MKTG 6640 Final Project - Group 1

Text Analytics of Yelp Reviews – LA Restaurants: Learning from Extremes

ReadMe: *https://github.com/leahekblad/MKTG-6640-Final-Project---Group-1*

Objective:

This project investigates what differentiates top-ranked restaurants on Yelp by analyzing customer reviews from Los Angeles restaurants. Specifically, we compared the Top 50 and Bottom 50 of the top 240 (based on Yelp rank) to identify common themes, sentiment patterns, and operational differences that contribute to restaurant success. The goal is to provide actionable insights to help restaurants improve their rankings and customer experience.

Methodology:

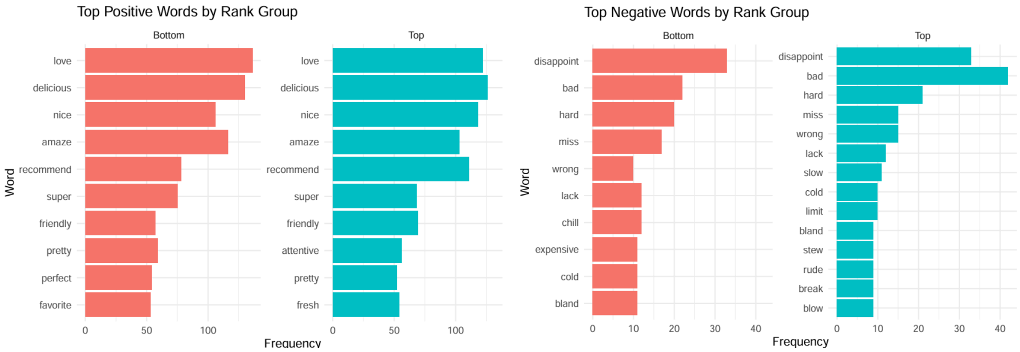
1. Data Collection

- **Source:** Kaggle – *Top 240 Recommended Restaurants in LA 2023* by Lorentz Yeung.
- **Data:** This dataset was compiled via web scraping from Yelp and includes detailed information for 240 restaurants across the Los Angeles area.
- **Key Variables:** Rank, CommentDate, Date, RestaurantName, Comment, Address, StarRating, NumberOfReviews, Style, Price.

2. Data Cleaning & Preprocessing

a. Initial Cleaning

- **Date Formatting:** Converted CommentDate and Date to Date type.
- **Categorical Conversion:** Converted RestaurantName and Style to factor variables.
- **Rank Grouping:**
  - Labeled restaurants as "Top" (Rank 1–50) and "Bottom" (Rank 191–240).
  - Additional groups were created for "Top25" and "Bottom25" based on rank.

- 

  - This analysis yielded results consistent with the initial combined top word counts: 8 of the top 10 positive words were shared across both the Top 50 and Bottom 50 groups. Similarly, many of the top negative words appeared in both groups. *Love* and *delicious* were the most common positive terms, while *disappoint* and *bad* were the most frequent negative ones.

b. Text Cleaning & Tokenization

- **Preprocess text:** Converted text to lowercase and removed punctuation and digits.
- **Tokenization:** Used unnest_tokens() to split text into individual words.
- **Stop Word Removal:** Used tidytext::stop_words to filter out common English stop words.
- **Lemmatization:** Applied lemmatize_words() to convert words to their base form.

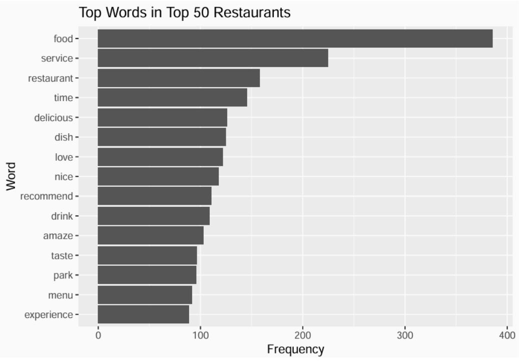
c. Document Structuring

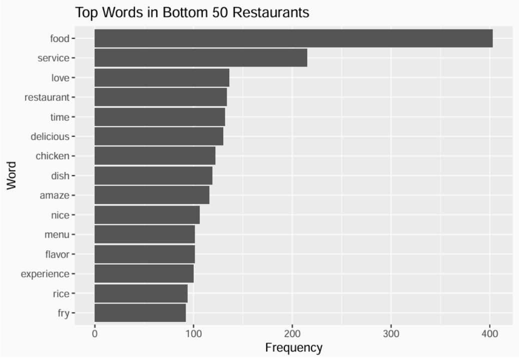
- Assigned unique doc_id to each review (for topic modeling).

3. Analysis Techniques

a. Frequency Analysis

- **Objective:** Identify most frequent lemmatized words in Top vs. Bottom restaurants.

- 



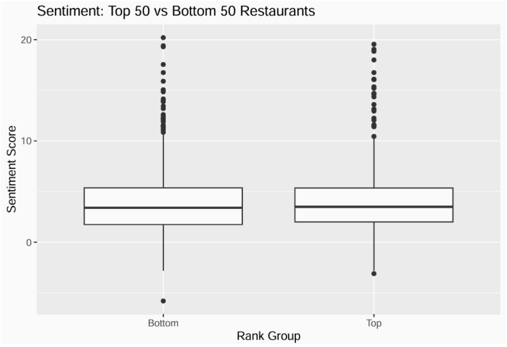
    - The word frequency chart shows notable similarities between reviews of Top 50 and Bottom 50 restaurants, with six of the most common terms appearing in both groups. Some terms, such as recommend, drink, taste, and park, appear only in one group, indicating subtle differences in customer focus. Since all restaurants fall within Yelp’s Top 240, some overlap in language is expected. To extract more meaningful distinctions, we segmented the reviews by sentiment - positive and negative - to explore emerging patterns across rank groups.
- **Tools:** dplyr::count(), ggplot2::geom_col() for bar plots.

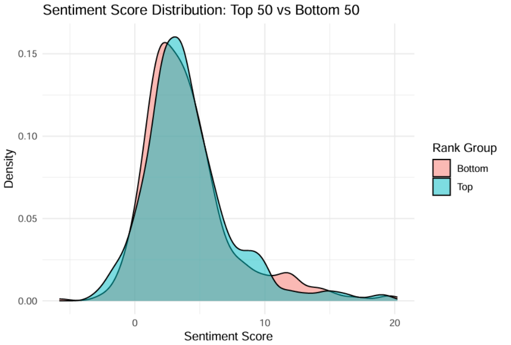
b. Sentiment Analysis

- **Lexicons Used:**
    - bing (tidytext) for word-level positive/negative classification.
    - syuzhet package for review-level sentiment scores.

- **Visualizations:** We used a comprehensive set of visualizations to explore and compare review patterns between the top and bottom ranked restaurants, such as boxplots and density plots of sentiment scores across groups.

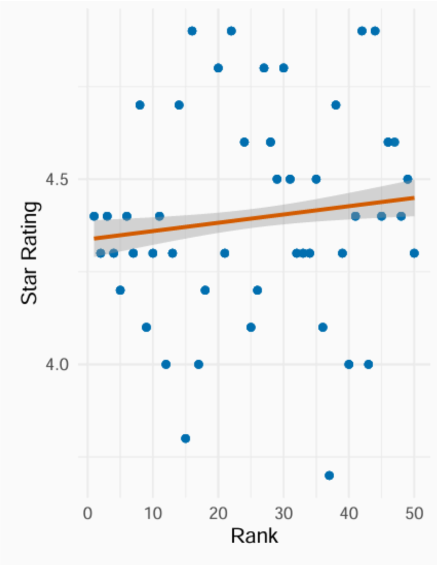
- **Output:** Summary statistics and comparison of sentiment distribution.

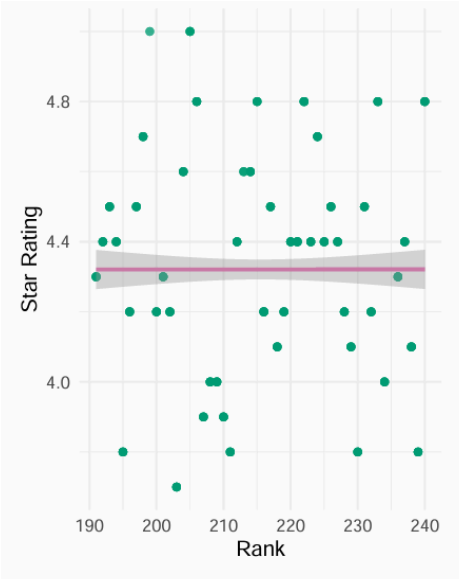
- 



    - The boxplot shows a wider sentiment range for Bottom 50 restaurants, with a lower minimum outlier compared to the Top 50 group. The density plot reveals that sentiment scores peak higher for top-ranked restaurants, with each group showing distinct local and global maxima.

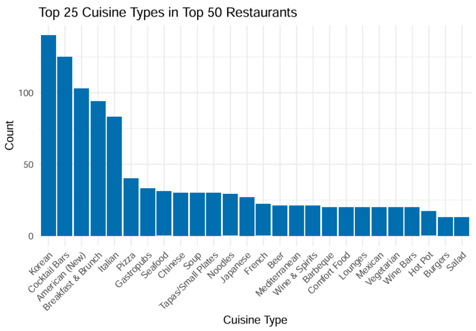
c. Star Rating Analysis

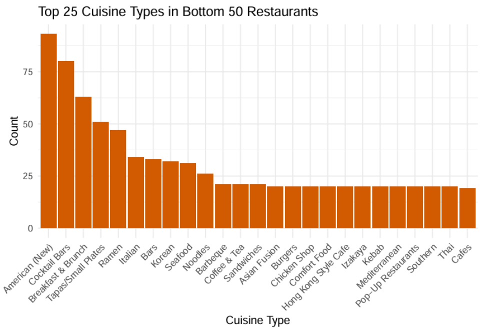
- Star Ratings vs. Rank scatterplots with regression lines for each group.
- Per-restaurant average star rating with dot plots for Top and Bottom groups.
- 



    - We plotted star ratings against Yelp rank for both the Top 50 and Bottom 50 restaurants. Among the Top 50, the trend line shows a slight increase in star ratings as rank worsens, while the Bottom 50 displays a flat trend, indicating no relationship between rank and rating. These results suggest that star ratings alone may not strongly influence a restaurant's overall rank within the top 240.

d. Cuisine Type Analysis

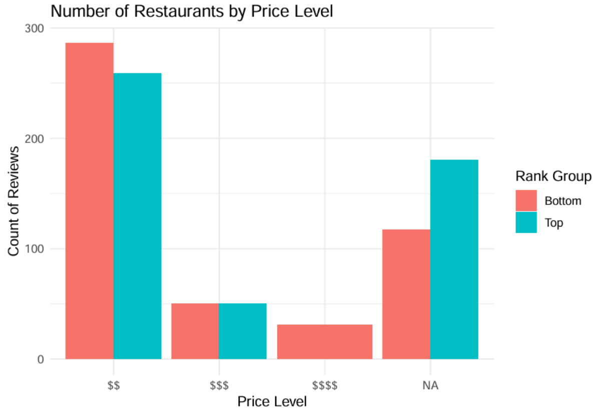
- **Multi-label Processing:** Used separate_rows() to split multiple cuisine styles.
- **Count Comparisons:** Counted and visualized frequency of cuisine types in both groups.
- 



  - The most frequent cuisine type among the Top 50 restaurants was Korean, whereas it ranked eighth among the Bottom 50. In contrast, the most common cuisine type in the Bottom 50 was New American. Coincidentily, cocktail bars were the second most frequent cuisine type in both groups. These differences may reflect distinct customer preferences or market positioning.

e. Price Level Analysis

- Converted Price to ordered factor.
- Compared distribution of restaurants by price level between Top and Bottom groups using bar plots.

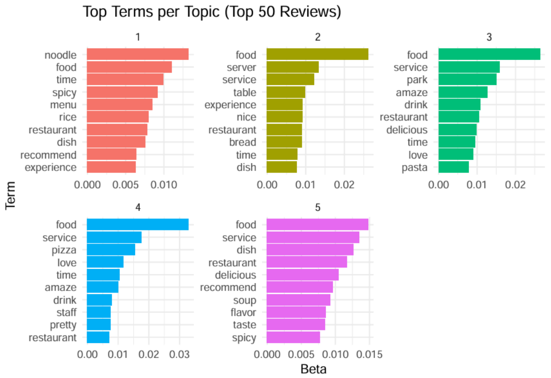


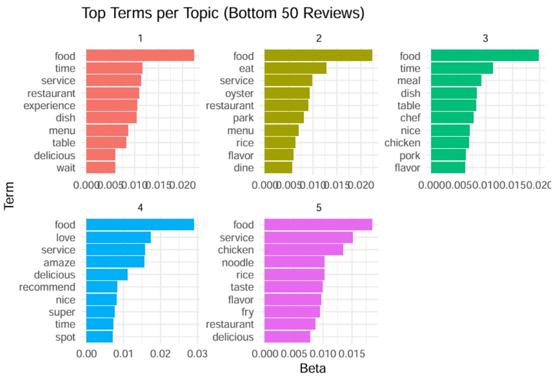
-
  - Restaurants with a lower price tier ($$) were more common in the Bottom 50 group. Notably, a higher proportion of restaurants in the Top 50 group were labeled as “N/A” for price, compared to the Bottom 50. Further investigation into the cause of these “N/A” labels may help clarify whether the missing data reflects reporting inconsistencies or a characteristic of higher-ranked restaurants.

f. Topic Modeling (LDA)

- **Technique:** Latent Dirichlet Allocation (LDA)
- **Tools Used:**
  - tidytext::cast_dtm() to create Document-Term Matrices.
  - topicmodels::LDA() with k = 5 topics.
  - tidy() from broom to extract top terms per topic.

- Separate Models for Top 50 and Bottom 50 reviews.
- **Output:** Plotted top 10 terms for each topic using faceted bar charts.

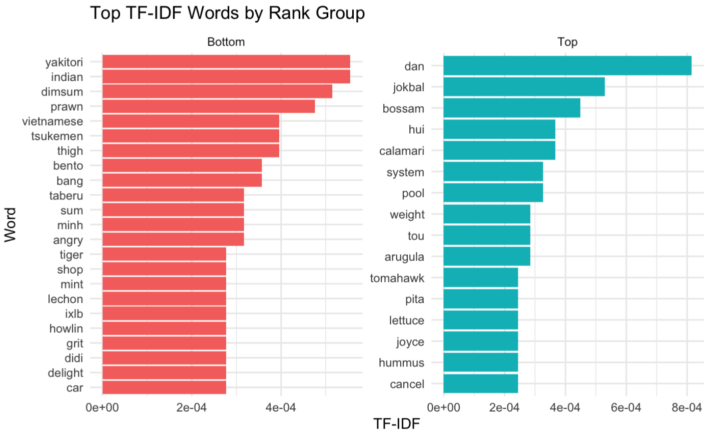




-
  - There is some overlap - for example, the word *food* appears as the most or second most frequent term across all topics. Both the Top 50 and Bottom 50 restaurants include positive terms in their reviews, suggesting that even lower-ranked establishments receive favorable feedback, and that the overall quality among the top 240 restaurants is generally high.

g. Distinctive Keywords by Rank Group (TF-IDF)

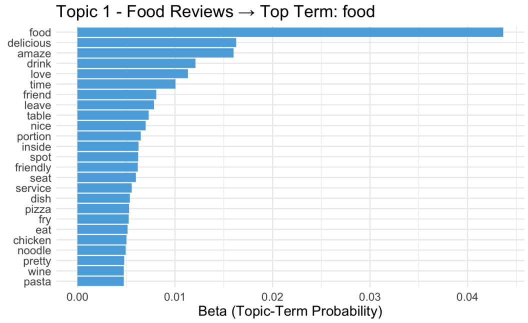
- **Technique:** TF-IDF (Term Frequency–Inverse Document Frequency)
- Highlighted terms that are uniquely important in a group of documents. We calculated TF-IDF separately for reviews of the Top 50 and Bottom 50 LA restaurants.

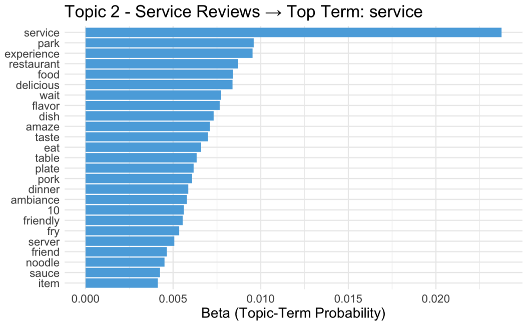


-
  - This chart shows the top TF-IDF words by group. **Bottom 50 reviews** highlight a variety of dishes and cuisines (e.g., *yakitori*, *dimsum*, *vietnamese*), suggesting diversity but a lack of cohesive identity. **Top 50 reviews** include more niche or branded terms (e.g., *jokbal*, *arugula*, *Joyce*), pointing to memorable offerings and stronger brand presence. This supports the idea that top restaurants leave a clearer, more distinctive impression in customer language.

h. **Keyword based filtering and LDA topic modeling**

- **Technique:** Keyword-based filtering and LDA topic modeling
- Separately analyzed reviews centered on *food* vs. *service*.
- **Output:** Each model uncovered the dominant themes reviewers associate with those aspects of the restaurant experience.

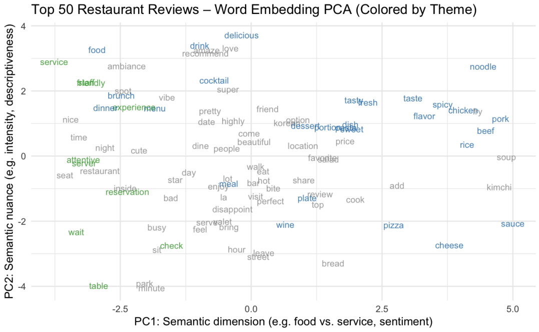

Topic 1 - Food Reviews → Top Term: food

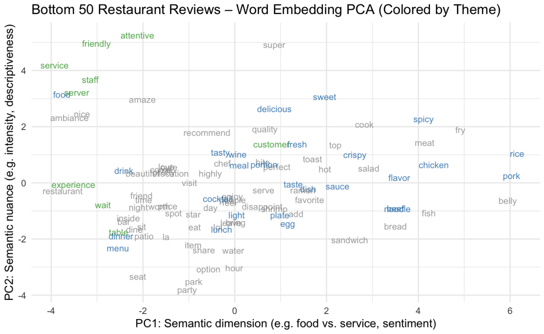

Topic 2 - Service Reviews → Top Term: service

-
  - We split reviews into food- and service-focused groups using keyword filters, then applied LDA to each. We didn't separate by rank due to data limitations and overlapping themes. Here is an example of one topic for food and one for service. **Food reviews** used emotional, sensory language (*delicious*, *love*, *ambiance*), while **service reviews** focused on logistics and issues (*wait*, *server*, *complain*).This shows that food evokes , while service prompts functional feedback.

i. **Word Embedding**

- **Technique:** We trained separate GloVe models on Top 50 and Bottom 50 reviews
- **Output:** 2D PCA plots with words color-coded by theme: blue (food), green (service), and gray (other).


Top 50 Restaurant Reviews – Word Embedding PCA (Colored by Theme)


Bottom 50 Restaurant Reviews – Word Embedding PCA (Colored by Theme)

-
  - **Top 50** reviews show clear clusters—food and service terms group by theme and emotion. **Bottom 50** reviews are more scattered, with less distinction and more neutral or vague terms. This reflects stronger, more consistent language in top reviews and weaker focus in bottom ones.

External Context & Industry Trends:

Before analyzing the review data, we wanted to research the current landscape of consumer behavior in the restaurant industry. Yelp's 2025 State of the Restaurant Industry Report notes a clear shift in dining preferences: customers are increasingly drawn to upscale, experience-driven establishments. Searches for restaurants in the "$$$" and "$$$$" price ranges rose by 10% and 17%, respectively, compared to pre-pandemic levels. At the same time, interest in more budget-friendly options slightly declined. Rather than simply seeking affordability, diners appear to prioritize uniqueness, quality, and atmosphere.

This trend helps explain some of the patterns we later observe in our text analytics. For instance, the prominence of Korean cuisine among top-ranked restaurants may reflect both local demographic influence and growing interest in culturally distinctive dining. Similarly, the emotionally expressive language found in many Top 50 reviews aligns with the industry's broader movement toward creating memorable, shareable experiences. With this context in mind, we now turn to our results to explore how these themes emerge from customer feedback.

Results:

Sentiment & Language:

- Top 50 reviews are emotionally rich and focused—especially around food.
- Bottom 50 reviews are less cohesive, often blending food and service issues.

TF-IDF Keywords:

- Top 50: Terms like prawn, yakitori, dimsum reflect upscale experiences.
- Bottom 50: Words like lettuce, hummus, cancel suggest casual or problematic themes.

Topic Modeling:

- Top 50: Themes center on food quality, ambiance, and recommendations.
- Bottom 50: Focused on wait times, inconsistencies, and vague service comments.

Word Embeddings:

- Top 50 reviews show tight semantic clusters by theme (food, service).
- Bottom 50 language is scattered, reflecting weaker thematic structure.

Cuisine & Pricing:

- Top 50 = more fusion/New American, higher price tiers ($$$+).
- Bottom 50 = broader mix, lower price points.

Recommendations:

- **Enhance Emotional Impact**: Focus messaging on food experience and ambiance to match language used in Top 50 reviews.
- **Fix Service Pain Points**: Address delays, inconsistent service, and unclear staff interactions.
- **Use Review Analytics**: Monitor sentiment trends and topic shifts for early warning signals.
- **Refine Comparison Groups** *(Next Step)*: Future analysis should compare the Top 50 against LA restaurants with *low star ratings* (e.g., <3.5), not just the next 190 in the top 240. This would better highlight what truly separates top from struggling restaurants.

References

Yelp. (2025). *2025 State of the Restaurant Industry Report*. Yelp Trends & Insights. https://trends.yelp.com/state-of-the-restaurant-industry-2025

Yeung, L. (2023). *Top 240 Recommended Restaurants in LA 2023* [Data set]. Kaggle. https://www.kaggle.com/datasets/lorentzyeung/top-240-recommended-restaurants-in-la-2023