Advanced topics in Machine Learning

SHIREL ALIMI 318968369

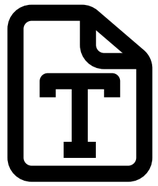
The problem:

In today's digital age, online reviews play a crucial role in shaping consumer opinions and influencing business decisions, particularly within the restaurant industry. The challenge of extracting valuable insights from a large volume of unstructured textual data in the form of restaurant reviews. However, analyzing and making sense of these reviews can be daunting due to their sheer volume and unstructured nature.

Project Goals :

- Analyzing restaurant reviews dataset.
- Performing sentiment analysis to categorize reviews.
- Clustering similar reviews to uncover common themes.
- Training and evaluating machine learning models.
- Predicting sentiment for new reviews in real-time.

Methodology



- **Text Preprocessing:**

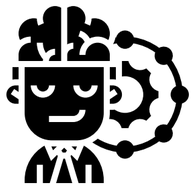
Method involving basic text cleaning techniques like lowercasing and removal of non-alphabetic characters, but may overlook nuanced linguistic nuances.

- **Clustering Analysis:**

Utilizing unsupervised machine learning techniques like KMeans clustering to group similar reviews, which can provide valuable insights but may require tuning of parameters.

- **Sentiment Analysis:**

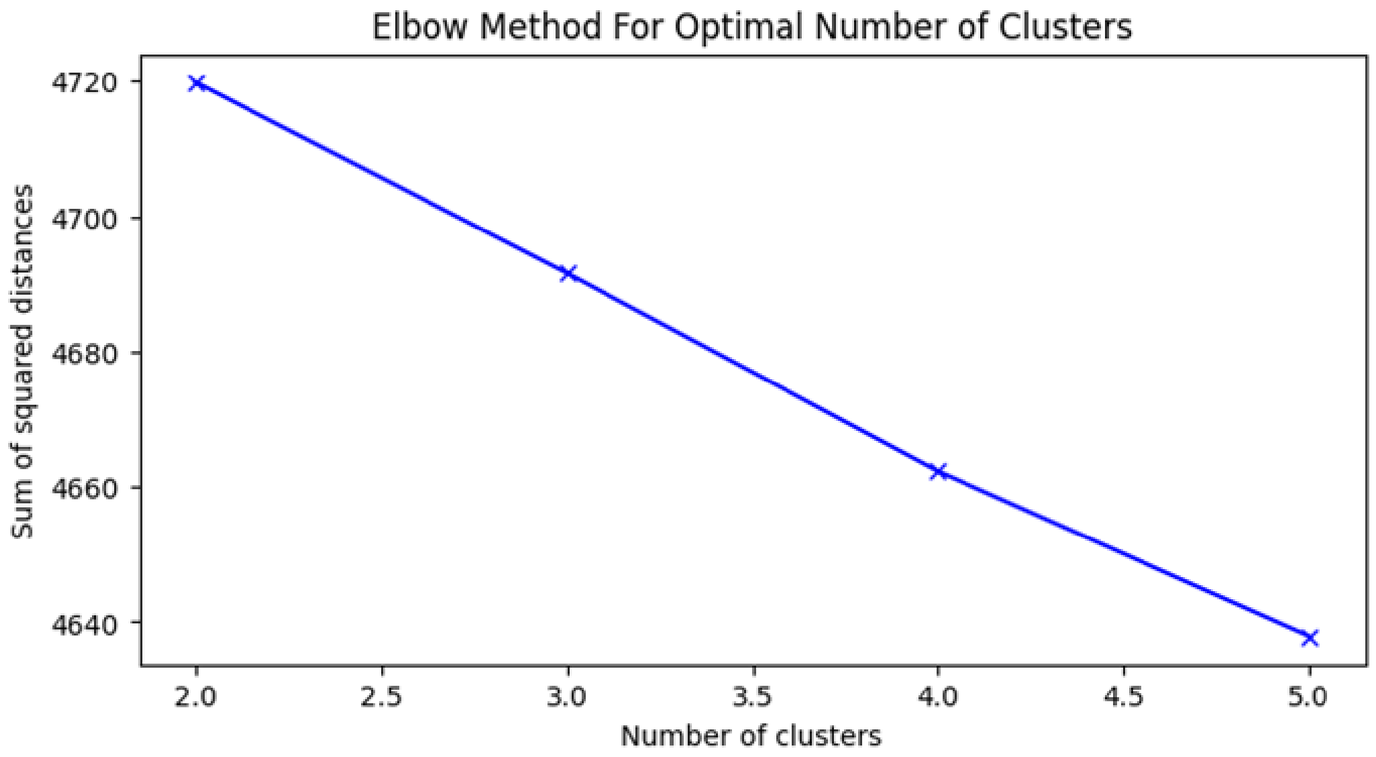
Extracting sentiment from reviews based on ratings, offering a straightforward approach but potentially overlooking subtle nuances in customer sentiment.



- **Deep Learning Models:**

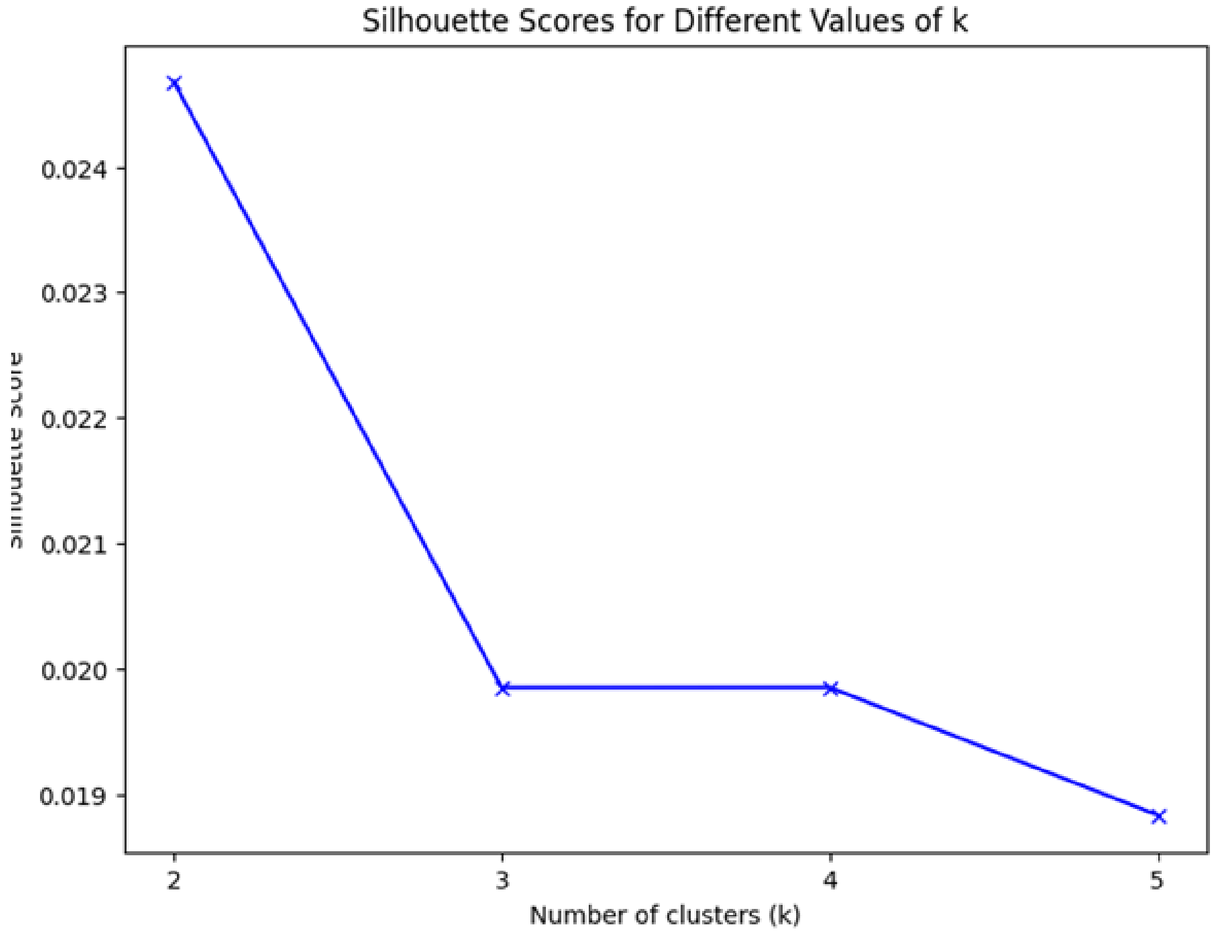
Exploring complex neural network architectures for sentiment analysis and clustering, promising sophisticated pattern recognition but may demand significant computational resources and data volumes

Elbow Method Plot



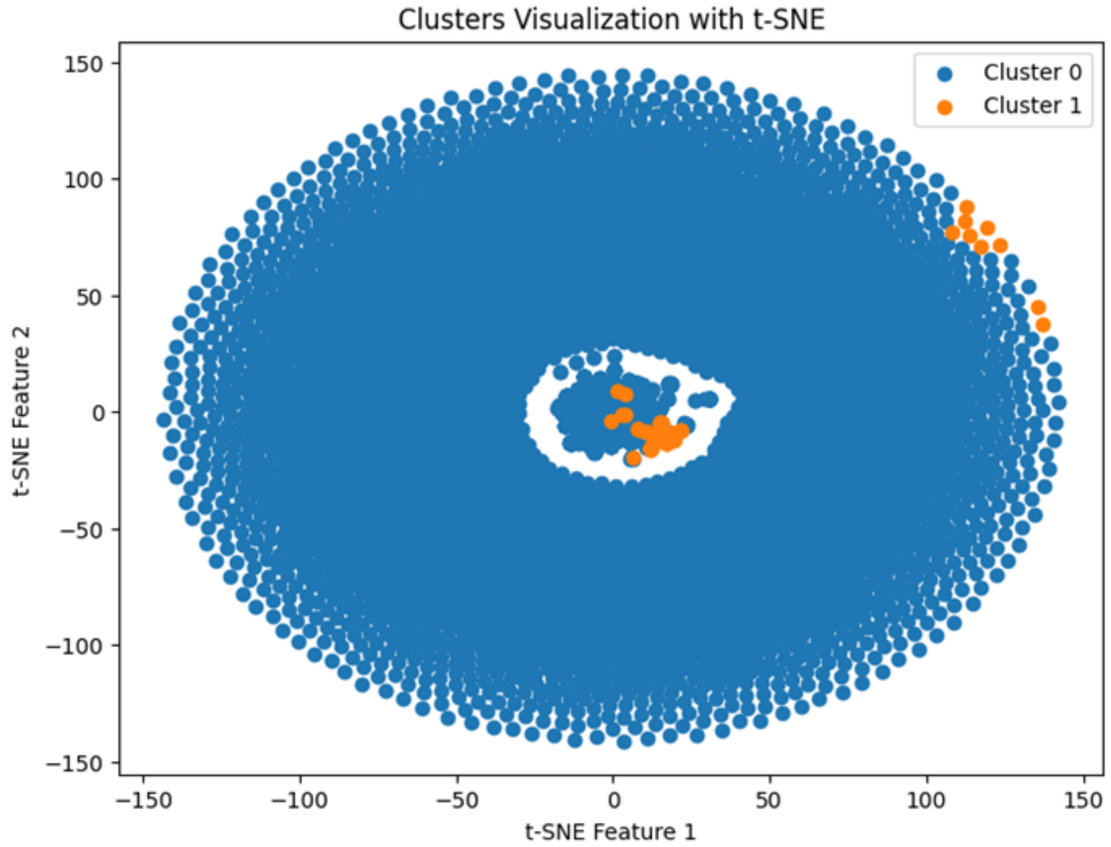
In implementing the KMeans algorithm, the choice of the parameter 'k' - representing the number of clusters - was a decision of utmost importance. I used the elbow method to
Visually identify the optimal 'k'. The graph depicted a clear elbow at k=2, indicating a natural distribution of the data set.

Silhouette Score Analysis



Silhouette Scores to evaluate the coherence of the clusters formed by KMeans . This metric, ranging from -1 to +1, measures how close each point in one cluster is to points in the neighboring clusters. Our analysis yielded scores that were positive and closer to +1, especially pronounced at k=2, underscoring well-defined and distinct clusters. The positive scores across the board bolstered our confidence in the clusters' reliability and the model's utility.

t-SNE Cluster Visualization



Another qualitative verification is provided by t-SNE simulation, which allows observing the distribution of clusters in two-dimensional space. The visualization shows two distinct clusters with minimal overlap, confirming the choice of k=2 from the elbow method and silhouette scores.

Sentiment Analysis using a Logistic Regression model:

Logistic regression predicts the probability of a binary outcome, such as sentiment. In this case, it classifies restaurant review sentiment as positive or negative based on preprocessed text features. After training on labeled data, the model predicts sentiment for new reviews.

Conclusions

I created a comprehensive framework for analyzing restaurant reviews using machine learning techniques. By combining NLP and clustering algorithms, I constructed and processed review data, yielding insightful evaluation metrics. This project improves decision making for restaurant owners and offers valuable insights into customer preferences, paving the way for future advances in restaurant analytics.


Thank You.