March 5, 2025

Presented by: Team 011

Ayush Trivedi
Dheeraj Pamnani
Dominic Darrah
Riya Agarwal
Sravani Bolla

YELP DATA ANALYSIS

CIS 509 : Analytical Unstructured data

TABLE OF CONTENTS

BUSINESS PROBLEM

For restaurants, customer satisfaction and loyalty are shaped by various factors, including food quality, service, pricing, and ambiance.

By exploring the relationship between reviews, customer recommendations (tips), and star ratings, we aim to identify key themes that drive positive dining experiences or lead to dissatisfaction.

By understanding what aspects of a restaurant customers frequently recommend, businesses can leverage these insights to enhance their offerings, improve reputation, and attract more guests.

DATASET SCOPE

Data Source

- Review Dataset
- Business Dataset
- Tip Dataset

Filtering Criteria

- State filtering (States: PA and FL)
- Category filtering (Categories of restaurant that serve American, Chinese, and Italian cuisine)

Source: https://business.yelp.com/data/resources/open-dataset/

SUMMARY STATISTICS

Total Review Count: 965,641

Number of Tokens	98,334,233
Number of Unique words	297,193
Number of Unique customers	351,921
Number of Businesses	8,642
Number of Reviews for PA	566,833
Number of Reviews for FL	398,808
Number of Reviews for American Cuisine Restaurants	676,072
Number of Reviews for Chinese Cuisine Restaurants	100,164
Number of Reviews for Italian Cuisine Restaurants	189,405

ANALYSIS PROCESS

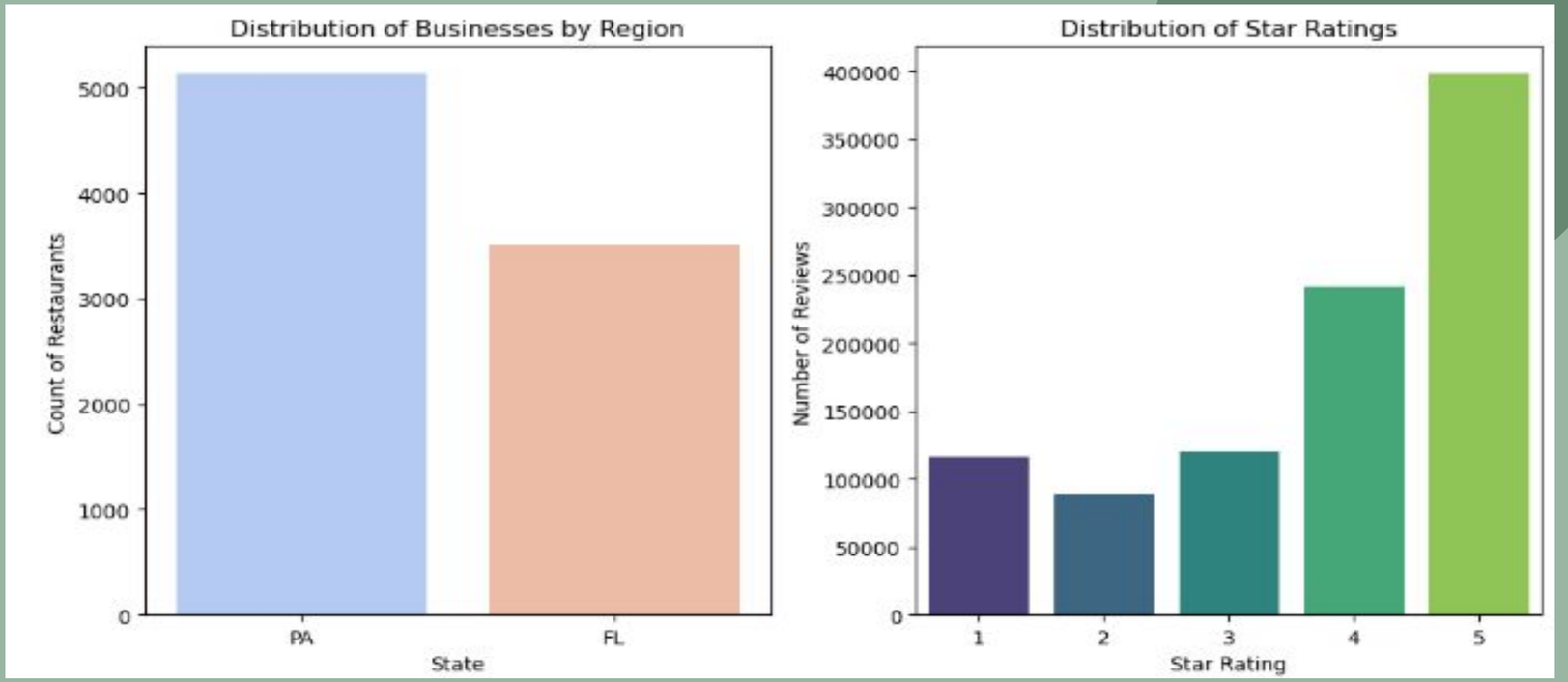
Sentiment Analysis - Word Cloud EDA among the star ratings and tips data

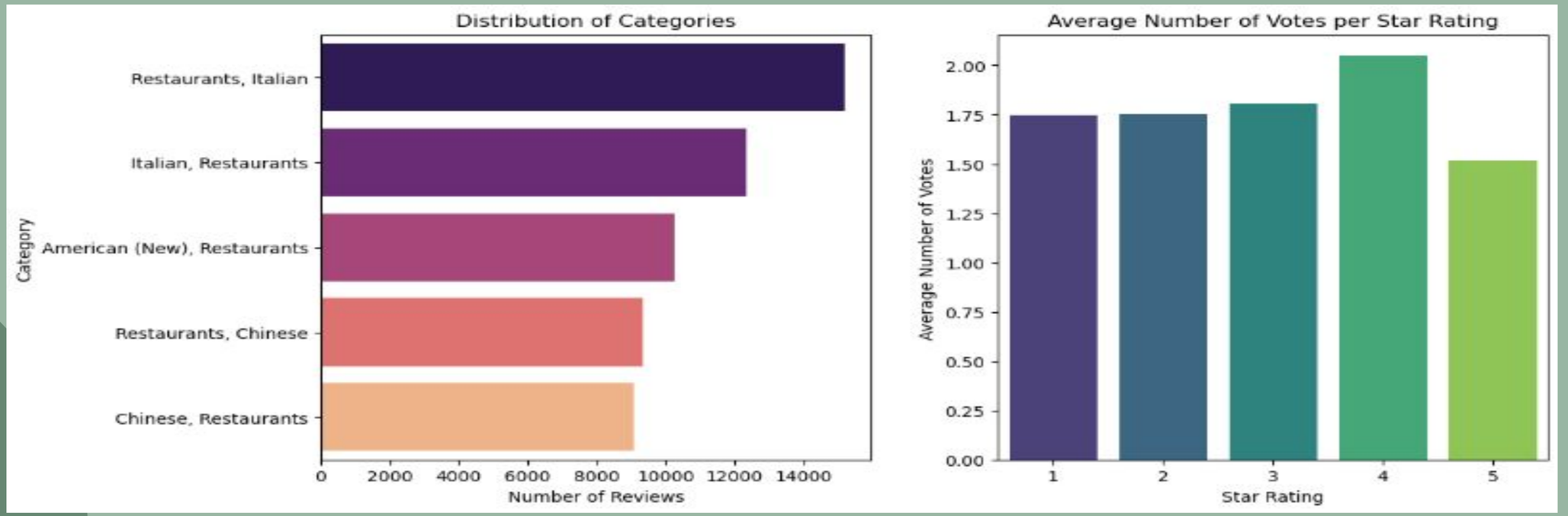
Topic Modeling for Review and Tips dataset

Sentiment Classification and Topic Modeling

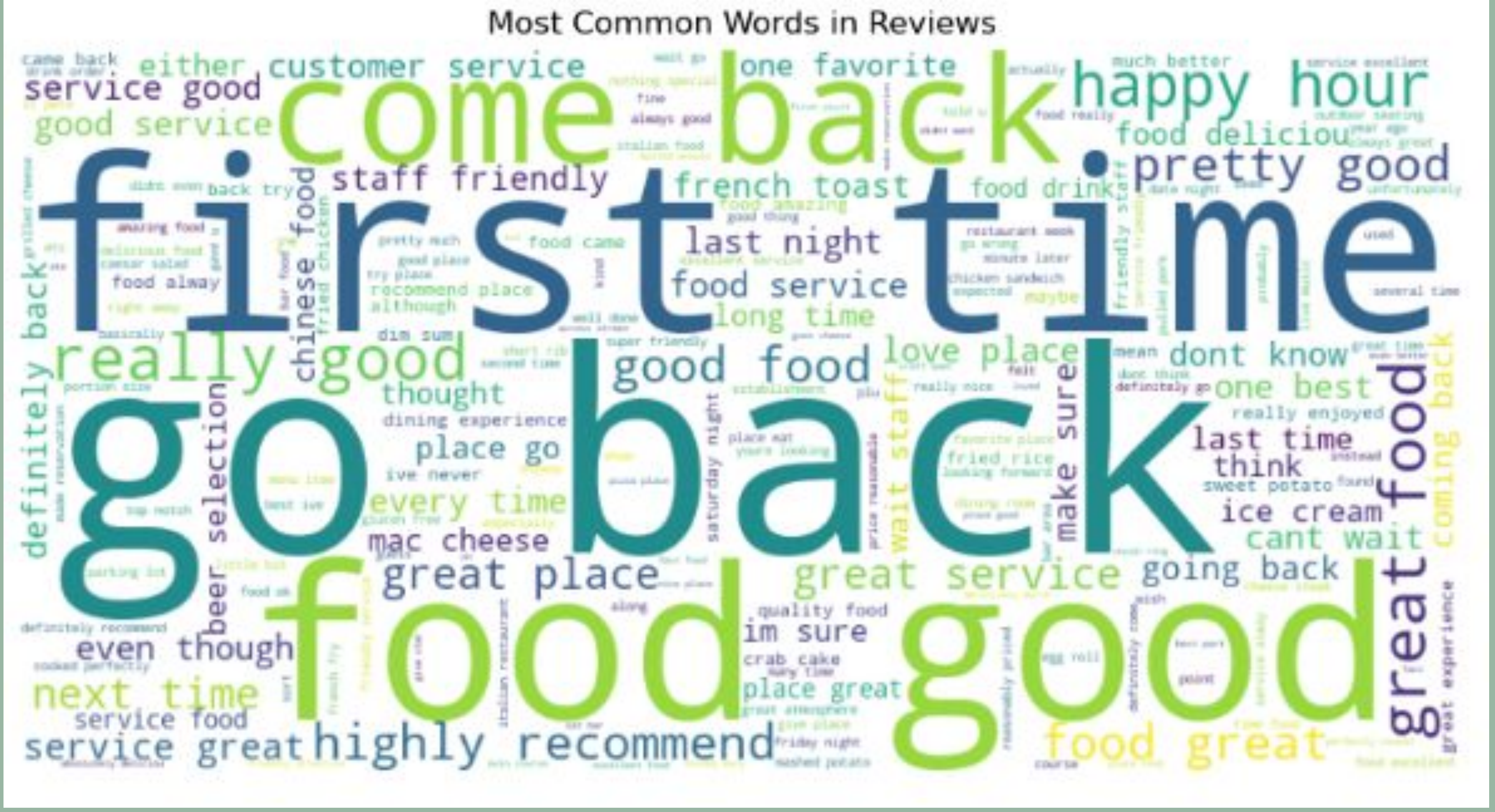
Regional and Cuisine Comparison using Topic Modeling

EXPLORATORY DATA ANALYSIS





EDA Continued...



Most Common Bigrams: [('go back', 50134), ('first time', 44208), ('food good', 41493), ('come back', 38685), ('really good', 36543), ('great food', 35663), ('happy hour', 34840), ('highly recommend', 31812), ('food great', 30288), ('good food', 29699)]

Most Common Trigrams: [('wait go back', 8055), ('cant wait go', 7612), ('cant go wrong', 6117), ('food great service', 6071), ('definitely go back', 5993), ('definitely come back', 5903), ('sweet potato fry', 5683), ('would go back', 5609), ('great food great', 5429), ('mac n cheese', 4906)]

Most Common Bigrams in Tips: [('in the', 26547), ('of the', 20792), ('and the', 19807), ('on the', 18690), ('the best', 17724), ('for a', 17660), ('food and', 17203), ('for the', 15973), ('is a', 14960), ('is the', 14002)]

Most Common Trigrams in Tips: [('This place is', 6884), ('Great food and', 4681), ('of the best', 3886), ('is the best', 3421), ('The food is', 3085), ('Great place to', 2897), ('the food is', 2812), ('I love this', 2768), ('one of the', 2467), ('One of the', 2451)]

EDA Continued...

1-Star Reviews: reveal complaints about food, service, delays, and overall experience.

5-Star Reviews: highlight praise for "food," "great," "delicious," "service," and "good," reflecting excellent dining experiences.

Topic Modeling - Review

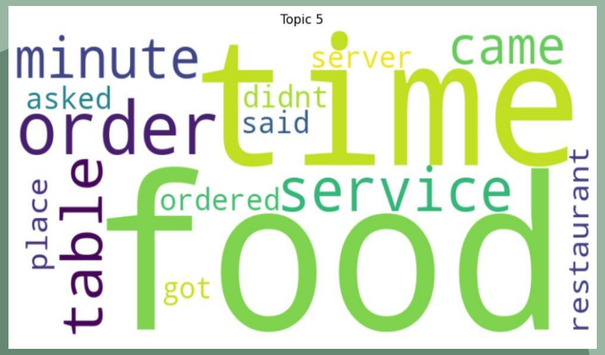
```
Topic 1:
amazing delicious beer time friendly service good place food great

Topic 2:
menu salad got food sauce delicious dish ordered chicken good

Topic 3:
food room time table night drink restaurant place like bar

Topic 4:
time fry sandwich cheese burger food like good place pizza

Topic 5:
server asked ordered came minute table service order time food
```



Topic Modeling - Tip

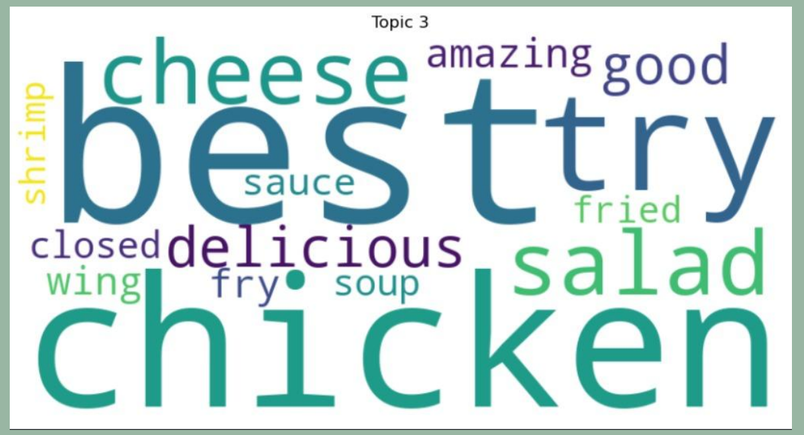
```
Topic 1:
happy menu lunch place bar hour drink beer great good

Topic 2:
amazing staff friendly excellent love place good service great food

Topic 3:
wing fry amazing good delicious salad cheese try chicken best

Topic 4:
got place dont service free bad eat order burger pizza

Topic 5:
im sandwich like pork philly time place best make dont
```



NLP METHODOLOGY

BERT Topic Modelling with UMAP

1. Data Preprocessing and Merging

Merged review ratings (sentiment based on stars) with tip and business files, cleaned the text in tip file(using NLTK stop words, standardization, removing alphanumeric characters), and filtered data by states, cuisines, and sentiment.

2. Class Balancing

Balanced all classes by -
Step1: grouping them and finding minimum group size.
Step2: Equally dividing the inter-class ratios for each class based on min-group size

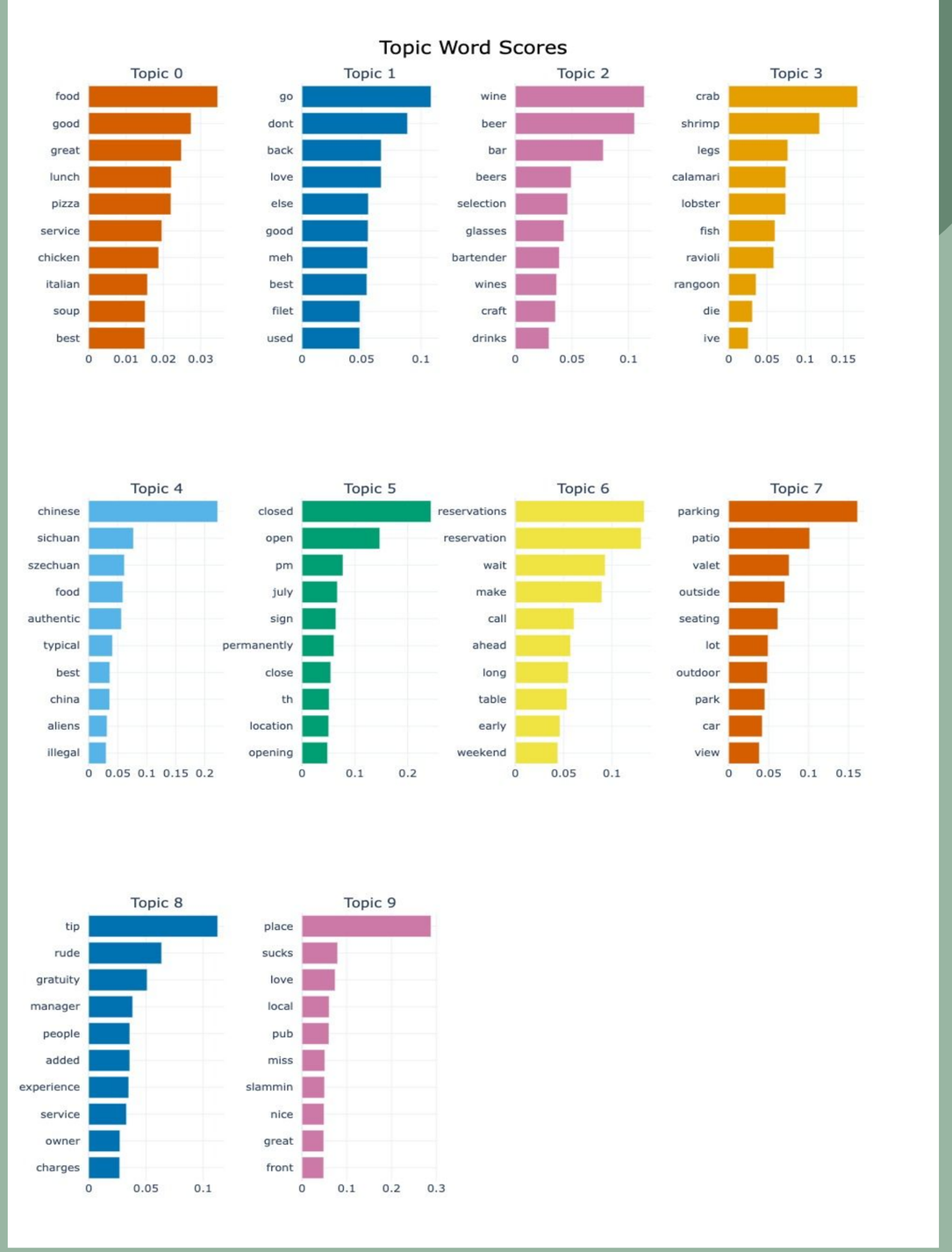
3. Topic Modeling and Visualization

BERTopic for topic modeling on cleaned tip text in merged file with UMAP reducing dimensions of embedding, further reduced model to 30 topics, and assigned topic labels and probabilities. Visualised topics generated from tip text per class (sentiment, cuisine, state)

Topic Word Scores

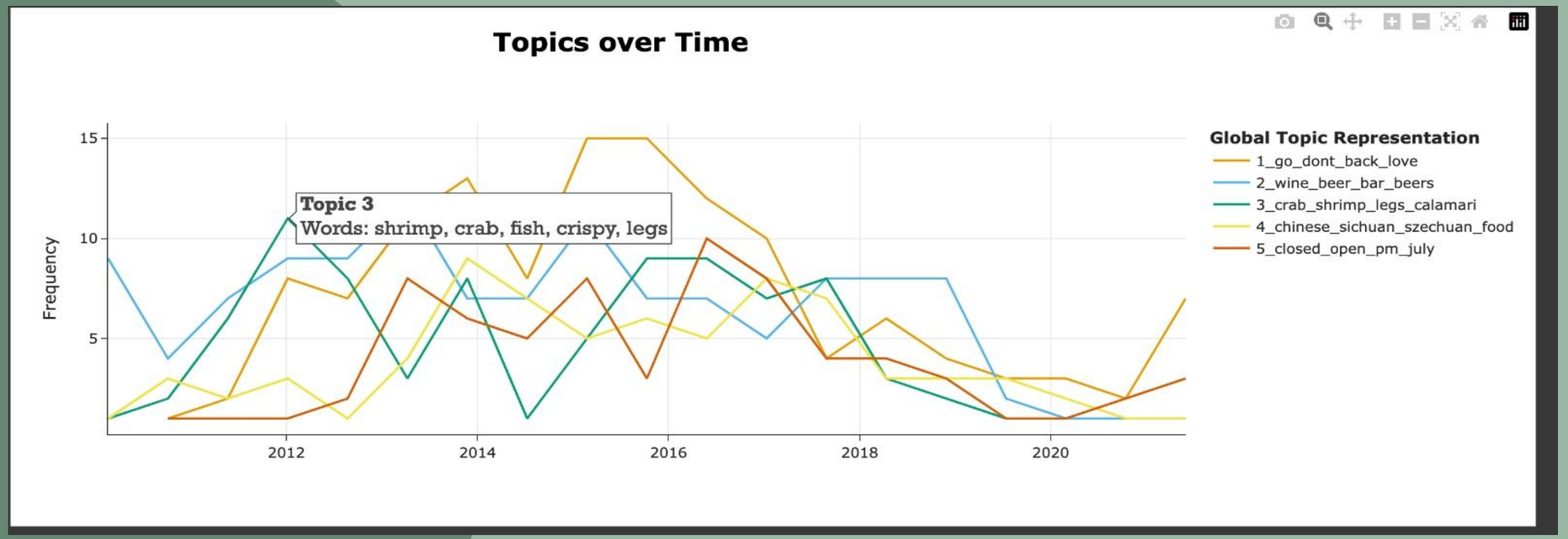
1. **Topic 0 - General Food Quality & Service**
   - **Insight:** Positive food and service experiences drive high ratings.
2. **Topic 1 - Mixed Customer Sentiments**
   - **Insight:** Polarized opinions suggest inconsistent experiences.
3. **Topic 2 - Drinks & Alcohol Selection**
   - **Insight:** A strong beverage program enhances customer satisfaction.
4. **Topic 3 - Seafood & Shellfish**
   - **Insight:** Seafood quality is a critical factor in customer reviews.
5. **Topic 4 - Chinese & Sichuan Cuisine**
   - **Insight:** Authenticity is a major driver of satisfaction in ethnic cuisine.
6. **Topic 5 - Restaurant Closures & Openings**
   - **Insight:**Closures & inconsistent availability (-ve)ly impact perception.
7. **Topic 6 - Reservations & Wait Times**
   - **Insight:** Long wait times frustrate customers, efficient reservations matter.
8. **Topic 7 - Parking & Outdoor Seating**
   - **Insight:** Accessibility & seating options influence customer decisions.
9. **Topic 8 - Tipping & Gratuity Issues**
   - **Insight:** Unclear tipping policies and rude service harm ratings.
10. **Topic 9 - General Negative Sentiment**

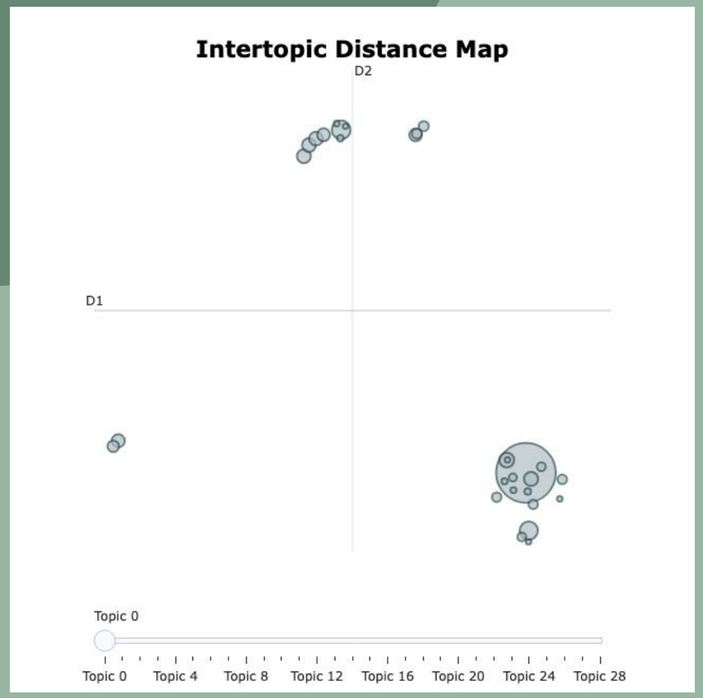
- **Insight:** Negative experiences lead to strong dissatisfaction in reviews.

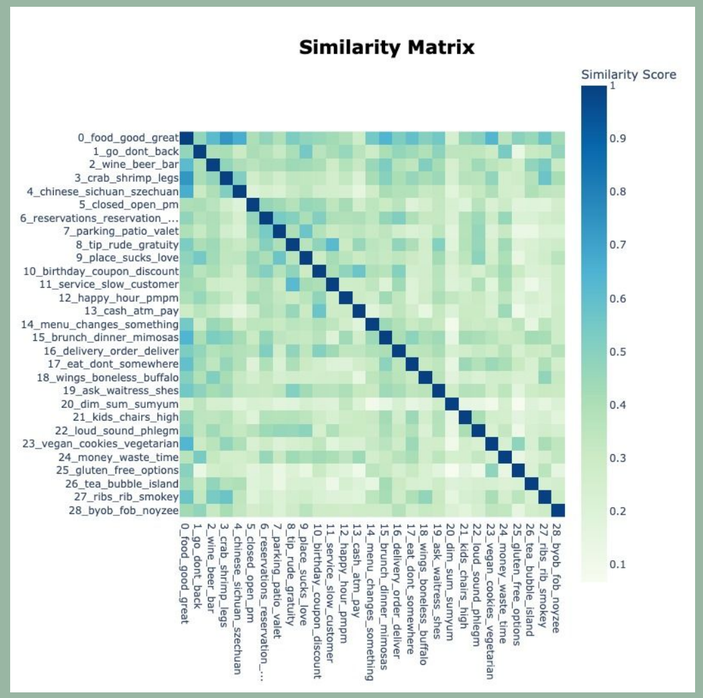


Insights/ Results

- Topics clusters, Correlation and Trend over time







x: 3_crab_shrimp_legs
y: 0_food_good_great
Similarity Score: 0.7373859

x: 23_vegan_cookies_vegetarian
y: 0_food_good_great
Similarity Score: 0.6341547

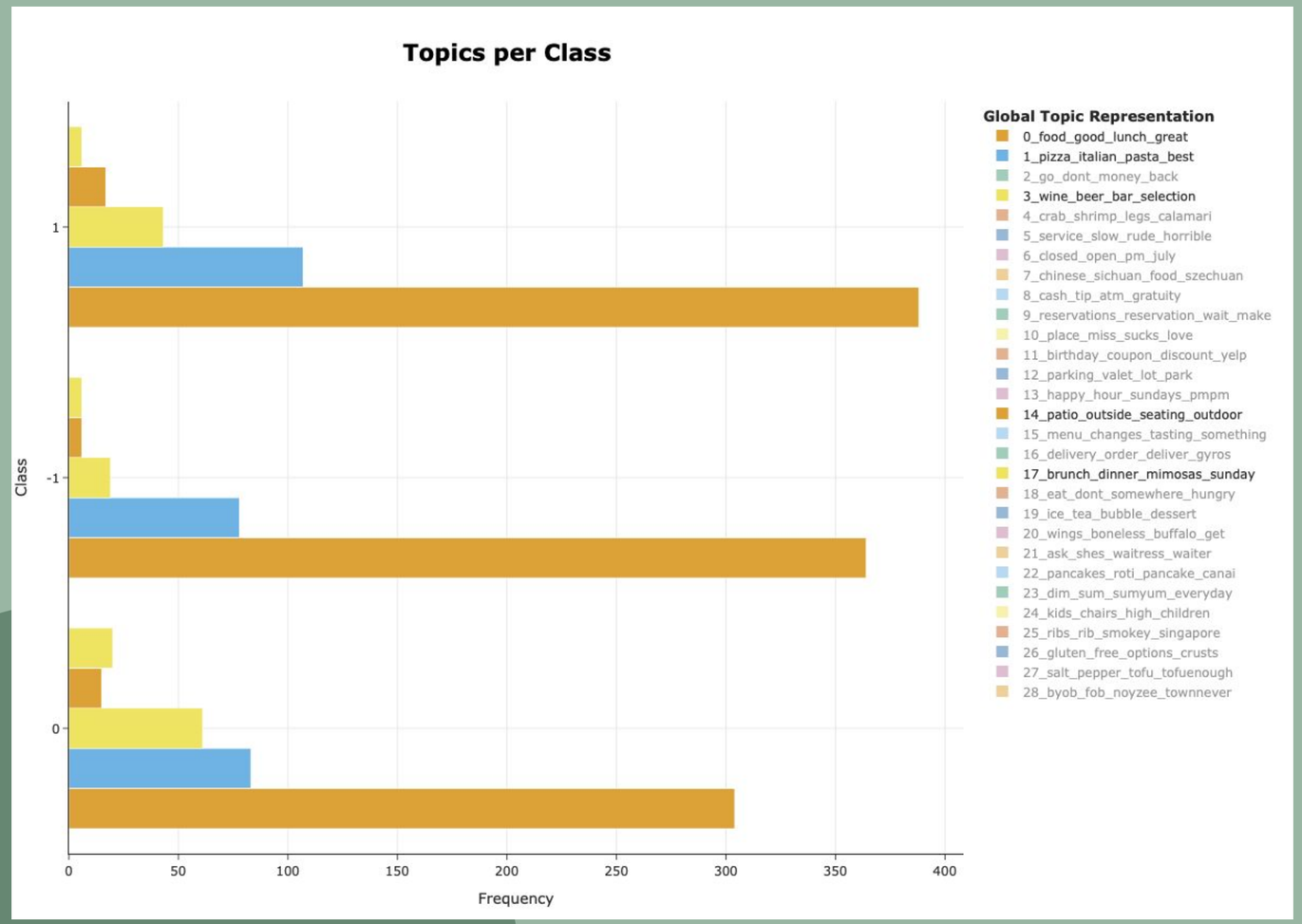
x: 0_food_good_great
y: 15_brunch_dinner_mimo
Similarity Score: 0.6405111

x: 2_wine_beer_bar
y: 18_wings_boneless_buffalo
Similarity Score: 0.5148593

x: 8_tip_rude_gratuity
y: 11_service_slow_customer
Similarity Score: 0.6159207

Insights/Results

- Sentiment Classification Topic Modeling



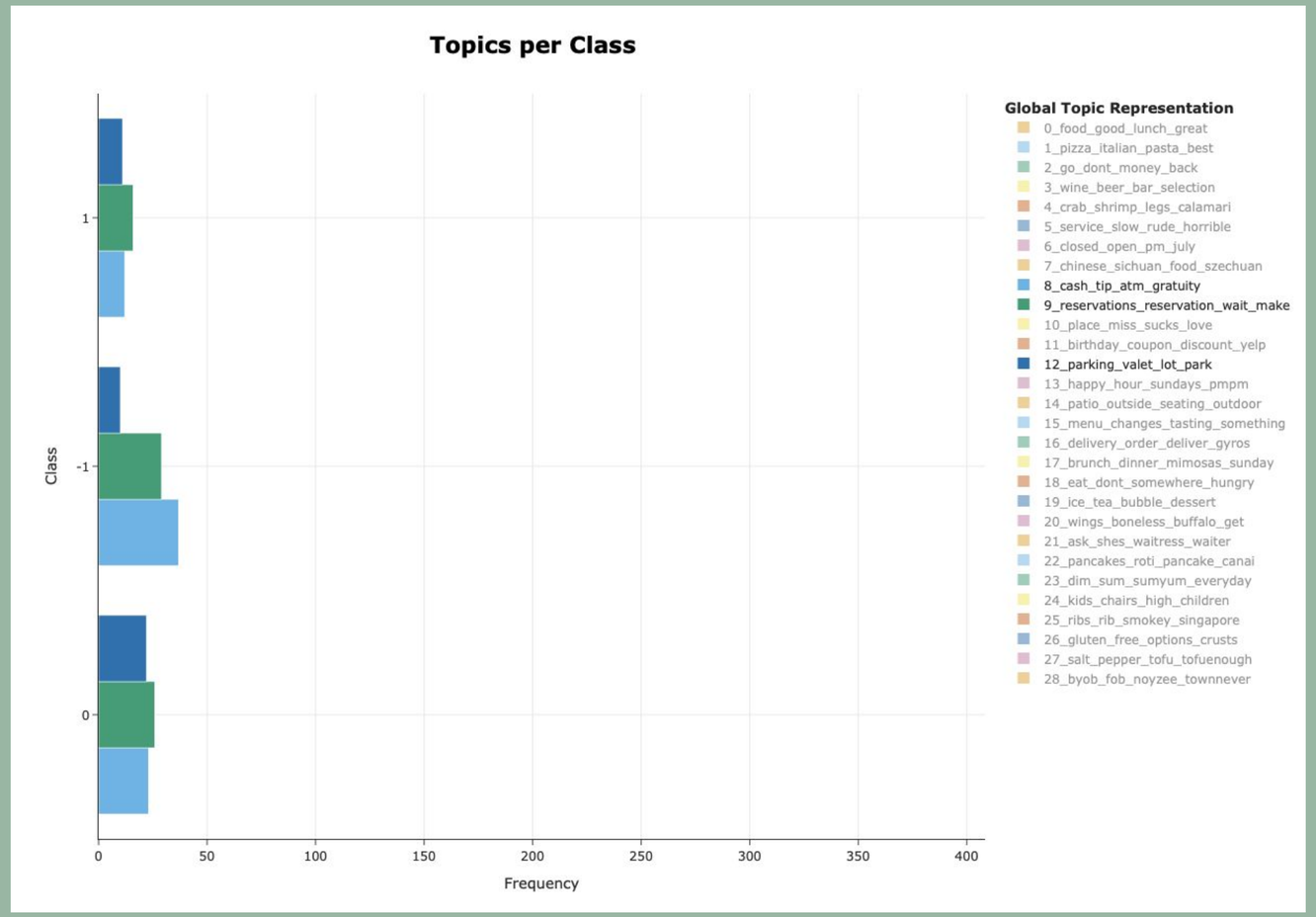
- **All sentiments:**
  - *Topics:*
    - food_good_lunch_great pizza_italian_past_best
  - Food quality is key indicator for overall restaurant experience

- **Positive sentiment:**
  - *Topics:*
    - brunch_dinner_mimosas_sunday
    - wine_beer_bar_selection
    - patio_outside_seating_outdoor
  - Dining experience correlate with customer satisfaction
    - Ambiance and specific dishes

Insights/Results

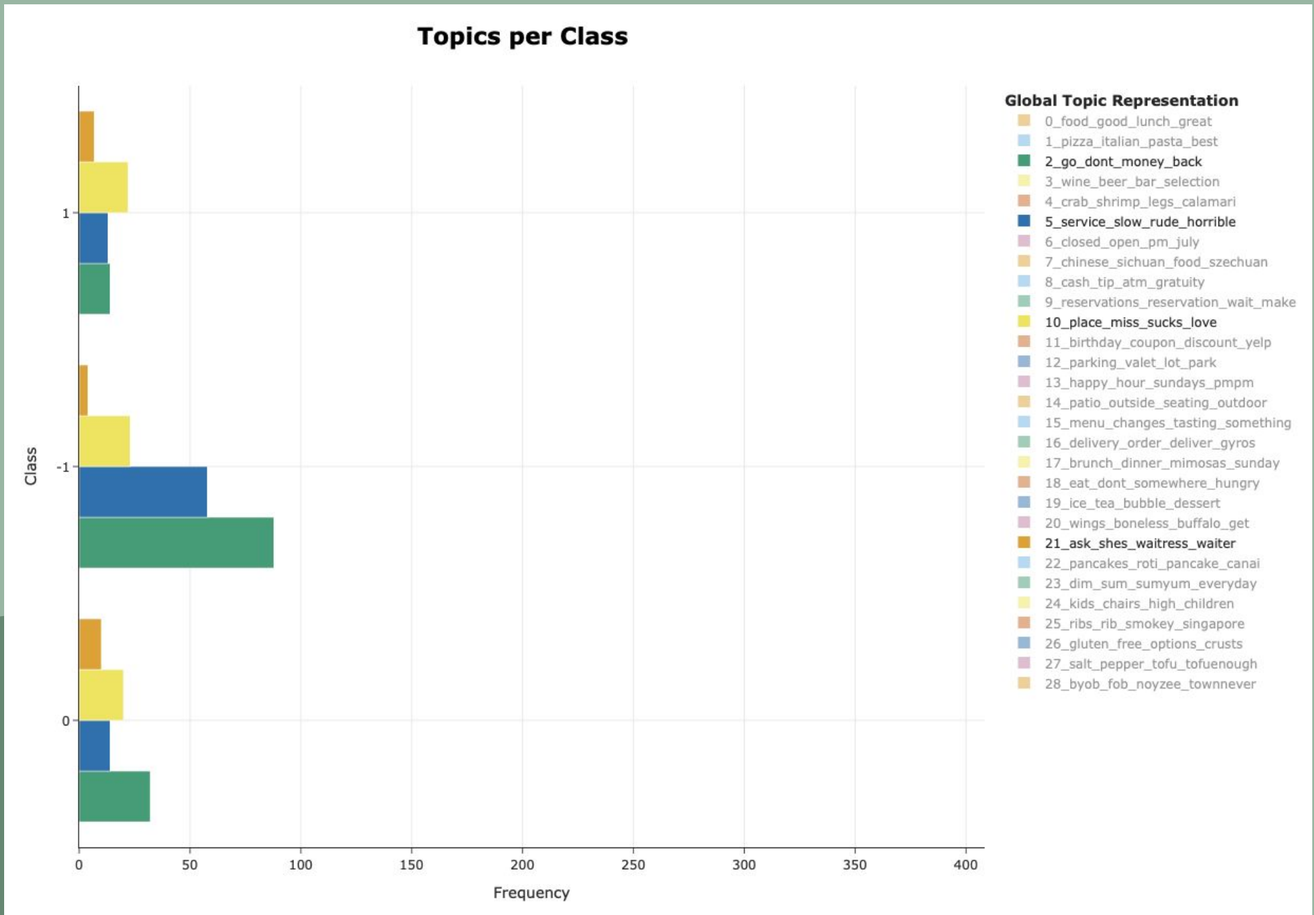
- Sentiment Classification Topic Modeling

- **Neutral sentiment:**
  - *Topics:*
    - cash_tip_atm_gratuity
    - reservations_reservation_wait_make
    - parking_valet_lot_park
  - Non-emotional administrative procedures of restaurants (informational)
    - Reservations
    - Payment
  - Logistical topics of restaurants



Insights/Results

- Sentiment Classification Topic Modeling

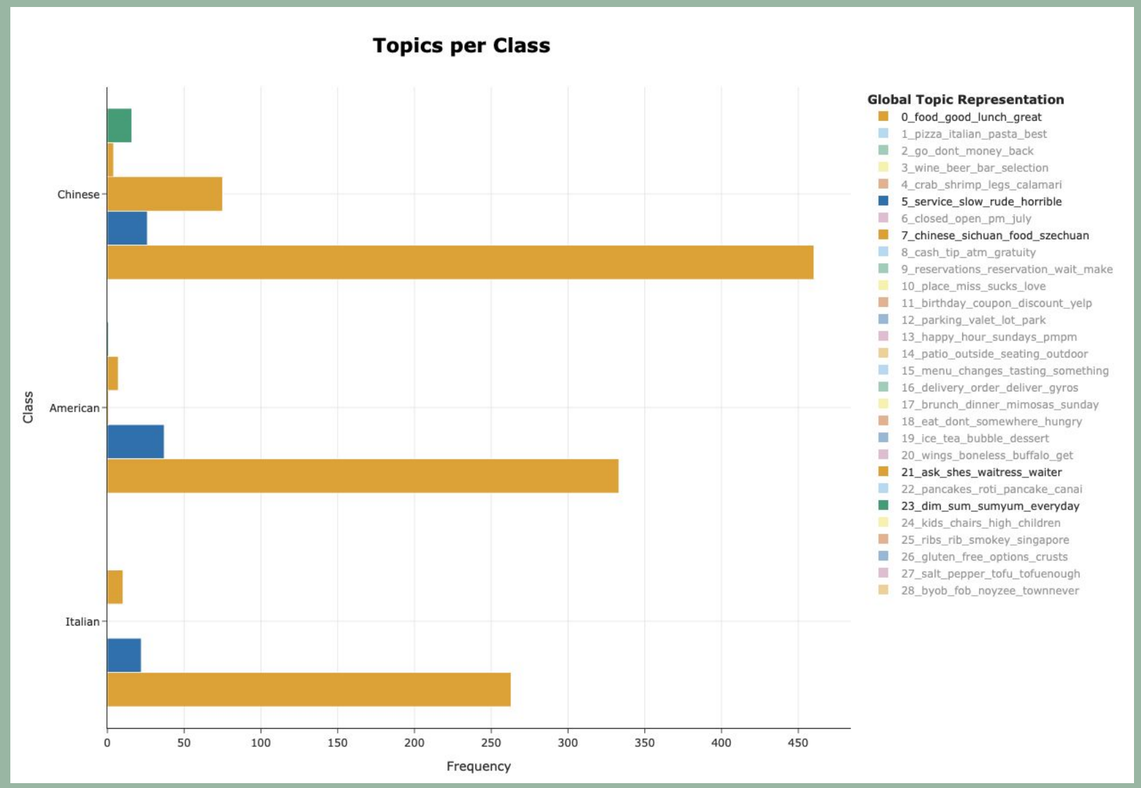


- **Negative sentiment:**
  - *Topics:*
    - *service_slow_rude_horrible*
    - *go_don’t_money_back*
    - *place_miss_sucks_love*
    - *ask_shes_waitress_waiter*
  - Poor service and refund problems
    - Operating hours
    - Negative overall experience
    - Waiting Staff issues

Insights/Results

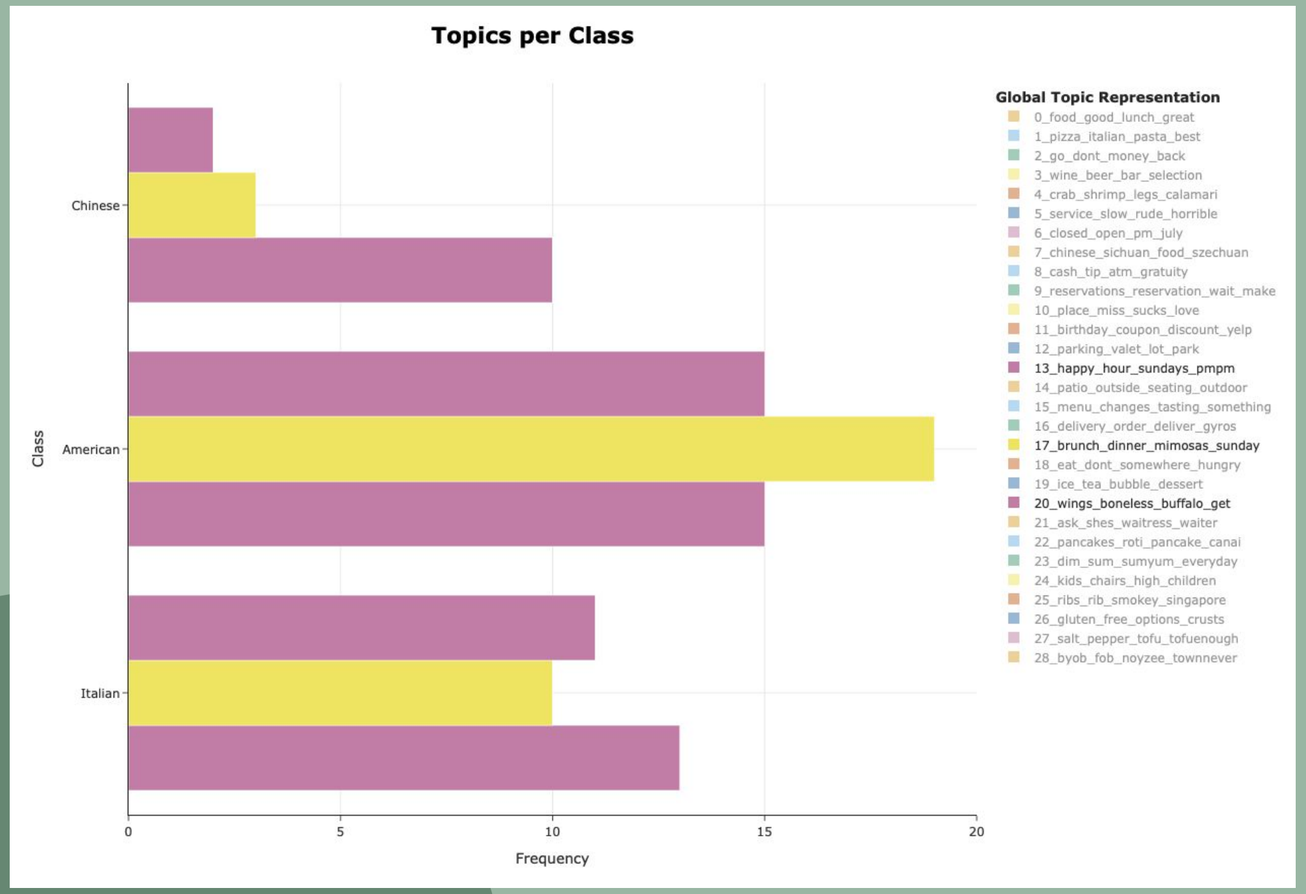
- Cuisine Classification Topic Modeling

- **All cuisines:**
  - *Topics:*
    - food_good_lunch_great
    - service_slow_rude_horrible
  - Food quality across all types of restaurants
  - Service issues among all cuisines
- **Chinese Cuisine:**
  - *Topics:*
    - chinese_sichuan_food_szechuan
    - dim_sum_sumyum_everyday
    - ask_shes_waitress_watier
  - Main dishes are important
  - Service complaints



Insights/Results

- Cuisine Classification Topic Modeling

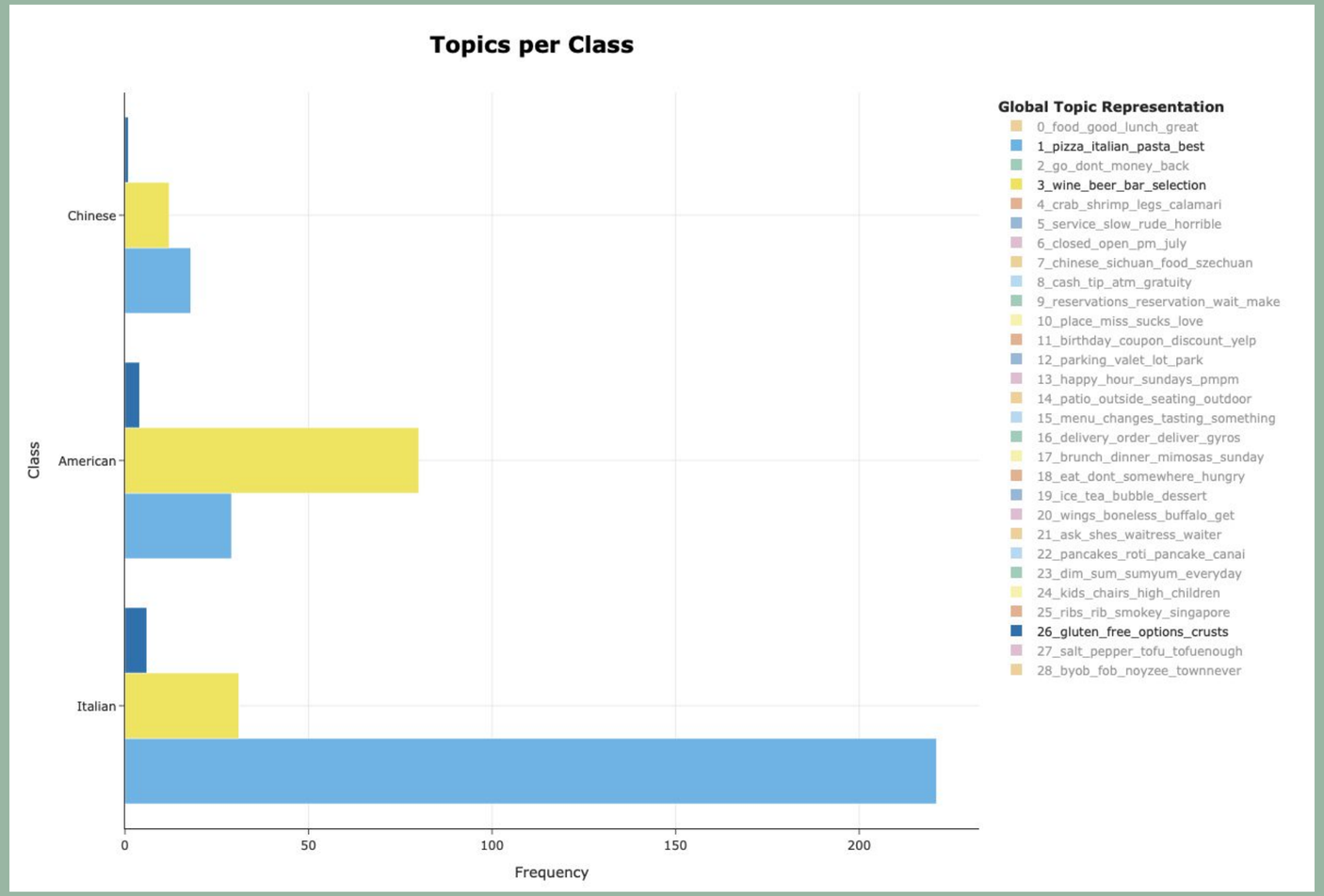


- **American Cuisine:**
  - *Topics:*
    - wings_boneless_buffalo_get
    - brunch_dinner_mimosas_sunday
    - happy_hour_sunday_ppmp
  - Main dishes are important
  - Brunch is popular meal
  - Happy hour is apart of American culture

Insights/Results

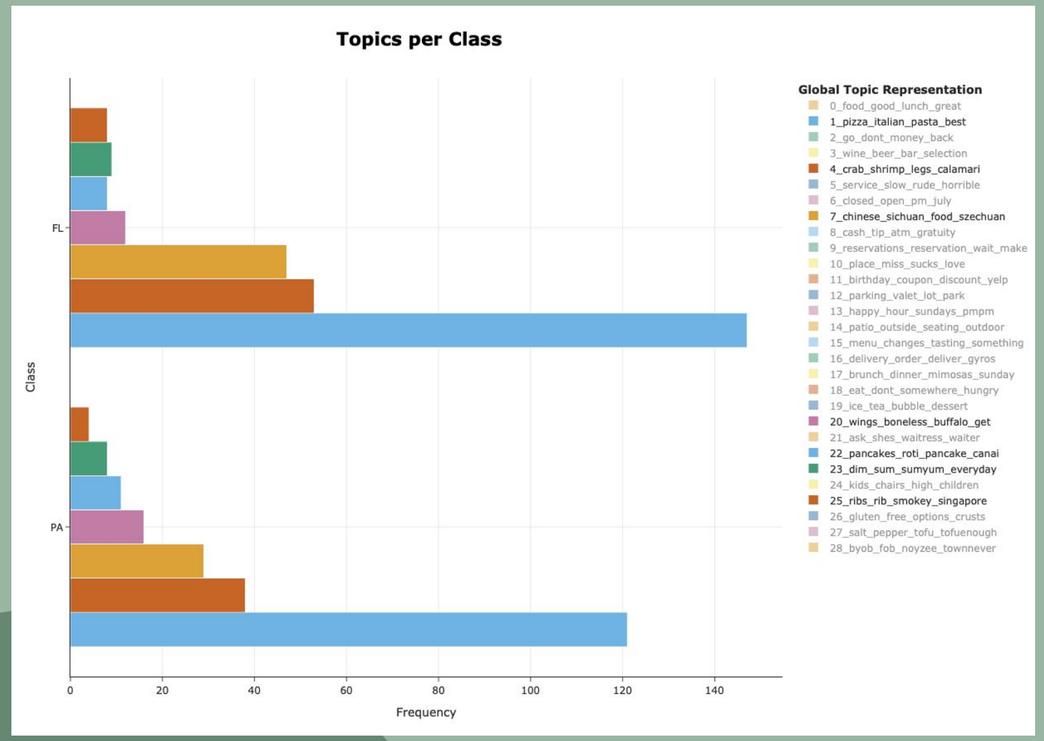
- Cuisine Classification Topic Modeling

- **Italian Cuisine:**
  - *Topics:*
    - pizza_italian_pasta_best
    - wine_beer_bar_selection
    - gluten_free_options_crusts
  - Pasta most important dish of cuisine
  - Dining paired with wine and beer
  - Gluten free options (pizza and pasta)



Insights/Results

- State Food Preferences Classification Topic Modeling



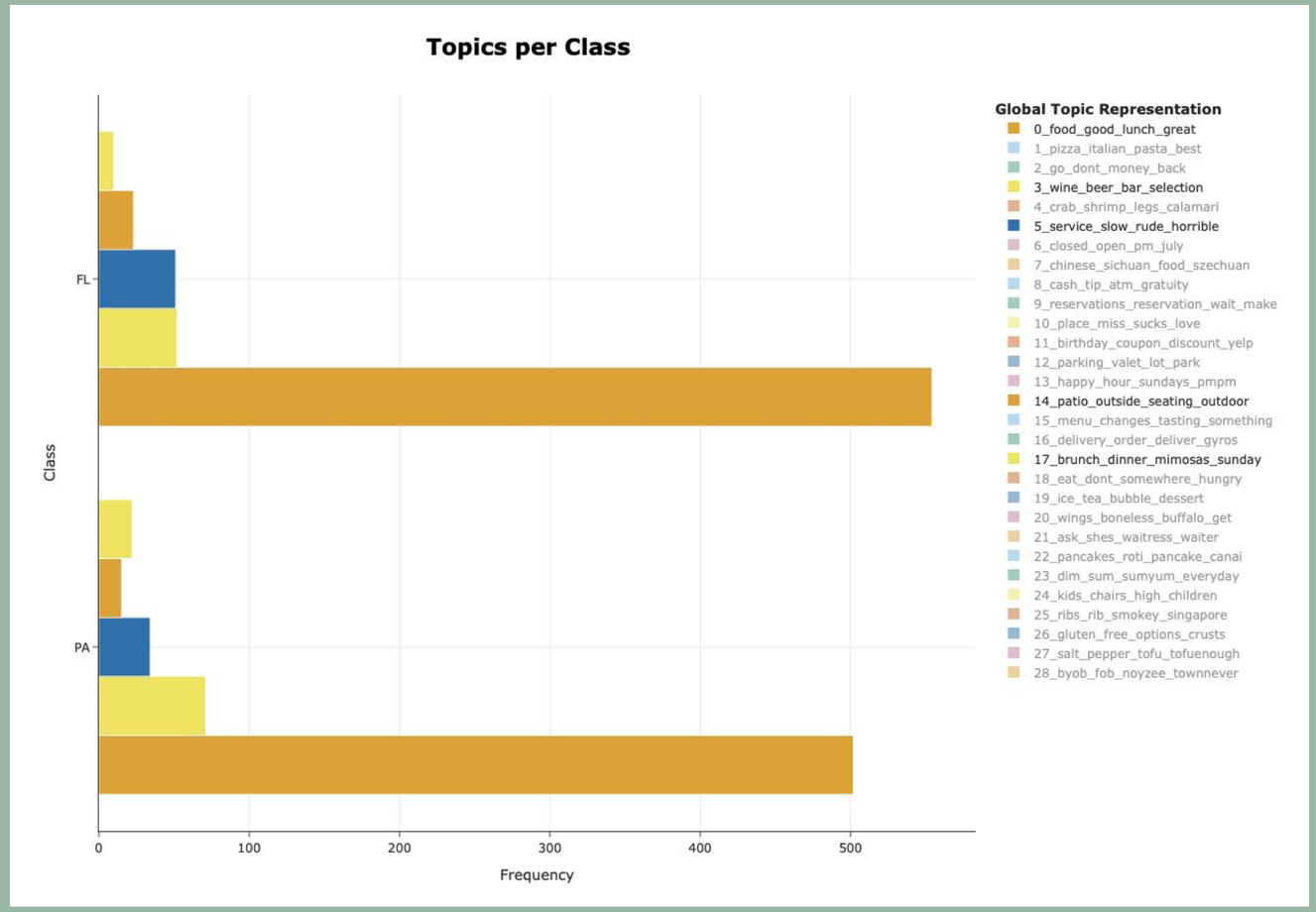
- **<u>Florida:</u>**
  - *Food preferences*
    1. Italian (Higher)
    2. Seafood (Higher)
    3. Chinese (Sichuan)
    4. Wings
    5. Pancakes
    6. Barbecue (Higher)
- **<u>Pennsylvania:</u>**
  - *Food preferences*
    1. Italian
    2. Seafood
    3. Chinese (Dim sum)
    4. Wings (Higher)
    5. Pancakes (Higher)
    6. Barbecue

Mainly same order, but frequency differs showing preference in taste and specific dishes

Insights/Results

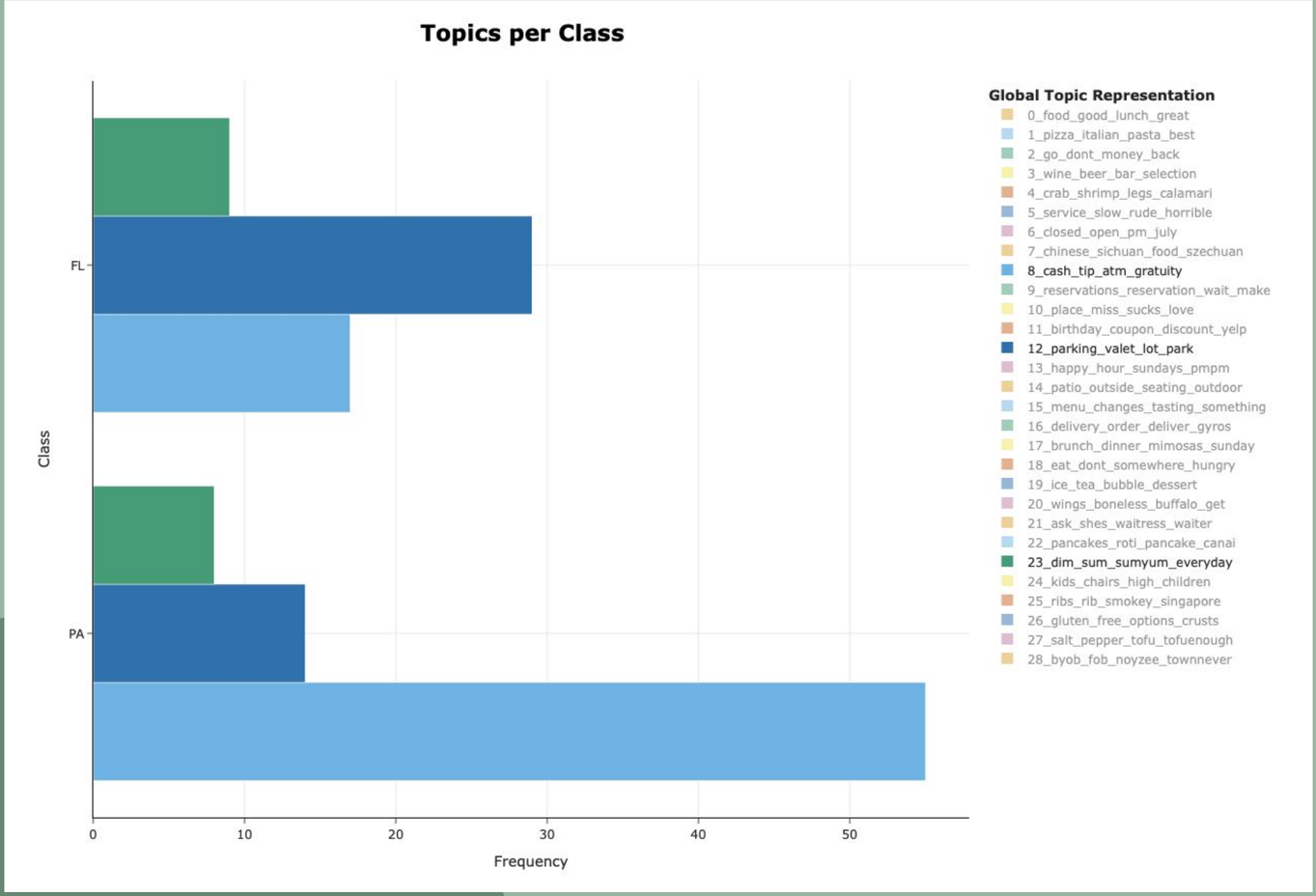
- State FL Classification Topic Modeling

- **Both states:**
  - *Topics:*
    - food_good_lunch_great
    - service_slow_rude_horrible
  - Food quality important in both states
  - Service issues
- **Florida:**
  - *Topics:*
    - patio_outside_seating_outdoor
    - brunch_dinner_mimosas_sunday
    - wine_beer_bar_selection
  - Outdoor dining and alcohol
  - Brunch is main meal



Insights/Results

- State (PA) Classification Topic Modeling



- **Pennsylvania:**
  - *Topics:*
    - parking_valet_lot_park
    - cash_tip_atm_gratuity
    - dim_sum_sumyum_everyday
  - Parking issues
  - Service (tips)

Conclusion/Recommendations

Sentiment Classification

- Food quality is main driver for overall satisfaction
- Service-related issues correlate highly with negative reviews
- Neutral review are operational issues and logistics

Cuisine Classification

- Cuisine emphasize food quality and specific dishes
- Customers interact differently depending on cuisines
- **American:**

Emphasize brunch specials and happy hour

- **Italian:**

Expand gluten-free options

- **Chinese:**

Improve quality of service

State Classification

- Cuisine preference differ by state. Restaurant critique business strategies to meet customer preferences in different locations
- **Florida:**

Improve outdoor seating options

Offer more brunch specials

- **Pennsylvania:**

Parking situation (urban restaurants)

Service quality

THANK YOU!