# Problem Statement

Analyze text from social media platforms to highlight the topics/issues that community is complaining or not happy with. The approach entails two steps which are as below:

- Build a Classifier that classifies tweets as complaints/extremely negative and non-complaints.
- Topic modelling on negative tweets to highlight topics/issues that a community is talking about.

# Data

Multiple datasets were used for this study.

- Sentiment140 _http://help.sentiment140.com/for-students/
- Consumer Complaints https://www.kaggle.com/cfpb/us-consumer-finance-complaint
- Amazon Reviews http://sifaka.cs.uiuc.edu/~wang296/Data/index.html

Each dataset is a huge dataset with enough observations to train a ML model. However, the goal of this project goes beyond just classifying text polarity to classify text as grievance or non-grievance. Hence we used multiple datasets as each dataset served a different purpose. Sentiment 140 is important because the end goal of this project is to extract grievances from tweets and hence twitter data must be used for model to understand twitter lingo and style. Consumer complaints was used because it comprises of complaints only and helps model focus on grievances. However, these complaints are against financial product and might make the model biased towards financial products. To address this problem, amazon reviews across different categories were also included so that the model is able to classify grievances in general and is not biased towards certain products.

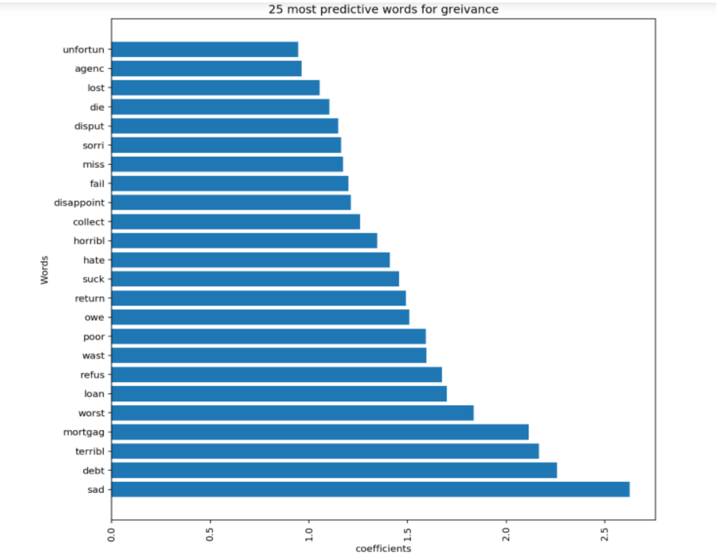
# Preprocessing and Cleaning

- Null values and duplicates were removed in each dataset.
- Post preprocessing, each dataset had only a text column and label column. For Sentiment 140, labels were changed to 1, grievance, for tweets with negative polarity and 0, non-grievance, for tweets neutral or positive tweets.
- All observations in Consumer complaints were labelled as 1 it comprises of complaints only.
- Labels were generated for Amazon reviews base on review rating. Observations with ratings less than or equal to 2 were labelled as grievances,1, and rest were labelled as 0, non- grievances.
- Data Cleaning comprised of steps below:

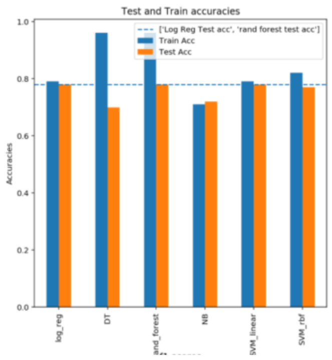
    - Remove user handles - e.g. @broskiii OH SNAP YOU WORK......
    - Remove hashtags - e.g. #at&amp;t is complete fail
    - Remove urls.
    - Address contractions e.g. 'didn't' by expanding contractions via contraction map.
    - Replace punctuations with whitespace.
    - Remove numbers.
    - Trim repeated letters in a word to successive repetitions e.g. ‘I am soooooooooo hungryyyyyyyyyyyyyyyyy’ - ‘I am soo hungryy’.
- Dataset used fit models had the composition below:
    - Total observations = 50,000
    - 0 = 25,000, 15,000 from sentiment 140, 10,000 from amazon reviews
    - 1 = 25,000, 10,000 from sentiment 140, 5000 from consumer complaints, 10000 from amazon reviews
    - All observations from sentiment 140_neutlabels

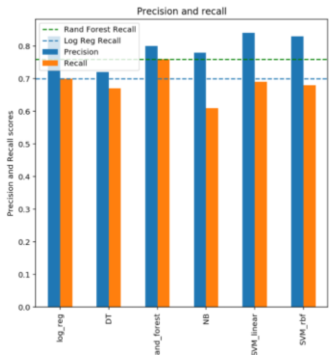
## Vectorization and fits

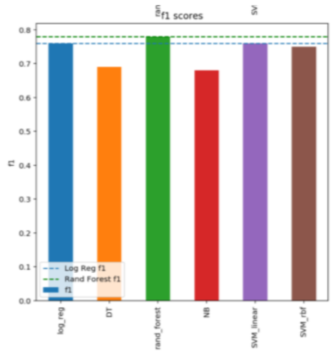
A custom tokenizer was created to be used with CountVectorizer. The custom tokenizer eliminated stop words and also stemmed words. Initial fit was a Logistic Regression with default parameters that had an accuracy of 78% on train set. However top predicting words i.e. words with highest coefficients for grievance, positive class, indicate that the model was biased towards financial product names, see figure below.



Hence a list with names of financial products was included in custom tokenizer to filter product names which gave more generalized results. Multiple models were fit and evaluated on vectorized data and the results are as below.







Logistic Regression and Random forest stood out in initial evaluation and were tuned in aws sagemaker. Although showed no improvement post tuning, it still had better Recall and F1 score than Logistic Regression. However, words with highest coefficients for random forest were biased toward product names. Hence, to evaluate if Random forest can generalize and maintain its performance, it was evaluated on a dataset that comprised of tweets only. The results were not in favor as it underperformed significantly. Hence, it was decided to use Logistic Regression and mixed dataset for the rest of the study.

## Topic modelling

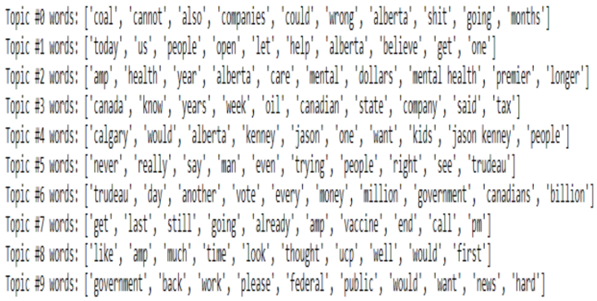
Twitter data was scraped for the city of Toronto, with a minimum retweet filter of 50, and transformed into vectors in the same manner as training data. Optimized Logistic regression model was then used to predict labels and a grievance dataframe, i.e. observations with label 1 only, was created for topic modelling. LDA was used for topic modelling with both Counvectorizer and Tf-idf vectorization. Former gave a lower model perplexity and hence was chosen as the preferred method.

Spacy entity recognition was evaluated as an alternative to topic modelling. Although it provided much focused entity recognition, as opposed to a lot of generic topic keywords from topic modelling, it failed to highlight topics and was ruled out as an alternative to topic modelling but rather to be used along with topic to highlight sub topics within a modelled topic.

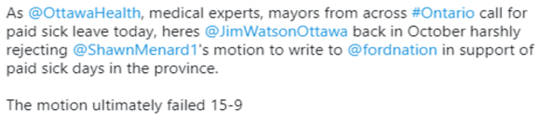
## Results

Entire process was run on twitter data from 2 Candian cities – Ottawa and Calgary. Below are the top 10 modelled grievance topics for each city along with some original/source tweets.

*Calgary*

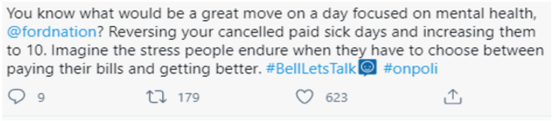


```
Topic #0 words: ['coal', 'cannot', 'also', 'companies', 'could', 'wrong', 'alberta', 'shit', 'going', 'months']
Topic #1 words: ['today', 'us', 'people', 'open', 'let', 'help', 'alberta', 'believe', 'get', 'one']
Topic #2 words: ['amp', 'health', 'year', 'alberta', 'care', 'mental', 'dollars', 'mental health', 'premier', 'longer']
Topic #3 words: ['canada', 'know', 'years', 'week', 'oil', 'canadian', 'state', 'company', 'said', 'tax']
Topic #4 words: ['calgary', 'would', 'alberta', 'kenney', 'jason', 'one', 'want', 'kids', 'jason kenney', 'people']
Topic #5 words: ['never', 'really', 'say', 'man', 'even', 'trying', 'people', 'right', 'see', 'trudeau']
Topic #6 words: ['trudeau', 'day', 'another', 'vote', 'every', 'money', 'million', 'government', 'canadians', 'billion']
Topic #7 words: ['get', 'last', 'still', 'going', 'already', 'amp', 'vaccine', 'end', 'call', 'pm']
Topic #8 words: ['like', 'amp', 'much', 'time', 'look', 'thought', 'ucp', 'well', 'would', 'first']
Topic #9 words: ['government', 'back', 'work', 'please', 'federal', 'public', 'would', 'want', 'news', 'hard']
```



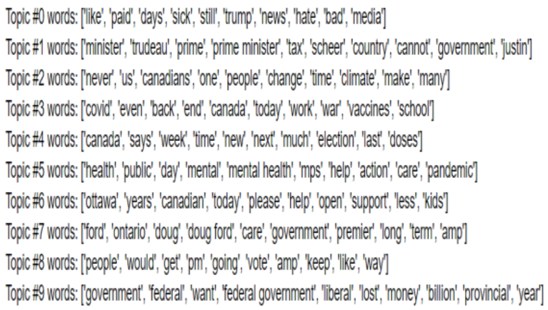
As @OttawaHealth, medical experts, mayors from across #Ontario call for paid sick leave today, heres @JimWatsonOttawa back in October harshly rejecting @ShawnMenard1's motion to write to @fordnation in support of paid sick days in the province.

The motion ultimately failed 15-9

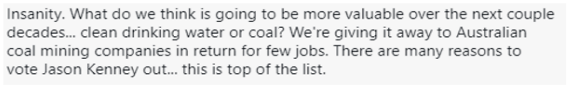


You know what would be a great move on a day focused on mental health, @fordnation? Reversing your cancelled paid sick days and increasing them to 10. Imagine the stress people endure when they have to choose between paying their bills and getting better. #BellLetsTalk #onpoli

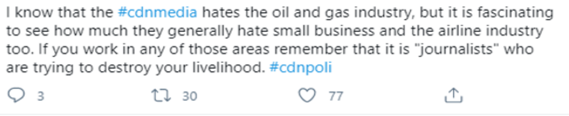
*Ottawa*



```
Topic #0 words: ['like', 'paid', 'days', 'sick', 'still', 'trump', 'news', 'hate', 'bad', 'media']
Topic #1 words: ['minister', 'trudeau', 'prime', 'prime minister', 'tax', 'scheer', 'country', 'cannot', 'government', 'justin']
Topic #2 words: ['never', 'us', 'canadians', 'one', 'people', 'change', 'time', 'climate', 'make', 'many']
Topic #3 words: ['covid', 'even', 'back', 'end', 'canada', 'today', 'work', 'war', 'vaccines', 'school']
Topic #4 words: ['canada', 'says', 'week', 'time', 'new', 'next', 'much', 'election', 'last', 'doses']
Topic #5 words: ['health', 'public', 'day', 'mental', 'mental health', 'mps', 'help', 'action', 'care', 'pandemic']
Topic #6 words: ['ottawa', 'years', 'canadian', 'today', 'please', 'help', 'open', 'support', 'less', 'kids']
Topic #7 words: ['ford', 'ontario', 'doug', 'doug ford', 'care', 'government', 'premier', 'long', 'term', 'amp']
Topic #8 words: ['people', 'would', 'get', 'pm', 'going', 'vote', 'amp', 'keep', 'like', 'way']
Topic #9 words: ['government', 'federal', 'want', 'federal government', 'liberal', 'lost', 'money', 'billion', 'provincial', 'year']
```



Insanity. What do we think is going to be more valuable over the next couple decades... clean drinking water or coal? We're giving it away to Australian coal mining companies in return for few jobs. There are many reasons to vote Jason Kenney out... this is top of the list.



I know that the #cdnmedia hates the oil and gas industry, but it is fascinating to see how much they generally hate small business and the airline industry too. If you work in any of those areas remember that it is "journalists" who are trying to destroy your livelihood. #cdnpoli

## Conclusion

- Optimized Logistic Regression model does a good job predicting grievances/complaints but gets confused when negative words like ‘missing’ are used in a positive context.
- Semantic must be included when training for a more robust model.
- Modelled topics from LDA, from grievance corpus, are topic keywords rather than summarized topics. Topics include a lot of generic keywords as well and are usually hard to interpret. Hence further investigation, like mapping back to and analyzing original tweets, has to be done to highlight community grievances.
- Text summarization is an ideal approach to interpret topics/grievances and must be included in future.
- Until text summarization is implemented, spacy must be used to highlight sub topics within modelled topics.
- Retweet filter is very essential because without any filter, modelled topic keywords are very generic and do not speak to the nature of communities.
- Twitter data with a minimum retweet filter is very limited for Canadian cities and hence this analysis must be extended to other platforms like Reddit, Quora etc.
- Multiple classes like expectations, queries, joy etc., have to identified to get a wholesome picture of a community as opposed to just grievance and non-grievance.