Yelp Reviews: Sentiment Analysis and Topic Modeling

Mike Rabayda, Judy Wang

May 6, 2023

GOOGLE DRIVE LINK:
https://drive.google.com/drive/folders/1a5qeBzPbNFUcVos53tb7b-mPsID7-rVL?usp=sharing

1 Introduction

In today's digital age, businesses and customers alike rely heavily on online reviews to make informed decisions. Online review platforms like Yelp have become a popular source of information for people looking for restaurants, hotels, and other services. However, the vast amount of data available on these platforms can be overwhelming. To make sense of this data, researchers can use datasets like Yelp's to extract insights that can help businesses better understand their customers' needs and preferences. We will discuss two research questions that use the Yelp dataset to get reviews from restaurants and user data to identify important topics.

The first research question aims to use the Yelp dataset to analyze reviews from restaurants and investigate if we predict the number of stars given the text of the reviews. This can provide valuable insights into what customers are looking for in a business, what aspects they value most, and what factors influence their decision to engage with a business. By analyzing reviews from different types of businesses, we can identify trends and patterns that can help business owners make informed decisions about their customer service and marketing strategy. For example, they may discover that customers are interested in vegan or vegetarian options for restaurants and/or that they prioritize cleanliness and hygiene in all of the businesses they visit.

The second research question focuses on using reviews to identify topics that are most prevalent in restaurants and thus the most important for users. This can be done by collecting data on users' reviews for a specific restaurant and seeing recurring topics in that specific restaurant. By analyzing this data, we can identify topics that users are most interested in at a specific business, which owners and managers can evaluate to make sure that the topics shown are representative of their business.

After using sentiment analysis using linear regression, it was found that the positive word that led

to a 5-star rating was based on authenticity, specific cuisines, quality, and uniqueness, and 1-star was based on bad service, texture and taste of food, and specific cuisines. From topic modeling, it was found that the top 5 restaurants with the most reviews had typical topics that corresponded to what the restaurant was known for in terms of items.

2 Dataset Collection

For this project, we collected data from the Yelp dataset. This dataset provides access to Yelp's database of businesses, user reviews, and ratings. This dataset contained two main files, business.json and reviews.json. We merged these two files based on the unique business id, which allowed us to match each review with its corresponding restaurant.

To create our corpus of reviews for analysis, we then preprocessed the data by removing stop words, punctuation, and other non-alphabetic characters, as well as stemming and lemmatizing the remaining words. We used tf-idf as the vectorizer to generate word embeddings, with a minimum document frequency (min_df) of 1000 to filter out infrequent and potentially irrelevant words. Additionally, for the first research question we further filtered the features to only include adjectives, which are likely to provide more informative and relevant information about the customer's dining experience. The resulting corpus consists of a set of reviews with associated adjectives, which we then used to train our linear regression model to predict the overall star rating of the restaurant. For the second research question, we filtered out the features to only be from nouns for topic modeling. This is so we got better results that were more representative of the restaurant.

The resulting corpus consists of over 4,000,000 documents, each representing a single review. Each document includes the following fields: review ID, user ID, business ID, star rating, number of useful, funny, and cool votes, review text, and date of review. The corpus is stored in a Google Drive folder and is available upon request.

Since the dataset initially included over 4 million reviews and ended up being too large for our programs to process and we chose reviews from businesses that had the most number of reviews for their business, as we didn't want to be making predictions on new businesses or very small businesses that might not have enough data for us to make an accurate prediction. Therefore, we filtered the dataset to only include the top 10 percent of restaurants that had the most reviews. This resulted in 170,775 reviews from 12,035 unique restaurants. For our first research question, we used all of the available data to perform sentiment analysis on the reviews. However, for our second research question, we narrowed our focus and only analyzed the top 5 most reviewed restaurants individually using topic modeling. This allowed us to explore the underlying topics and themes that customers were discussing

in their reviews for each of these highly popular restaurants.

3 Computational Text Analysis Methods

For our first research question, which aims to analyze reviews from restaurants and extract topics that are related to restaurants, we used a linear regression model. Specifically, we performed a sentiment analysis on the adjectives in the reviews to extract features that are related to the sentiment of the reviews. We choose adjectives because before this we would get people's names in the sentiment analysis. We didn't think this was representative or had meaning to what we were analyzing and found that adjectives gave us the best, most coherent results. In addition, we only included features that were shown in more than 1000 reviews so we would get words that were prevalent throughout the reviews.

We then used these features as inputs to a linear regression model, which was trained to predict the star rating of each review. This model can provide valuable insights into what aspects of a restaurant are most important to customers and what factors influence their decision to visit a restaurant. By analyzing the coefficients of the linear regression model, we identified which features are most strongly associated with higher or lower ratings. Below is a representation of the sentiment analysis. To do this, we obtained the coefficients of the model and the corresponding feature names. We then sorted the feature names by their coefficients, taking the top and bottom 100. Next, we created two separate strings for the positive and negative words, where the words were repeated by the magnitude of their coefficients. We then passed these strings into the WordCloud package, along with some parameters to adjust the appearance of the word cloud, such as the maximum font size and the background color. Finally, we plotted the resulting word clouds using the matplotlib package. The end result was a visually appealing representation of the most influential words in the linear regression model.

For our second research question, which focuses on using user data to identify topics that are most important to them, we used LDA as our method of analysis. LDA is a topic modeling algorithm that can identify latent topics in a corpus of text. We used LDA to identify the most frequently occurring topics in the reviews, which can help us understand what topics are most important to Yelp users. LDA algorithm is applied to the vectorized text data to identify latent topics in the text. The number of topics is set by the user, and in this case, it is set to 3. We chose 3 topics because we found that it gave the most coherent topics. We also added a threshold to the features for topic modeling to only be included if it showed up in more than 50 reviews. Therefore every feature selected for a topic is a relevant word. We also decided to only include nouns as the features because prior to this we had a lot of words that had no meaning without context and a lot of positive adjectives. Since these restaurants

that were chosen were the top 5 most reviewed, we already know that positive adjectives would be associated with the restaurants. After placing these specifications, the algorithm identifies the top words that are most closely associated with that topic. Finally, each review in the preprocessed text data is assigned to one of the identified topics using the LDA model. The topic distribution for a specific restaurant is calculated by counting the number of reviews assigned to each topic and normalizing the counts. This provides insight into the topics that are most frequently mentioned in reviews for that restaurant. By analyzing these topics, we can gain insights into what types of restaurants users prefer and what factors influence their ratings and reviews.

4 Results, Analysis, & Interpretation

4.1 Sentiment Analysis

Top 50 Words:

vegan, cute, incredible, unique, delicious, wine, phenomenal, best, flavorful, amazing, vietnamese, nashville, delightful, sandwich, cafe, authentic, creative, glad, southern, healthy, pulled, rich, small, knowledgeable, favorite, fresh, popular, interesting, outstanding, vegetarian, perfect, fabulous, fantastic, excellent, worth, creamy, true, lovely, wonderful, exceptional, super, traditional, sweet, indian, korean, top, crispy, local, friendly, baked

Bottom 50 Words:

worst, horrible, terrible, awful, slow, poor, frozen, decent, chinese, bad, empty, greasy, low, negative, cold, rude, average, typical, fine, salad, soggy, general, ok, prime, late, least, live, last, pretty, cheap, half, past, basic, standard, consistent, le, attentive, okay, happy, total, disappointing, large, better, friday, sour, rare, mexican, polite, table, wait

Top Feature Names:

The top features suggest that customers value high-quality, unique, and authentic cuisine, as well as a positive dining experience. For example, the presence of "vegan" in the top features suggests that restaurants that offer vegan options may receive positive reviews. The presence of "cute" and "delightful" suggests that customers appreciate a pleasant dining atmosphere and experience. The words "incredible", "unique", and "amazing" suggest that customers value high-quality and creative dishes. The words "wine" and "knowledgeable" suggest that customers appreciate a good wine selection and knowledgeable staff.

The bottom features suggest that customers are sensitive to poor quality food and service, as well as long wait times and unappealing menu items. For example, the presence of "worst", "horrible", "terrible", and "awful" suggest that poor quality food and service can lead to negative reviews. The

words "slow" and "late" suggest that customers may be dissatisfied with long wait times or slow service. The words "frozen", "greasy", "soggy", and "sour" suggest that customers may be dissatisfied with the quality of food.

One potentially problematic finding from this data is the presence of some negative feature words in the top 50, such as "slow" and "rude". While these words may not be as strongly negative as others in the bottom 50, their presence in the top 50 suggests that some customers may still experience issues with service at highly rated restaurants. Additionally, some words in the top 50, such as "knowledgeable", may be indicative of a higher price point or level of formality, which may limit the accessibility of highly rated restaurants to certain customers.

Another problematic finding is the presence of cuisine-specific words in both the top and bottom features, such as "Vietnamese" and "Chinese". While these words may suggest that certain cuisines are more likely to receive positive or negative reviews, they may also reflect biases or preferences among customers that do not necessarily reflect the quality of the restaurant or its food.

In addition, we analyzed how well our model was able to do sentiment analysis, which was done through the Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). The values provide a measure of the performance of a regression model. In this case, we got a MSE of 0.24 which suggests that the model has an average error of 0.24 units, while the RMSE of 0.49 means that the model's predictions deviate from the actual values by approximately 0.49 units on average. Therefore, the model's performance may be considered moderate to good. This would make sense since the presence of some neutral or ambiguous feature words in both the top and bottom 50, such as "fine" and "typical", may make it difficult to draw clear conclusions about what factors contribute to positive or negative reviews. These words may indicate that some customers are neutral or ambivalent towards certain aspects of the dining experience, or that they may be using these words as a "catch-all" to describe an overall average or unremarkable experience.

Figure 1: A word cloud representation of the top 100 positive and negative coefficients

4.2 Topic Modeling

Looking at the topic modeling results for the five restaurants, some interesting patterns emerge. Acme Oyster House is dominated by Topic 2 with a weight of 0.398, which is characterized by words like "food," "service," and "time." However, Topic 1 and Topic 0 also have significant weights of 0.333 and 0.269, respectively. Top words for Topic 1 include "place," "oyster," and "wait," while top words for Topic 0 include "oyster," "seafood," and "shrimp." On the other hand, Topic 3 is dominated by words like "oyster," "orleans," and "dozen."

For Oceana Grill, Topic 1 is the dominant topic with a weight of 0.404, which is characterized by words like "oyster," "orleans," and "seafood." Topic 2 and Topic 0 also have significant weights of 0.337 and 0.259, respectively. Top words for Topic 2 include "food," "service," and "place," while top words for Topic 0 include "restaurant," "menu," and "experience." Topic 3, on the other hand, is characterized by words like "time," "cake," and "crab."

At Hattie B's Hot Chicken - Nashville, Topic 1 and Topic 2 have similar weights of 0.401 and 0.379, respectively, while Topic 0 has a lower weight of 0.220. Top words for Topic 1 include "line," "place," and "chicken," while top words for Topic 2 include "chicken," "mac," and "hattie." Topic 3 is characterized by words like "food," "time," and "chicken."

For Mother's Restaurant, Topic 2 is the dominant topic with a weight of 0.438, which is characterized by words like "food," "service," and "order." However, Topic 0 and Topic 1 also have significant weights of 0.327 and 0.235, respectively. Top words for Topic 1 include "mother," "orleans," and "bean," while top words for Topic 0 include "restaurant," "food," and "menu." Topic 3 is characterized by words like "place," "line," and "ham."

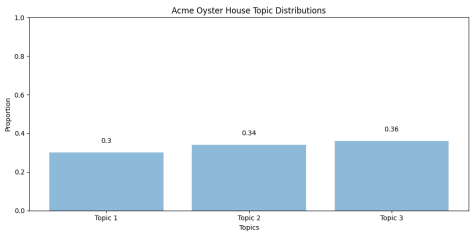
Finally, at Reading Terminal Market, Topic 2 is the dominant topic with a weight of 0.371, which is characterized by words like "place," "everything," and "food." Topic 1 also has a significant weight of 0.360, which is characterized by words like "time," "food," and "place," while Topic 0 has a lower weight of 0.270 and is characterized by words like "market," "cheese," and "meat." Topic 3 is characterized by words like "food," "market," and "place."

In addition, it is possible that some of the topics could be confused with one another. For example, at Acme Oyster House, Topic 1 and Topic 2 could potentially be confused, as they both include the top words "place" and "food". Similarly, at Oceana Grill, Topic 1 and Topic 2 could be confused, as they both include the top words "food" and "service". At Hattie B's Hot Chicken - Nashville, Topic 1 and Topic 2 could also potentially be confused, as they both include the top word "chicken". Finally, at Reading Terminal Market, Topic 1 and Topic 2 could be confused, as they both include the top word "place". Further analysis and exploration of the data could provide a clearer understanding of the topics and their relationships.

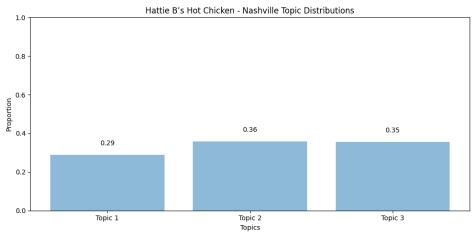
In summary, the dominant topics and their associated top words vary across the five restaurants, indicating differences in customers' perceptions and experiences. However, certain themes such as "food," "service," and "place" are common across all restaurants, highlighting their importance in customers' evaluations of their dining experiences.

Table 1: Topic Analysis for Selected Restaurants

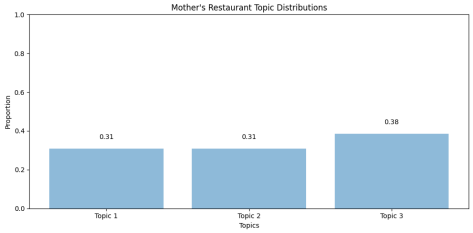
Restaurant	**Top Words for Topic 1**	**Top Words for Topic 2**	**Top Words for Topic 3**
Acme Oyster House	place, oyster, wait, seafood	food, service, time, oyster	oyster, orleans, dozen, crab
	gumbo, line, bar, order, minute	order, line, minute, bar	acme, bar, line, time
	service, food, time, orleans	place, wait, orleans, dozen	place, minute, order, service
Oceana Grill	oyster, orleans, seafood, taste	food, service, place, shrimp	time, cake, crab, restaurant
	place, food, restaurant, shrimp	server, restaurant, time, orleans	pasta, place, food, shrimp
	service, time, cake, pasta	pasta, oyster, seafood, crab	orleans, service, server, seafood
Hattie B's Hot Chicken	line, place, chicken, wait	chicken, mac, hattie, cheese	food, time, chicken, order
	side, food, order, mac	side, wait, food, order	wait, spicy, line, side
	time, hattie, cheese, spicy	line, time, place, spicy	mac, place, hattie, cheese
Mother's Restaurant	mother, orleans, bean, line	food, service, order, time	place, line, ham, order
	restaurant, time, food, order	place, line, restaurant, mother	restaurant, time, food, orleans
	place, ham, service	orleans, bean, ham	mother, service, bean
Reading Terminal Market	time, food, place, market	place, everything, food, market	food, market, place, everything
	everything	everything	time



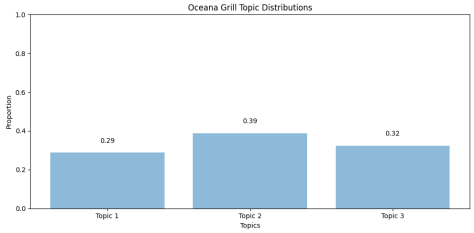
(A) Acme Oyster House



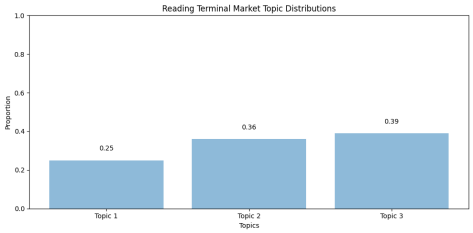
(B) Hattie B's Hot Chicken - Nashville



(C) Mother's Restaurant



(D) Oceana Grill



(E) Reading Terminal Market

Figure 2: Topic Distributions for Top 5 Most Reviewed Restaurants

5 Conclusions & Future Work

In this research, we use of Yelp's dataset to extract insights that could help businesses better understand their customers' needs and preferences. By analyzing reviews from restaurants, the researchers were able to identify the factors that led to positive and negative ratings, such as authenticity, specific cuisines, quality, uniqueness, bad service, texture, and taste of food.

In addition, the study used topic modeling to identify the most prevalent topics in restaurants and those that were most important for users. This information can be useful for restaurant owners and managers to evaluate and improve their business strategies based on what their customers are interested in.

Researching this is impactful because it provides valuable insights to businesses that can help them improve their customer experience and overall success. For example, by identifying the factors that lead to negative ratings, businesses can focus on improving those areas and potentially increase their ratings and revenue. Similarly, by understanding what topics are most important to their customers, businesses can tailor their offerings and marketing strategies to better meet their needs and preferences.

In conclusion, this study demonstrates how you can use natural language processing and data analysis to inform business decisions through reviews.