Data Mining for Cuisine Mapping, Dish Recognition, and Restaurant Recommendation

Ashok Sharma

January 30, 2025

1 Introduction

This paper investigates the application of data mining, natural language processing (NLP), and machine learning to analyze restaurant reviews. The key objectives include cuisine similarity visualization, dish recognition, popularity-based ranking, restaurant recommendation, and hygiene prediction. Various methodologies such as topic modeling, clustering, sentiment analysis, and classification models were employed to extract meaningful insights from textual data.

2 Methodology

2.1 Cuisine Similarity Mapping

- Used NLP techniques (TF-IDF, LDA, LSI) to model cuisine similarities.
- Applied cosine similarity and topic modeling to categorize cuisines based on text reviews.

2.2 Dish Recognition and Expansion

- Implemented SegPhrase and ToPMine to identify dish names from reviews.
- Filtered and refined dish labels through manual validation and automated clustering.

2.3 Dish Popularity Ranking

- Employed multiple ranking methodologies: frequency count, restaurant mention count, average ratings, and sentiment scores.
- Used Seaborn visualizations to compare ranking effectiveness.

2.4 Restaurant Recommendation System

- Developed ranking models based on sentiment and user ratings.
- Excluded low-review restaurants to ensure reliability.

2.5 Hygiene Prediction Model

- Preprocessed textual reviews and extracted key features using CountVectorizer and TF-IDF.
- Applied classification models (SVM, Naïve Bayes, Random Forest, Gradient Boosting).
- Used ensemble learning to improve accuracy.

3 Key Findings and Contributions

3.1 Usefulness of Results

- Provided a structured methodology to analyze restaurant reviews for meaningful insights.
- Developed a cuisine similarity map for user-friendly exploration of different cuisines.
- Created a robust dish and restaurant ranking system to aid in dining decisions.

3.2 Novelty of Exploration

- Combined various NLP techniques for better cuisine classification.
- Integrated sentiment-based dish ranking, moving beyond simple count-based rankings.
- Applied ensemble learning to improve hygiene rating predictions.

3.3 Contribution to Knowledge

- Highlighted effective clustering techniques for cuisine classification.
- Demonstrated the impact of different text representations on classification performance.
- Provided a comparative analysis of machine learning classifiers for restaurant hygiene prediction.

4 Future Work & Practical Implementation

- **Development of an Interactive Platform:** Create a web or mobile interface for user interaction.
- **Integration with Real-time Data Sources:** Enhance prediction accuracy by integrating live restaurant review feeds.
- **Optimization of Predictive Models:** Fine-tune classifiers with external datasets for improved hygiene predictions.
- **Customizable User Preferences:** Allow users to filter results based on dietary restrictions, cost preferences, and taste profiles.

5 Technologies and Tools Used

- **NLP & Text Processing:** NLTK, Gensim, TextBlob
- **Machine Learning Models:** Scikit-learn, XGBoost, Naïve Bayes, SVM, Random Forest
- **Data Visualization:** Seaborn, Matplotlib
- **Clustering & Topic Modeling:** LDA, LSI, TF-IDF, Cosine Similarity
- **Feature Engineering:** Tokenization, Stopword Removal, Stemming, Lemmatization

6 Conclusion

This research provides a comprehensive approach to data mining for restaurant analytics, leveraging NLP and machine learning to enhance decision-making in the dining industry. The findings contribute valuable insights into cuisine clustering, dish recognition, sentiment-based ranking, and hygiene prediction, paving the way for more intelligent and consumer-friendly restaurant recommendation systems.

7 References

- Scikit-learn Column Transformer
- Gensim Preprocessing
- Kaggle Sklearn Pipelines
- MLens Ensembling
- No Free Lunch Theorem in Data Science