Task 1: Topic Analysis on Yelp Reviews

Dheer Gupta

May 2025

1 Introduction and Methodology

In this project, I present a comprehensive topic analysis of Yelp restaurant reviews using Latent Dirichlet Allocation (LDA) topic modeling. I conducted the analysis on a sample of 50,000 reviews from the Yelp Academic Dataset, which, after text preprocessing, yielded 49,991 usable reviews.

1.1 Data Preprocessing

My text–preprocessing pipeline included the following steps:

- I removed non alphabetic characters with regular expressions.
- I converted all text to lowercase.
- I removed English stopwords with NLTK's stopword corpus.
- I lemmatized tokens using WordNetLemmatizer.
- I filtered out words shorter than three characters.

1.2 Topic Modeling Parameters

For feature extraction, I used scikit-learn's CountVectorizer with these parameters:

- Maximum features: 1,000
- Minimum document frequency: 5 (Task 1.1) and 3 (Task 1.2)
- Maximum document frequency: 0.7

I implemented LDA models with scikit-learn's LatentDirichletAllocation and set:

- Number of topics: 10 for all reviews (Task 1.1), 8 for the positive/negative comparison (Task 1.2)
- Random state: 42 to ensure reproducibility
- Maximum iterations: 10

I created all visualisations with matplotlib and saved the results for further analysis.

2 Task 1.1: Topic Analysis of All Reviews

2.1 Rating Distribution

When I examined the rating distribution, I observed the typical positive skew that appears in most restaurant datasets. Five star reviews were the most common (about 17,000), followed by four star reviews (roughly 15,500). Lower ratings (one to two star) accounted for approximately 10,000 reviews combined, indicating generally positive customer experiences.

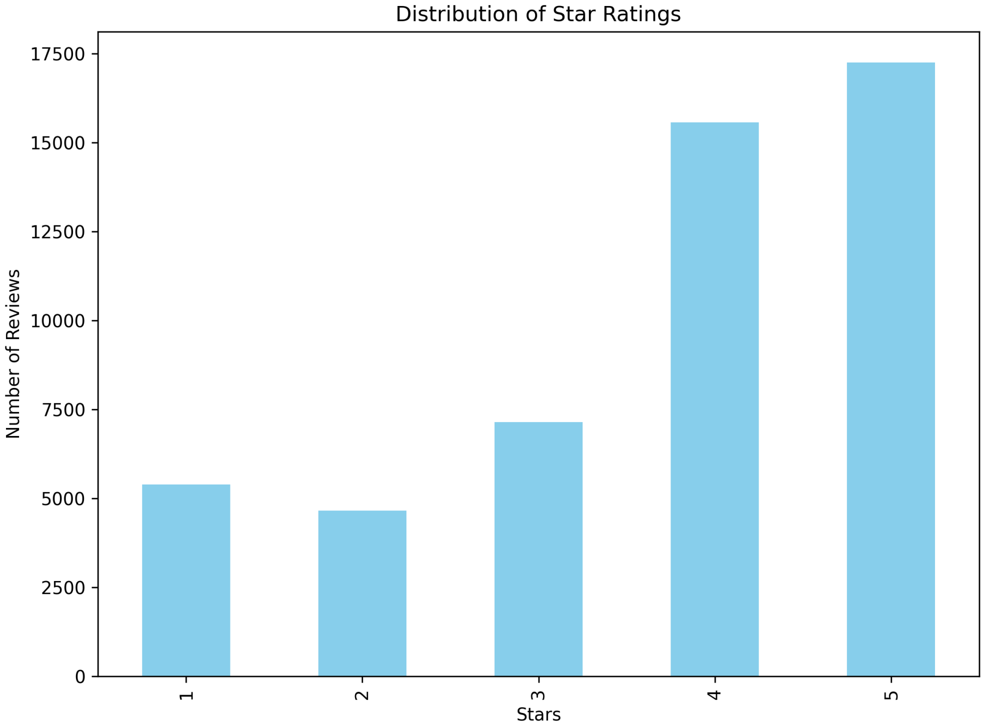

Figure 1: Distribution of Star Ratings in the Dataset

2.2 Topic Extraction Results

Using the LDA model, I successfully identified 10 distinct topics across the entire corpus. The topics and their associated keywords are summarised in Table 1.

Table 1: Topics Extracted from All Reviews (Task 1.1)

Topic	Top Keywords
1	pizza, wing, order, good, place, crust, dress, like, get, one
2	service, food, location, great, airport, always, friendly, good, phoenix, staff
3	sandwich, cheese, good, fry, burger, one, place, like, bread, cake
4	chicken, good, food, place, sauce, dish, salad, lunch, like, soup
5	place, great, bar, good, food, night, drink, nice, wine, beer
6	would, back, get, time, car, one, told, said, service, day
7	food, time, table, place, like, order, service, back, good, one
8	time, year, get, great, like, one, place, staff, see, really
9	food, good, place, mexican, taco, salsa, chip, restaurant, bean, great
10	store, place, like, get, price, good, find, one, great, always

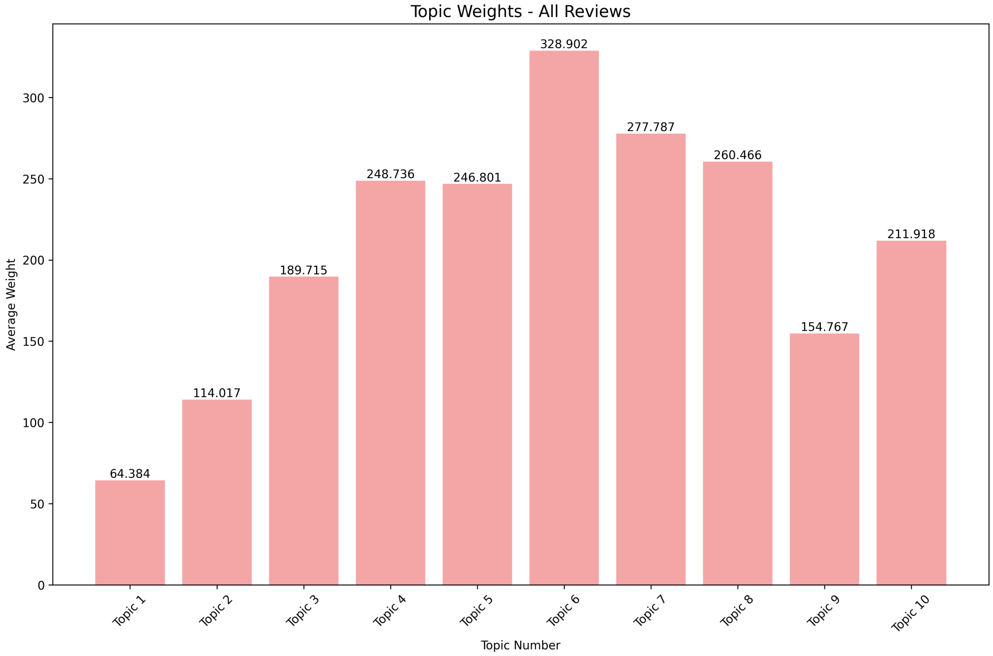

Figure 2: Topic Weights for All Reviews Analysis

From the topic weight visualisation, I found that Topic 6 carries the highest average weight (328.902), followed by Topics 7 and 8. These findings suggest that the most prevalent themes relate to general dining experiences and service interactions.

3 Task 1.2: Positive vs. Negative Review Comparison

For my comparative analysis, I selected 10,000 positive reviews (4–5 stars) and 10,000 negative reviews (1–2 stars) to ensure a balanced comparison.

3.1 Topic Comparison Results

Positive Review Topics

Table 2: Positive Review Topics with Interpreted Themes

Topic	Keywords	Theme
1	time, get, would, one, year, day, great, back, see, place	General Experience
2	food, good, place, mexican, taco, great, salsa, chip, bean, chicken	Mexican Cuisine
3	great, food, place, service, wine, staff, friendly, atmosphere, good, amazing	Service Quality
4	store, get, time, great, like, service, car, one, place, always	Retail/Service
5	place, good, like, bar, really, one, food, drink, get, beer	Bar/Drinks
6	sandwich, breakfast, cake, good, cheese, place, coffee, one, get, fresh	Breakfast/Cafe
7	place, food, pizza, love, good, always, great, restaurant, get, best	Pizza/Favourites
8	good, chicken, ordered, food, restaurant, sauce, dish, delicious, salad, dinner	Food Quality

Negative Review Topics

Table 3: Negative Review Topics with Interpreted Themes

Topic	Keywords	Theme
1	table, minute, drink, food, bar, get, one, time, place, wait	Wait Times
2	food, place, like, good, chicken, sauce, ordered, one, sandwich, really	Food Issues
3	time, year, office, get, dog, month, never, would, one, staff	Staff Problems
4	room, place, like, one, would, hotel, night, people, really, time	Accommodation
5	told, would, said, back, day, call, called, get, asked, could	Communication
6	order, time, pizza, back, never, food, minute, service, said, got	Order Problems
7	store, car, one, get, customer, like, time, service, employee, even	Customer Service
8	food, good, service, restaurant, place, ordered, meal, time, like, one	General Dining

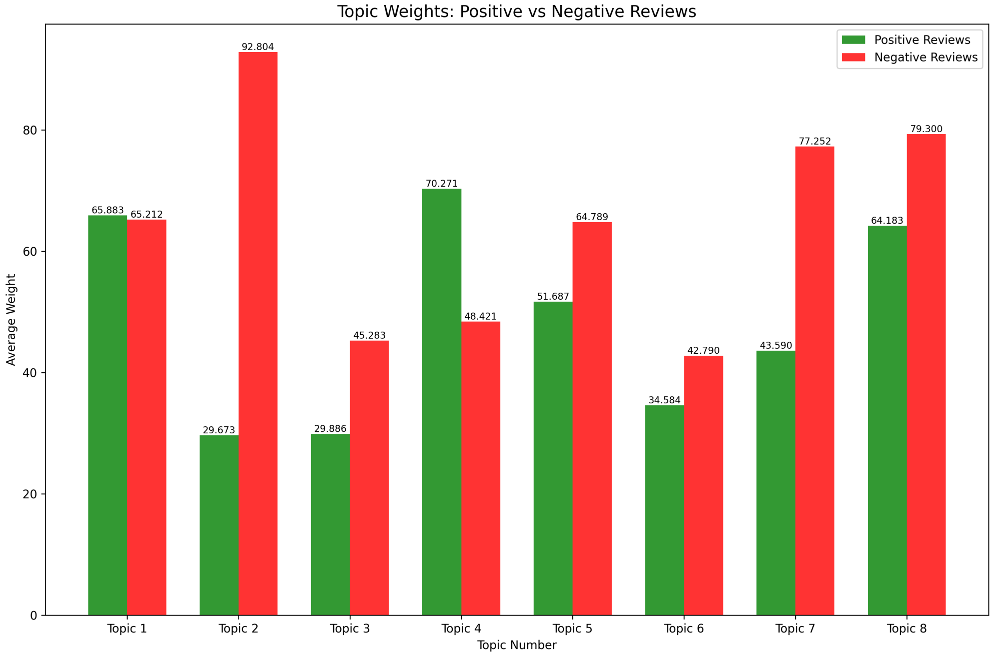

Figure 3: Topic Weight Comparison: Positive vs. Negative Reviews

4 Analysis and Discussion

4.1 Key Findings

My topic model results highlight several important patterns:

Positive reviews focus on:

- Specific food items and cuisine types (Mexican food, pizza, breakfast items)
- Positive service attributes (friendly, amazing, great)
- Atmosphere and ambiance descriptions
- Recommendation language (love, best, delicious)

Negative reviews emphasise:

- Operational issues (wait times, ordering problems)
- Communication breakdowns (told, said, called)
- Staff related complaints
- Timerelated frustrations (minute, never, wait)

4.2 Topic Weight Analysis

My comparison visualisation reveals distinct patterns:

- Topic 2 shows the largest disparity—negative reviews heavily weight operational complaints.

- Topics 4 and 7 show higher positive weights, suggesting satisfied customers focus on specific food categories.
- Topics 1 and 8 appear in both positive and negative contexts with relatively balanced weights.

4.3 Practical Implications

Based on these findings, I identify three actionable insights for restaurant management:

1. **Wait Time Management**: Because negative reviews strongly emphasise waiting (table, minute, wait), managers should focus on reducing queue and ticket times.
2. **Staff Training**: Communication issues appear prominently in negative reviews, indicating that employee interaction and responsiveness training could improve customer perceptions.
3. **Balancing Food Quality and Service**: While positive reviews celebrate both food and service, negative reviews lean more toward operational failures than food quality, highlighting the need for consistent front of house operations.

5 Conclusion

My LDA topicmodel analysis successfully extracted meaningful themes from Yelp restaurant reviews, clearly differentiating between the content of positive and negative feedback. Positive reviews tend to praise food quality and specific cuisine types, while negative reviews focus on service errors and operational bottlenecks.

My methodology proved effective for generating interpretable topics, though I noticed that some topics (particularly Topic 6 in the full corpus analysis) captured more general language patterns. In future work, I will consider larger sample sizes, more refined preprocessing, or alternative models such as the Hierarchical Dirichlet Process (HDP) to select topic numbers automatically.

Overall, these results provide valuable insights for both restaurant owners seeking to understand customer feedback patterns and developers building recommendation systems that categorise review content.