**Restaurant Recommendation System Using Yelp Data**

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

# **Section 1 – Introduction**

The city's fast-paced lifestyle has led to a rise in the demand for takeout and dining out. Because various people have different needs, it can be challenging to determine the best option for someone when there are so many possibilities available. As a result, intelligent recommendation systems that are suited to people's needs have been developed. These technologies improve the dining experience by suggesting restaurants based on factors like location, personal preferences, and customer reviews.

The goal of this project is to develop a restaurant recommendation system using Yelp data that leverages real user reviews and business information to generate personalized restaurant recommendations. Yelp dataset provides a comprehensive collection of user reviews, business information, and user interactions, making it a valuable resource for building an effective recommendation system [1]. The proposed system is an asset for creating an effective recommendation system as it offers an extensive compilation of user reviews, business details, and user interactions. The suggested system seeks to provide personalized suggestions that go beyond traditional rating-based sorting by examining trends in user preferences and restaurant features.

In order to anticipate customer preferences and improve restaurant discovery, this project will make use of machine learning approaches such as collaborative filtering, content-based filtering, and hybrid recommendation models. Furthermore, reviews from customers will undergo sentiment analysis to extract insightful information and enhance recommendation precision. In addition to serving frequent users, the system will also solve issues like the cold-start problem, which impacts new users who have never used the platform before.

In the food and hospitality sector, the use of this recommendation system has useful applications that benefit customers as well as restaurant owners. Users will get recommendations that are specific to their preferences, and businesses can use data-driven insights to draw in the right customers. By making restaurant selection more effective and customized, the idea eventually seeks to close the gap between customers and restaurants.

# **Section 2 – Problem**

## **2.1 Problem Definition**

Customers have a difficult time finding restaurants that suit their tastes because there are so many options available. Personalized recommendations are not efficiently provided by the rating-based sorting systems. By providing personalized recommendations based on user behavior, ratings, and restaurant features, this project seeks to create a machine-learning-based restaurant recommendation system that improves user experience.

## **2.2 Significance of the Problem**

# The problem of choosing a restaurant involves more than just convenience; it also involves increasing customer satisfaction and boosting business efficiency. Users frequently make incorrect judgments due to the growing number of restaurant options, which results in poor dining experiences. A well-designed recommendation system offers substantial benefits by assisting users in finding restaurants that suit their dietary requirements, tastes, and budget constraints in a timely manner.

# Restaurants can gain from focused customer engagement from a business perspective by addressing the relevant demographic and boosting foot traffic and revenue. Improved retention and devotion to customers can also result from personalized advice. If properly deployed, this system can maximize restaurant owners' marketing efforts and lessen user fatigue while making decisions.

# By leveraging machine learning and sentiment analysis, the proposed system enhances recommendation accuracy beyond simple ratings. The system will address challenges like the cold-start problem, making it more inclusive for new users. This project has practical implications in the food and hospitality industry, demonstrating the power of data-driven insights to improve consumer experience and business outcomes.

# **Section 3 – 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.

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 [6]. 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 [7]. 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 4 – Proposed Methodology**

## **4.1 Plan**

There will be several stages to this project, each of which will concentrate on a crucial component of creating a successful restaurant recommendation system. Data collection, preprocessing, exploratory data analysis (EDA), sentiment analysis, feature engineering, model creation, evaluation, and deployment will all be part of the method's structured workflow.

The Yelp Open Dataset serves as the project's data source and includes a variety of JSON files, such as business.json, review.json, user.json, checkin.json, and photo.json [1]. The dataset offers user-generated reviews, ratings, interactions, and business details—all of which are essential for creating a reliable recommendation system.

To make sure the dataset is clean and organized for analysis, data preprocessing is an essential step. Handling missing values, eliminating duplicate records, standardizing formats, and removing unnecessary data are all included in this. Tokenization, stemming, and stop word removal will also be used to process user review language to get it ready for sentiment analysis and machine learning models. To learn more about the dataset, exploratory data analysis (EDA), will be used. We will examine trends in restaurant popularity, rating distributions, and user behavior patterns using visualization tools like Matplotlib and Seaborn.

A crucial component of this project is sentiment analysis, which will be used to extract insights from customer reviews. We will utilize TF-IDF, Word2Vec, and VADER sentiment scoring to quantify sentiment polarity and analyze customer perceptions of different restaurants. Sentiment scores will be integrated into the recommendation engine to provide context-aware restaurant suggestions based on customer sentiment rather than just numerical ratings.

Collaborative Filtering will be used to create the recommendation system, and matrix factorization techniques will be used to convert user-restaurant interactions into embeddings. These embeddings will be learned using neural networks, which will optimize for Mean Squared Error (MSE) to predict user ratings. Other models, including hybrid filtering, which combines content-based and collaborative methods, will also be created and evaluated.

The system will ultimately be made available as an interactive application that allows users to enter their preferences and get personalized restaurant recommendations. With real-time feedback, the model will keep getting better, guaranteeing scalability and improved performance over time. The ultimate objective is to develop a highly customized and flexible restaurant recommendation system that improves user choice and dining experiences.

## **4.2 Challenges or Barriers**

One of the primary challenges in developing this restaurant recommendation system is handling the large volume of data in the Yelp dataset. The dataset consists of millions of records, including user interactions, restaurant details, and reviews, requiring significant computational power to process efficiently. To overcome this, I will utilize the college-provided high-performance computing resources, which will allow for parallel processing and optimized memory management. Additionally, using Pandas and Dask will help handle large datasets more efficiently.

Another significant challenge is ensuring data quality and completeness. The dataset may contain missing values, duplicate records, or inconsistencies, which could negatively impact model accuracy. To mitigate this, data cleaning techniques such as missing value imputation, duplicate record removal, and outlier detection will be implemented before training the models.

Another issue is sentiment analysis's accuracy. Slang, sarcasm, and ambiguous language are common in user-generated evaluations, which makes it challenging to gather insightful information. Advanced Natural Language Processing (NLP) methods like Word2Vec, TF-IDF, and BERT embeddings will be employed to overcome this and provide more precise sentiment classification.

Another difficulty is evaluating and generalizing the model. If the model learns patterns in the training data instead of generalizable trends, overfitting may happen. In order to avoid this, the model training procedure will include cross-validation, dropout regularization, and hyperparameter tuning to enhance the model's capacity to generalize to new data.

## **4.3 Project Deliverables**

The final report will include a comprehensive overview of the project, detailing the methodologies applied, the results obtained, conclusions drawn from the analysis, and recommendations for future work. Alongside the report, a presentation will be prepared to summarize the key findings concisely for quick reference. The dataset used in the project, along with the processed analysis and models, will also be provided as part of the deliverables.

The following are the key deliverables:

1. Capstone Report
2. Dataset
3. Analysis of Data
4. Defense Presentation
5. All Code

**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.