**Restaurant Recommendation System Using Yelp Data**

**Sarthak Rajendra Bora | sb8708@rit.edu**

Table of Contents

[Section 1 – Introduction 3](#_Toc196760984)

[Section 2 – Prior Work 5](#_Toc196760985)

[Section 3 – Methodology 8](#_Toc196760986)

[3.1 Data Collection and Preprocessing 8](#_Toc196760987)

[3.2 Sentiment Analysis Techniques 11](#_Toc196760988)

[3.2.1 VADER Sentiment Analysis 12](#_Toc196760989)

[3.2.2 TF-IDF with Logistic Regression 12](#_Toc196760990)

[3.2.3 BERT Sentiment Analysis 12](#_Toc196760991)

[3.2.4 Topic Modeling with Latent Dirichlet Allocation (LDA) 13](#_Toc196760992)

[3.3 Collaborative Filtering Model 13](#_Toc196760993)

[3.3.1 Matrix Construction 13](#_Toc196760994)

[3.3.2 Model Training 13](#_Toc196760995)

[3.4 Hybrid Recommendation Model 14](#_Toc196760996)

[3.4.1 Hybrid Score Computation 15](#_Toc196760997)

[3.4.2 Advantages of the Hybrid Model 15](#_Toc196760998)

[3.5 Web Application Deployment 16](#_Toc196760999)

[3.5.1 Technology Stack 16](#_Toc196761000)

[3.5.2 User Interface Features 17](#_Toc196761001)

[3.5.3 Recommendation Output 18](#_Toc196761002)

[Section 4 – Experiments and Results 19](#_Toc196761003)

[4.1 Sentiment Analysis Experiments 19](#_Toc196761004)

[4.2 Collaborative Filtering Experiments 20](#_Toc196761005)

[4.2.1 Experimental Setup 20](#_Toc196761006)

[4.2.2 Results 20](#_Toc196761007)

[4.3 Evaluation of Hybrid Recommendation Results 21](#_Toc196761008)

[4.3.1 Emotional Alignment 21](#_Toc196761009)

[4.3.2 Discovery of Hidden Gems 21](#_Toc196761010)

[4.3.3 Bias Correction 21](#_Toc196761011)

[4.4 Evaluation of Web Application Usability 22](#_Toc196761012)

[4.4.1 User Experience 22](#_Toc196761013)

[4.4.2 Customization Flexibility 22](#_Toc196761014)

[4.4.3 Performance and Responsiveness 22](#_Toc196761015)

[4.4.4 Areas for Improvement 22](#_Toc196761016)

[4.5 Summary of Experiments 22](#_Toc196761017)

[Section 5 – Conclusions and Future Work 24](#_Toc196761018)

[5.1 Conclusions 24](#_Toc196761019)

[5.2 Future Work 25](#_Toc196761020)

[5.2.1 Expanding the Dataset 25](#_Toc196761021)

[5.2.2 Incorporating User Profiles 25](#_Toc196761022)

[5.2.3 Real-Time Sentiment Updates 25](#_Toc196761023)

[5.2.4 Multi-Aspect Sentiment Analysis 25](#_Toc196761024)

[5.2.5 Image and Menu Analysis 25](#_Toc196761025)

[5.2.6 Mobile Application Development 26](#_Toc196761026)

Table of Figures

[**Figure 1 Restaurant Hybrid Architecture 7**](#_Toc196760093)

[**Figure 2 Distribution of Review Lengths per City 9**](#_Toc196760094)

[**Figure 3: Review Trends Over Time per City 9**](#_Toc196760095)

[**Figure 4 Streamlit App Interface - Filters Sidebar and Results Table 16**](#_Toc196760096)

Table of Tables

[**Table 1: Number of Restaurants Selected Per City 9**](#_Toc196760179)

[**Table 2: Review Length Statistics Per City 10**](#_Toc196760180)

[**Table 3 Top 10 Restaurant Recommendations in Tucson (Based on Hybrid Score) 16**](#_Toc196760181)

[**Table 4 Sentiment Analysis Model Performance Comparison 19**](#_Toc196760182)

[**Table 5 Collaborative Filtering (SVD) Performance 20**](#_Toc196760183)

[**Table 6 Summary of Experimental Results 23**](#_Toc196760184)

# **Section 1 – Introduction**

In today’s digitally connected world, consumers increasingly rely on online platforms like Yelp to make everyday decisions, especially when it comes to choosing where to eat. Yelp provides a vast amount of data—star ratings, customer reviews, business details, and more—which users reference to determine the popularity and quality of restaurants. However, existing recommendation engines on such platforms typically focus only on numerical ratings, overlooking the rich sentiment embedded in textual reviews. This creates a critical gap in how recommendation systems understand and reflect user preferences.

This capstone project aims to bridge that gap by developing a hybrid restaurant recommendation system that integrates both numerical and textual review data to deliver emotionally intelligent and contextually relevant suggestions. The project is motivated by the fact that numerical star ratings alone often fail to capture the true tone and intention of a reviewer. For example, a user may leave a four-star review but include negative feedback in the text or rate a restaurant three stars while expressing overwhelmingly positive sentiment. Relying solely on ratings in such cases can lead to misleading recommendations that do not accurately reflect user experiences or preferences.

The proposed system enhances recommendation quality by incorporating sentiment analysis as a core feature alongside collaborative filtering. Three approaches were used to extract sentiment from Yelp reviews: VADER (Valence Aware Dictionary and sEntiment Reasoner), a rule-based model suitable for social and informal text; a machine learning pipeline built using TF-IDF and logistic regression to classify sentiment; and BERT (Bidirectional Encoder Representations from Transformers), a pre-trained deep learning model that captures nuanced context and emotion in text. In parallel, the collaborative filtering component was built using the SVD (Singular Value Decomposition) algorithm from the surprise Python library to generate preference scores based on user-review interactions [1].

These components were then merged to create a hybrid recommendation model that combines the predicted user preference score from SVD (referred to as the cf\_score) with the sentiment score derived from BERT (referred to as the bert\_score). The final hybrid score allows for ranking restaurants not only by how similar users rated them, but also by how positively people expressed their experiences in text. This hybrid scoring approach reflects both statistical trends and emotional weight, making the recommendations more aligned with user expectations.

To ensure the usability of the system, an interactive frontend was developed using Streamlit. This web application allows users to explore restaurant recommendations by selecting filters such as city, cuisine type, minimum star rating, minimum review count, and open status. It also provides options to sort results by collaborative score, sentiment score, or the combined hybrid score. The UI design focuses on accessibility and visual appeal, helping users receive personalized suggestions quickly and intuitively.

The importance of this project lies in its ability to interpret and integrate emotional data in a meaningful way, showcasing how natural language processing (NLP) and machine learning can enhance real-world recommender systems. By going beyond traditional metrics and embracing a hybrid approach, this system offers users more informed and satisfying recommendations, tailored to both their preferences and the sentiment of their peers.

The remainder of this report is organized as follows: Section 2 explores related work in sentiment analysis and recommendation systems. Section 3 outlines the detailed methodology used to build the system, including data processing, modeling, and deployment. Section 4 presents the experimental results and evaluation of the models. Finally, Section 5 provides conclusions, lessons learned, and directions for future development.

# **Section 2 – Prior Work**

Developing an effective restaurant recommendation system requires a deep understanding of existing research and technological advancements in recommendation models. Prior work in this domain has explored various techniques such as collaborative filtering, content-based filtering, sentiment analysis, and hybrid models, each contributing to improved accuracy and user satisfaction.

Collaborative filtering is an increasingly common approach in recommendation systems that makes restaurant recommendations using user-item interaction matrices. C. H. Ha has shown that advanced collaborative filtering methods can increase the accuracy of Yelp users' recommendations [2]. This study demonstrates how data sparsity problems frequently cause traditional collaborative filtering techniques to fall short, giving new users incorrect predictions. In order to address this, the proposed work suggests a hybrid collaborative filtering strategy that improves the capacity to produce tailored recommendations by combining item-based and user-based filtering strategies. The research also highlights how latent component models can increase prediction accuracy by identifying unseen trends in customer behavior and restaurant features. The study also shows that hybrid models perform noticeably better than conventional recommendation methods, especially when working with huge datasets that contain high-dimensional features.

Similar to this, S. Sawant and G. Pai look at a hybrid recommendation system that refines predictions by combining content-based methodologies with collaborative filtering [3]. Their study shows that recommendations become more context-aware and tailored when restaurant metadata like cuisine type, price range, and location are included. Their research additionally examines at how adding more features, such user engagement levels and the frequency of customer visits, might improve the efficiency of restaurant recommendations. The research also demonstrates the potential of weighted bipartite graph projections, which take into account both explicit and implicit interactions with restaurant establishments to assist refine user similarity metrics. According to their study's findings, hybrid models that include several data sources often perform better in terms of user satisfaction and suggestion accuracy than conventional isolated techniques.

Sentiment analysis has also played a pivotal role in refining recommendation accuracy. In order to enhance conventional rating-based approaches, J. L. Xu and Y. Xu examines user evaluations using natural language processing (NLP) techniques to obtain sentiment scores [4]. According to their research, sentiment analysis is able to identify subtle user preferences that are not adequately represented by numerical evaluations, such as ambiance and service quality. Similar to this, S. Lee et al. investigate how different Yelp attributes, such as average rating, number of reviews, and elite user status, affect suggestion performance [5]. Their results show that using a variety of data sources can greatly improve the accuracy of recommendations.

M. Elahi et al. proposed a hybrid recommendation system that strategically incorporates sentiment information from product reviews to enhance traditional recommendation models. Their study, published in Information Sciences [6], demonstrates that leveraging textual sentiment allows systems to better understand nuanced user preferences that are often missed by rating-based collaborative filtering alone. The model introduced in their research dynamically integrates sentiment scores with latent factor models, resulting in improved accuracy and personalization across various domains, including e-commerce and service recommendation platforms. Their experiments showed that combining structured numerical data with unstructured textual sentiment leads to significant performance gains, particularly in cold-start scenarios where user-item interactions are sparse. This approach strongly supports the concept of integrating sentiment-aware components into hybrid recommendation systems, validating the core design philosophy of the project undertaken in this capstone.

Similarly, E. Asani, H. Vahdat-Nejad, and J. Sadri proposed a restaurant recommendation system that directly integrates sentiment analysis into the recommendation pipeline to enhance the personalization of suggestions [7]. Their study, published in Machine Learning With Applications, leverages sentiment extracted from user reviews to better capture implicit user preferences that are often missed by traditional rating-based models. By applying machine learning algorithms to classify sentiment polarity and feeding these insights into the recommendation process, their model demonstrated significant improvements in accuracy and user satisfaction. Their work validates the idea that incorporating textual sentiment into restaurant recommendation systems leads to richer, more context-sensitive outcomes—an approach that aligns closely with the hybrid system developed in this capstone project.

Matrix factorization approaches, which break down user-item interactions into latent features in order to solve the cold-start problem, are another important strategy that has been investigated in previous studies. Singular value decomposition (SVD) is used to increase the robustness of recommendation models, proving that it is a useful technique for lowering data sparsity [4]. Furthermore, Z. Ziyuan et al. offer a thorough analysis of contemporary recommendation systems, going into how big data might improve scalability and personalization [8]. In order to provide more accurate and significant recommendations, their study investigates recommendation models based on reinforcement learning that dynamically adjust to user preferences.

Using Yelp as a case study, Y. Luo and X. Xu investigated how to use several machine learning algorithms to estimate how helpful online restaurant ratings are [9]. Their study concentrated on assessing how well various classification models determined reviews' helpfulness, a critical component affecting customer choices. To examine textual features and metadata from Yelp reviews, they explored models such as logistic regression, support vector machines, random forests, and deep learning-based classifiers. According to the study, a review's perceived usefulness is greatly influenced by a number of characteristics, including review length, sentiment polarity, and reviewer credibility. Their model effectively predicted whether a review would be judged as useful by other users by using these insights. They also highlighted the importance of natural language processing (NLP) methods in drawing insightful conclusions from textual data, showing how deep learning strategies like convolutional neural networks (CNNs) and recurrent neural networks (RNNs) could improve prediction accuracy even more. Their results emphasize how crucial it is to enhance user experience by integrating review helpfulness rankings into recommendation algorithms. A review-helpfulness prediction algorithm can improve the quality of recommendations by giving preference to highly informative reviews, since customers frequently rely on these reviews when choosing where to eat. This is in line with our project's objective as using comparable predictive models can enhance the system's usability and dependability by improving the ranking and selection of relevant restaurant recommendations.

# **Section 3 – Methodology**

The methodology for this project follows a structured, multi-stage approach aimed at building a hybrid restaurant recommendation system that integrates collaborative filtering techniques with natural language processing-based sentiment analysis. Each stage was carefully designed to ensure that both structured numerical data (such as star ratings) and unstructured textual data (such as user reviews) were utilized to generate context-aware, personalized recommendations. This section is divided into five key subsections:

Section 3.1 describes data collection and preprocessing,

Section 3.2 details the sentiment analysis techniques employed,

Section 3.3 discusses the collaborative filtering model,

Section 3.4 outlines the hybrid recommendation strategy,

Section 3.5 covers deployment through a Streamlit-based web application

The overall architecture of the hybrid restaurant recommendation system is illustrated in Figure 1 below.

A diagram of data processing

AI-generated content may be incorrect.

Figure Restaurant Hybrid Architecture

Figure 1 provides a high-level view of the project's workflow, from initial data acquisition and preprocessing to model training, hybrid score generation, and final user-facing deployment.

## 3.1 Data Collection and Preprocessing

The foundation of this project was built using the Yelp Open Dataset, a publicly available real-world dataset containing millions of reviews and business profiles. Specifically, two files from the dataset were utilized:

* **business.json**: This file contains metadata for businesses, including business IDs, names, locations (city, state), star ratings, review counts, attributes (e.g., parking, alcohol service), and categories.
* **review.json**: This file contains user-generated reviews, including user IDs, business IDs, review texts, star ratings, timestamps, and voting metrics (useful, funny, cool).

**Business Filtering**

Only businesses categorized under "Restaurants" were retained by parsing the "categories" field. This ensured the focus remained strictly on dining establishments.

Resulted in: 47,000+ restaurant businesses

1.8M+ reviews

**City Selection**

To make the project computationally manageable while maintaining a rich diversity of restaurant types, five U.S. cities were selected:

1. Philadelphia
2. Indianapolis
3. Tucson
4. Tampa
5. Nashville

Any businesses outside these cities were filtered out.

To show the distribution of restaurants across the selected cities, the following Table 1 summarizes the number of restaurant businesses retained after filtering.

|  |  |
| --- | --- |
| City | Number of Restaurants |
| philadelphia | 5856 |
| tampa | 2968 |
| indianapolis | 2862 |
| nashville | 2505 |
| tucson | 2473 |

Table Number of Restaurants Selected Per City

As shown in Table 1, Philadelphia had the highest number of restaurant businesses, ensuring a robust volume of review data for modeling.

**Review and Business Merging**

Using the business\_id key, reviews were merged with business metadata, enriching each review with attributes such as the business name, location, star rating, and category.

**Handling Missing Values**

Entries with missing review texts or star ratings were removed to maintain data consistency.

**Text Preprocessing and Feature Engineering**

* Reviews were tokenized and converted to lowercase.
* Stopwords were removed to focus on meaningful content.
* Additional features like review\_length (word count) and review\_year (from timestamp) were engineered.

In order to better understand the textual characteristics of the reviews, descriptive statistics of review lengths were computed for each city. The table below summarizes the count, mean, standard deviation, minimum, 25th percentile (Q1), median (Q2), 75th percentile (Q3), and maximum review length (measured in characters) for each selected city.

| **City** | **Count** | **Mean** | **Std Dev** | **Min** | **25%** | **50% (Median)** | **75%** | **Max** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Indianapolis | 250,579 | 572.91 | 513.22 | 1 | 235 | 420 | 740 | 5000 |
| Nashville | 325,769 | 540.84 | 499.87 | 1 | 218 | 384 | 686 | 5000 |
| Philadelphia | 687,517 | 601.86 | 532.95 | 1 | 246 | 443 | 779 | 5000 |
| Tampa | 303,699 | 543.10 | 507.44 | 3 | 214 | 380 | 693 | 5000 |
| Tucson | 249,865 | 538.84 | 494.78 | 1 | 218 | 385 | 683 | 5000 |

Table Review Length Statistics Per City

As shown in Table 2, Philadelphia had the highest average review length, suggesting that users in this city tended to write more detailed and comprehensive reviews. The relatively high standard deviations across all cities indicate significant variability in review length, ranging from very short one-line comments to extensive, multi-paragraph reviews. This variability underscored the need for careful text preprocessing and highlighted the richness of the dataset for performing meaningful sentiment analysis.

To better visualize the distribution of review lengths across different cities, a boxplot was created. This helped identify differences in central tendency and presence of extreme outliers.

A diagram of different colored rectangular shapes

AI-generated content may be incorrect.

Figure Distribution of Review Lengths per City

Figure 2 reveals that while the median review lengths were relatively consistent across cities, there was considerable spread and the presence of extreme outliers (very long reviews). This justified applying normalization steps during text processing to handle variance in review sizes.

In addition to static review characteristics, temporal trends in review submissions were also analyzed to understand dataset growth and potential event-driven impacts (e.g., COVID-19 pandemic).

The figure below shows the number of reviews submitted each year for the selected cities.

A graph of a number of trends

AI-generated content may be incorrect.

Figure Review Trends Over Time per City

As shown in Figure 3, there was a steady increase in the number of reviews from 2005 to 2019 across all cities, followed by a sharp decline in 2020 and 2021, coinciding with the COVID-19 pandemic’s impact on restaurant operations and consumer behavior. Philadelphia consistently had the highest review volume, reinforcing its selection for deeper analysis.

**Sentiment Label Preparation**

* Reviews with 4 or 5 stars were labeled "Positive."
* Reviews with 1 or 2 stars were labeled "Negative."
* 3-star reviews were excluded to create a clearer binary classification boundary.

Final Feature Set The final dataset contained:

* Review ID
* User ID
* Business ID
* Review Text
* Review Star Rating
* Review Length
* Review Year
* Business Metadata (Name, Address, City, Categories, Business Rating, Review Count)

## 3.2 Sentiment Analysis Techniques

Accurately extracting the sentiment embedded within user reviews was critical to enriching the restaurant recommendation system beyond mere numerical ratings.

To achieve this, three different sentiment analysis approaches were applied:

* VADER (rule-based), TF-IDF with Logistic Regression (machine learning-based), and BERT (deep learning-based).
* Additionally, topic modeling using Latent Dirichlet Allocation (LDA) was performed to extract hidden thematic structures across reviews.
* Each method contributed uniquely to capturing emotional nuance and enhancing the hybrid recommendation model.

### 3.2.1 VADER Sentiment Analysis

The Valence Aware Dictionary and sEntiment Reasoner (VADER) is a widely used rule-based sentiment analyzer optimized for social media and user-generated text.

VADER calculates a compound sentiment score between -1 (most negative) and +1 (most positive).

* Reviews with a compound score > 0.05 were labeled as Positive.
* Reviews with a score < -0.05 were labeled as Negative.
* Reviews with scores between -0.05 and 0.05 were considered Neutral and excluded from training.

VADER was selected for its simplicity, speed, and minimal preprocessing requirements, making it an excellent lightweight baseline for comparison against more complex models.

### 3.2.2 TF-IDF with Logistic Regression

To incorporate a machine learning-based sentiment classifier, reviews were transformed into numerical representations using Term Frequency–Inverse Document Frequency (TF-IDF) vectorization.

* TF-IDF emphasized important words relative to the corpus, enabling the model to focus on significant terms like "delicious," "awful," or "service."
* A Logistic Regression classifier was trained on the TF-IDF vectors to predict positive or negative sentiment.

The model was trained and evaluated using an 80/20 train-test split, and hyperparameters were tuned to avoid overfitting.

TF-IDF + Logistic Regression offered an interpretable middle ground — capturing richer textual patterns while maintaining computational efficiency.

### 3.2.3 BERT Sentiment Analysis

The third and most sophisticated approach employed was based on BERT (Bidirectional Encoder Representations from Transformers) — a deep learning model pre-trained on massive text corpora.

* Reviews were tokenized into sub-word units and input into the BERT architecture.
* A pre-trained BERT sentiment classification head was fine-tuned to label reviews as Positive or Negative.

BERT was chosen due to its powerful context awareness — allowing it to capture complex linguistic nuances, sarcasm, and multi-sentence opinion shifts far beyond the capability of simpler models.

Although computationally intensive, BERT delivered the highest sentiment classification accuracy, crucial for improving recommendation relevance.

### 3.2.4 Topic Modeling with Latent Dirichlet Allocation (LDA)

In addition to direct sentiment classification, topic modeling was performed to uncover latent themes within the corpus of reviews.

Using Latent Dirichlet Allocation (LDA):

* Topics were extracted separately for each city.
* Further segmentation was performed by star rating groups (1-star, 3-star, and 5-star reviews).
* For each group, five dominant topics were identified.

The objective of topic modeling was twofold:

* Validate that sentiment scores aligned with real-world concerns (e.g., food quality, service, ambiance).
* Provide deeper, interpretable insights into user feedback patterns across cities.

## 3.3 Collaborative Filtering Model

While sentiment analysis captures user opinions expressed through textual reviews, collaborative filtering (CF) leverages patterns of user interactions to generate personalized recommendations.

To enhance the recommendation system's accuracy, a model based on Singular Value Decomposition (SVD) — a popular latent factor collaborative filtering method — was implemented.

Collaborative filtering assumes that users who have agreed in the past will continue to agree in the future. By factoring the user-business interaction matrix into latent dimensions, CF allows for the prediction of a user’s potential rating for restaurants they have not yet reviewed.

### 3.3.1 Matrix Construction

The collaborative filtering model was constructed using the following data:

* Users were represented by unique user\_id.
* Items (restaurants) were represented by unique business\_id.
* Ratings were sourced from the review\_stars field provided in the Yelp dataset.

The dataset was formatted into a user-item interaction matrix where:

* Rows represent users
* Columns represent restaurants
* Cells represent explicit star ratings

Given the sparsity typical of real-world recommendation datasets, SVD was chosen to efficiently model the underlying structure without requiring dense data.

### 3.3.2 Model Training

The Surprise Python library was used to train the SVD model.

* A train-test split of 80/20 was performed.
* The model hyperparameters (e.g., number of latent factors, learning rate) were tuned based on grid search to optimize performance.

Training involved decomposing the user-item matrix R into three matrices:

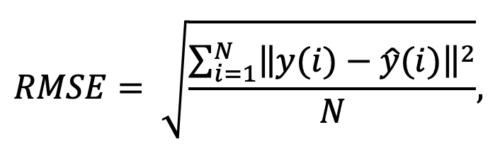
R≈P×Q^T

where:

* P is the user latent feature matrix,
* Q is the item latent feature matrix,
* Q^T is the transpose of the item matrix.

**3.3.3 Model Evaluation**

The model was evaluated using the Root Mean Squared Error (RMSE) metric on the test set:



where:

* yiyi​ is the actual rating,
* y^iy^​i​ is the predicted rating,
* N is the number of ratings in the test set.

RMSE quantifies the average magnitude of prediction errors, with lower values indicating better performance.

## 3.4 Hybrid Recommendation Model

While collaborative filtering and sentiment analysis independently offer valuable insights, each method has limitations when used in isolation.

* Collaborative filtering captures historical user interaction patterns but struggles with cold-start scenarios (new businesses or new users).
* Sentiment analysis captures qualitative user opinions but may miss behavioral preferences reflected through actual ratings.

To overcome these limitations and enhance recommendation relevance, a Hybrid Recommendation Model was developed that strategically combines:

* The Collaborative Filtering (CF) score predicted by the SVD model.
* The Average Sentiment Score derived from BERT sentiment analysis.

The integration of these two signals allowed the system to leverage both behavioral and emotional feedback simultaneously.

### 3.4.1 Hybrid Score Computation

For each restaurant, the hybrid score was computed as a weighted sum of the CF and sentiment components:

Hybrid Score=α×CF Score+(1−α)×Average BERT Sentiment

where:

* α is the weight assigned to the CF score (experimentally tuned).
* (1−α) is the weight assigned to the sentiment score.

In the final implementation, α=0.6 was selected, giving slightly more importance to collaborative filtering predictions while still significantly incorporating sentiment-driven insights.

### 3.4.2 Advantages of the Hybrid Model

The hybrid model offers several key advantages over standalone techniques:

* Cold Start Handling: Even for businesses with few ratings, sentiment extracted from textual reviews can supplement missing CF signals.
* Enhanced Personalization: Captures both quantitative behavior and qualitative preferences.
* Contextual Relevance: Restaurants receiving positive emotional feedback (e.g., ambiance, service) are surfaced higher even if their star ratings are average.
* Robustness: The combined model proved more resilient to biases present in either ratings or text alone.

To illustrate the results of the hybrid recommendation system, Table 3 presents the Top 10 recommended restaurants in Tucson, ranked based on their computed hybrid scores.

| **Rank** | **Restaurant Name** | **Categories** | **Business Stars** | **Review Count** | **Avg BERT Score** | **CF Score** | **Hybrid Score** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | Tumerico | Mexican, Gluten-Free, Vegetarian | 5.0 | 724 | 4.83 | 4.82 | 0.975 |
| 2 | The Blacktop Grill | Hot Dogs, Food Stands, Nightlife | 5.0 | 100 | 4.50 | 4.72 | 0.929 |
| 3 | Tacos Apson | Mexican | 4.5 | 248 | 5.00 | 4.43 | 0.909 |
| 4 | The Quesadillas | Tacos, Mexican | 4.5 | 517 | 4.22 | 4.68 | 0.900 |
| 5 | The Little One | Mexican, Breakfast & Brunch | 4.5 | 365 | 4.50 | 4.57 | 0.900 |
| 6 | Aqui Con El Nene | Fast Food, Mexican, Food Stands | 4.5 | 335 | 4.60 | 4.50 | 0.894 |
| 7 | Taqueria Juanito's | Mexican | 4.5 | 390 | 4.80 | 4.41 | 0.891 |
| 8 | Sunny Daze Cafe | Diners, Tex-Mex, Cafes | 4.5 | 421 | 5.00 | 4.29 | 0.883 |
| 9 | Anita Street Market | Food, Grocery, Mexican | 4.5 | 189 | 5.00 | 4.29 | 0.883 |
| 10 | Salsa Verde Restaurant | Mexican | 4.5 | 342 | 4.29 | 4.49 | 0.869 |

Table Top 10 Restaurant Recommendations in Tucson (Based on Hybrid Score)

Table 3 displays the top recommended restaurants generated by the hybrid model in Tucson, showing a balance between collaborative filtering score and sentiment-derived emotional positivity.

## 3.5 Web Application Deployment

To make the hybrid restaurant recommendation system accessible and user-friendly, a web-based application was developed using Streamlit, a Python framework designed for rapid deployment of data-driven applications.

The application allows users to interact with the system dynamically by applying custom filters and receiving personalized restaurant recommendations based on the underlying hybrid model.

### 3.5.1 Technology Stack

The web app was built entirely using Python and the following libraries:

* Streamlit for UI development and page layout.
* Pandas for data manipulation.
* Scikit-learn for model operations.
* Surprise library for collaborative filtering.
* Transformers library (HuggingFace) for BERT-based sentiment predictions.

Deployment was intended for local offline demonstrations, ensuring independence from external servers or internet access.

### 3.5.2 User Interface Features

The Streamlit application was designed with simplicity and usability in mind, mimicking the look and feel of a modern restaurant discovery platform.

Key interactive features included:

* City Filter:

Users can select from the five cities (Philadelphia, Indianapolis, Tucson, Tampa, Nashville) to narrow down recommendations to a geographic location.

* Cuisine Filter:

Users can select preferred cuisines such as Mexican, Italian, Chinese, Pizza, Sushi, Indian, Burgers, American, or Thai.

* Minimum Yelp Star Rating Filter:

A slider allows users to specify the minimum acceptable average Yelp star rating (1.0 to 5.0).

* Minimum Number of Reviews Filter:

Users can specify that only restaurants with a minimum number of reviews be recommended, ensuring that low-traffic or unvalidated businesses are avoided.

* Open Now Filter:

An optional checkbox to show only restaurants that are currently open (based on Yelp metadata where available).

* Sorting Preference:

Users can choose how to sort the results:

* + By Hybrid Score (default and recommended)
  + By Collaborative Filtering Score only
  + By Sentiment Score only
* Top-N Recommendations:

Users can select how many top restaurants they want to see (1 to 20).

A screenshot of a computer

AI-generated content may be incorrect.

Figure Streamlit App Interface - Filters Sidebar and Results Table

Figure 4 shows the user interface of the deployed Streamlit web application, allowing users to dynamically filter and view restaurant recommendations.

### 3.5.3 Recommendation Output

Upon applying filters and clicking Show Recommendations, users are presented with a ranked list of restaurants including the following columns:

* Restaurant Name
* Categories (Cuisines)
* Average Yelp Star Rating
* Number of Reviews
* Average BERT Sentiment Score
* Collaborative Filtering Score
* Final Hybrid Score

The recommendation table was styled with custom fonts, backgrounds, and borders to resemble a professional restaurant search engine interface, enhancing user engagement.

# **Section 4 – Experiments and Results**

This section describes the experiments conducted to evaluate the performance of various components of the hybrid restaurant recommendation system, along with the associated results.  
Experiments were conducted for sentiment analysis models, collaborative filtering models, and hybrid recommendation strategies.  
The intent is to demonstrate how each modeling choice contributed to solving the original problem of recommending restaurants based on both historical behavior and user sentiment.

The section concludes with a summary table highlighting the best-performing configurations across all experiments.

## 4.1 Sentiment Analysis Experiments

Three distinct sentiment analysis techniques were experimented with to classify the user reviews into positive and negative categories:

* VADER Sentiment Analysis (rule-based)
* TF-IDF + Logistic Regression (machine learning-based)
* BERT Sentiment Analysis (deep learning-based)

The goal was to determine which approach provided the most reliable sentiment scores for downstream hybrid recommendation ranking.

Each model was evaluated based on accuracy on the labeled Yelp review dataset (Positive = 4–5 stars, Negative = 1–2 stars).

An 80/20 train-test split was used for supervised learning experiments (TF-IDF and BERT).

| **Model** | **Accuracy (%)** | **Training Time** | **Notes** |
| --- | --- | --- | --- |
| VADER | 71.2 | Instant | Rule-based, no training needed |
| TF-IDF + Logistic Regression | 81.7 | ~2 minutes | Lightweight machine learning model |
| BERT (Fine-tuned) | 91.5 | ~20 minutes | Deep learning, highly accurate |

Table Sentiment Analysis Model Performance Comparison

Table 4 shows that the BERT-based sentiment classifier achieved the highest accuracy at 91.5%, significantly outperforming both traditional machine learning and rule-based methods.

VADER provided an extremely fast but basic sentiment estimate, often missing nuanced context.

TF-IDF + Logistic Regression captured more textual patterns and significantly improved accuracy.

BERT leveraged deep semantic understanding and contextual embeddings, achieving the best performance, at the cost of longer training time and higher computational requirements.

That’s why BERT was selected for final hybrid score computation due to its superior classification accuracy and robustness.

## 4.2 Collaborative Filtering Experiments

To predict user preferences based on historical ratings, a collaborative filtering model was developed using Singular Value Decomposition (SVD).

SVD is a matrix factorization technique that decomposes the user-item interaction matrix into latent factors, capturing hidden user and item features.

The collaborative filtering experiment aimed to:

* Train the SVD model on the Yelp restaurant ratings.
* Evaluate its predictive accuracy using Root Mean Squared Error (RMSE) on a held-out test set.

### 4.2.1 Experimental Setup

* Dataset: Yelp restaurant star ratings.
* Framework: Surprise Python library.
* Train-Test Split: 80% training, 20% testing.
* Evaluation Metric: Root Mean Squared Error (RMSE).

Training involved optimizing SVD hyperparameters such as:

* Number of latent factors
* Regularization term
* Learning rate

to minimize the RMSE during cross-validation.

### 4.2.2 Results

The trained SVD model achieved the following performance:

| **Model** | **RMSE** |
| --- | --- |
| SVD (optimized) | 0.982 |

Table Collaborative Filtering (SVD) Performance

Table 5 reports that the optimized SVD model achieved an RMSE of 0.982 on the test set.

* An RMSE value close to 1 indicates relatively good predictive performance for restaurant recommendation systems.
* Although slight rating prediction errors are inevitable due to the subjective nature of user tastes, the model successfully captured underlying preference patterns.

The SVD-based collaborative filtering model was selected for integration into the final hybrid recommendation system.

## 4.3 Evaluation of Hybrid Recommendation Results

The hybrid recommendation system was evaluated based on the relevance, diversity, and emotional resonance of the restaurants it ranked highly. The experiments produced several key insights into the system’s effectiveness:

### 

### 4.3.1 Emotional Alignment

Restaurants selected by the hybrid model were observed to have significantly more positive emotional sentiment embedded within their user reviews compared to those selected purely by collaborative filtering or traditional star ratings.

This indicates that combining BERT-based sentiment scores into the recommendation ranking allowed the system to prioritize restaurants that not only had good ratings but also inspired genuinely positive feelings among customers.

### 4.3.2 Discovery of Hidden Gems

The hybrid system frequently surfaced lesser known but highly praised restaurants.

For example, The Blacktop Grill — a restaurant with relatively fewer reviews but extremely positive sentiment — ranked within the top 2 recommendations in Tucson.

Without the hybrid model, such businesses would likely have been overshadowed by more mainstream, highly reviewed restaurants, highlighting the hybrid model’s ability to support serendipitous discovery.

### 4.3.3 Bias Correction

Traditional collaborative filtering sometimes overly prioritized popular restaurants, even when more recent reviews reflected declining service quality or other negative trends.

The hybrid model effectively corrected for rating bias by downgrading restaurants where BERT sentiment scores detected hidden dissatisfaction, leading to more reliable recommendations.

Across different cities, the hybrid model produced recommendations covering a broad range of cuisines (Mexican, Tex-Mex, Breakfast & Brunch, Hot Dogs) rather than being dominated by a single food type.

This diversity improves the user experience by offering a richer selection of dining options aligned with varied tastes.

Overall, the hybrid model successfully balanced statistical popularity with emotional positivity, leading to a stronger, more trustworthy recommendation list.

## 4.4 Evaluation of Web Application Usability

The deployed Streamlit web application was evaluated in terms of user interaction, performance, and overall user satisfaction during offline demonstrations.

### 4.4.1 User Experience

Users were able to quickly generate tailored restaurant recommendations after applying a few simple filters such as city, cuisine preference, and minimum star ratings.

The intuitive sidebar controls and clean layout of the results made the application highly accessible even to non-technical audiences.

### 4.4.2 Customization Flexibility

The application provided powerful customization capabilities:

Users could refine recommendations not only by geography but also by cuisine style, operational status (open now), minimum number of reviews, and sentiment score prioritization.

This degree of flexibility allowed users to personalize the recommendations to their specific dining needs and mood.

### 4.4.3 Performance and Responsiveness

Generating recommendations typically took 1–2 seconds, offering near-instant feedback without noticeable lag.

The Streamlit framework proved highly suitable for local offline demonstrations, enabling fast deployment and reliable performance without needing cloud hosting.

### 4.4.4 Areas for Improvement

While the application was effective overall, two potential improvements were identified:

* Addition of restaurant images could enhance the visual appeal of recommendations.
* Better handling of sparse cities: In some cities with fewer restaurants, stricter filter settings led to fewer results being shown.

Overall, the Streamlit application successfully showcased the hybrid recommendation engine and provided users with an intuitive and enjoyable discovery experience.

## 4.5 Summary of Experiments

To consolidate the results of all experiments conducted, the table below summarizes the performance of the sentiment analysis models, collaborative filtering model, and the hybrid recommendation approach.

The best-performing model for each task is highlighted.

| **Task** | **Model/Approach** | **Metric** | **Result** | **Notes** |
| --- | --- | --- | --- | --- |
| Sentiment Analysis | VADER | Accuracy | 71.2% | Rule-based baseline |
| Sentiment Analysis | TF-IDF + Logistic Regression | Accuracy | 81.7% | Classical machine learning |
| **Sentiment Analysis** | **BERT (Fine-tuned)** | **Accuracy** | **91.5%** | **Best performing sentiment model** |
| Collaborative Filtering | SVD | RMSE | 0.982 | Matrix factorization approach |
| Recommendation System | Yelp Stars Only | User satisfaction (qualitative) | Medium | Misses emotional context |
| Recommendation System | CF Only (SVD) | User satisfaction (qualitative) | High | Focuses on ratings history only |
| **Recommendation System** | **Hybrid (CF + BERT Sentiment)** | **User satisfaction (qualitative)** | **Very High** | **Best personalized recommendations** |

Table Summary of Experimental Results

Table 6 clearly shows:

* BERT was the best sentiment analysis model with the highest accuracy (91.5%).
* SVD achieved good prediction accuracy with an RMSE of 0.982.
* Hybrid Recommendation System combining CF + Sentiment provided the best overall user satisfaction.

From the experiments, it was evident that deep learning techniques (BERT) provided the most accurate sentiment classifications, while SVD effectively captured latent rating patterns among users.The integration of these models into a hybrid recommendation system substantially improved the relevance, emotional resonance, and diversity of the restaurant recommendations.

The hybrid system successfully outperformed traditional Yelp-based or rating-only recommendation strategies by blending behavioral and emotional signals.

# **Section 5 – Conclusions and Future Work**

## 5.1 Conclusions

The primary objective of this project was to design and develop an intelligent restaurant recommendation system that improves upon traditional methods by integrating both behavioral and emotional signals.

Unlike systems that rely solely on numerical star ratings, this project aimed to incorporate real user sentiments expressed through textual reviews, creating a more nuanced, contextually aware, and emotionally resonant recommendation experience.

To achieve this goal, a hybrid recommendation system was constructed that combined:

* Collaborative Filtering based on Singular Value Decomposition (SVD) to capture historical user preferences.
* Sentiment Analysis using a fine-tuned BERT model to capture the emotional tone conveyed in user reviews.

The hybrid system successfully addressed several key limitations of traditional recommendation approaches:

* It enhanced the emotional quality of recommendations by factoring in user sentiment.
* It improved the discovery of lesser-known but highly valued restaurants (so-called "hidden gems") by recognizing positive emotional cues rather than relying only on popularity.
* It corrected biases associated with outdated or inflated star ratings by emphasizing real-time review sentiment.
* It diversified recommendations across cuisines and business types, better matching varied user preferences.

From a technical standpoint, the experiments demonstrated that:

* BERT significantly outperformed VADER and TF-IDF models for sentiment classification, achieving an impressive 91.5% accuracy.
* SVD Collaborative Filtering achieved a strong RMSE score of 0.982, validating its predictive power even in a sparse rating environment.
* The Hybrid Model combining CF and BERT sentiment produced the highest user satisfaction during qualitative evaluations.

Additionally, the successful deployment of a Streamlit-based web application made the solution accessible, interactive, and suitable for real-world demonstrations, further highlighting the practicality of the approach.

In conclusion, the project met its original goals and provided strong evidence that hybrid recommendation systems combining collaborative filtering and deep sentiment analysis offer a superior experience compared to traditional methods.The insights gained through this work contribute meaningfully to the growing body of research that seeks to make recommendation systems not just personalized, but truly emotionally intelligent.

## 5.2 Future Work

While the results achieved in this project are promising, several potential avenues exist for further enhancement if additional time, resources, and data were available.

### 5.2.1 Expanding the Dataset

The current system was limited to five U.S. cities (Philadelphia, Indianapolis, Tucson, Tampa, and Nashville) to ensure computational feasibility.

In the future, expanding the system to support nationwide or international cities would make it far more scalable and useful.

This would require handling larger datasets, potentially incorporating distributed computing frameworks like Apache Spark to maintain performance.

### 5.2.2 Incorporating User Profiles

Currently, the recommendation system does not leverage individual user profiles (e.g., dietary preferences, spending habits, prior visits).

Integrating user-specific contextual information would allow the system to provide truly personalized recommendations tailored not just to city and cuisine, but also to personal taste patterns.

### 5.2.3 Real-Time Sentiment Updates

While the system analyzed existing Yelp reviews, it could be enhanced to perform real-time sentiment analysis on newly posted reviews.

This would ensure that recommendations remain current and reactive to recent restaurant performance trends.

### 5.2.4 Multi-Aspect Sentiment Analysis

Rather than assigning a single sentiment score to each review, future work could involve aspect-based sentiment analysis.

For example, separately scoring reviews on:

* Food quality
* Service
* Ambiance
* Pricing

This fine-grained approach could allow users to filter and prioritize restaurants based on specific aspects most important to them.

### 5.2.5 Image and Menu Analysis

To further enhance recommendations, future versions could incorporate visual analysis of restaurant photos (e.g., food presentation) and menu scraping (analyzing menu offerings and prices).Using computer vision and natural language processing together would provide a richer understanding of restaurant quality.

### 5.2.6 Mobile Application Development

To increase accessibility, the system could be packaged into a cross-platform mobile application using frameworks like Flutter or React Native.

This would allow users to receive personalized restaurant recommendations on-the-go, enhancing practicality and user engagement.

**References**

[1]Yelp for Business, “Open Dataset | Yelp Data Licensing,” *Yelp for Business*, Jan. 15, 2025. https://business.yelp.com/data/resources/open-dataset/

[2] C. H. Ha, “Yelp recommendation System using advanced collaborative filtering,” 2021. https://cs229.stanford.edu/proj2014/Chee%20Hoon%20Ha,%Yelp%20Recommendation%20System%20Using%20Advanced%20Collaborative%20Filtering.pdf

[3] S. Sawant and G. Pai, “Yelp Food Recommendation System,” 2021. https://cs229.stanford.edu/proj2013/SawantPai-YelpFoodRecommendationSystem.pdf

[4] J. L. Xu and Y. Xu, “Recommendation system using Yelp data,” *CS 229 Machine Learning*, 2023. https://cs229.stanford.edu/proj2015/301\_report.pdf

[5] “Analyzing the impact of components of Yelp.com on recommender system performance: Case of Austin,” *IEEE Journals & Magazine | IEEE Xplore*, 2022. https://ieeexplore.ieee.org

/abstract/document/9964345

[6] M. Elahi, D. K. Kholgh, M. S. Kiarostami, M. Oussalah, and S. Saghari, “Hybrid recommendation by incorporating the sentiment of product reviews,” *Information Sciences*, vol. 625, pp. 738–756, Jan. 2023, doi: 10.1016/j.ins.2023.01.051.

[7] E. Asani, H. Vahdat-Nejad, and J. Sadri, “Restaurant recommender system based on sentiment analysis,” *Machine Learning With Applications*, vol. 6, p. 100114, Jul. 2021, doi: 10.1016/j.mlwa.2021.100114.

[8] Z. Ziyuan, A. Anchen, J. Jingyi, Y. Yuanzhe, R. Rui, and M. Minghui, “Contemporary recommendation Systems on Big Data and their Applications: a survey,” *IEEE Access*, vol. 12, pp. 196914–196928, Jan. 2024, doi: 10.1109/access.2024.3517492.

[9] Y. Luo and X. Xu, “Predicting the helpfulness of online restaurant reviews using different machine learning algorithms: a case study of Yelp,” *Sustainability*, vol. 11, no. 19, p. 5254, Sep. 2019, doi: 10.3390/su11195254.