Machine Learning Project
Avraham Reissberg, Shivani Tayade, Sravya Bhaskara

Aspect-Based Restaurant Recommendation System

1. Introduction

1.1 Executive Summary

This report presents an aspect-based sentiment analysis approach to enhance restaurant recommendations using the Yelp Academic Dataset. Traditional rating systems fail to capture aspect-specific preferences (food, service, ambiance). By extracting sentiment for each aspect from user reviews, our model provides personalized recommendations based on user priorities.

1.2 Business Goal and Objective

Business Goal: Improve user experience by providing personalized restaurant recommendations.

Objective:

- Extract and analyze aspect-based sentiments from user reviews.
- Match user preferences with businesses topping in preferred aspects.
- Optimize recommendation accuracy by evaluating sentiment prediction methods.

---

2. Background

- **Dataset**: The Yelp Academic Dataset contains five key datasets: Reviews, Users, Businesses, Check-ins, and Tips. Our focus is on the Review, User, and Business datasets to extract sentiment from a review given by a user to the restaurant.
- **Models Used**: We employ Aspect-Based Sentiment Analysis (Text-Blob and BERT Models) and Recommendation Systems (Content-Based and Collaborative Filtering) to derive insights from user reviews.
- **Input**: User-written reviews containing sentiments about various aspects (food, service).
- **Output**: Personalized restaurant recommendations based on user preferences.

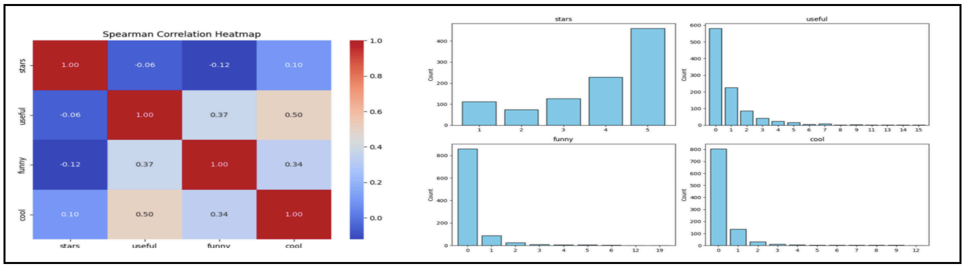
Example
Input: User favors food quality in reviews. Output: Recommends restaurants liked by similar users for food quality.

---

3. Exploratory Data Analysis (EDA)

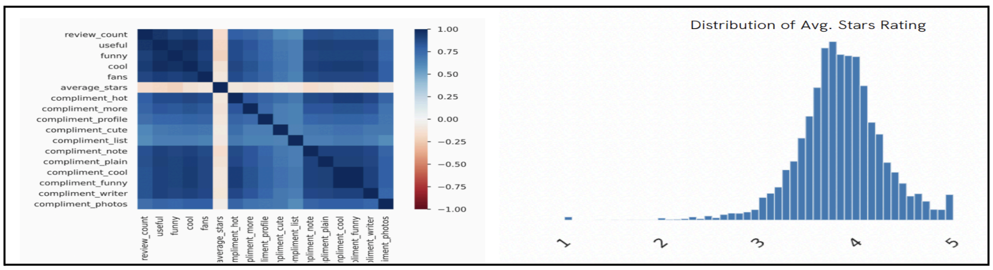
Review Dataset: Contains user reviews with star ratings and text data.

- Strong correlations between review metrics (useful, funny, cool) and user engagement indicate influential reviewers drive interactions.
- Highly skewed distributions suggest a small number of users dominate engagement.



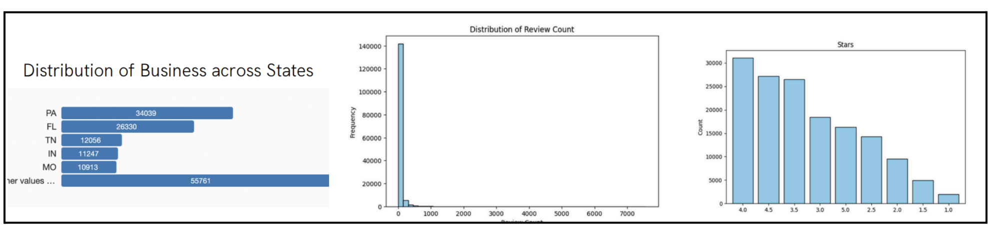
User Dataset: Captures user preferences and interactions.

- Strong correlations between review metrics (useful, funny, cool) and user engagement indicate influential reviewers drive interactions.
- Mostly positive star ratings- skewed distributions



Business Dataset: Provides business details such as location, category, and overall ratings.

- Businesses concentrated in major cities, ideal for targeted marketing.
- Wide variance in star ratings and review counts, reflecting diverse customer experiences.
- Incomplete attribute and operational hour data highlight areas businesses can improve.



---

4. Project Overview

4.1 Text Extraction:

- <u>Preprocessing</u>: Clean the review text by removing noise, punctuation, stopwords, and applying lemmatization.

- Aspect Identification: Extract key aspects such as food, service, ambiance, and value from reviews using NLP techniques.
- Context Extraction: Capture relevant phrases surrounding each aspect to understand the sentiment context (Review: "The food was fresh and delicious" → Aspect: "Food", Context: "Delicious")

4.2 Sentiment Analysis:

- Aspect-Based Sentiment Scoring: Assign sentiment scores (on a scale of 1 to 5) to each identified aspect using a sentiment analysis model.
- Weighted Aggregation: Apply predefined weights to different aspects based on their importance to the user to calculate an overall sentiment score.
- Comparison with Ratings: Compare computed sentiment scores with user-provided star ratings to validate accuracy and refine the model.

4.3 Recommendation:

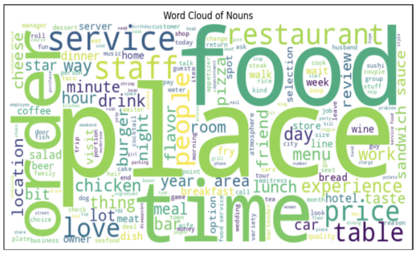
- Hybrid Recommendation Model: Use implicit (user interactions, reviews) and explicit (star ratings, aspect preferences survey) factors for personalized recommendations.
- User Similarity Matching: Identify users with similar preferences based on their aspect importance (e.g., a user who prioritizes food quality will be matched with others who highly rate food).
- Business Recommendation: Suggest restaurants that align with the user's preferred aspects, ensuring a more personalized and relevant recommendation.

---

5. Aspect Based Sentiment Analysis

5.1 Word Count & Aspect Identification:

- Combine Reviews: Merged all user reviews to analyze word frequency.
- Word Cloud: Visualized the most frequently occurring words in reviews.
- Bi-grams & Tri-grams: Identified common word pairs ("food service") and triples ("place food service").
- TF-IDF (Term Frequency-Inverse Document Frequency): Ranked words based on their importance in distinguishing reviews. (food, place were highest)
- Final Aspects Identified: "Food", "Service", "Place", and "Time"




```
Top 10 Most Frequent Bigrams:
food service: 39
customer service: 28
service food: 26
place food: 23
ice cream: 21
love place: 21
time time: 18
order food: 16
food drink: 14
place order: 13

Top 10 Most Frequent Trigrams:
place food service: 4
food love place: 4
hair hair hair: 4
parking lot staff: 3
ice cream cone: 3
food service food: 3
roast beef roast: 3
order fish taco: 3
prix fixe menu: 3
service beer selection: 3
```

```
Top 10 Most Important Words by TF-IDF Score:
good        40.777769
food        36.418233
place       35.288902
great       34.092139
time        26.499395
service     25.935970
come        22.277575
order       22.039387
love        21.482448
like        20.836829
dtype: float64
```

5.2 Context Extraction:

Context from Aspects: Look for the words which are describing the each of the aspect

Review	→	Our favorite is the lamb curry and korma
Aspect	→	Food
Context	→	"Our favorite"

- Keyword Matching: Initial approach using predefined keywords but lacked flexibility.
    - Text: "Amazingly amazing wings and homemade bleu cheese"
    - Food: "No Context Found"
- NLP-Based Semantic Mapping: Used similarity models to better associate words with relevant aspects. Uses pre-trained language model spaCy, which processes the input sentence into an embedding and similarity is computed using cosine metric
    - Text: "Amazingly amazing wings and homemade bleu cheese"
    - Food: "Amazingly wings and homemade bleu cheese"
- No Context: If no relevant description is found, we return "No Context Found"

5.3 Sentiment Analysis:

- Aspect-Specific Sentiment Scoring: Each extracted content was assigned a sentiment score from 1 to 5. If context is not found for, we assigned default sentiment value as 2

- Models:
    - TextBlob: Provided a quick polarity-based sentiment score. It's a rule based approach which is simple but lacks context understanding.
    - Hugging Face NLP Model: Used a pre-trained deep learning model for sentiment analysis, enabling better contextual understanding and handling of complex and sarcastic sentiment expressions.
    - Model Used: NLP based model was chosen as they outperform traditional sentiment analysis methods It applies a pre-trained model- DistilBERT model and provides binary labels (positive or negative) and scores (-1 to 1). These scores are scaled to get ratings on scale of 1 to 5

5.4 Evaluations:

- Sentiment Aggregation: Applied weighted averaging to combine aspect-based sentiments into a final sentiment score.

- Performance Metrics:
  - Compared predicted sentiment scores (using different weight combinations) with actual user ratings.
  - Used Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) to determine the best-performing aggregation with lowest MSE and RMSE.

```
Weights: {'Place': 0.3, 'Food': 0.3, 'Time': 0.2, 'Service': 0.2}
 MSE: 1.52, RMSE: 1.23
```

5.5 Results:

- Stars - Actual Ratings
- Final Sentiment - weighted average of predicted aspect scores (Predicted)

	review_id	user_id	business_id	text	stars	final_sentiment	Place	Food	Time	Service
0	KU_O5udG6zpxOg-VcAEodg	mh_-eMZ6K5RLWhZyISBhwA	XQfwVwDr-v0ZS3_CbbE5Xw	If you decide to eat here, just be aware it is...	3	3.0	5.0	5.0	1.0	1.0
1	BiTunyQ73aT9WBnpR9DZGw	OyoGAe7OKpv6SyGZT5g77Q	7ATYjTIgM3jUlt4UM3lypQ	I've taken a lot of spin classes over the year...	5	3.0	3.0	2.0	2.0	5.0

---

6. Recommendation Systems

This recommendation system helps users discover restaurants that align with their preferences. It leverages both content-based filtering and collaborative filtering

6.1 Content Based Filtering:

The content-based approach suggests restaurants by analyzing user-specified preferences, such as food, service, or ambiance.

Approach

- This approach recommends restaurants based on user preferences (provided as input).
- For new users, the system takes inputs like city and important aspects and then identifies the top-rated restaurants for those aspects.
- These recommendations are sorted based on the user's primary and secondary preferences, making it an effective solution to the cold start problem.

Results:

City: Philadelphia
Aspect 1: Place
Aspect 2: Food

```
Top 20 restaurants in Philadelphia based on Place and Food, sorted
```

	business_name	Place	Food	final_sentiment
261	Franzone's Pizzeria & Restaurant	5.0	5.0	5.0
485	El Camino Real	5.0	5.0	5.0
932	Mac's Tavern	5.0	5.0	5.0
451	Gennaro's Tomato PIE	5.0	5.0	4.0

6.2 Collaborative Filtering:

Collaborative filtering recommends restaurants based on the preferences of similar users. We use aspect-based ratings to determine user similarity. Users are considered similar if they prioritize the same aspects (e.g., ambiance or food quality).

- Example: A food aspect for a user has sentiment as 5 and, service as 3, but still gives 5 stars as overall rating (actual) → Food is the most important aspect

Approach

- We considered Aspect-based ratings for each user and restaurant.
- Identify users with matching aspect preferences (food, service).
- Suggest restaurants that are highly rated by users with similar preferences.
- Compute user similarity using Cosine Similarity to compare aspect-based rating patterns.

Results:

```
Top 5 recommended restaurants for user John:

Smiths Restaurant and Bar
Taqueria La Hacienda
Rittenhouse Grill
Jasmine Rice - Rittenhouse
Bridesburg Pizza
```

---

7. Challenges

- Scalability & Processing Efficiency: Handling large volumes of review data required extensive computational resources, addressed by leveraging GPUs for faster processing.
- Sentiment Aggregation: Determining the best method for combining aspect-level sentiment scores into an overall rating, optimized using weighted averaging.
- User Preference Variability: Different users assign varying importance to aspects, making personalized recommendations complex and requiring dynamic weighting strategies.

---

8. Future Scope

- Enhanced Context Understanding: Implement advanced NLP models (e.g., transformers) for better aspect-context extraction.
- Hybrid Recommendation Models: Combine collaborative and content-based filtering to improve personalization.
- Real-time Processing: Optimize computation to enable real-time restaurant recommendations.
- Integrate AI Assistant: Develop methods to provide users with AI powered recommendations.