Sinead O'Connor
Springboard Data Science Career Track

CAPSTONE 2
MILESTONE REPORT 2
SENTIMENT ANALYSIS OF YELP FAST FOOD REVIEWS

The Problem:

What do customers think of your fast food restaurant? What are they saying about your restaurant online?

The Client:

In the age of the Internet, people are constantly looking online to see what others think of restaurants, venues, hotels, etc. to try to make informed decisions about where they want to spend their money. With this in mind, negative reviews and comments on the internet can greatly affect a restaurant's bottom line. It is important for restaurant owners or managers to have an understanding of what customers are saying about their establishment online.

Rather than sifting through thousands upon thousands of tweets or comments, social media departments can use the results from predictive analytics models to quickly determine the public's perception of their restaurant, and use these results to inform actions to take to improve their services. For instance, businesses would be able to quickly identify negative reviews and respond to unhappy reviewers in a timely and professional manner. This will be especially useful for fast food establishments, many of which are regional, national, or even glainos with many franchises.

The Data:

I will be using data from the Yelp Challenge. The Yelp Dataset consists of several files, including a business file, a review file, a user file, and a photo file. For the purposes of this project, I will only be using the business and review files. The business file contains information such as name, address, number of reviews, and the categories under which the business falls, for 192,609 businesses in the United States and Canada. The review file includes the user id of the reviewer, the business id, the rating (out of 5 stars), and the text of the review for 6,685,900 Yelp reviews.

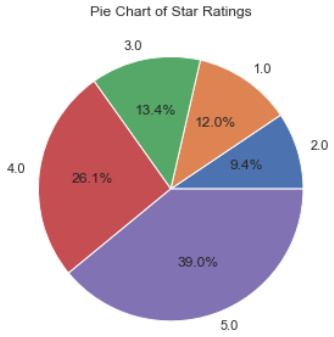
I first downloaded the Yelp Dataset tar file and extracted the individual json files. Due to the size of the review json file, I installed Java Development Kit and pyspark, and loaded the business and review json files. I filtered the business data frame on the "categories" column to obtain only businesses with "Restaurant" included in the list, and I further filtered out any categories that seemed to be mislabeled, for example one of the remaining restaurants was also listed under the "Laser Hair Removal" category and turned out to be CanadaMed Laser. The resulting data frame had information on 53,324 restaurants. Using the business_id column which was common

to the business and review datasets, I filtered the review data frame to obtain the 3,643,450 reviews for the restaurants.

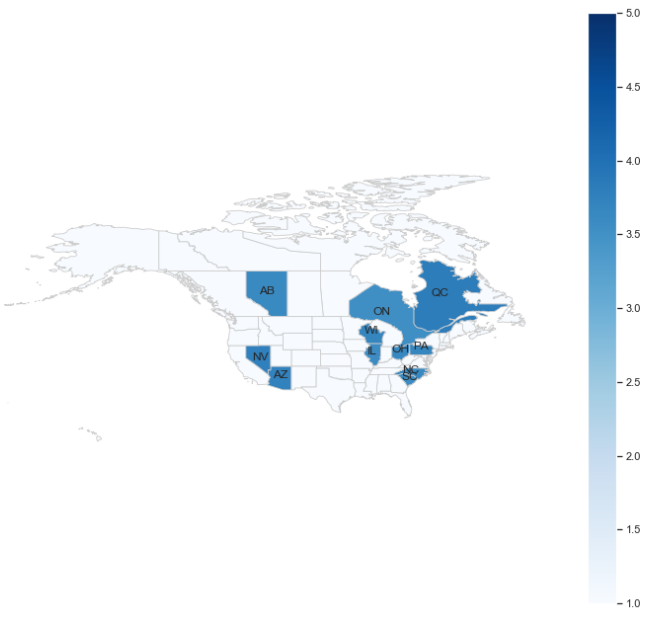
Exploratory Data Analysis - Data Storytelling

To initially explore the data, I examined various questions about all of the restaurants (both fast food and not.

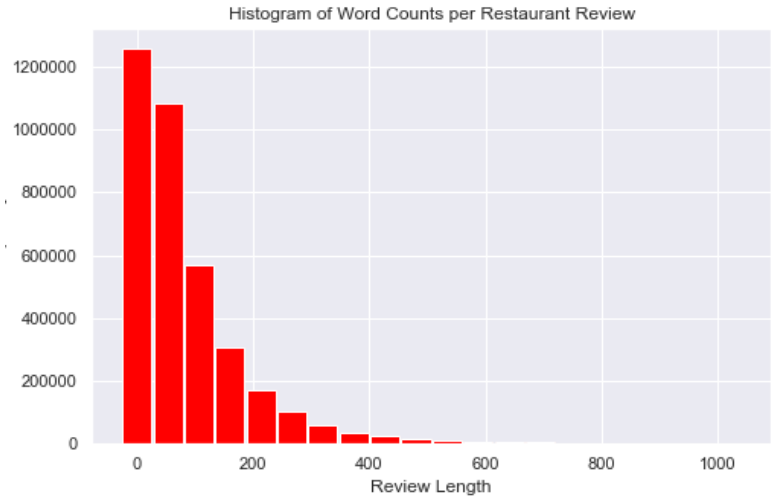
As seen in the below pie chart, of all the restaurant reviews, 65.1% were 4- and 5-star ratings. It is possible that restaurant patrons want to reserve 1- or 2-star ratings for restaurants that are particularly bad. Another reason the ratings may tend to be more favorable is because people are less likely to go to bad restaurants, so there would in turn be fewer reviews.



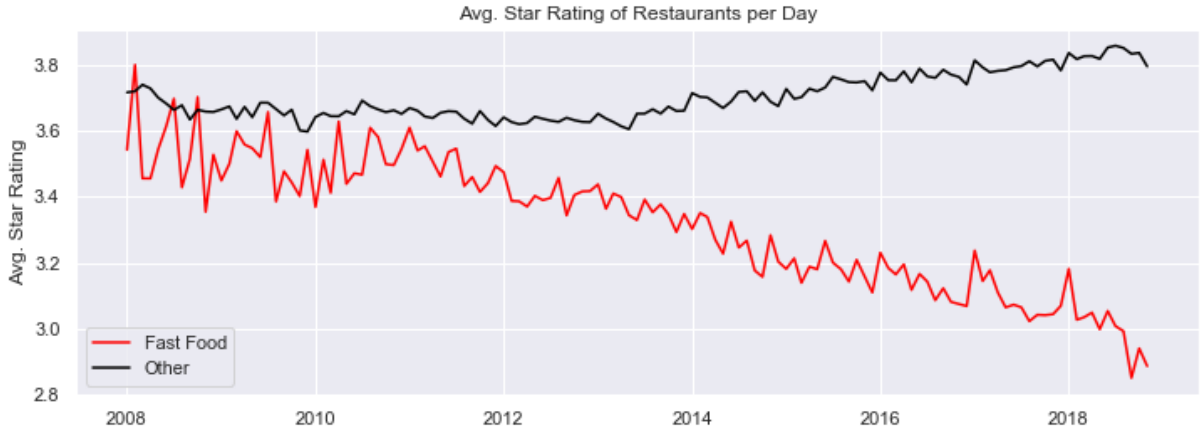
In examining the average star-rating by US state and Canadian province, we see that average rating does not appear to differ greatly from state to state. The lowest possible rating is 1 and the highest possible is 5, and we see on the below map, which was made using geopandas, that it is difficult to distinguish between the shades of blue. However Nevada and Quebec seem to have higher ratings, while areas like Ontario and Illinois appear to have lower ratings. The lowest average rating is 3.54 (Ontario) and the highest is 3.82 (Quebec).



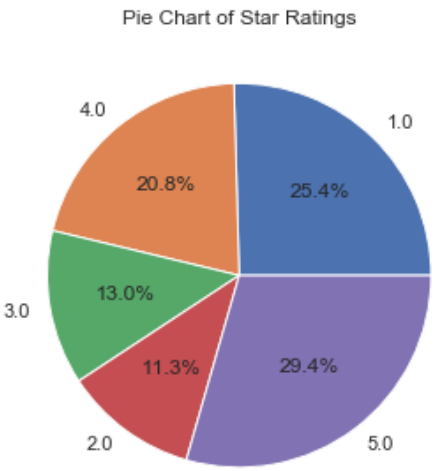
In this dataset, the smallest review length is 0 words, and the highest review length is 1,067 words. We see a strong right-skew in the below histogram with most reviews being under about 100 words. The frequency decreases as the length of the review (in words) increases. This makes sense because most people just write a few sentences on Yelp reviews, rather than paragraphs and paragraphs.



As we begin to compare fast food establishments to the non-fast food, we see that the average rating for the 220,507 fast food reviews (3.175 stars) is about 0.56 stars lower than the average rating of other 3.4 million other restaurant reviews (3.742 stars). The below time series plot also illustrates that non fast food restaurants have consistently been higher than those of fast food restaurants. In addition, fast food ratings seem to have decreased overall from about 3.6 in 2011 to about 2.9 by the end of 2018. This may be because Americans have realized that fast food is not healthy. For other restaurants, we see a slight overall increase in average reviews per month from about 3.6 in 2013 to about 3.8 at the end of 2018. We are seeing more volatility in the fast food monthly ratings because the sample size is much smaller.



The most frequent rating of fast-food restaurant is 5-stars as it is with the all restaurants, however, fast food restaurant reviews were also given 1-star ratings 25.4% as seen in below pie chart.



As a final step in this portion of Expolartory Data Analysis, I made word clouds for each star-rating to explore the different words used by reviewers based on their opinion of the restaurant.

In the word-cloud for 1-star fast food reviews, we see words such as "customer service", "employee", "staff", and "manager." This indicates that people who are giving 1-star ratings are often unhappy with service. We also see the word "manager" in only the 1-star review, which makes sense because customers only typically need to speak with a manager if they are very dissatisfied. We also see the word "service" in the 2-star cloud. We also see the word "cold" in the one-star and two-star couds. Fast food customers typically expect hot food.

The 3-star cloud, has the word "great", but keep in mind it could be "not great." Note that I removed the word "good" because all 5 clouds had it prominently featured, but before I removed it one of the words was "pretty good." There are also the words "ok" and "bad." The word "though" appearing in the word cloud also indicates that the customer may have liked some things and disliked others. The 4-star cloud has many positive words prominently featured including "love", "delicious", "nice", and “best."

The 5-star could clearly has the most positive words. It has all the same words I mentioned in the 4-star cloud as well as adjectives like "amazing", "awesome" and "favorite", and "perfect".
It is also interesting to note that the 5-star word cloud has the phrases "customer service" and "great service." It shows that good or bad service may really make or break a review. It may be one of the things that brings a 3-star review down to a 1- or 2-star review, or what brings a 4-star review to a 5-star review.

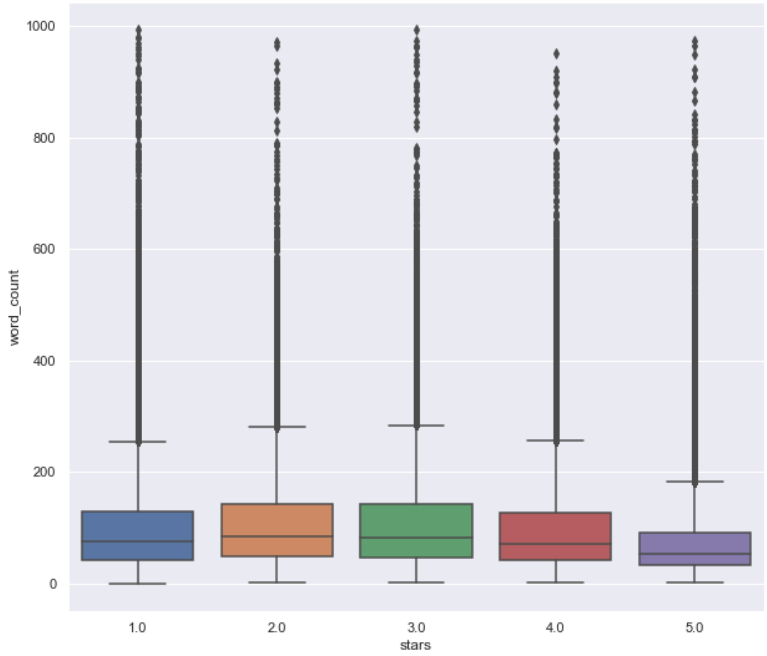
Going forward, I will only be using the fast food reviews and businesses.

Exploratory Data Analysis - Inferential Statistics

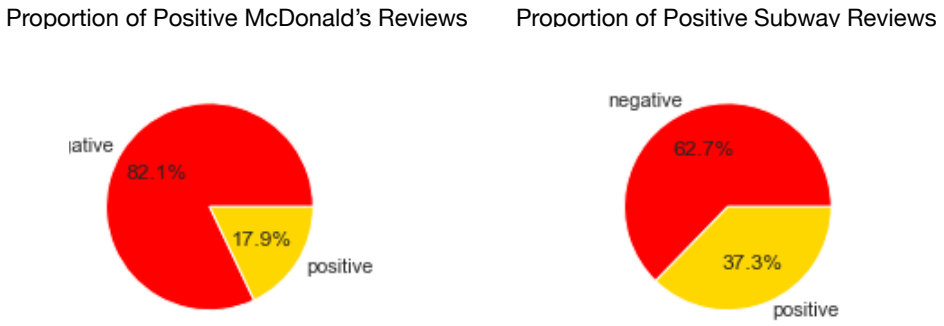
In the Inferential Statistics notebook, I tested several hypotheses about the fast food reviews.

Here is a summary of the findings:

- We observed that the average rating of 29,284 reviews of Canadian fast food restaurants was 3.338. while the average of 191,223 US fast food reviews was 3.15. Using a two-sample z-test, we saw that this difference was statistically significant.
- There is a weak positive correlation of 0.23 between the length of a review in words and the number of useful votes the review received. The correlation is statistically significant, but it is important to remember that it could be statistically significant due to the large sample size of 220,507.
- From the below boxplots we see that, the five-star reviews seem to have a lot less spread than the others, and that, overall, the reviews tend to be shorter. The two-star and 3-star reviews have the highest average review lengths, 110.11 and 109.02, respectively. It makes sense that 2-star and 3-star reviews tend to be longer because there may be more to explain if did not have a great time at the restaurant or had a mixed review. Through an ANOVA test, we concluded that there is a statistically significant difference in at least one of the star-ratings. We rejected the null hypothesis that the average review length for one-star, two-star, three-star, four-star, and five-star reviews were all the same.



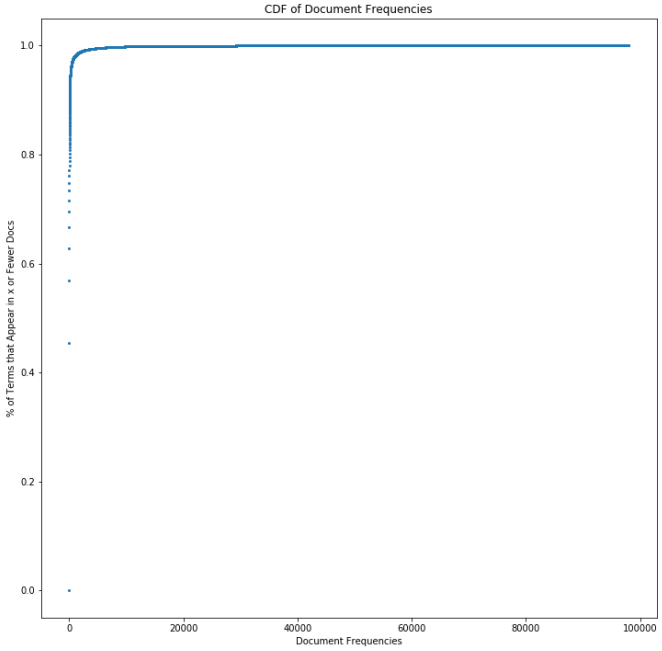
- In the fast food dataset, the proportion of positive (4- or 5-star) reviews for McDonald's (17.9%) was less than half the proportion of positive Subway reviews (37.9%), as seen in the below pie charts. There is a statistically significant difference in these proportions.

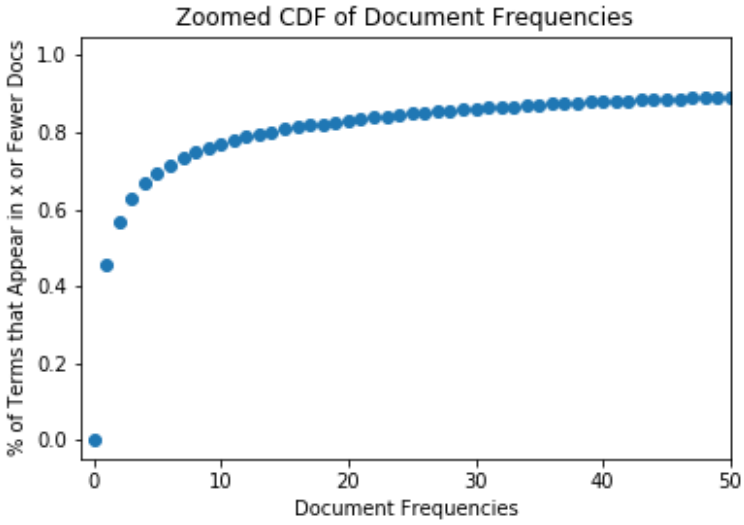


Baseline Modeling

Prior to feature engineering and training the baseline models, I preprocessed the reviews text. I tokenized the reviews using NLTK's wordpunct_tokenize. I also noticed that not all of the reviews were in English, so I decided to use the langdetect library to filter out non-English reviews. Given my computer was not powerful enough to use langdetect on the full set of 220,507 reviews, I first filtered reviews that had non-ascii characters and that contained French, Spanish, or German stopwords. This left us with 23,180 reviews to further examine, and using langdetect, I filtered out 843 non-English reviews and dropped them from the fast food data frame.

I first tried using Multinomial Naive Bayes. To engineer the features, I used a CountVectorizer on the tokenized reviews, supplying stop words, but without providing a minimum or maximum document frequency. The training accuracy was 0.653 and the test accuracy was 0.595 and there was imbalance. Given the 0.058 difference between the training and test accuracy it seems there is overfitting, so I tried tuning hyper parameters, using a CDF of document frequencies to choose a min_df and max_df. The word that appeared in the most documents was "food" which makes sense given they are fast food reviews.





Based on the above Cumulative Distribution Function and zoomed version of it, I chose min_df = 3 and max_df = 70,000. I then trained both a Naive Bayes and Logistic Regression model, with the below results:

Naive Bayes:

[Training Classification Report:]

	precision	recall	f1-score	support
1.0	0.73	0.82	0.77	42004
2.0	0.42	0.35	0.38	18620
3.0	0.50	0.44	0.47	21441
4.0	0.55	0.52	0.54	34079
5.0	0.73	0.78	0.75	48604
micro avg	0.64	0.64	0.64	164748
macro avg	0.59	0.58	0.58	164748
weighted avg	0.63	0.64	0.63	164748

Training Accuracy: 0.6413188627479545

[Test Classification Report:]

	precision	recall	f1-score	support
1.0	0.72	0.80	0.76	13981
2.0	0.34	0.27	0.30	6207
3.0	0.40	0.35	0.37	7093
4.0	0.48	0.46	0.47	11552
5.0	0.70	0.74	0.72	16083
micro avg	0.59	0.59	0.59	54916
macro avg	0.53	0.53	0.53	54916
weighted avg	0.58	0.59	0.59	54916

Test Accuracy: 0.5944169276713526

Logistic Regression:

[Training Classification Report:]

	precision	recall	f1-score	support
1.0	0.77	0.91	0.84	42004
2.0	0.61	0.38	0.47	18620
3.0	0.60	0.46	0.52	21441
4.0	0.60	0.52	0.56	34079
5.0	0.72	0.85	0.78	48604
micro avg	0.69	0.69	0.69	164748
macro avg	0.66	0.62	0.63	164748
weighted avg	0.68	0.69	0.68	164748

Training Accuracy: 0.6946184475684075

[Test Classification Report:]

	precision	recall	f1-score	support
1.0	0.75	0.88	0.81	13981
2.0	0.43	0.27	0.33	6207
3.0	0.44	0.35	0.39	7093
4.0	0.51	0.43	0.46	11552
5.0	0.68	0.82	0.74	16083
micro avg	0.63	0.63	0.63	54916
macro avg	0.56	0.55	0.55	54916
weighted avg	0.60	0.63	0.61	54916

Test Accuracy: 0.627831597348678

From the above classification reports, we see that the Naive Bayes model with min_df = 3 and max_df = 70,000 decreases overfitting slightly, however test performance is also slightly worse than when using the default parameters. The Logistic Regression model has a higher training and test accuracy, of 0.605 and 0.628, respectively, and has better precision, recall, and f1-scores. However, there is still overfitting. In both models, we see there is imbalanced data, and the models are much better at classifying the larger categories (1-star and 5-star reviews) than the other reviews. This could also be because it's easier to identify extreme reviews.

Logistic Regression Confusion Matrix

		PREDICTED RATING				
		1 Star	2 Star	3 Star	4 Star	5 Star
TRUE RATING	1 Star	12307	949	292	143	290
	2 Star	2736	1660	1114	366	331
	3 Star	869	945	2479	1895	905
	4 Star	345	219	1365	4914	4709
	5 Star	260	100	333	2272	13118

From the above confusion matrix for the Logistic Regression model, we also see that most of the 1-star reviews that are misclassified, are misclassified, as 2-star review, and vice versa. Similarly, most of the 5-star reviews that are misclassified, are misclassified as 4-star reviews and vice versa. This is good because it's less concerning to misclassify a 4-star rating as a 3- or 5-star rating than as a 1-sat rating, for example.

I also tried running a Logistic Model on the same features with the parameter class_weight = 'balanced' to address the class imbalance, however it performed worse in each metric than the original logistic regression model, and the imbalance was even slightly more extreme.

What's Next?

I will try a binary classification with positive and negative reviews as the target variable rather than star ratings to see if this improves the imbalance.

Other techniques I plan to use to extend and improve upon the baseline include using a Tf-Idf Vectorizer, a Random Forest Model, and a different value for n-gram.