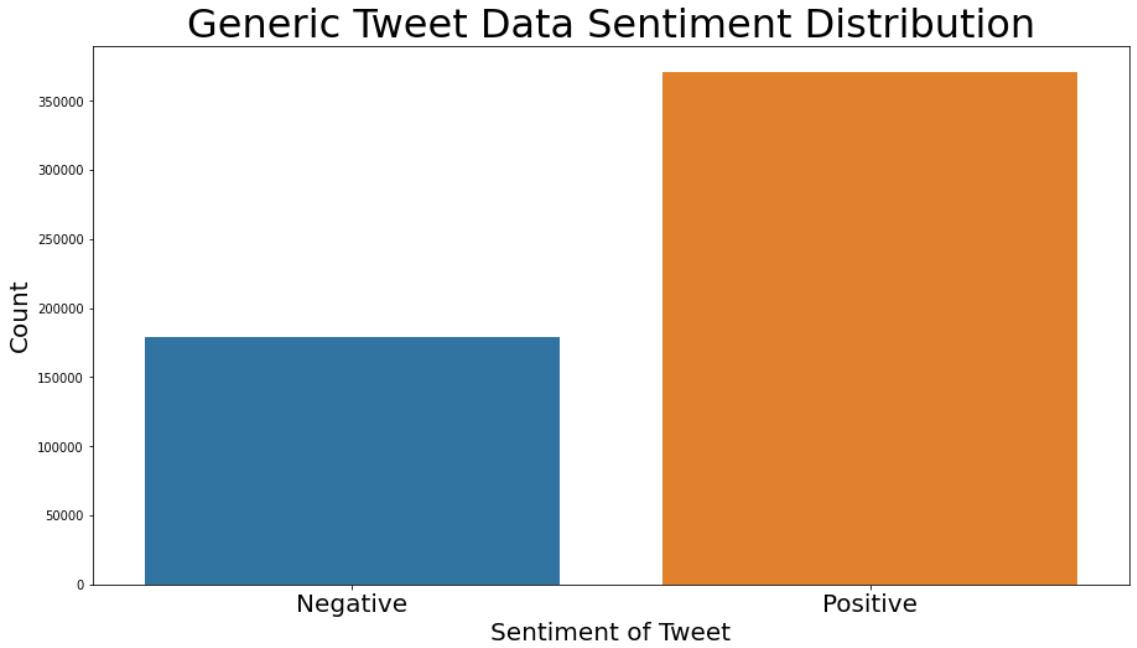
Canadian Elections 2019 – Sentiment Analysis

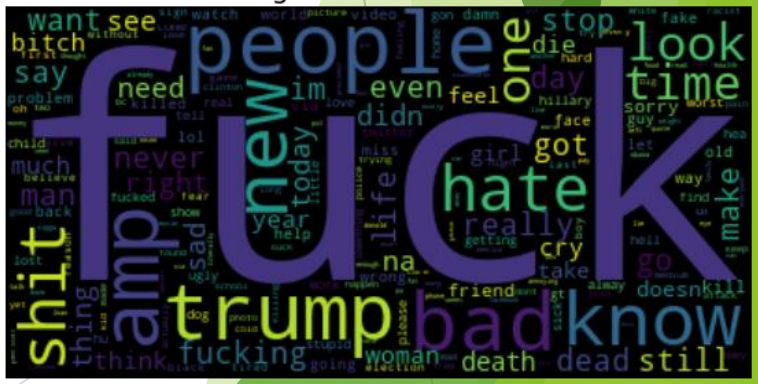
Exploratory Analysis Generic Tweets Dataset

- 550391 rows x 3 columns
  - Columns: ID, Text, Label
- 371341 Positive (67%), 179050 Negative (33%)
  - Majority positive sentiment tweets
- Word Clouds
  - Lots of strong positive and negative keywords



Word Cloud of Positive Sentiment Generic Tweets

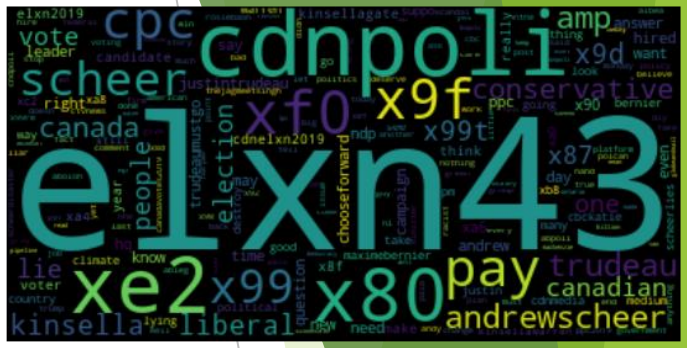
Word Cloud of Negative Sentiment Generic Tweets



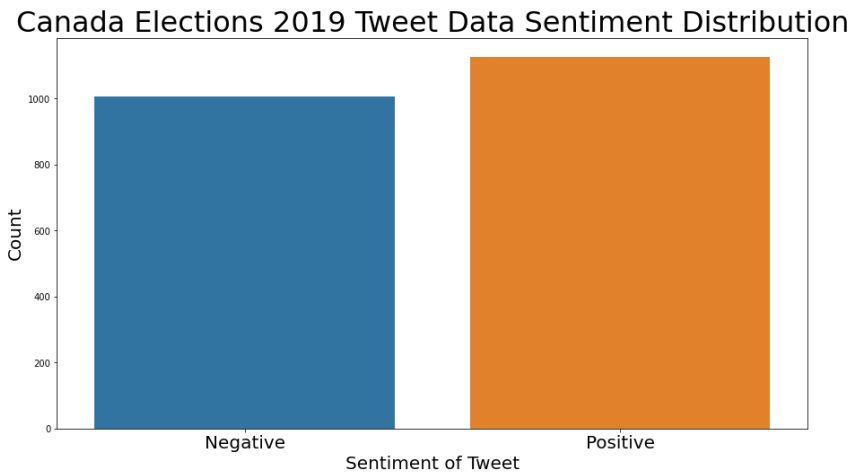
Exploratory Analysis Elections Tweets Dataset

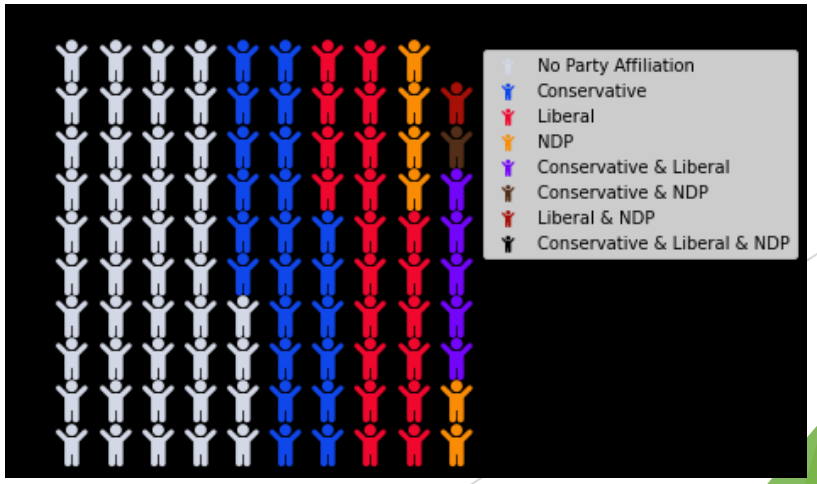
- 2133 rows x 3 columns
  - Columns: sentiment, negative reason, text
- 1006 Positive (47%), 1127 Negative (53%)
  - Evenly split sentiments
- Word Clouds
  - Few words that are easily identifiable to positive or negative sentiment.
  - Repeat words for both sentiments (elxn43)
- Party Affiliation
  - Most tweets are not associated with a political party.
  - Most popular Conservative > Liberal > NDP
  - Tweets associated with multiple parties
    - Conservative & Liberal > Conservative & NDP > Liberal & NDP

Word Cloud of Positive Sentiment Election Tweets



Word Cloud of Positive Sentiment Election Tweets





Model Preparation and Feature Engineering

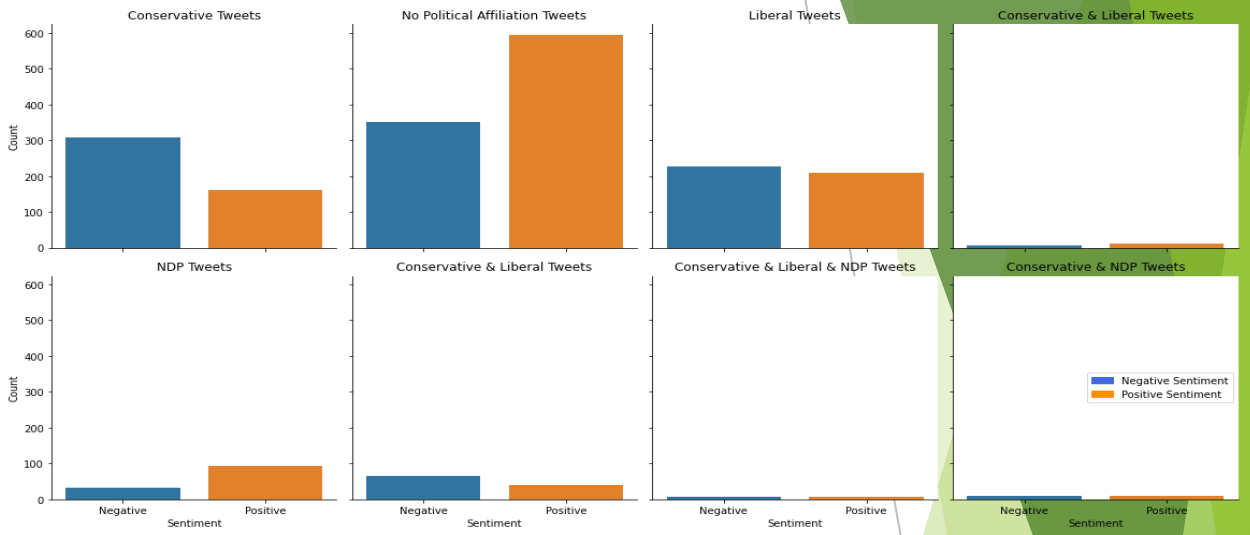
- Applied data cleaning to remove HTML tags, character codes, URLs, stop words then made lower case, lemmatized, tokenized.
  - Before: RT @MianUsmanJaved: Congratulations Pakistan on becoming #No1TestTeam in the world against all odds! #JI_PakZindabadRallies https://t.co/1o…
  - After: `['mianusmanjaved', 'congratulation', 'pakistan', 'becoming', 'no1testteam', 'world', 'odds', 'ji', 'pakzindabadrallies']`
- Split 70% training 30% test
- Vectorized using Bag of Words and TF-IDF methods
  - Set max_features = 1000
    - Improve model performance while reducing computational time
    - Reduce model noise of low frequency features/words.
- Top 10 Features Generic Tweets:
  - `['amazing', 'amp', 'best', 'bihday', 'day', 'good', 'great', 'happy', 'love', 'not']`
- Model Results (accuracy scores) predicting generic tweet sentiment on test set

	Log. Reg.	K-NN	Naïve Bayes	Decision Trees	Random Forest	XG Boost
Bag of Words	93.6%	91.2%	94.1%	92.7%	94.1%	91.4%
TF-IDF	92.3%	86.8%	94.1%	92.9%	94.4%	93.7%

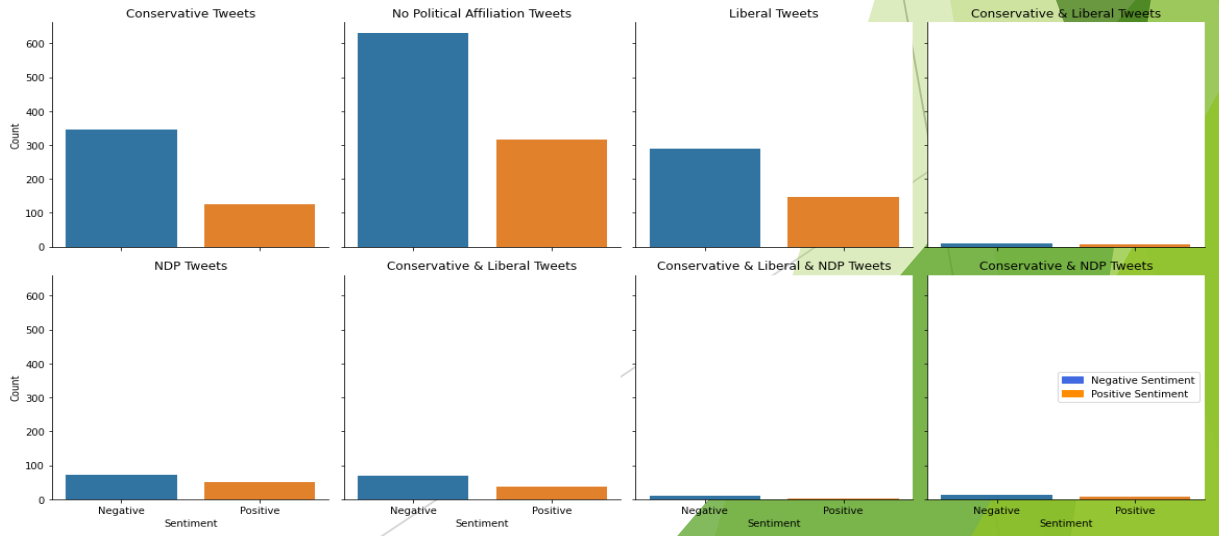
Election Sentiment Predictions Visualized

- Sentiment prediction more accurate for tweets affiliated with a political party
- Prediction of tweets with no political affiliation is poor and over predicts negative sentiment.
- Model Performance Takeaways:
  - Keywords associated with positive and negative sentiments are not clear in election tweets versus generic tweets.
  - Election tweet keywords are specific to the Canadian elections and don't show up in the generic tweet dataset. (i.e., SNC has a negative connotation but is not in generic tweets)
  - Better to train on more election data or tweets that are politically motivated.
- Election Takeaways:
  - Conservative party had the largest proportion of negative sentiment tweets and these were largely related to Women's Healthcare (abortion) and Racism which were hot topics during their campaigns.
  - Liberal Party also had a significant number of negative sentiment tweets and these were primarily targeted at Justin Trudeau and the SNC 'scandal'.
  - NDP was the only party which had more positive sentiment tweets however their overall number of tweets was less than both parties.
  - This provides us with a lot of insight of the public sentiment towards the parties on twitter and the general reasons for these sentiments which can be extracted from keywords and a word cloud. Further data exploration of the elections tweet data can be done.

True Sentiments by Party Affiliation



Predicted Sentiments by Party Affiliation



Model Results: Elections Sentiment Prediction

- Best model used for sentiment prediction: TF-IDF Random Forest

	Accuracy	Precision	Recall
Random Forest (TF_IDF)	60.2%	0.703	0.431

Confusion Matrix	
True Negative (798)	False Positive (208)
False Negative (641)	True Positive (489)

- The model does not perform as well on the Canada Elections 2019 data compared to the generic tweet data.
- Low recall indicating the model does not predict positive sentiment tweets.
- Further evidence: high number of false negatives

Model Results: Elections Negative Reason Prediction

- 3 models used: Log. Reg., Naïve Bayes, Random Forest with hyperparameter tuning via GridSearchCV
  - Grouped negative reasons:
    - Others, Separation, Privilege :0, Scandal:1, Tell lies:2, Economy:3, Women Rep. Right and Racism, Healthcare, Healthcare & Marij: 4, Climate:5,
  - Overall model accuracy is not high compared to previous results
  - Negative reasons with low data had lower accuracy which suggests we need more data and/or better grouping

Group	0	1	2	3	4	5
Accuracy	78.2%	58.4%	43.3%	23.5%	23.1%	55.6%

	Accuracy
Logistic Regression	49.0%
Naïve Bayes	51.3%
Random Forest	58.6%