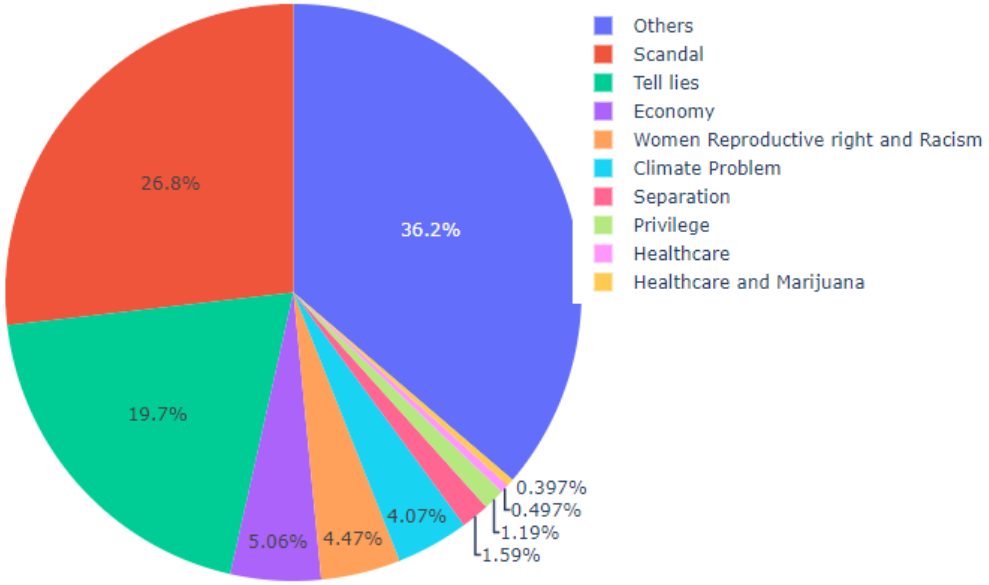
- As we can see, most of the tweets belong to conservative and liberal parties
- For liberal party, higher proportion of tweets are positive
- For conservative higher proportion of tweets are negative.
- NDP has a smaller number of tweets in comparison to the other two parties
- The outcome of the election was Liberal Party in 2019 that is aligned with our result.
- According to our analysis, Liberal tweets was found to be the most numerous with highest percentage of positive sentiment, while the conservative that went second in the election had highest rate of negative tweets
- Word could also show most n egative public's opinion were t oward major parties (Liberal an d Conservatives)

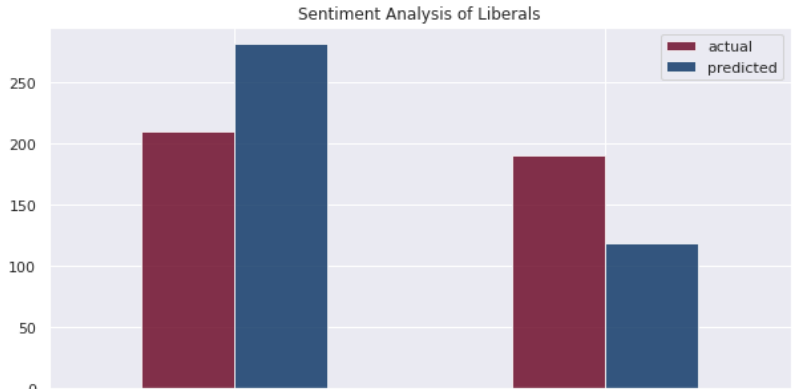


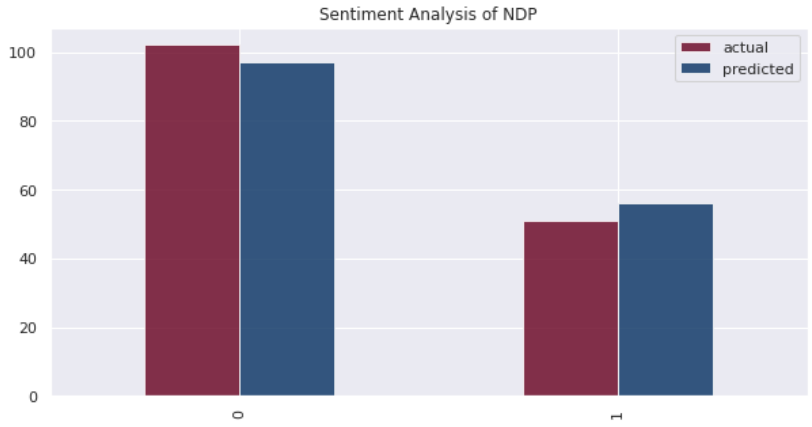


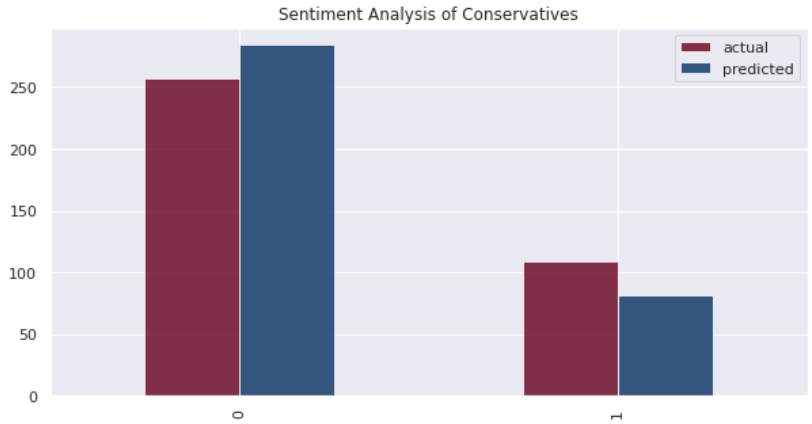


ANALYSIS OF NEGATIVES REASONS AND PREDICTION FOR PARTIES









- Models predict positive tweet for liberals and conservative less than actual, while it predicts positive for NDP more.
- According to pie plot, five major reasons are mentioned for negative tweets, which are scandal, tell lies, economy, Women Reproductive right and Racism and Climate Problem

FEATURE IMPORTANCE AND MODEL RESULTS

	LosgisticRegression	KNNClasifier	NaiveBayes	SVM	DecisionTree	RandomForest	XGBoost
Bag Of Words	94.404	90.055	90.366	93.826	91.073	61.598	83.721
TF-IDF	94.420	81.992	88.487	93.881	90.920	60.237	83.743

- As we can see from the results, the highest accuracy belongs to the logistic regression classifier with the TF-IDF features, and the other models are Logistic Regression - SVM - Decision Tree in sequence
- We take a look at the features that mainly contribute to negative sentiment versus the ones that mainly contribute to the positive ones using feature importance on our logistic regression model
- Based on feature importance, the words that are extracted as most important features are totally relevant to positive and negative tweets in consequence

	TF_IDF_features	coefficients
809	happy	24.122728
773	great	19.499455
50	amazing	18.817661
1060	love	17.465300
1760	thank	17.082163
157	best	15.852614
587	excited	14.348990
146	beautiful	14.036502
1370	proud	13.717582
760	good	13.695143

	TF_IDF_features	coefficients
814	hate	-9.430379
120	bad	-8.311460
1481	sad	-8.046466
1965	worst	-7.469636
717	fuck	-6.334693
881	hurt	-6.276247
451	death	-6.224464
1964	worse	-6.191434
1357	problem	-6.178231
1688	stupid	-6.139757