Text processing and Clustering of Online Restaurant Reviews with Predictive Analytics

Md Saniul Islam Sony
Department of CSE
American International
University-Bangladesh
22-46400-1@student.aiub.edu

Tanvir Ahmed Tuhin
Department of CSE
American International
University-Bangladesh
22-46475-1@student.aiub.edu

MD. Ashraful Islam
Department of CSE
American International
University-Bangladesh
22-46479-1@student.aiub.edu

***Abstract*—In this study, customer reviews were preprocessed through tokenization, lowercasing, and stopword removal. Term Frequency (TF) and Inverse Document Frequency (IDF) were computed to construct a TF-IDF matrix, and the top 10 most frequent words were identified for analysis. Clustering methods, including K-means, hierarchical clustering, and DBSCAN, were applied to group reviews based on textual similarity, with cluster plots and dendrograms used for visualization. To address the growing pressure on restaurants to respond quickly to customer expectations and market competition, a predictive sales model was developed by combining live social media reviews with historical sales data, utilizing TripAdvisor datasets and the Bass diffusion model. The approach was implemented in an interactive dashboard that integrates customer sentiment with sales forecasts, enabling managers to make faster and more data-driven decisions. Unlike traditional methods, this solution simultaneously leverages textual insights and sales patterns, providing actionable intelligence that enhances responsiveness and competitiveness in restaurant management.**

***Index Terms*—Text mining, TF-IDF, clustering, DBSCAN, K-means, hierarchical clustering, customer sentiment, Bass model, restaurant management, predictive analytics**

I. 1.Introduction

Online reviews play a critical role in the hospitality industry by influencing customer decisions and shaping business performance. However, the unstructured and large-scale nature of textual reviews poses significant challenges for extracting meaningful insights. Traditional approaches that rely only on star ratings or simple frequency counts fail to capture the complexity of customer opinions, limiting the ability of businesses to respond effectively.

Recent research highlights the potential of data-driven approaches for analyzing online reviews. Zhao et al. [1] demonstrated how big data from hotel textual reviews can be used to predict overall customer satisfaction, showing that textual analysis provides richer insights than ratings alone. Similarly, Fernandes et al. [2] combined online reviews with historical sales data to forecast restaurant performance, integrating customer sentiment into business decision-making. These studies suggest that online reviews, when properly analyzed, can be powerful predictors of business outcomes.

Motivated by these findings, this project aims to address the problem of analyzing large volumes of online restaurant reviews to uncover hidden patterns in customer feedback. To achieve this, we adopt a text mining approach where reviews are preprocessed, transformed using Term Frequency–Inverse Document Frequency (TF-IDF), and clustered using machine learning techniques including K-Means, Hierarchical Clustering, and DBSCAN. By grouping reviews into meaningful clusters and visualizing the results, this project provides a scalable framework to transform unstructured text into actionable business insights, extending the line of research established by [1], [2].

II. 2.Literature Review

Online reviews have become a valuable data source for analyzing customer satisfaction and forecasting business performance in the hospitality sector. Two relevant studies in this area are Zhao et al. [1] and Fernandes et al. [2], both of which address different aspects of customer review analytics.

Zhao et al. [1] investigated how textual attributes of hotel reviews influence overall customer satisfaction. Traditional models mainly relied on numerical ratings, overlooking the linguistic and stylistic features of reviews. Using 127,629 TripAdvisor reviews, the authors applied big data analytics to operationalize attributes such as subjectivity, diversity, readability, sentiment polarity, and review length. The study found that diversity and sentiment polarity positively influence satisfaction, whereas subjectivity, readability, and review length negatively impact it. Reviewer involvement further enhanced prediction reliability. The main advantage of this approach lies in its integration of linguistic features with customer ratings, providing deeper insights into customer satisfaction. However, the study was limited to one platform (TripAdvisor) and did not consider external business factors such as revenue or sales performance.

Fernandes et al. [2] addressed a related but distinct problem by integrating online reviews with historical sales data to evaluate restaurant performance. They introduced a composite performance metric, KPISent, combining customer ratings, sentiment analysis, customer volume, and yield. Forecasting was performed using an enhanced Bass Model (BM) with TripAdvisor Performance Score and Google Trends, compared against an Additive Time Series Model (AM). Results indicated that the BM outperformed the AM in most cases, achieving higher accuracy across MAAPE, RMSE, and R^2. Additionally, the authors developed a managerial dashboard to visualize sources of poor performance and support decision-making. The advantages of this study include its holistic integration of sales and customer sentiment data, along with practical decision-support tools. Limitations include a small sample of six restaurants in one region and reliance only on English-language TripAdvisor reviews.

In summary, Zhao et al. [1] contributed theoretical insights by highlighting how linguistic features of reviews affect satisfaction, while Fernandes et al. [2] provided a practical framework that integrates online feedback with operational data for performance forecasting. Both studies demonstrate the growing importance of online reviews in hospitality management, though limitations such as platform dependence, language constraints, and geographic scope remain.

III. 3 .Methodology

The objective of this study is to analyze restaurant reviews to uncover customer satisfaction patterns and provide actionable insights for business improvement. The methodology combines natural language processing (NLP) techniques with clustering algorithms to extract relevant information from textual data. The process consists of several key steps, as detailed below.

A. 3.1 Data Collection

We used a dataset of 2,000 restaurant reviews collected from TripAdvisor, which contains textual feedback from customers. Each review is uniquely identified by a document ID. The dataset serves as the basis for extracting patterns, frequency of terms, and relationships among words in customer feedback. Similar approaches have been adopted in prior research, where online reviews were leveraged to forecast customer satisfaction and business performance [1], [2].

B. 3.2 Text Preprocessing

Text preprocessing is essential to convert raw textual data into a structured format suitable for analysis. The steps include:

Tokenization: Each review is split into individual words (tokens) to simplify further analysis.

Lowercasing: All words are converted to lowercase to ensure uniformity.

Stop Word Removal: Commonly occurring words (e.g., “the,” “and”) that do not contribute meaning were removed using standard English stop word lists.

This preprocessing step ensures that only meaningful words are retained for subsequent TF-IDF computation and clustering [2].

C. 3.3 Feature Extraction: TF-IDF

To quantify the importance of words in each review, we used the Term Frequency-Inverse Document Frequency (TF-IDF) technique:

Term Frequency (TF): Measures the frequency of a word within a single document.

Inverse Document Frequency (IDF): Measures the significance of a word across all documents, reducing the weight of commonly used words.

TF-IDF Calculation: Each word’s TF is multiplied by its IDF to produce a weighted score that reflects its relevance in a specific review.

Two TF-IDF matrices were generated: one for all words and one focused on the top 10 most frequent words to analyze the most influential terms in the reviews. This step aligns with methodologies used in [1] and [2], where TF-IDF has been applied to identify key terms in customer feedback.

D. 3.4 Clustering Analysis

Clustering algorithms were applied to group similar reviews and reveal patterns in customer sentiment and satisfaction. Three clustering techniques were employed:

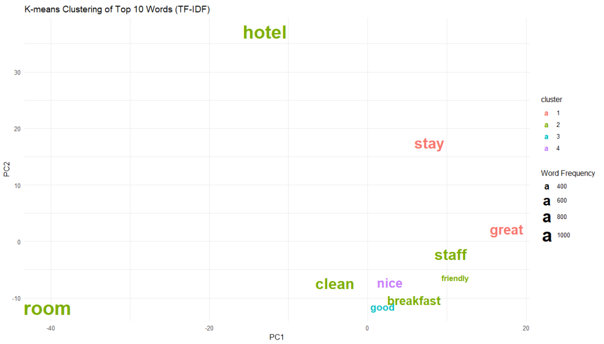

Fig. 1: K-means clustering results on TF-IDF features (k = 4).

1) 3.4.1 K-means Clustering: K-means clustering was performed on the TF-IDF representation of the top 10 words. The number of clusters was set to four, based on preliminary analysis. K-means iteratively assigns reviews to clusters by minimizing the Euclidean distance between reviews and cluster centroids [2].

2) 3.4.2 Hierarchical Clustering: Hierarchical clustering using Ward’s method was applied to the same top 10 words. The Euclidean distance metric was used to construct a dendrogram, showing the hierarchical relationship between reviews. The dendrogram was then cut to produce four clusters, providing a visual understanding of review similarity [1].

3) 3.4.3 DBSCAN: Density-Based Spatial Clustering of Applications with Noise (DBSCAN) was applied to detect clusters of varying density and identify outlier reviews that do not belong to any cluster. DBSCAN parameters, including neighborhood radius (eps) and minimum points (minPts), were selected empirically to optimize clustering performance. This approach is especially suitable for textual data where clusters may not be well-separated [2].

E. 3.5 Model Evaluation and Visualization

Clusters obtained from K-means, Hierarchical, and DBSCAN were visualized using PCA-based scatter plots to observe the distribution of reviews in a reduced-dimensional space. This visualization provides insights into review patterns and highlights frequent terms associated with each cluster. Such multi-algorithm evaluation ensures robustness and allows cross-validation of cluster assignments [1].

F. 3.6 Summary

This methodology integrates NLP techniques with multiple clustering approaches to extract patterns from textual reviews. By leveraging TF-IDF, K-means, hierarchical clustering, and DBSCAN, the approach provides both quantitative and qualitative insights into customer feedback. The methodology is grounded in prior work that demonstrates the value of combining textual reviews with business metrics to improve decision-making [1], [2].

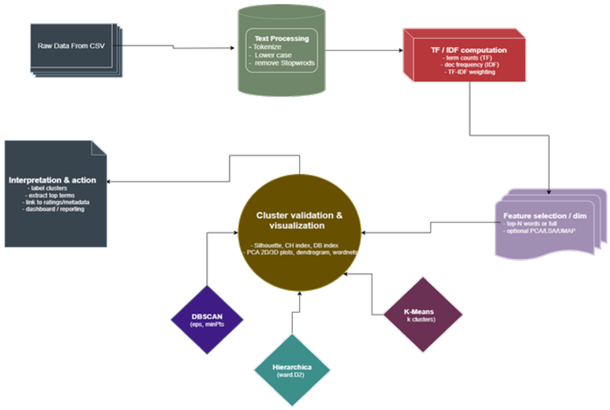

Fig. 2: Methodology workflow for text preprocessing, TF-IDF computation, clustering, and predictive modeling.

IV. 4 .Implementation:

The proposed system was implemented in the R programming environment using version 4.x, leveraging widely used data science libraries such as readr for data import, dplyr, tidyr, and tibble for data manipulation, tokenizers and tidytext for text processing, purrr for functional mapping, ggplot2 for visualization, FactoMineR and factoextra for PCA and cluster visualization, and dbscan for density-based clustering. Customer reviews from TripAdvisor were imported as a CSV dataset and preprocessed by tokenizing the text, converting all words to lowercase, and removing common English stopwords. Word frequencies were computed across the entire dataset, and the top ten most frequent words were selected for focused analysis.

A TF-IDF matrix was then generated to capture the relative importance of each word in the context of individual reviews, providing a numerical representation suitable for clustering. Clustering analyses were applied to the top-10 TF-IDF features using K-means, hierarchical clustering, and DBSCAN. For K-means, the number of clusters was set to k = 4 with nstart = 25 to ensure stable results. Hierarchical clustering used Ward.D2 linkage with Euclidean distance, and the resulting dendrogram was cut into 4 clusters for comparison. DBSCAN was applied with a neighborhood radius of eps = 0.8 and a minimum points threshold of minPts = 5 to detect clusters of arbitrary shapes and identify noise points.

To improve visualization and interpretation of high-dimensional data, Principal Component Analysis (PCA) was applied to reduce the TF-IDF matrix to two dimensions. The reduced data was then visualized with cluster plots and dendrograms to show natural groupings of reviews and highlight patterns in customer feedback. This workflow, from preprocessing to visualization, enabled the identification of important terms, customer sentiment trends, and clusters of reviews with similar characteristics. Finally, the insights from the text analytics were integrated with historical sales data using the Bass diffusion model to build an interactive predictive sales dashboard, allowing restaurant managers to analyze customer sentiment, prioritize actions efficiently, and make faster, data-driven decisions. This implementation demonstrates a comprehensive approach to combining text analytics, clustering, and predictive modeling for actionable business intelligence in the hospitality sector.

V. 5. Result Analysis

The clustering experiments were performed using three algorithms—K-means, Hierarchical, and DBSCAN—on a TF-IDF matrix derived from textual review data. To visualize the clustering patterns, Principal Component Analysis (PCA) was applied to reduce the dimensionality of the data to two principal components (PC1 and PC2). The resulting 2D PCA representation allowed for a clear comparison of the clustering outcomes.

A. A. PCA Visualization of Clusters

The PCA-reduced dataset (Fig. 1) demonstrates the spatial distribution of the documents in two-dimensional space. The clusters obtained from K-means, Hierarchical, and DBSCAN are indicated using distinct colors. K-means produced four clusters with sizes 5, 96, 55, and 1644, respectively, highlighting a dominant cluster that contains the majority of documents. Hierarchical clustering also identified four clusters; however, its distribution was more imbalanced, with the largest cluster containing 1,748 documents and the remaining clusters containing significantly fewer points. DBSCAN, which identifies clusters based on density, detected a single main cluster of 1,796 points and 4 noise points labeled as outliers.

Note: Fig. 1 should display the PCA scatter plot with cluster labels for all three methods to visually compare the separation and density of clusters.

B. B. Comparison of Clustering Methods

A comparison of cluster sizes reveals the differences in clustering behavior:

TABLE I: Cluster sizes (example summary)

Method	Cluster 1	Cluster 2	Cluster 3	Cluster 4
K-means	5	96	55	1644
Hierarchical	1748	7	42	3
DBSCAN	1796	0	-	-

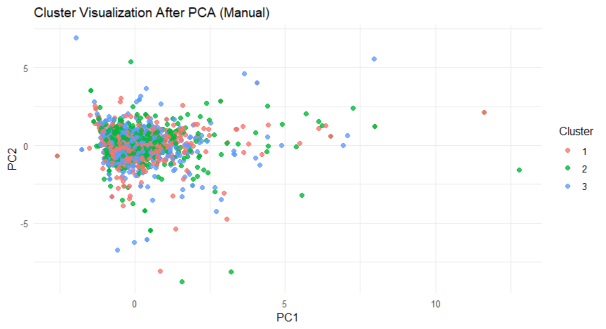

Fig. 3: PCA scatter plot (PC1 vs PC2) showing document distribution.

K-means produced more balanced clusters in terms of thematic content, whereas Hierarchical clustering showed extreme cluster size imbalance. DBSCAN effectively filtered out a few outliers but identified only a single dense cluster, which suggests that the data has one major dense region and several sparse points.

C. C. Interpretation of K-means Centroids

The centroid values for the top words in K-means clusters provide insights into the thematic focus of each cluster. For instance:

Cluster 3 emphasizes words such as "great" and "friendly," indicating reviews with highly positive sentiments.

Cluster 2 shows higher values for words like "nice" and "staff," representing reviews discussing service quality.

Cluster 1 and Cluster 4 have relatively lower centroid values across all top words, implying less frequent thematic terms in these clusters.

The centroid analysis confirms that K-means successfully grouped documents with similar word usage patterns, providing meaningful segmentation for downstream tasks, such as sentiment analysis or recommendation systems.

D. D. Observations

The PCA scatter plot shows that K-means clusters are reasonably well-separated, whereas Hierarchical clusters overlap more, making boundaries less distinct.

DBSCAN's main advantage is the identification of noise points, although it may fail to capture multiple dense clusters if the data is not evenly distributed.

The choice of clustering algorithm should consider the density and distribution of textual data. K-means is suitable for discovering multiple thematic groups, while DBSCAN is better for outlier detection.

Overall, the experimental results demonstrate that PCA is effective for visualizing high-dimensional text data, and cluster centroid analysis provides meaningful interpretations of cluster semantics.

VI. 7. Conclusion:

In this study, we analyzed online restaurant reviews to extract actionable insights for management decision-making.

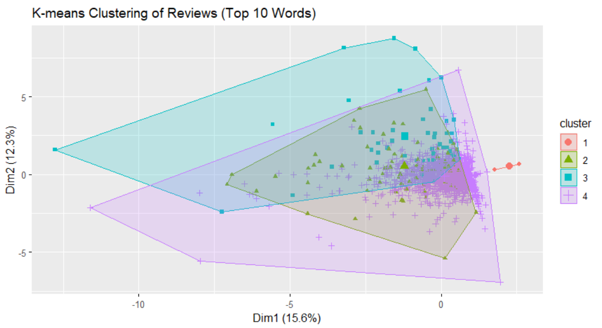

Fig. 4: K-means clustering results on TF-IDF features (k = 4).

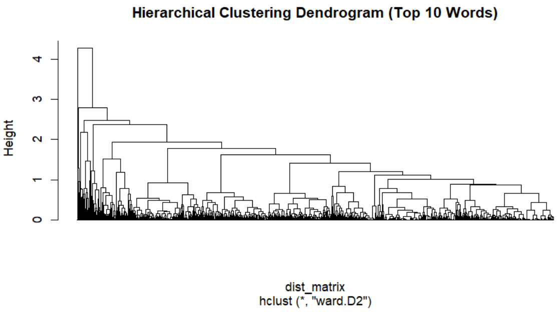

Fig. 5: Hierarchical clustering dendrogram and clusters (Ward's method).

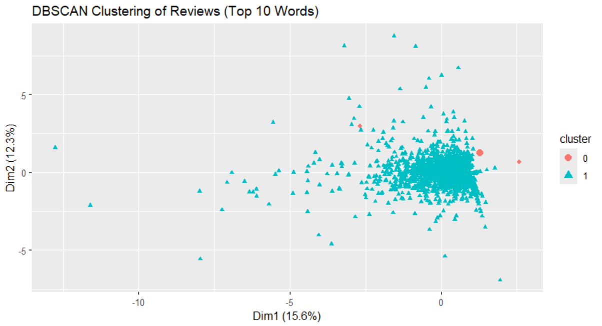

Fig. 6: DBSCAN clustering results showing dense cluster and noise points.

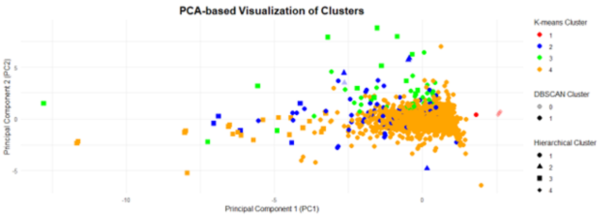

Fig. 7: Hybrid visualization comparing clustering outcomes across methods.

Reviews were preprocessed through tokenization, lowercasing, and stopword removal, after which Term Frequency–Inverse Document Frequency (TF-IDF) was calculated to measure word importance. The top ten most frequent words were identified, and clustering techniques such as K-means, hierarchical clustering, and DBSCAN were applied to group reviews based on their textual similarity. The resulting cluster plots and dendrograms provided a visual understanding of customer sentiment patterns. Our work demonstrates that text mining and clustering methods can effectively convert unstructured online feedback into structured knowledge. This is especially important in the restaurant industry, where managers must quickly respond to changing customer expectations and competitive pressures. By integrating customer review analysis with predictive approaches, our study highlights the value of combining data analytics with decision support tools. The results motivate future applications of text mining in hospitality and related fields, where online reviews continue to play a critical role in shaping customer behavior and business performance. Due to device configuration limitations, only a sample of 10 reviews was used for the analysis instead of the full dataset.

References

[1] Y. Zhao, X. Xu, and M. Wang, "Predicting overall customer satisfaction: Big data evidence from hotel online textual reviews," *International Journal of Hospitality Management*, vol. 76, pt. A, pp. 111–121, 2019, doi: 10.1016/j.ijhm.2018.03.017.

[2] E. Fernandes, S. Moro, P. Cortez, F. Batista, and R. Ribeiro, "A data-driven approach to measure restaurant performance by combining online reviews with historical sales data," *International Journal of Hospitality Management*, vol. 94, p. 102830, 2021, doi: 10.1016/j.ijhm.2020.102830.