



IMPROVING THE YELP REVIEW EXPERIENCE WITH ML

Beheshteh Mostaghni
Hee Kang
Angela Detweiler
Alex Lam

May 7th, 2018

How would you rate this review?

“I went here through a friendsgroupon coupon and had a nice time. The treats were good and the company was good. It was nice and cosy.”



How would you rate this review?

“Luv the atmosphere, great setting for breakfast or lunch, cozy, homey. Had coffee, 2 eggs with a burger, toast, potatoes. Coffee, eggs, toast unremarkable, potatoes...why bother putting them on the plate?, burger good, Other than the atmosphere, I'd search for a better breakfast elsewhere, but atmosphere may be worth it if average is ok for you. I should mention the person serving my table, was superb.”



How would you rate this review?

“ Dingy atmosphere, and the food was mediocre. The menu items sound more exciting than they are, and the whole restaurant just didn't feel very clean.”

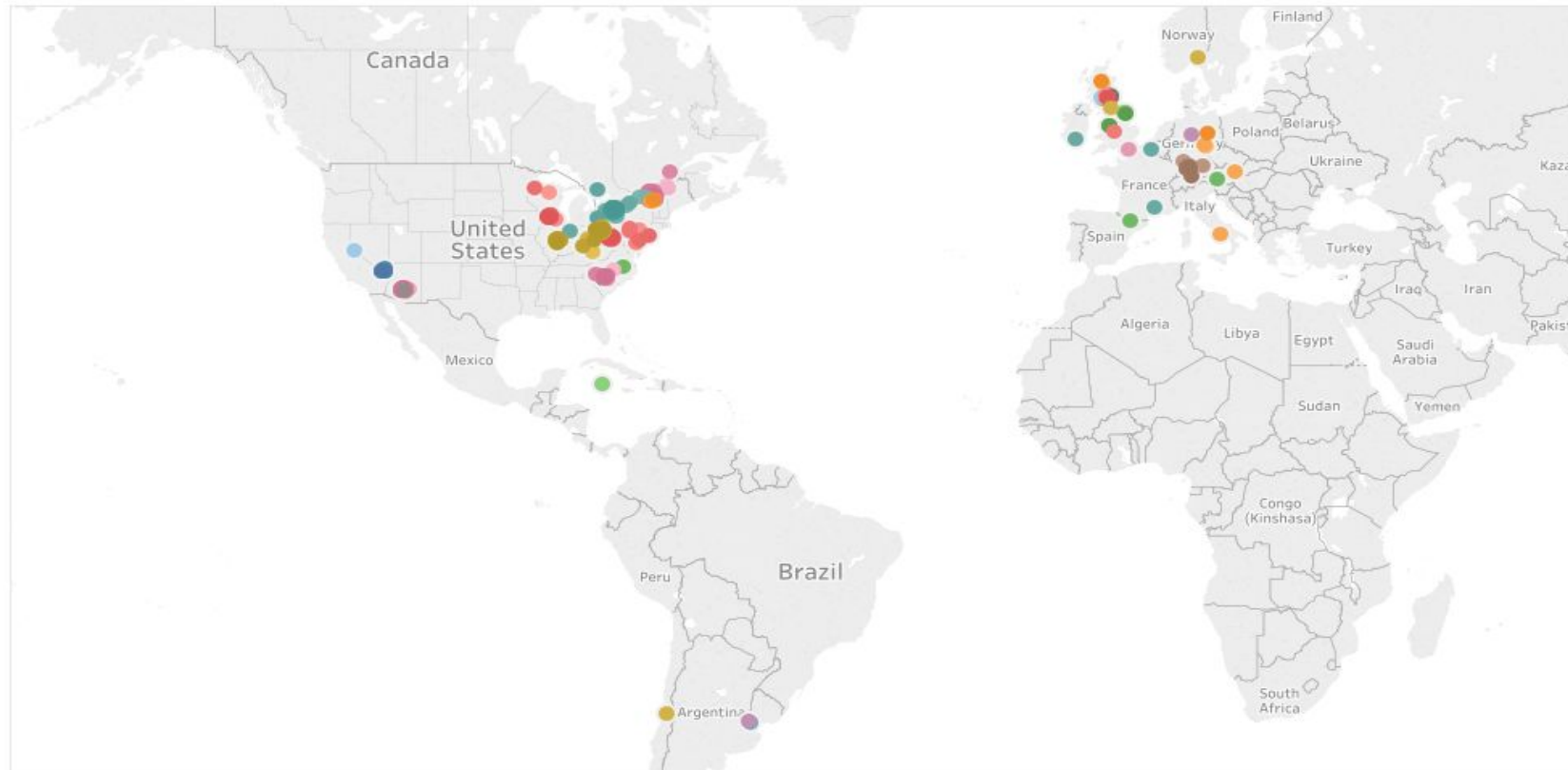


- **Problem:** When you are researching restaurants on Yelp, do you look at the star rating or do you read the review? Do you look at both? Given that reviews are highly subjective, and star ratings can be influenced by various aspects of business performance, can we use machine learning to standardize the interpretation of reviews?
- **Goal:** Our goal is to apply Natural Language Processing (NLP) and other features from Yelp reviews into a model to reduce the discrepancy between reviews and star ratings.

Source of Data

The data used in this project was downloaded from [Kaggle Yelp Dataset](#) (yelp_business.csv and yelp_review.csv). As you can see this data is skewed geographically.

The locations of the businesses with yelp reviews on our data file



Data Screening and Selection

■ Subset

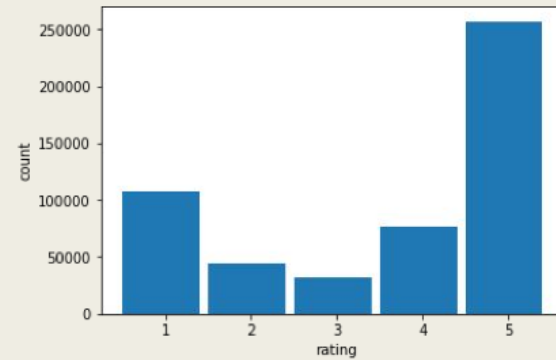
- Limit scope to Restaurants, Bars, Other Food and Drink Related Establishments
- Focus on Reviews, Text, Star Ratings

	Entire Dataset	Restaurants, Bars & Other	Users with 50-100 Reviews
Businesses	174,567	72,103	50,642
Users	1,326,101	960,561	4,676
Reviews	5,261,668	3,636,641	323,100
Files	7	2	1
Size	5.53 GB	3.82 GB	273 MB

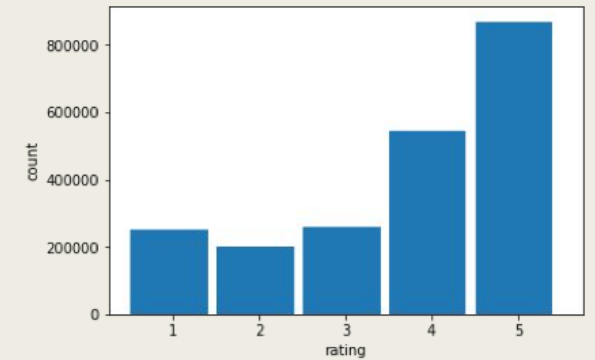
Narrowing the Dataset Based on User Review Numbers

As Users write more reviews, the distribution of star ratings given become more and more normal. Otherwise, ratings are more commonly given on the extremes

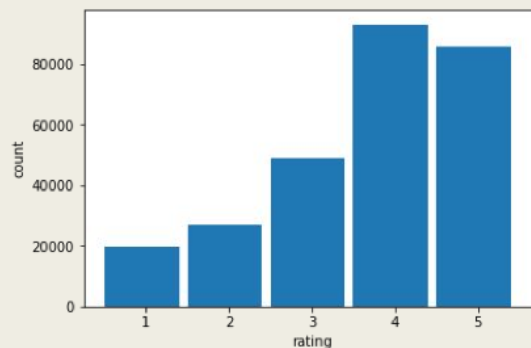
1 Review Only



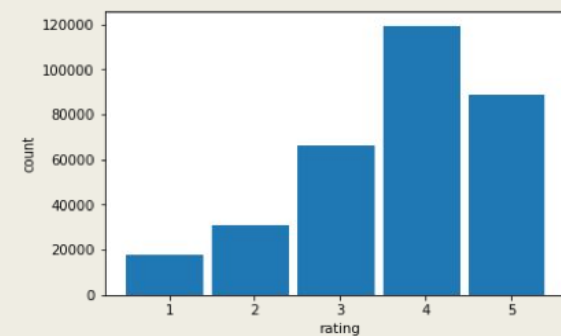
1-30 Reviews



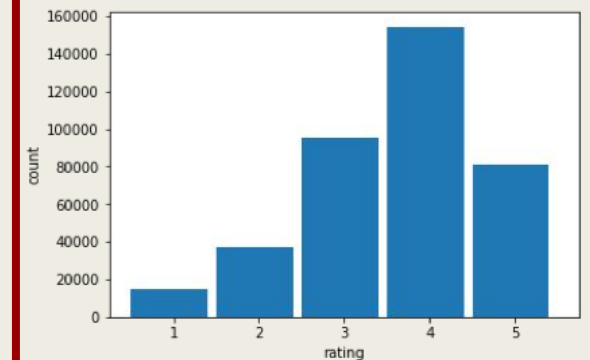
30-50 Reviews



50-100 Reviews

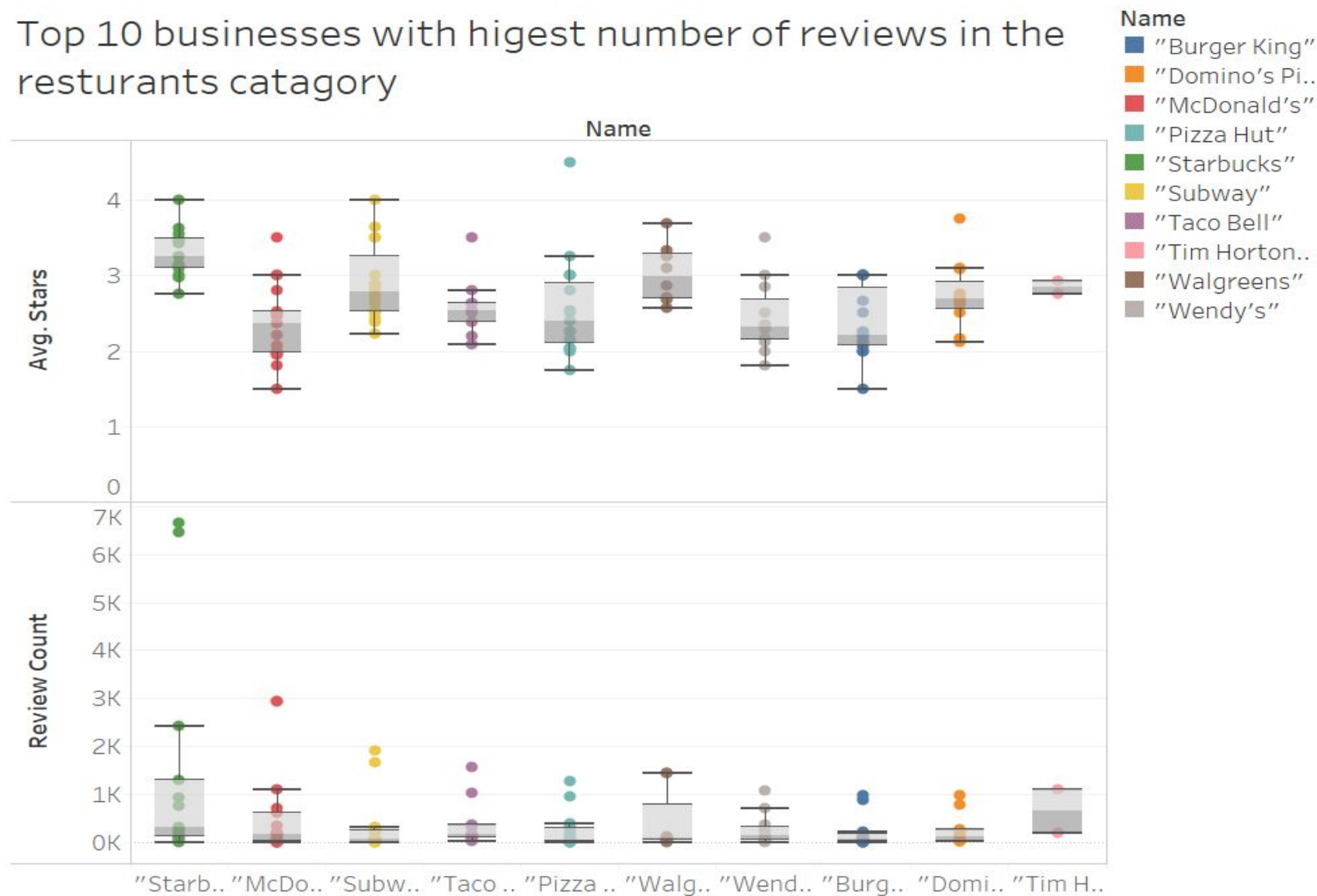


100-1000 Reviews

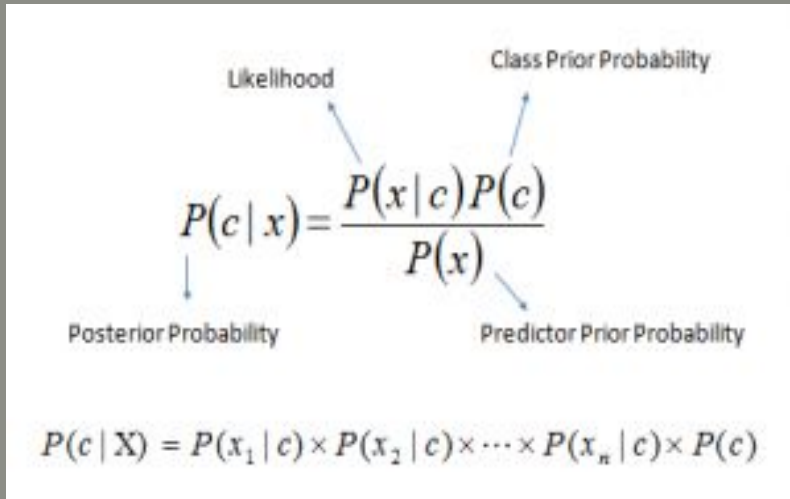


Top 10 Businesses with Highest Review Counts

Top 10 businesses with highest number of reviews in the restaurants category



Naïve Bayes Models



The diagram shows the Naive Bayes formula with labels for its components:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

Labels and arrows:

- Likelihood** points to $P(x|c)$
- Class Prior Probability** points to $P(c)$
- Posterior Probability** points to $P(c|x)$
- Predictor Prior Probability** points to $P(x)$

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \dots \times P(x_n|c) \times P(c)$$

Naive Bayes uses method to predict the probability of different class based on various attributes. This algorithm is mostly used in text classification and with problems having multiple classes.

Pros:

- It is easy and fast to predict class of test data set. When assumption of independence holds, this model performs better compared to other models like logistic regression and you need less training data.
- It perform well in case of categorical input variables compared to numerical variables. For numerical variables, normal distribution is assumed (bell curve, which is a strong assumption).

Cons:

- If categorical variable has a category (in test data set), which was not observed in training data set, then model will assign a 0 (zero) probability and will be unable to make a prediction. This is often known as “Zero Frequency”. To solve this, we can use the smoothing technique. One of the simplest smoothing techniques is called Laplace estimation.
- On the other hand naive Bayes is also known as a bad estimator, so the probability outputs from `predict_prob` are not to be taken too seriously.
- Another limitation of Naive Bayes is the assumption of independent predictors. In real life, it is almost impossible that we get a set of predictors which are completely independent.

Data Pipeline

Raw text

Tokenization

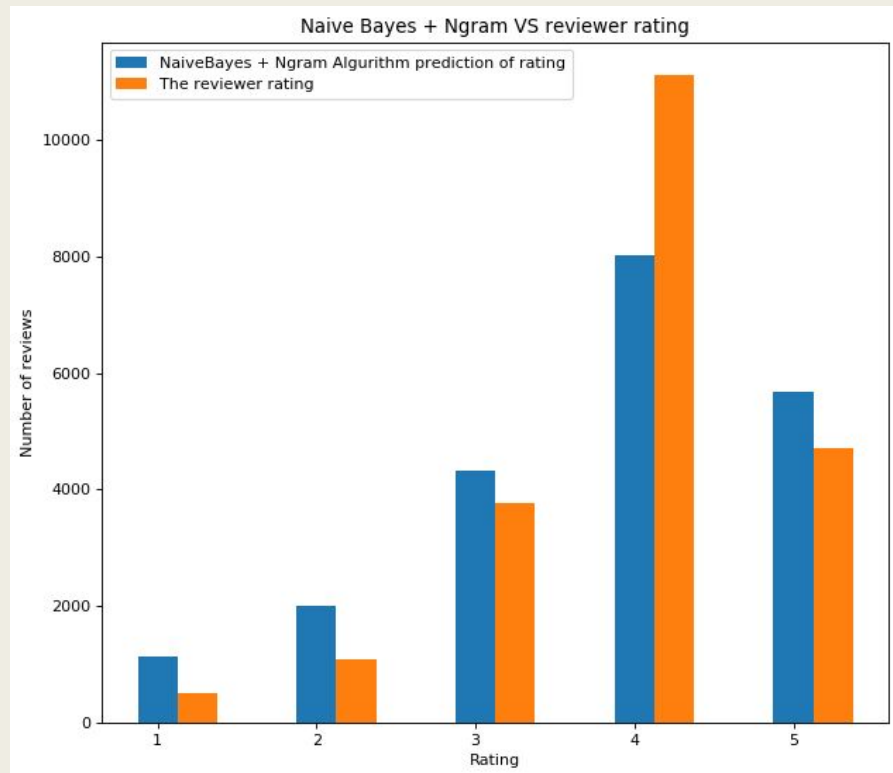
2-gram

Stop words filtering

TF_IDF

Naive Bayes

Accuracy of model at predicting reviews was: 0.4418003231554448



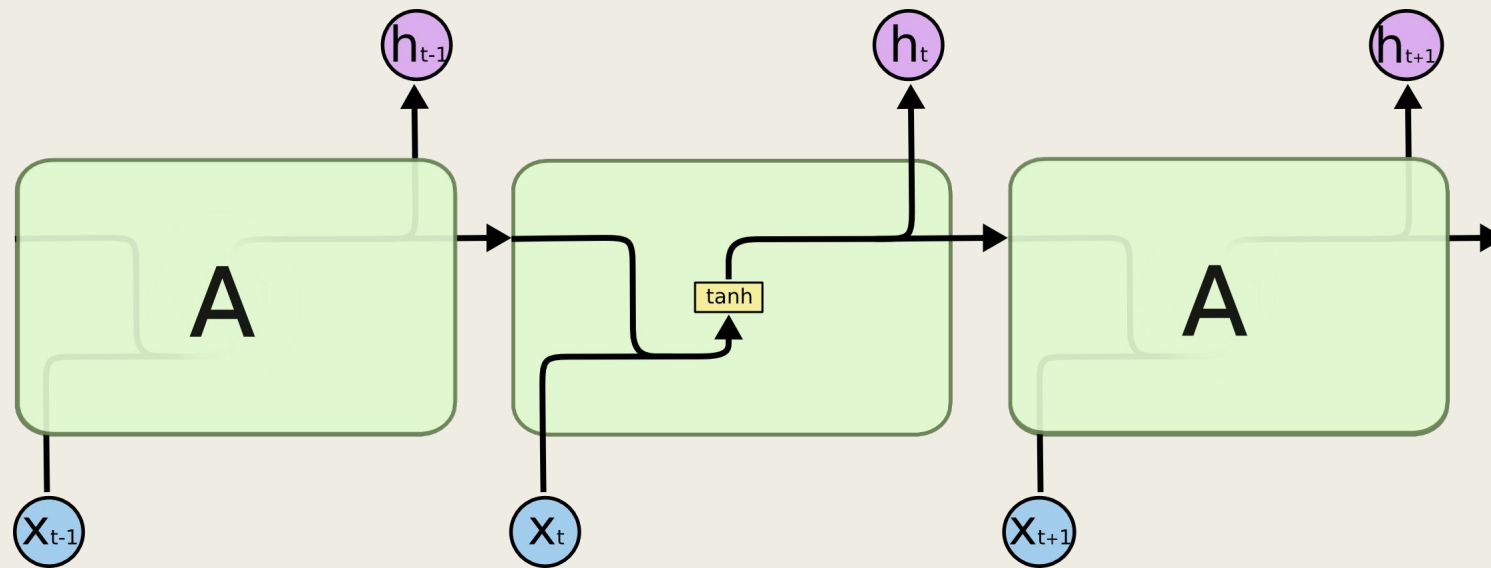
Neural Network Models (Long Short term Memory layer)

Regular model w/ kept words: [Being, Evil, Is, Great] : Sentiment: neutral (Evil/Great)

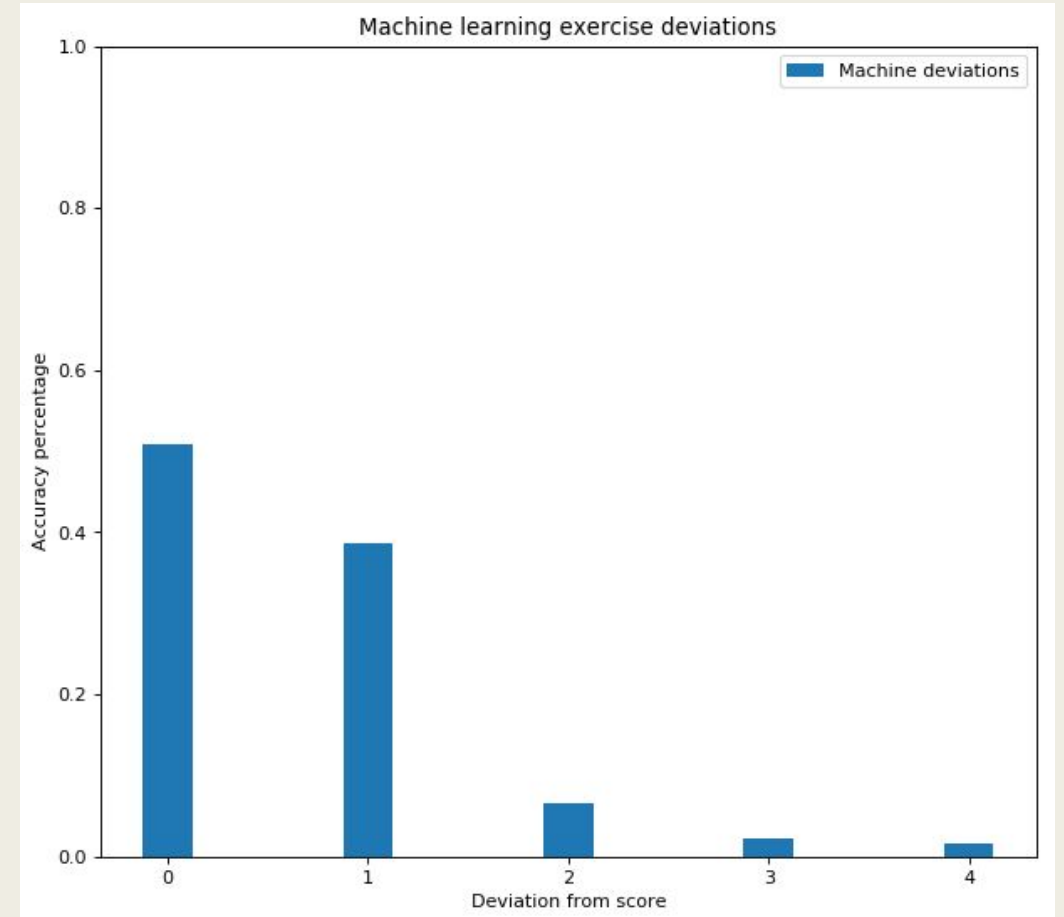
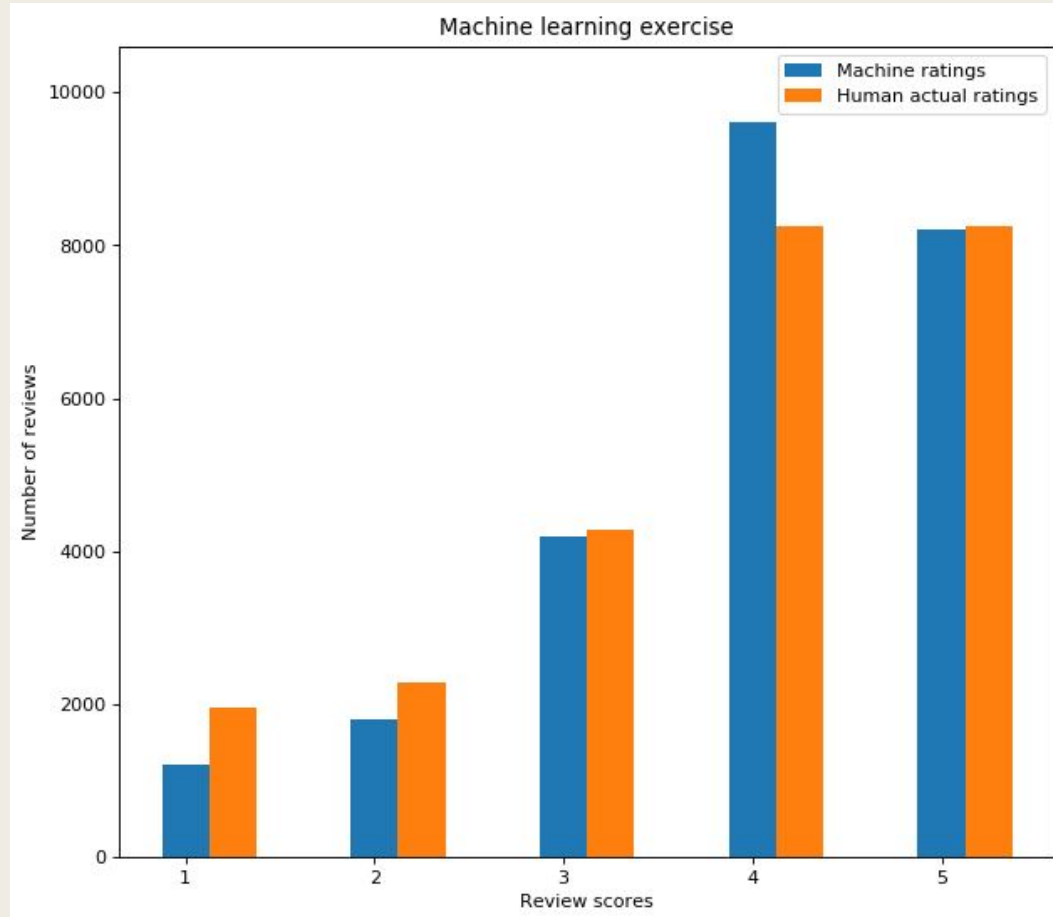
By giving the neural net some short term memory using the LSTM model (Gives it context)

LSTM(4)

LSTM model w/ kept words: [Being, Evil, Is, Great] : Sentiment: negative (Being Evil is Great)



Neural Network Models



How Can We Improve Star Ratings?

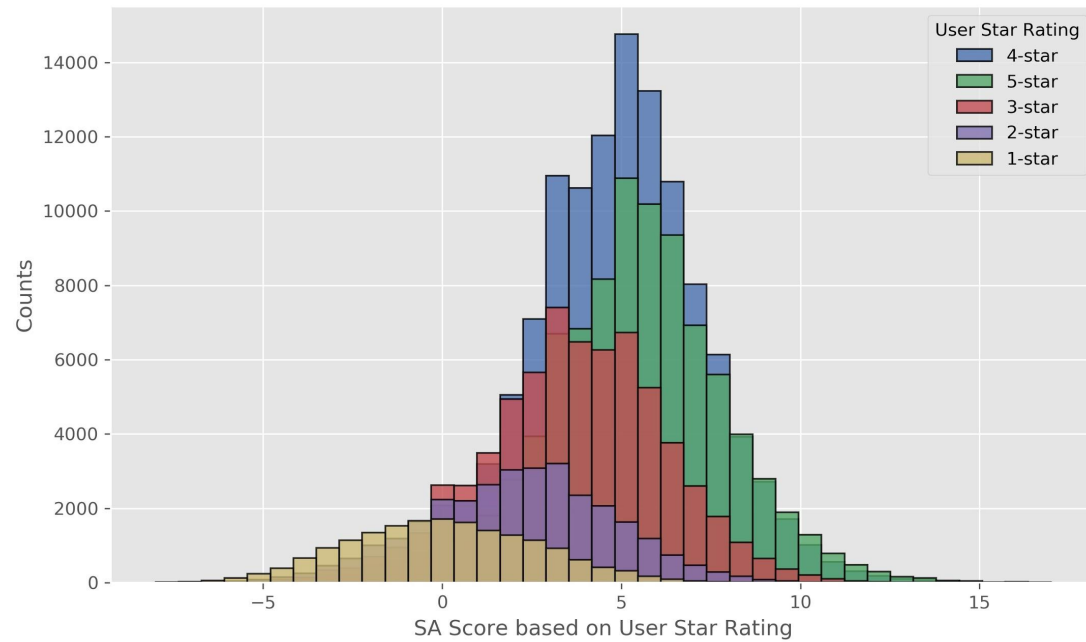
WORKFLOW (one approach)

1. Yelp Review text (323K records)
2. Tokenized review (NLTK)
3. Words compared against AFINN word list to get valence score (-5 to 5)
 - AFINN
 - AFINN-en-165.txt - list of 3,382 words;
 - AFINN-emoticon-8.txt - list of 96 emoticons
4. $SA_Score\ normalized = \frac{\text{sum scores}}{\sqrt{N}}$ to account for review length

gr8	3	
great	3	
greater	3	
greatest		3
greed	-3	
greedy	-2	
green wash		-3

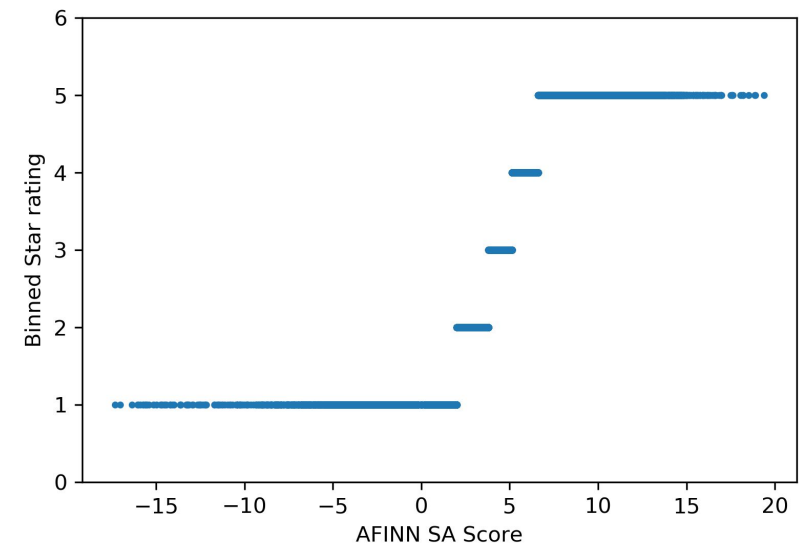
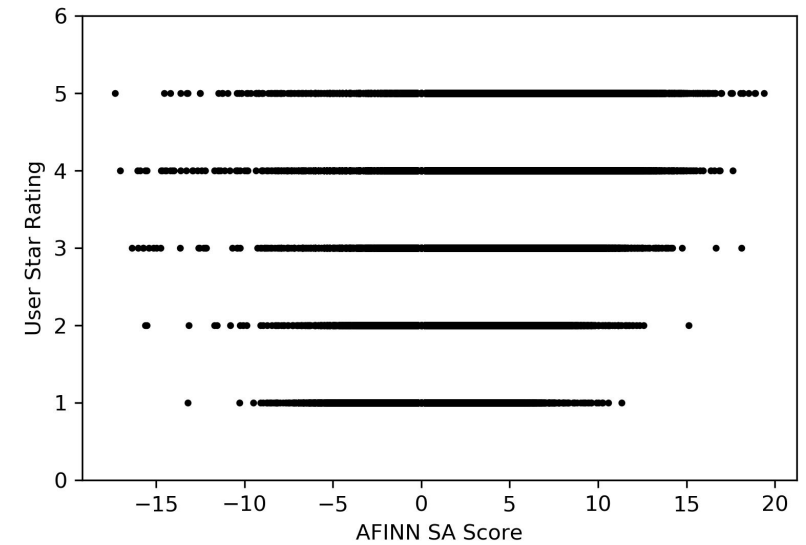
:-?	-1
:->	2
:-*	2
:-D	3
:-P	3
:-S	-2
:-p	3
:-/	-2
:D	3

SA Score Frequency Based on Star Rating



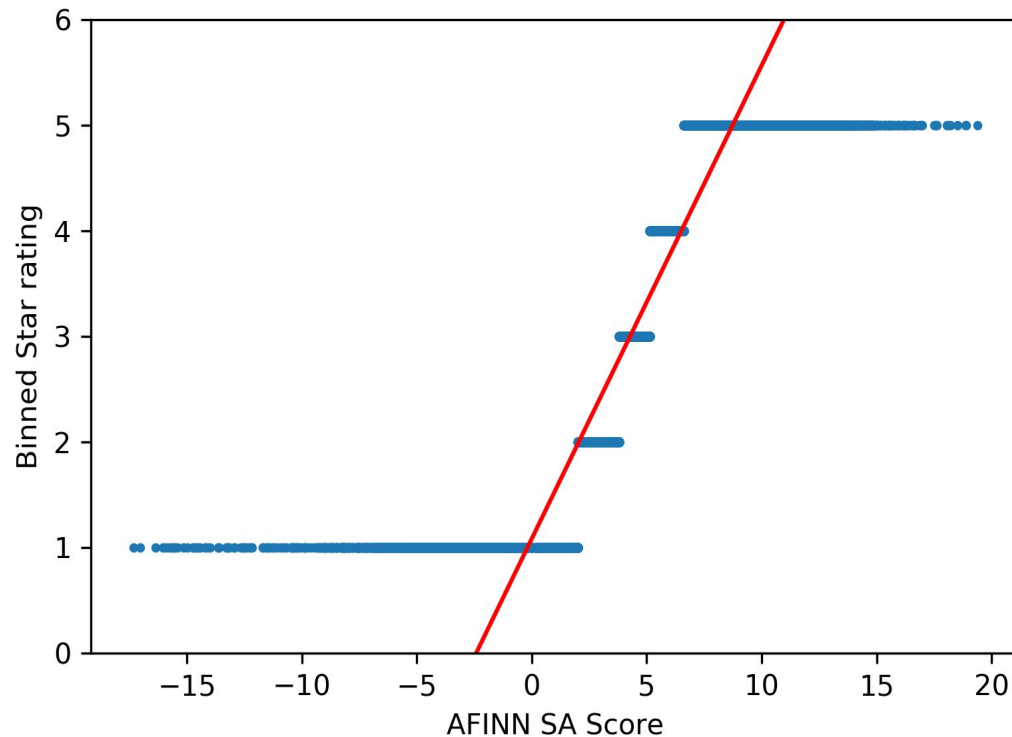
User-star-rating

**New-star-rating
based on 5-quantiles
(0.2, 0.4, 0.6, 0.8)**



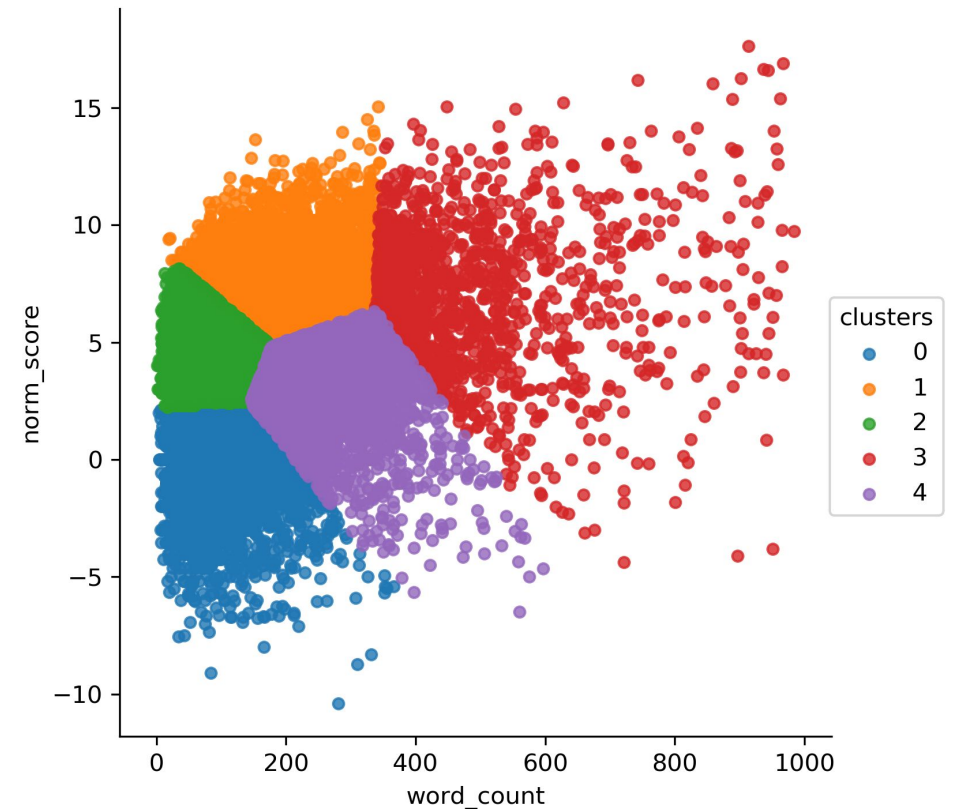
Linear Regression Model (1-variable)

Model accuracy: $R^2 = 0.849$ /MSE = 0.301



K-Means (2-variables)

Incorporate new features and reclassify star-rating categories based on new clusters (N-clusters = 5)



Features to Take into Consideration for Future Model

FEATURES

- Star Rating Distribution; Percentages of 1-star reviews, 2-star reviews, etc.
- Review Counts
- Rating Distribution Statistics per User: Mean, Median, Variance, etc.
- Use of unsupervised machine learning to identify groups based on user behavior
 - *K-Means Clustering*
 - *Perhaps there are users who only rate when extremely happy or upset (rate at extremes), users who rate moderately, etc.*

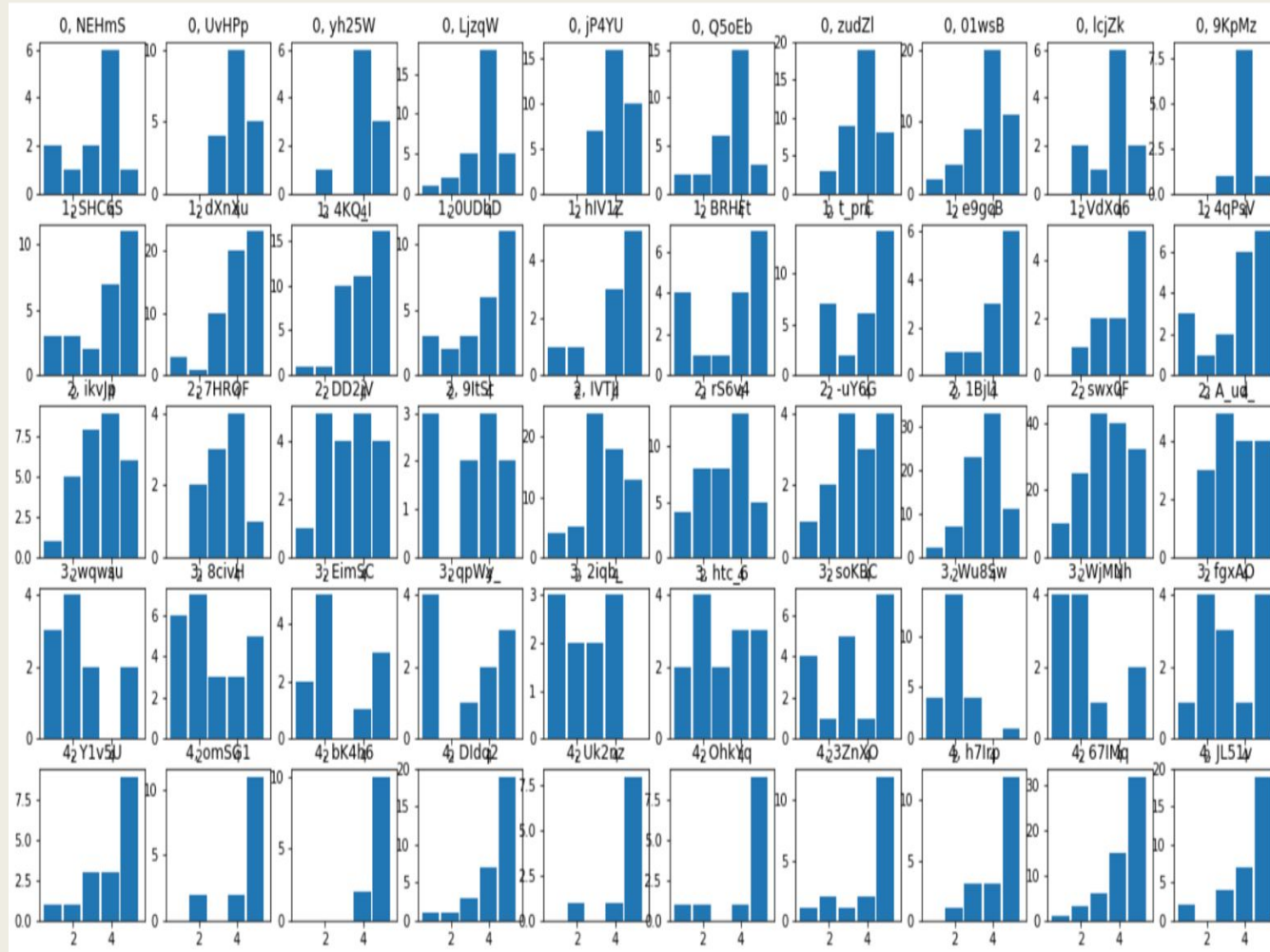
```
['1_x',  
'2_x',  
'3_x',  
'4_x',  
'5_x',  
'1_y',  
'2_y',  
'3_y',  
'4_y',  
'5_y',  
'10th_quantile',  
'25th_quantile',  
'75th_quantile',  
'90th_quantile',  
'max',  
'mean_avg',  
'med',  
'min',  
'review_count',  
'std',  
'var']
```



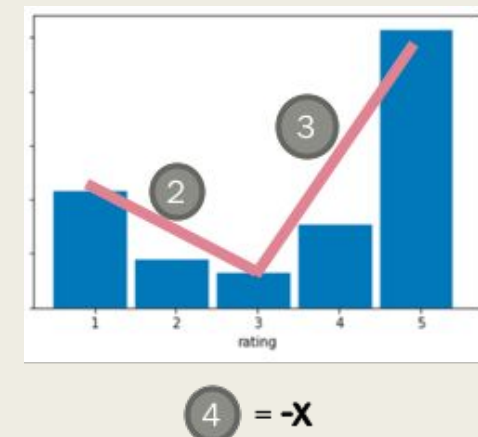
```
['1',  
'2',  
'3',  
'4',  
'5',  
'review_count',  
'mean-med',  
'min_med_slope',  
'med_max_slope',  
'inflection']
```

K-Means Clustering

Each Row is a random sample of user's rating distribution from each of the clusters determined by K-Means



Key	Metric	Definition	Intent
1	Mean-Med	Difference between Mean and Median Value	Rough Approx. of Skew
2	Min-Med Slope	Slope in Percentage Points from Min to Med	
3	Med-Max Slope	Slope in Percentage Points from Med to Max	
4	Inflection	Product of <2> and <3>	Reveals if trend is constant or bell-shaped



Inflection is negative if the two slopes are in opposite directions

Next Steps

- Further analysis of features to incorporate in model - i.e. user review/rating behavior.
- New rating input - In order to make our model more robust, we will have humans rate text reviews, and we will incorporate these ratings into our test model to improve star rating output based on reviewer experience.
- Game - proof-of-concept.

Closing

```
Enter review here: This bootcamp would not be as great as it was if not for the people inside. The teacher was great, the TAs were great, but most of all, the students were great. I have enjoyed my adventure with you guys and hopefully our paths cross once again. I don't know what my model will output, but I give this experience 5 stars overall
Loading tokenizer...
Got tokenizer for vocab size: 200 in 0:00:00.109260
2.6% confident of rating being 1,
7.5% confident of rating being 2,
19.4% confident of rating being 3,
37.4% confident of rating being 4,
33.0% confident of rating being 5.
Therefore, I predict the rating to be 4
```

- LSTM will only remember a few words, but our team will remember you guys for a lifetime.

p.s. Model gave this a 4, it must've deviated 1 down

Challenges

- K-Means - Since n features > 2 , data cannot be well represented graphically to confirm if clusters make sense
 - *Instead, samples plotted from each predicted cluster to qualitatively check for similarities*
- Overall difficulty determining if ML model is accurate or if hypothesis can be truly accepted or rejected
- Keeping the big picture in perspective

References

- Project Github Repo - https://github.com/minas26902/UCB_Project3_ML
- AFINN Repo - <https://github.com/fnielsen/afinn>
- Yelp Review Dataset via Kaggle - <https://www.kaggle.com/yelp-dataset/yelp-dataset>