# MIE1624 Assignment 2 - Predicting Kaggle Data Scientist Compensations

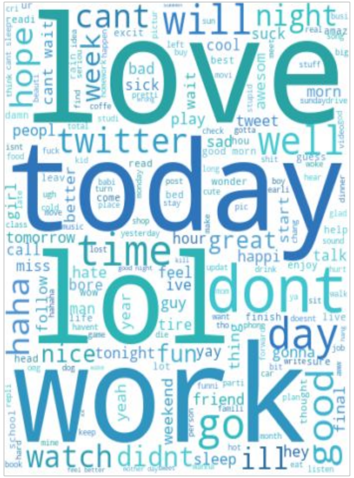
Alex Kwan - 1001559057

# Introduction & Exploratory Data Analysis

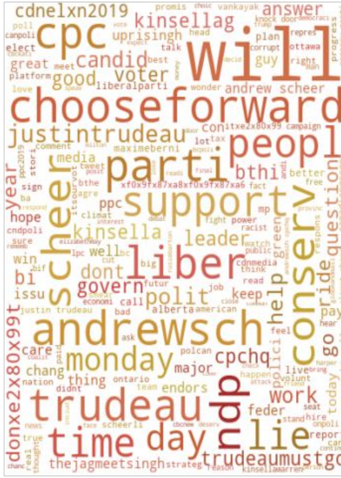
- Goal:
  - Use Sentiment Analysis to understand how public opinion on Twitter can tell us about the 2019 Canadian Election
  - Evaluate how a model trained on generic tweets predicts the sentiment of political tweets
  - Evaluate NLP models to predict the reason for negative sentiment political tweets

- 2 Raw Datasets:
  - Generic Tweets with sentiment from 2009, n = 200,000
  - Political Tweets with sentiment and reason, n =2,133

Generic Tweets



Election Tweets

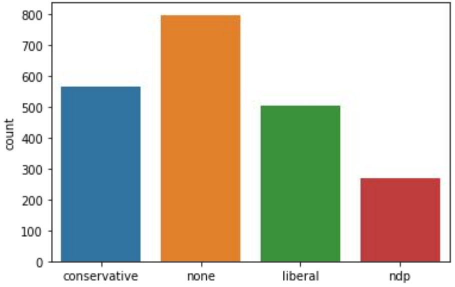


# Data Cleaning Rules and Feature Engineering

For both datasets, cleaned by lower case, stemming, remove all links, remove nltk stopwords, remove HTML tags

For political tweets:

- Remove stop words with specific words and hashtags for the Canadian Election ("Canada", "#elxn43", "PM", "Vote2019"
- Predicted Party in mentioned of tweet using count of keywords associated with each party
    - Liberal: #ChangeForward, #TrudeauMustGo, @JustinTrudeau, black face, SNC Lavalin, #LPC
    - Conservative: #ScheerLies, @AndrewScheer, #CPC
    - NDP: @JagmeetSingh, orange, #NDP



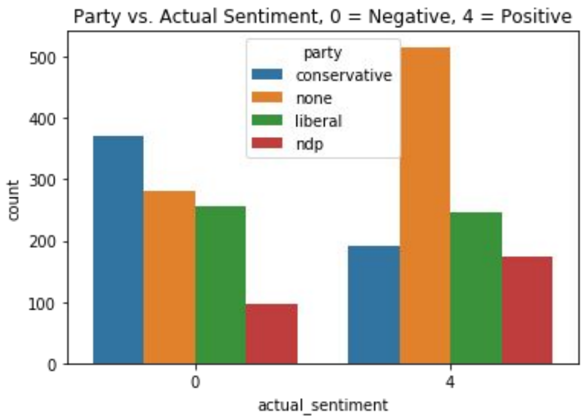
# Visualizations

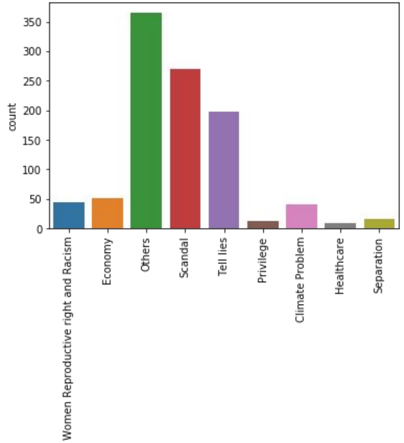
The actual (not predicted) number of positive and negative sentiment tweets correlates to the election results of Liberals 1st, Conservatives 2nd, NDP 3rd:

- Conservatives have the most tweets, but >60% are negative
- Liberals are second in tweets, with # positive = # negative
- NDP have the fewest tweets

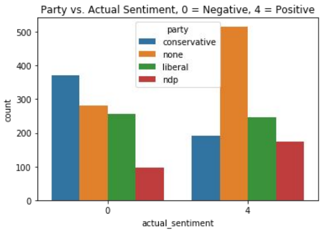
Common negative sentiment reasons are

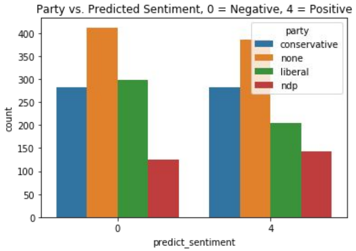
- unclear (Other)
- scandal (black face, SNC Lavalin)
- lies (Trudeau failure to keep promises and Scheer making false claims)





# Model Results





**Generic Tweets Sentiment Prediction (5k features)**

| Algorithm | WF Accuracy | TF-IDF Accuracy |
|---|---|---|
| **Logistic Regression** | 0.7358 | 0.7336 |
| **K-NN** | N/A | N/A |
| **Naive Bayes** | 0.727 | 0.649 |
| **SVM** | 0.733 | 0.732 |
| **Decision Trees** | 0.584 | 0.591 |
| **Random Forest** | 0.606 | 0.602 |
| **XG Boost** | 0.690 | 0.690 |

**Political Tweets Sentiment Prediction (5004 features, extra 4 from party classification)**

Evaluating the Logistic Regression with WF (Word Frequency) Features for Sentiment Prediction of Political Tweets: Accuracy = 0.503 :(
→ Model trained on generic tweets unable to classify political tweets

Using multi-class Logistic Regression to predict the negative sentiment reason, Accuracy = 0.5478 :(
→ Small # of training samples for most classes
→ Overfitting since #features > # training examples

# Discussion and Future Work

Sentiment Analysis is useful, confirms that higher proportion critical of the Conservatives since twitter demographics skew to a younger crowd:

- Conservatives want to cut education, OSAP, don't support gay marriage and birth control which conflicts with the values of the younger population (who have debt, support taxes, support LGBT)

But, poor performance of models trained on generic tweets to predict sentiment of political tweets which could be misleading since it over-predicts the # of positive sentiment Conservative tweets and also poor performance on predicting the reason for negative sentiment due to small sample sizes.

**Future Work:**

- **Train models on actual political tweets not generic tweets from 10 years ago**
- **Better cleaning to remove common contractions and emojis, use counts of hashtags**
- **Word embeddings(fastText), deep learning**