Group 17 Yelp Business Executive Summary

Introduction

With increased use of social media and internet usage, user ratings are becoming more important for businesses, and most people are used to choosing restaurants based off of ratings on Yelp. According to existing research, a rating improvement of one star led to an increase in revenue of between 5 and 9 percent for certain businesses. (Graves, 2021). In this project, we will analyze the relationship between ratings and customer's comments on Yelp and give recommendations to these restaurants.

Data Processing

We are interested in the restaurants which serve sandwiches, and we would like to focus our attention on the Yelp user reviews that contain important information about quality of menu items and customer service. We focus on five brands: Subway, Jimmy John's, Jersey Mike's Subs, Quiznos, Potbelly Sandwich Shop, and Firehouse Subs, which are all popular chain franchises across the United States. First, we merged the "business.json" and "review.json" by joining two sets with business ID. Then, we extracted all these restaurants' names and corresponding average ratings, user rating, and user reviews.

For the two statistical models in the methodology section, we standardized each word or group of words by creating a TF-IDF word matrix. This is done by multiplying the term frequency of a word in each document by the log of the inverted fraction of document frequency, or fraction of documents containing the term. The formula is shown below in Equation 1:

$$TFIDF = \frac{term\ i\ frequency\ in\ document\ j}{total\ number\ words\ in\ document\ j} \cdot \log\left(\frac{total\ number\ of\ documents}{number\ of\ documents\ with\ word\ i}\right)$$

Equation 1: TF-IDF score definition

This TF-IDF word matrix was merged with the cleaned restaurant data, and this dataset was used to build our linear and logistic regression models. For terms, we extracted each unigram and bigram out of the corpus of all reviews.

Statistical Methodology

One statistical method we use is sentiment analysis, which is a natural language processing (NLP) technique used to determine whether data is positive, negative, or neutral. The goal is to provide a useful indication of how customers felt about their experience in each restaurant, perform analysis on textual data in order to help businesses monitor brand and product sentiment in customer feedback, and understand customer needs (*Sentiment Analysis Guide*). We will filter the keywords in the evaluation of each restaurant for analysis. Based on the customer sentiment reflected in these keywords, we will make recommendations to restaurants to help them improve their ratings on Yelp.

First, we process the comments with lemmatization to break down the comments into single words. Second, we use 'SentimentIntensityAnalyzer()' in the Python nltk package to generate a sentiment score for each user review. The scores are in the range of (-1, 1), where –1 indicates

strongly negative sentiment and 1 indicates strongly positive sentiment. Next, we use a technique called “Bigram Language Model”, where we combine all the words in each comment in pairs and list all combinations. Each combination is considered one bigram. Then we focused on the bigram pairs with the most occurrences and picked the ones relating to food items and service qualities. For each chosen pair, the corresponding median of the sentiment scores of reviews containing the pair is calculated.

We also implemented multiple linear regression (MLR) and multinomial logistic regression models. For both models, our *X* matrix features were manually chosen by team members based on frequency of word occurrence and its relevancy to our topics. The linear model is shown below:

$$y = \beta_{0j} + \beta_{1j}X_1 + \cdots + \beta_{pj}X_p$$

Equation 3: Linear Regression

Logistic regression allows us to group stars into categories with 0.5 difference from 1-5, and this is more representative of the data. The formula for calculating the probability of our predicted values is:

$$P_j(x) := P\left[Y = j \mid X_1 = x_1, \ldots, X_p = x_p\right]$$
$$= \frac{e^{\beta_{0j}+\beta_{1j}X_1+\cdots+\beta_{pj}X_p}}{1+\sum_{l=1}^{J-1} e^{\beta_{0l}+\beta_{1l}X_1+\cdots+\beta_{pl}X_p}}$$

Equation 4: Multinomial Logistic Regression

Where *x* is the column of each feature or n-gram in the TF-IDF matrix, β is the coefficient corresponding to each feature, *p* is the number of coefficients or columns used for the features input in this model, and *j* is the number of classes of the response variable.

Analysis and Findings

Based off our sentiment analysis scores for certain choices of n-grams, there were several points that stood out to us.

Pairs	**Sentiment**	**Pairs**	**Sentiment**
('customer', 'service')	0.6285	('roast', 'beef')	0.9314
('freaky', 'fast')	0.5041	('meat', 'cheese')	0.9246
('staff', 'friendly')	0.9186	('italian', 'sub')	0.92415
('drive', 'thru')	0.5829	('fresh', 'bread')	0.92625
('great', 'service')	0.9316	('lettuce', 'tomato')	0.9287
('10', 'minute')	0.21035	('cheese', 'steak')	0.93555
('15', 'minute')	-0.0078	('hot', 'pepper')	0.9136
('wont', 'back')	-0.2783	('meatball', 'sub')	0.92015
('order', 'wrong')	-0.54985	('bag', 'chip')	0.6379

Table 0: Bigram Paris and Median Sentiment Scores

It's interesting to observe that the phrase 'order wrong' have a weak sentiment of about -0.5, while reviews that mention "staff friendly" have a sentiment score of 0.9. It is also worth noting that most mention of food items relate to high sentiment scores.

For the linear model, our feature words and corresponding coefficients are shown in Table 1.

Features	Coefficient
bread	-1.54970724
staff	2.1517492
beef	2.87647606
lunch	0.44702267
cheese	1.07377112
fresh	1.30268824
fast	1.6430803
flatbread	-1.64876142
teriyaki	-2.42673023
meatball	-0.29871609
italian	2.15272223
tomato	2.69492333
steak	2.21383941

Table 1: Coefficients for Multilinear Regression

With a resulting R-squared of 0.03, this linear model does not seem to account for variation in the data well. When we calculate the accuracy of predictions on a test set, the model accuracy is about 0.41. Although this is not an accurate evaluation because linear regression is giving us continuous output that we need to round towards an actual rating, this model can allow us to interpret some of the coefficients. For example, if a store has reviews that mention "Italian" (most likely referring to the sandwich name), on average the ratings would increase by about 2.2 stars given that the other features were included in the reviews. Meanwhile, restaurants that have reviews mentioning "teriyaki" will decrease the star rating by an average of 2.4. Along with teriyaki, reviews about bread, flatbread, and meatball will have some negative impact on star rating, while mention of freshness, cheese, tomato, and steak have a positive influence. Because "meatball" seems to have a slightly negative coefficient that's small compared to others, we went back to look at reviews mentioning meatball sub orders. Restaurants such as Firehouse have mostly positive reviews for this sub, but Subway has some mixed reviews that may be resulting in the negative coefficient value, since our dataset does have most observations for Subway. Words related to service such as "staff" and "fast" are also positively correlated with ratings, and reviewers who bought their lunch at these places generally left ratings on the higher end.

For the multinomial logistic regression model, we decided to fit and train the model with the same 13-word vectors as was used in the previous model. With cross-validation, we calculated an accuracy of about 0.42. When we split the data into 80% training and 20% testing sets, the accuracy of our model on our test set is about 0.43. While there is much room for improvement, this model seems to perform better on our test data than MLR.

Improvements

One possibility for most of the model's predictions being 3-star is because the vast majority of actual training data is within the 2.5-3.5 range. Since this is a multi-class model, one method of improvement would be to find appropriate weights for the different response values in order to possibly reduce the influence of the 3-star ratings on predictions. These weights could be based off the percentage of ratings for each score. Another solution is to use machine learning up-sampling methods to fix the data imbalance issue.

Business Recommendations

Based off our sentiment analysis, we can give several useful recommendations regarding food menu items and restaurant service for sandwich shops in the US. From the sentiment analysis aspect, some of the most popular food items that businesses can develop or advertise includes roast beef, cheese steak, Italian sub, meatball sub, and hot pepper. For service, friendly staff and good service is always a plus.

From the linear regression coefficients, we can conclude that sandwich shops in general should focus more their bread quality, especially the flatbread that they are serving. Whether it's freshness or taste, the user reviews reflect the need for improvements in this area of the menu. Condiments such as cheese and tomato along with meat options like beef and steak are quite popular and are included in some positive reviews, so we suggest doing some promotional sales based off subs or combos that come with these ingredients. Lastly, we don't recommend a franchise such as Jimmy Johns that doesn't have a meatball sub or teriyaki chicken option to add these to their menu, since reviews from other stores don't seem to reflect positively on these menu items.

On the customer service side, speed is also positively correlated with star rating, but our sentiment analysis results show that there isn't a particularly strong positive or negative sentiment associated with wait time in minutes, so we don't recommend focusing on this aspect of staff training. Instead, staff could bring up Yelp ratings by having a considerate and caring attitude towards customers.

Conclusion

Although we're able to give many recommendations based on statistical analysis, there are many areas of further research. To improve our predictive models, we suggested putting some weights on the logistic regression model. We could also use statistical methods such as chi-squared goodness of fit testing on each word feature vector, in order to better select word features for both our models. One last area would be to look at combining words that have similar meaning, so that we can merge some columns that have very sparse word counts or TF-IDF scores. This might help our models' prediction accuracy and lessen the amount of word features we need to consider for our models.

References

Graves, A. (2021, September 20). *Why every restaurant should care about customer ratings*. Bloom Intelligence. Retrieved December 6, 2022, from https://bloomintelligence.com/blog/care-about-customer-ratings/

Sentiment Analysis Guide. MonkeyLearn. (n.d.). Retrieved December 6, 2022, from https://monkeylearn.com/sentiment-analysis/

Anand, A. (n.d.). *What is a bigram language model?* Educative. Retrieved December 7, 2022, from https://www.educative.io/answers/what-is-a-bigram-language-model

Member's Contribution:

1) AQ built linear and logistic regression models, wrote methodology, analysis and recommendations parts of the executive summary and presentation. Created the Github repository.
2) MZ worked on data wrangling and processing, performed sentiment analysis by writing Python code. Worked on the corresponding part of the executive summary and presentations. Built the R Shiny application.
3) CZ Worked on data cleaning, executive summary and presentation.