Final Report: Data Mining Project

2. Usefulness of Results

Task 1: Topic Modeling

I used Latent Dirichlet Allocation (LDA) for topic modeling, which revealed clear patterns in user reviews. Positive reviews often highlighted quality and positive experiences, while negative reviews pointed to dissatisfaction with specific issues.

Usefulness:
Businesses can identify strengths and areas for improvement.
Marketing teams can tailor campaigns around aspects customers value most
Researchers can use the methodology for customer sentiment analysis in other domains.

Task 2: Similarity Matrix and Clustering

By clustering cuisines based on their TF-IDF representations and cosine similarity, we were able to group similar cuisines together and uncover meaningful relationships.

Usefulness:
Restaurant chains can identify overlapping offerings with competitors.
Researchers can apply the clustering approach to analyze other textual similarity.

Task 3: Dish Name Refinement and Expansion

The refinement and expansion of Italian dish names through techniques like SegPhrase and word2vec resulted in a more comprehensive and accurate list of dish names, including terms like "humble pie" and "bone marrow"

Usefulness:
Food bloggers, culinary researchers, and restaurant database developers benefit from a detailed list of dishes.
NLP models for cuisine-specific tasks can leverage this dataset for improved performance.

Task 4: Popular Dish Analysis

I identified the most popular dishes along with their sentiment profiles, on example of Margherita pizza, which showed a high degree of customer satisfaction.

Usefulness:
Helps restaurants prioritize dishes with high customer satisfaction.
Assists customers in making informed dining choices.
Task 5: Restaurant Recommendation

Based on sentiment scores for specific dishes, we developed a system to recommend restaurants that align with customers' preferences.

Usefulness:
Provides personalized dining recommendations, improving customer satisfaction.
Can be adapted for food delivery apps or travel guides.

Task 6: Sentiment-Based Classification

Logistic Regression model achieved a Macro F1 Score of 0.3558, balancing class weights for underrepresented classes.

Usefulness:
Can be used for automated review classification.
Businesses can categorize feedback quickly for internal analysis.

3. Novelty of Exploration

Combining LDA with pyLDAvis provided not only textual insights but also an interactive way to interpret the relationships between topics, making the analysis more actionable and integration of pattern mining as SegPhrase and semantic analysis as word2vec for dish name expansion is an effective approach that demonstrates good results. Combining sentiment analysis with frequency-based dish rankings allowed for a unique prioritization of dishes based on customer satisfaction

4. Contribution of New Knowledge

This project demonstrated that LDA can effectively distinguish between review subsets, providing a framework for domain-specific topic modeling and that clustering cuisines using TF-IDF and cosine similarity is feasible, revealing the dependency of clustering quality on the number of clusters chosen. Additionally, the project highlighted the effectiveness of combining SegPhrase and word2vec for identifying domain-specific terms, a method that can be applied to expand vocabulary in other fields. In future projects we can keep that sentiment analysis combined with popularity metrics offers a better understanding of customer preferences and demonstrated that aggregating sentiment scores can yield meaningful insights for dish-specific restaurant recommendations. Developing a dish-specific restaurant recommendation algorithm based on aggregated sentiment scores adds a novel perspective to personalized recommendation systems.