Yelp Data Mining Project: Final Report

Dheer Gupta

May 2025

1 Project Activities Summary

Over six sequential tasks, I applied data-mining techniques to Yelp's academic dataset to answer a variety of restaurant-related questions. Below is a compressed summary of each task's objectives, methods, and results.

1.1 Task 1: Topic Analysis on All Reviews

Objective: Extract latent themes from a random sample of 50,000 Yelp restaurant reviews. **Methods:**

- Preprocessed text (lowercasing, regex-based cleaning, stopword removal, lemmatization).
- Vectorized using `CountVectorizer` (max 1,000 terms, min_df = 5, max_df = 0.7).
- Applied Latent Dirichlet Allocation (LDA) with 10 topics.
- Inspected top keywords per topic and visualized topic-weight distributions.

Results:

- Identified ten coherent topics (e.g. *pizza/wing*, *service/staff*, *sandwich/cheese*, *Mexican cuisine*).
- Found Topic 6 (generic would/get/back language) carried the highest average weight (328.902).
- Concluded that LDA successfully separated food-centric topics from operational and service themes.

1.2 Task 1.2: Positive vs. Negative Review Comparison

Objective: Compare topics in 10,000 positive (4–5 star) vs. 10,000 negative (1–2 star) reviews. **Methods:**

- Preprocessed text as before.
- Vectorized with `CountVectorizer` (max 1,000 features, min_df = 3, max_df = 0.7).
- Applied LDA separately to positive and negative corpora (8 topics each).
- Interpreted themes and compared keyword distributions.

Results:

- Positive topics emphasized *specific food items and cuisine types* (Mexican food, pizza, breakfast items), *positive service attributes* (friendly, amazing, great), and *atmosphere and ambiance descriptions*.
- Negative topics highlighted *operational issues* (wait times, ordering problems), *communication breakdowns* (told, said, called), and *staff-related complaints*.
- Determined that positive reviews focus on praising food and service, whereas negative reviews focus on operational bottlenecks and time-related frustrations.

1.3 Task 2: Cuisine-Map Mining

Objective: Build an exploratory cuisine map showing relationships among 60 cuisines in Yelp's business and review data. **Methods:**

- Loaded and filtered `business.json` and `review.json` for Restaurant businesses.
- Exploded multi-category fields, retained only cuisines, and required 500,000 characters of text per cuisine, yielding 60 cuisines for analysis.
- Created three text representations: Raw TF (2,000-term counts, L2-normalized), TF–IDF, and LDA (10-component topic mixtures).
- Computed pairwise cosine similarities for each representation and visualized them as heat maps.
- Clustered the LDA vectors using K-Means and Agglomerative (average-link, cosine) at $k = 4$ and $k = 8$.
- Evaluated using silhouette scores.

Results:

- LDA similarity heatmap revealed block-diagonal patterns grouping *American/bar food*, *Asian (Sushi/Japanese/Thai/Korean)*, *Latin/Mexican*, and *Mediterranean/vegetarian* cuisines.
- Clustering at $k = 4$ grouped cuisines into four macro clusters; $k = 8$ provided finer separation with Sushi peeling away from general Japanese and a Caribbean + Hawaiian cluster appearing.
- K-Means with $k = 8$ achieved the highest silhouette score (0.614), followed by Agglomerative $k = 8$ (0.527).
- Insights: Potential for substitute recommendation (e.g. diners who like Thai may like Korean) and identifying cuisine relationships based on review language rather than geographic origins.

1.4 Task 3: Finding Indian Dishes

Objective: Build a recognizer of Indian dish names appearing in Yelp reviews. **Methods:**

- *Manual Tagging (3.1):* Started with 142 candidate phrases (`Indian.label`). Removed non-dishes (e.g. taj mahal, mount everest), flipped mislabels, and added missing staples (e.g. naan, dal, samosa). Result: 27 verified dish names.

- *Pattern-Based Mining (3.2):* Filtered Indian restaurants from `business.json`, sampled 5,000 of their reviews, and ran regex patterns (e.g. ordered X, tried X, had X) to extract phrases not in the 27. Kept phrases with 3 mentions.
- Ranked 20 new candidates by frequency; manually filtered noisy matches (e.g. and it, lunch buffet).

Results:

- Original dish labels: 48 → 27 after cleaning.
- Discovered 20 new candidates; top matches included lamb vindaloo (11 mentions), lunch buffet (10), garlic naan (8), chicken curry (8), palak paneer (5), etc.
- Final result: 27 verified dishes plus additional discovered candidates for downstream tasks.

1.5 Task 4: Popular Dish Mining

Objective: Rank Indian dishes by composite popularity using mention count, sentiment, star rating, and restaurant diversity. **Methods:**

- Combined 27 verified dishes + 12 newly discovered dishes = 37 total dishes analyzed.
- Loaded 8,000 Yelp reviews for Indian restaurants; recorded dish mentions with associated business, review stars, and sentiment.
- Computed for each dish: total mentions, average sentiment, average stars, and number of restaurants.
- Scored using composite function:

  $$\text{Dish Score} = 0.4\times\text{Total Mentions}+0.3\times(\text{Avg Sentiment}+1)\times10+0.2\times\text{Avg Review Stars}+0.1\times\text{Number of R}$$

- Ranked dishes by descending score.

Results:

- Top 3 dishes: naan (1,014.7 score, 2,485 mentions), masala (643.1, 1,565 mentions), curry (565.6, 1,362 mentions).
- Other top dishes: tikka masala (377.2), chicken tikka (369.8), etc.
- Insight: Generic but ubiquitous items (naan, masala, curry) dominated the ranking, reflecting both frequency and positive sentiment.

1.6 Task 5: Restaurant Recommendation

Objective: For each of the top 3 dishes (naan, masala, curry), recommend the best restaurants. **Methods:**

- For each dish, collected all restaurant-level mentions.
- For each restaurant, computed mention count, average sentiment, and average stars for that specific dish.

- Scored using:

  $$\text{Restaurant Score} = 0.3\times\text{Dish Mention Count}+0.4\times(\text{Avg Sentiment}+1)\times10+0.3\times\text{Avg Review Stars}$$

- Sorted and reported top restaurants for each dish.

Results:

- *Naan:* Top performers – Mount Everest India's Cuisine, Mint Indian Bistro, India Palace.
- *Masala:* Top performers – India Palace, Mint Indian Bistro, Curry Corner.
- *Curry:* Top performers – Mint Indian Bistro, Mount Everest India's Cuisine, Curry Corner.
- Insight: Mint Indian Bistro and Mount Everest India's Cuisine consistently ranked high across multiple dishes, indicating overall quality, while some restaurants excelled for specific dishes.

1.7 Task 6: Predicting Hygiene Failures

Objective: Predict whether a restaurant fails its health inspection using Yelp review text and auxiliary features. **Methods:**

- Loaded 546 labeled restaurants (273 pass, 273 fail) and 12,753 unlabeled test restaurants from `hygiene.dat` and auxiliary files.
- Engineered features:
  - *TF–IDF (max 50,000 features; unigrams + bigrams)* on concatenated review text.
  - *Average sentiment polarity* (TextBlob).
  - *Total review text length* (characters).
  - One-hot encoded *cuisines* and *zip codes*.
  - *Number of reviews* and *average star rating* (scaled).
- Trained SVM, Random Forest, and Logistic Regression via 5-fold cross-validation. Evaluated using macro-F1 score.
- Selected Logistic Regression (best CV F1: 0.638); tuned probability threshold to 0.500.
- Predicted on 12,753 test restaurants.

Results:

- **CV F1 Scores:** SVM 0.569, Random Forest 0.620, Logistic Regression 0.638.
- **Final predictions:** 4,435 out of 12,753 restaurants predicted to fail (failure rate = 0.348).
- **Feature dimension:** Final feature matrix had 3,056 features total.
- Insight: Review sentiment and star ratings correlated meaningfully with inspection outcomes; cuisine and location features provided additional predictive value.

2 Project Highlights

In this section, I reflect on the usefulness, novelty, and new knowledge gained from the project.

2.1 Usefulness of Results

Cuisine-Map for Recommendations. The cuisine-map (Task 2) groups cuisines by textual similarity of how users discuss them in reviews. This can power recommendation systems in multiple ways:

- **Cross-Cuisine Discovery:** If a user enjoys *Thai* food, the map shows *Japanese* and *Korean* as nearby—logical next exploration targets.
- **Substitute Recommendations:** By mapping user preferences onto the cuisine network, the system can suggest related cuisines when preferred options are unavailable (e.g. Caribbean to a Hawaiian fan).

Restaurant owners and food-tech apps can benefit by guiding diners to novel but related cuisines, potentially increasing customer retention and discovery.

Dish Discovery and Popularity Mining. Tasks 3–5 created a pipeline to extract, validate, and rank Indian dishes from free-text reviews. Practical applications include:

- **Menu Automation:** Given online reviews, the system can automatically generate lists of popular dishes for any cuisine—useful for aggregators or delivery apps when detailed menus are unavailable.
- **Culinary Analytics:** Chefs and restaurateurs can track dish popularity trends—identifying rising dishes like garlic naan and lamb vindaloo to inform menu decisions.
- **Targeted Marketing:** Marketing teams can highlight top-ranked dish-restaurant combinations (e.g. Mint Indian Bistro's masala dishes) in promotions to attract customers searching for specific items.

Health Inspection Risk Assessment. The hygiene-failure classifier (Task 6) demonstrates that publicly available Yelp data can predict official health inspection outcomes. Applications include:

- **Public Health Monitoring:** Health departments could use such models to prioritize inspection resources, focusing on restaurants with higher predicted failure risk.
- **Consumer Awareness:** Platforms could display hygiene risk indicators based on review sentiment patterns, helping customers make informed dining decisions.
- **Restaurant Self-Monitoring:** Business owners can track their review sentiment trends to identify early warning signs of potential hygiene issues before official inspections.

2.2 Novelty of Exploration

Comparative Topic Analysis by Rating. While LDA topic modeling is well-established, my Task 1.2 approach of *separately* analyzing positive versus negative review corpora revealed distinct thematic patterns. This comparative methodology illuminated that positive reviews focus on specific food items and service quality, while negative reviews emphasize operational failures—an insight not captured by standard unified topic modeling approaches.

Hybrid Pattern Mining with Manual Curation. Task 3's combination of regex-based pattern extraction (ordered X, tried X) with manual label cleaning represents a practical middle ground between fully automated NLP systems and purely manual approaches. This hybrid method proved *surprisingly effective* at surfacing rare multiword dish names (e.g. lamb vindaloo, palak paneer) while maintaining high precision through human oversight—achieving strong results without requiring sophisticated entity recognition frameworks.

Multi-Factor Dish Popularity Scoring. Task 4's composite scoring function balanced raw mention frequency with sentiment analysis, star ratings, and restaurant diversity using carefully weighted components. While popularity scoring exists, the specific combination and normalization approach (transforming sentiment from $[-1, 1]$ to contribute meaningfully alongside mention counts) provided an interpretable unified popularity metric that distinguished quantity from quality of mentions.

Cross-Domain Predictive Validation. Task 6's demonstration that aggregated Yelp review sentiment and text features can predict real-world health inspection outcomes represents novel cross-domain validation. Previous work typically focuses on review-to-review predictions (e.g. sentiment analysis); this restaurant-level aggregation approach with official inspection ground truth provides new evidence for the predictive value of user-generated content in regulatory contexts.

2.3 Contribution of New Knowledge

Method Performance Insights. Through systematic comparison across tasks, I discovered method-specific performance patterns:

- **Dimensionality and Clustering:** LDA's 10-dimensional topic representations (Task 2) produced clearer cuisine clusters than high-dimensional TF-IDF, suggesting that topic-based compression improves similarity computation for categorical clustering.
- **Linear Models for Text Classification:** Logistic Regression consistently outperformed Random Forest and SVM for the hygiene prediction task (Task 6), indicating that linear models remain competitive when text features are combined with meaningful auxiliary signals like sentiment scores.
- **Simple Patterns vs. Complex NLP:** Lightweight regex patterns (Task 3) achieved high precision for dish extraction, suggesting that domain-specific simple approaches can compete with sophisticated NLP pipelines for certain extraction tasks.

Sentiment as Hygiene Proxy. Task 6 established that average sentiment polarity across all reviews for a restaurant has significant predictive power for health inspection failures ($F1 = 0.638$). This finding validates the hypothesis that customer-perceived restaurant quality, as expressed in reviews, correlates with objective hygiene standards—providing empirical evidence for using crowd-sourced feedback as a complement to traditional inspection methods.

User Language Shapes Cuisine Relationships. The Task 2 cuisine-map revealed that cuisine similarities emerge from how users *discuss* food rather than from traditional cultural or geographic categorizations. For example, Mediterranean clustering with Vegetarian and Cafés indicates that health-conscious dining language transcends strict regional boundaries. This insight suggests that

recommendation systems should prioritize user-centric semantic similarity over predefined cultural taxonomies.

3 Next Steps for Practical Deployment

Transforming these research findings into a working system requires several key developments:

3.1 Real-Time Data Integration

- **Streaming Pipeline:** Replace static JSON processing with real-time review ingestion via Yelp Fusion API or similar data sources to ensure current information.
- **Incremental Model Updates:** Implement automated retraining schedules (e.g. weekly) so that topic models, popularity rankings, and hygiene classifiers adapt to evolving review patterns and restaurant landscapes.

3.2 User Interface Development

- **Interactive Cuisine Map:** Build a web application where users can explore cuisine relationships visually, click to discover similar cuisines, and find highly-rated restaurants within each category.
- **Dish-Focused Search:** Create a Menu Explorer interface that allows users to search for specific dishes (leveraging Task 4's popularity rankings) and discover where to find the best versions (using Task 5's restaurant recommendations).
- **Risk-Aware Dining:** Integrate hygiene risk indicators into restaurant listings, displaying predicted inspection outcomes alongside explanatory factors (sentiment trends, review patterns) for user transparency.

3.3 Production Model Management

- **Performance Monitoring:** Establish continuous accuracy tracking by comparing hygiene predictions against actual inspection results when available, enabling model performance assessment and improvement.
- **Threshold Optimization:** Implement dynamic calibration of classification thresholds based on prediction feedback to maintain optimal precision-recall balance as data distributions evolve.
- **Bias Auditing:** Regular analysis to ensure predictions don't unfairly discriminate against specific cuisines, neighborhoods, or restaurant types, maintaining algorithmic fairness in deployment.

3.4 System Extensibility

- **Multi-Cuisine Expansion:** Apply the dish mining methodology (Task 3) to Chinese, Mexican, Italian, and other cuisines, rapidly building comprehensive dish databases for broader coverage.

- **Alternative Prediction Targets:** Extend the review-based prediction framework beyond hygiene to anticipate service quality, food safety incidents, or business sustainability outcomes.
- **Geographic Adaptation:** Customize models for different cities or regions, accounting for local cuisine preferences and regulatory environments.

These enhancements would transform static analytical results into a dynamic, user-centric platform that helps diners discover new cuisines, locate outstanding dishes, and make informed decisions about restaurant safety—creating practical value from the data mining insights developed across all six tasks.