Nova Information Management School
BSc in Data Science

Text Mining 2024-2025

Group Project: *"Solving the Hyderabadi Word Soup"*

Project Report

Group 6

João Capitão | 20221863
Maria Rodrigues | 20221938
Vidhi Rajanikante | 20221982
Yehor Malakhov | 20221691

1. Introduction

1.1. Context and Project Outline

In today's data-driven world, Text Mining techniques have become essential, allowing organizations to extract meaningful insights from vast amounts of unstructured text. Hyderabad is a major city in India, well-known for its rich cultural heritage, vibrant markets and landmarks. Hyderabad's Tourism Board, recognizing the value of Text Mining, is exploring how these techniques can enhance their services and create a more engaging experience for visitors

Various published studies have demonstrated the effectiveness of Text Mining methods in improving service efficiency and extracting valuable insights from model analyses.

The primary objective of this project is to analyze Hyderabad's restaurants using a dataset of reviews from Zomato. The process begins with data preprocessing, followed by Data Exploration using visualization tools such as Word Clouds and Tree Maps to extract insigths. Feature extraction techniques, including Bag-of-Words (BoW), will then be applied to structure the text data for subsequent modeling.

The models will be trained using Multilabel Classification to categorize cuisine types and assign quality scores to future restaurants in the city. Additionally, Sentiment Analysis will be employed to predict restaurant ratings based solely on the polarity (positive or negative sentiment) of their reviews. Co-occurrence analysis, clustering, and topic modeling will provide deeper insights into related dishes, classify reviews by topics, and interpret the meaning behind each topic.

1.2. Methodology

To ensure the proper implementation of this project, we will follow the **CRISP-DM methodology**. This approach is chosen because it provides a structured and comprehensive framework for organizing and executing our text mining project effectively.

2. Literature Review

2.1. Multilabel Classification[1]

Multiclass Classification techniques were leveraged to enhance the response time of the messages sent to a university admissions campaign, in which times often concerned more than one topic. By using three datasets, A and B classified the messages into one class, with 15 and 17 classes respectively, whereas C is used for multilabel classification by associating the messages to a vector of length 15. Thus, the authors used 3 ways to classify messages, along with 3 types of text vectorization, namely TF-IDF, OHE and Embedding, and trained multiple models using both ML and DL techniques, namely Random Forest, CatBoost, XGBoost, SVM, Logistic Regression, LSTM, GRU and RuBERT. In the end, the best classifier used Logistic Regression and TF-IDF with a final f1-score of 70% (Dataset A) and 93% (Dataset C).

2.2. Sentiment Analysis[2]

The usefulness of Sentiment Analysis is gathered to get an overview of how tourist attractions are perceived by their reviews on *Tripadvisor*. The authors performed a lexicon-based sentiment analysis using the AFINN lexicon, a word score between -5 and 5, to get the final scores for the reviews by summing the scores of all its words. They were able to see what aspects provoked certain types of emotion from the reviewers (i.e. “sadness” for littering), thus, allowing improvement on those aspects and/or services.

2.3. Co-occurrence & Clustering[3]

Co-occurrence and clustering analysis are two techniques of great use to infer the strength and relationship of different words. These techniques are applied to *CCTV news* documents consisting of noun, verb and adjective keywords. The authors got the weight for each word in the documents using TF-IDF, followed by the K-core theory analysis. Through the analysis of the strength of the co-occurrence keywords and obtaining the cluster for the latter, it was possible to infer which words have a tendency to be related, therefore, create groups of these words. For example, words like ‘interconnected’ and ‘invigorated’ and ‘inclusive’ and ‘innovative’ are strongly related between themselves, and to the subject of the referred G20 summit.

2.4. Topic Modelling[2]

Topic Modelling is usually used in tandem with other analyses, as is this case with sentiment analysis. It was employed for grouping related keywords, using the Non-Negative Matrix Factorization method, where the final results consisted of the best *k* topics and the words that best represented those topics, for each of the 10 most popular attractions in India. The results categorized efficiently the popular topics in all locations, giving further insights into what aspects tourists find more fascinating.

3. Data Understanding

3.1. Data Description

3.1.1. Restaurants Dataset

- The "restaurants" dataset (*105_restaurants.csv*) includes 105 restaurants located in Hyderabad, India. It provides valuable details such as the cost per person (in Indian Rupee ₹), associated collections, cuisine types and operating hours for each restaurant.
- An initial examination of the dataset revealed some missing information in the "Collections" field. Additionally, certain variables such as "Links" and "Timings," appear to offer limited analytical value and can likely be discarded. This dataset could also be leveraged for tasks such as restaurant segmentation.

3.1.2. Reviews Dataset

- The "reviews" dataset (*10k_reviews.csv*) contains metadata about reviewers and their feedback on the restaurants listed in the "restaurants" dataset. Each review includes the restaurant and reviewer names, the review text, the rating provided, reviewer metadata (e.g., number of reviews and followers), the timestamp of the review and the number of images shared with it.
- Reviews without valid content (e.g., missing text or inappropriate ratings such as "Like") were dropped, as they would hinder meaningful analysis. This dataset holds potential for various applications, including sentiment analysis and label classification.

3.2. Data Exploration

Our initial exploration of the datasets revealed the following insights:

- **Restaurants Dataset**:
  - The most common cuisine types are North Indian and Chinese.
  - The average cost of dining at a restaurant is ₹861.43 per person.
- **Reviews Dataset**:
  - Ratings of 5, 4 and 1 dominate, suggesting a tendency for reviewers to express extreme opinions. This pattern could be due to the fact that individuals with strong opinions, whether positive or negative, are more likely to leave reviews. The mean rating is also 3.600261, which further suggests that people tend to overrate their experiences. Given that the rating scale ranges from 1 to 5, the mean would typically be expected to be around 3 if ratings were more evenly distributed.
  - Most reviews are submitted during late hours (20:00–24:00), with summer being the season that sees the highest number of reviews.

4. Data Preparation

4.1. Data Cleaning

In the *restaurants* dataset, ‘Links’ and ‘Timings’ columns were dropped and the commas from the *cost* column were removed. Additionally, we applied multi-hot encoding to the ‘cuisines’ and ‘collections’ columns, converting each possible value into a binary vector to represent multiple categories within those columns.

In the *reviews* dataset, metadata was dropped as it was not quite relevant to our objectives. Additionally, some columns were converted to more computationally efficient data types. Next, preprocessing the actual review text was focused upon. To address the various goals we had, such as multilabel classification, sentiment analysis, clustering, and topic modelling, we created a preprocessing function, *pipeline*, that could handle all these tasks. We simply needed to specify the desired hyperparameters. It took the reviews dataframe as input

and returned a dataframe with new columns, including one with the cleaned text and additional columns for each specific preprocessing step. The text cleaning was handled by another function, *text_cleaner*, which performed tasks like lowercasing, punctuation removal, emoji removal and other user-defined steps.

4.2. Additional exploration on the reviews

For further exploration of the reviews, feature extraction was applied to the preprocessed text, starting with Bag of Words (as TF-IDF wasn't applicable in our case due to having a single document). We visualized the most common words, also breaking them down by Part-of-Speech Tags. A network graph was constructed after examining the correlation between words.

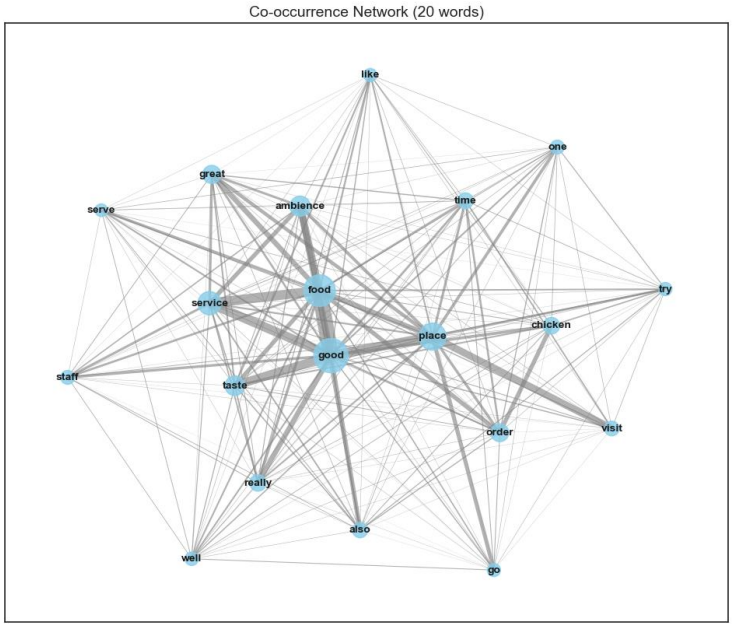


Figure 1: Network graph of the top 20 words in the reviews

To compare the most important words across different ratings, we first split the reviews dataset into five documents based on the rating categories (ignoring the minimal 0.5 ratings). After splitting, it made sense to use TF-IDF for feature extraction instead of Bag of Words, as it reduces the emphasis on common words across all reviews and gives more weight to words that appear uniquely in specific ratings.

From the word clouds below, we see a clear distinction between Rating=1 and the other ratings.

Figure 2: TF-IDF word clouds for reviews categorized by each of the five ratings.

4.3. Specific Data Preprocessing

4.3.1. For Multilabel Classification:

The two dataframes were merged into a unified dataset combining reviews and cuisine types. The cuisine types were binarized for easier classification, and stop words were removed to reduce noise. Lastly, words were lemmatized to their root forms, minimizing dimensionality and improving analysis efficiency.

4.3.2. For Sentiment Analysis:

Punctuation and UTF-8 emojis were retained as they convey sentiment. Stopwords were also kept as removing them could alter the sentence's meaning. For example, removing the stopword "not" could unintentionally flip the sentiment of the sentence.

4.3.3. For Co-occurrence Analysis and Clustering:

Stopwords were removed as they are not relevant to our objectives. Additionally, all types of punctuation, emojis, and other unnecessary elements were eliminated. The text was lemmatized and converted to lowercase, among other transformations.

4.3.4. For Topic Modelling:

When using Sklearn methods, we used the matrix generated from Bag of Words feature extraction and the lemmatized tokens from the clean_review_lemmatized column in Gensim. The Bag of Words matrix was reused from the review dataset exploration, while the lemmatized tokens had to be converted into a list using the split method. By using lemmatized words, we reduced the dimensionality, enabling the model to focus more on the meaning of the words, which improved topic coherence.

5. Modelling

5.1. Multilabel Classification

For multilabel classification, we have chosen a classifier chain that employs decision trees. This choice is based on the belief in the correlations among different cuisines, which justifies the use of a classifier chain, and highlights the superior classification capability of decision trees.

5.2. Sentiment Analysis

The sentiment of the reviews was calculated using two different analyzers: VADER and TextBlob. For each analyzer, we applied two approaches: full-text analysis and per-sentence analysis. The results were as follows:

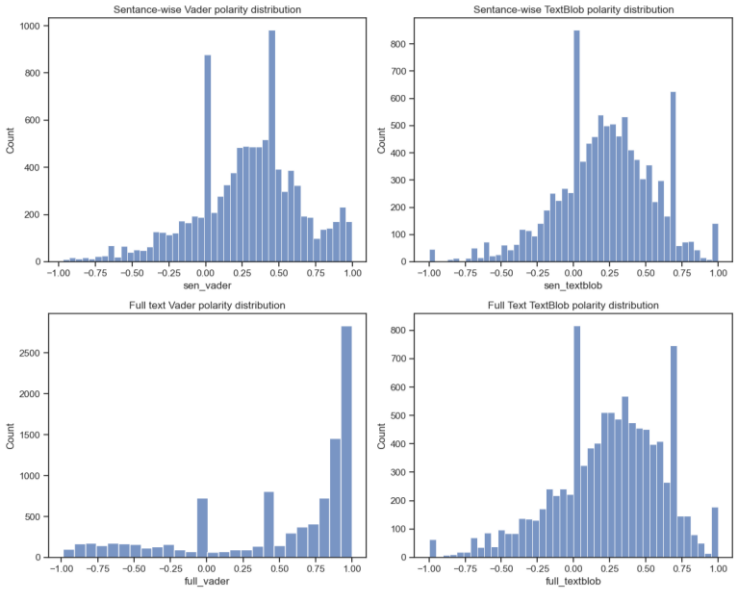


Figure 3: Sentiment distribution for two analyzers (VADER and TextBlob) using two approaches: full-text analysis and per-sentence analysis.

The following questions were addressed: Are the ratings and the sentiment of the sentences correlated? Can we predict the rating solely based on the sentence's sentiment? These questions shall be answered in the next section.

5.3. Co-occurrence & Clustering Analysis

We aimed to identify which dishes are mentioned together in the reviews. To achieve this, all unique nouns were extracted from the text and those corresponding to foods and drinks were selected. Using these selected words, a co-occurrence matrix and a network graph were constructed to analyze their relationships.

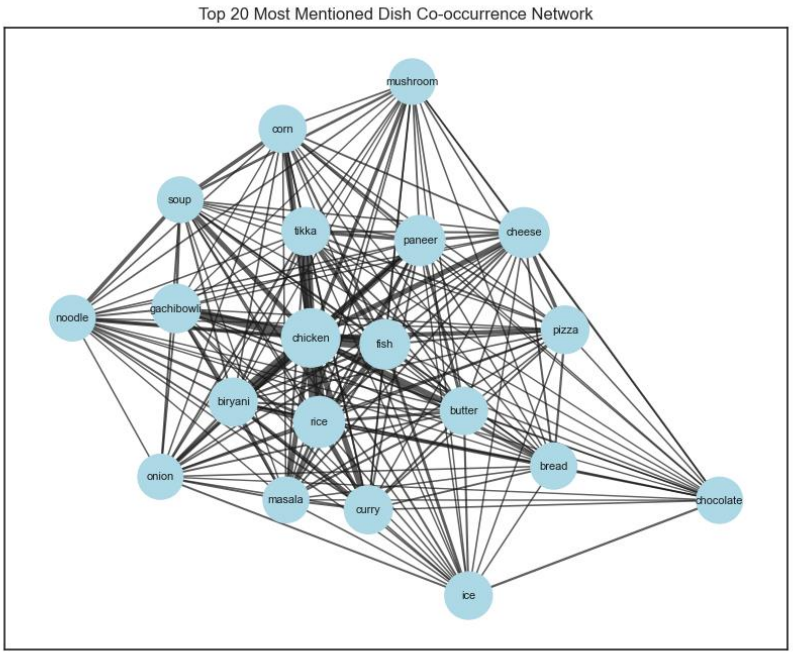


Figure 4: Network graph of the top 20 foods in the reviews

Next, we aimed to form clusters to determine whether cuisine types could be identified based on these clusters. After merging the reviews dataset with the "Cuisines" column from the restaurants dataset, two feature extraction methods were applied: Bag of Words (BoW) and Doc2Vec.

Before clustering with KMeans, we plotted the inertia and used the elbow method to determine the optimal number of *k* clusters. Once the clusters were generated, each cluster's name was based on its content and the most common cuisine types were displayed within each cluster.

However, since the clusters showed little distinction in terms of cuisine types, we explored a different approach. A new column in the reviews dataset containing only words related to foods or drinks was created and used for clustering, yet again using BoW as it provided better results).

We also tried using HDBSCAN to create clusters, but the results were not as effective

5.4. Topic Modelling

The purpose of applying topic modeling is to determine whether reviews can be classified into emergent topics and to identify what those topics consist of. To achieve this, we performed Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA), using both the Sklearn and Gensim libraries.

Starting with Sklearn's LSA implementation, we first obtained the document-topic matrix, where each document represents a review. Then, we extracted the components, which correspond to the topic-word matrix, showing the importance of each word to a topic. Using the Gensim library, we employed LSI (Latent Semantic Indexing), as its methods offer different insights. To implement the LSI model, we needed a dictionary of the reviews and the reviews in corpus format. For both libraries, these steps were carried out for 20, 15, and 10 topics. The final topics for each iteration were analyzed, and the best model was selected.

Although the LDA implementation is similar to LSA, LDA is generally more interpretable. This is because LDA ensures a coherent word distribution, leading to a more interpretable distribution of words within topics. LDA was performed with 20, 15, 10, and 8 topics. The results were then analyzed, and the optimal number of topics was chosen.

6. Evaluation

6.1. Multilabel Classification

After grid search, we end up with the following set of hyperparameters. The model built with these hyperparameters performed with the following results, primarily achieving an F1 score of 0.38, which can be considered quite high given the existence of over 40 different classes.

```
output:

{'classifier__base_estimator__class_weight': 'balanced',
'classifier__base_estimator__criterion': 'gini',
'classifier__base_estimator__max_depth': 15,
'classifier__base_estimator__max_leaf_nodes': 50,
'classifier__base_estimator__min_samples_leaf': 1,
'classifier__base_estimator__min_samples_split': 10}
```

```
output:

accuracy = 0.037
precision_weighted = 0.365
recall_weighted = 0.447
f1_weighted = 0.38
```

6.2. Sentiment Analysis

As mentioned earlier, VADER and TextBlob help us answer two key questions: whether the rating and sentiment are correlated, and if we can predict the rating based solely on the perceived sentiment. The results and corresponding answers to these questions are as follows:
From the plot below, we can conclude that the review sentiment is, in fact, fairly correlated with the ratings.

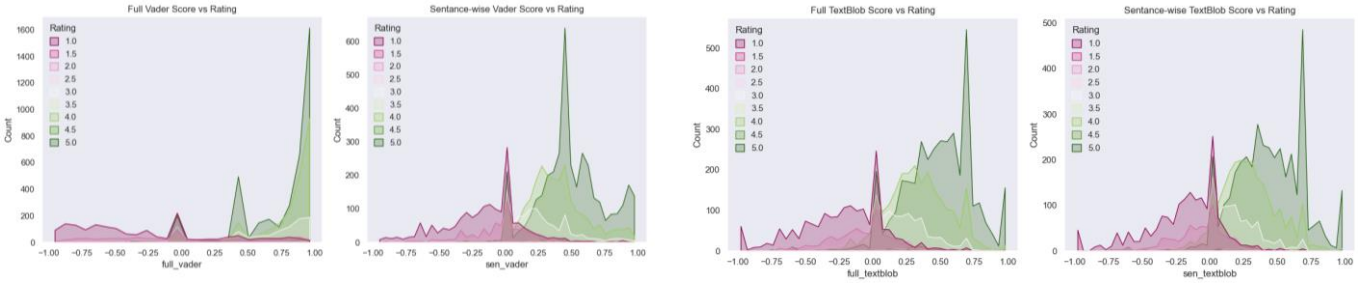


Figure 6: Sentiment distribution grouped by rating

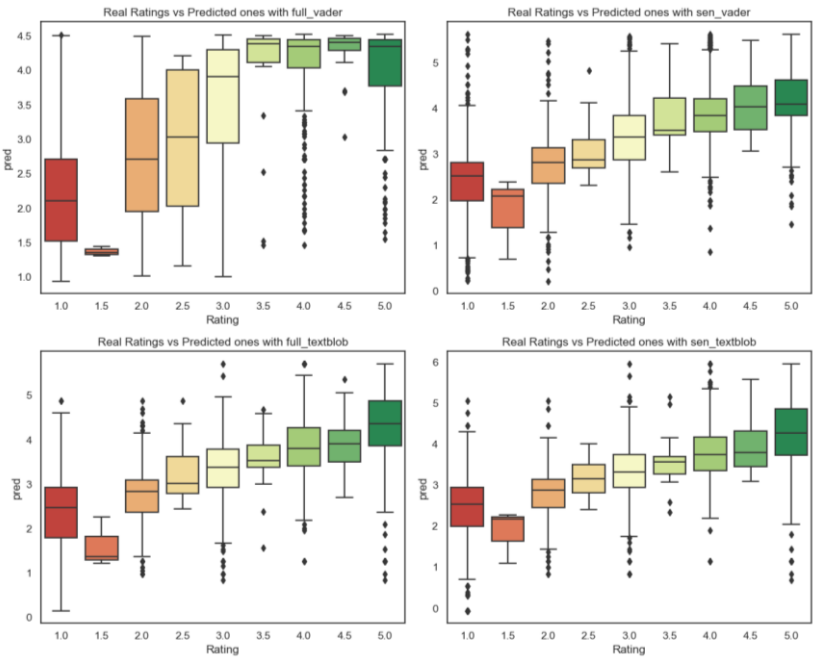


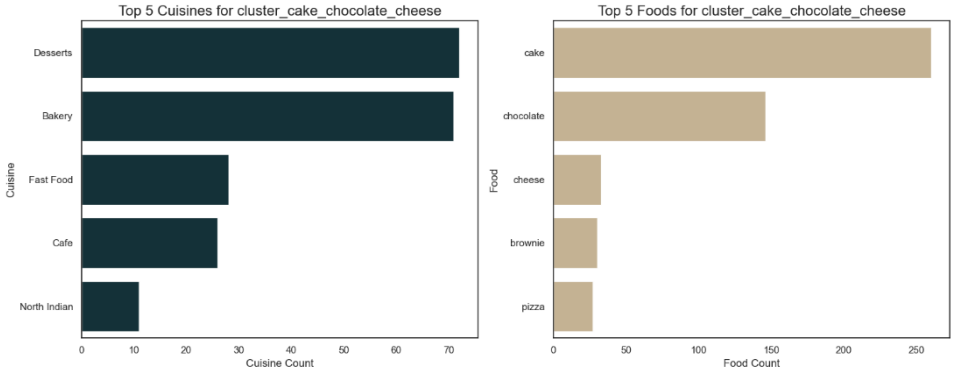
Figure 7: Distribution of predicted ratings based on review sentiment compared to actual ratings

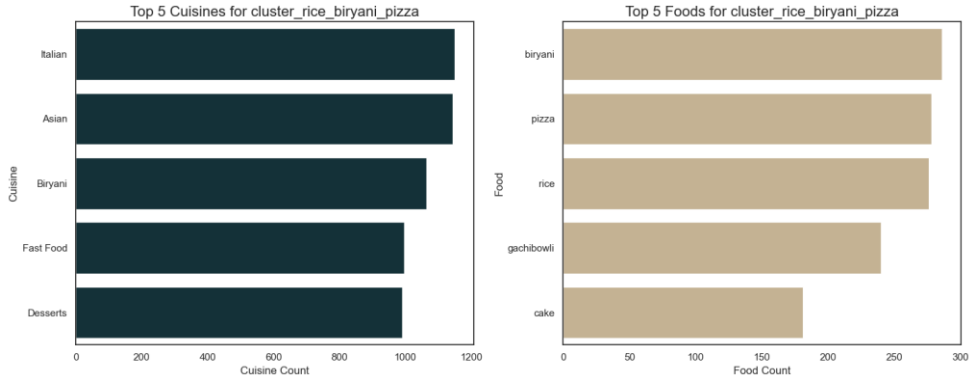
The highest R^2 obtained was 0.5 using the full TextBlob model, meaning that 50% of the variance in ratings can be explained by the reviews.
Additionally, the Mean Absolute Error (MAE) was 0.83, meaning that, on average, our predictions deviate from the actual ratings by 0.83 units.

6.3. Co-occurrence & Clustering Analysis

6.3.1. .

The final visualization showcases the most common cuisine types within each cluster, alongside the most frequently mentioned foods for each cluster.
This resulted in some interesting clusters, such as the Desserts and Italian clusters. Regarding the Italian cluster, note that this plot was generated after removing the 'North Indian', 'Chinese', and 'Continental' cuisine types, which dominated most clusters in the initial visualization. To highlight other patterns, these dominant cuisine types were excluded in a second visualization.





6.4. Topic Modelling

For LSA, the final number of topics selected was 15. For example, we can observe the weight of each topic for the word 'chicken' (Figure 9).

As an example, a review from the restaurant Cream Stones, an ice cream establishment (observation 983), was used. The most emergent topic for this review was Topic 4. We can observe what top 5 words hold more weight to Topic 4 (Figure 10).

```
array([ 0.27245342,  0.65700769, -0.04089384, -0.08240239, -0.49518027,
        0.06694   , -0.04006742,  0.13443615,  0.13038939, -0.19475652,
        0.11629341, -0.08371036,  0.07014007, -0.07958133,  0.08812318])
```

Figure 8: Weight of each Topic for 'chicken'

```
[('order', 0.5242419325931293),
 ('chicken', -0.4107104795254137),
 ('food', -0.4041135517619045),
 ('good', 0.18439785581230214),
 ('taste', 0.15654355839851994)]
```

Figure 10: Top 5 Words for Topic using Sklearn LSA

From the 15 topics created, we can observe the words that are most important to each and infer their meaning. For example, Topic 4 includes words suggesting a good place to visit with friends that sells sweets, while Topic 2 might refer to a restaurant that serves biryani, a mixed rice dish made with meat or seafood and spices.

However, this analysis is somewhat challenging to interpret due to the repetition of words across multiple topics. It could be useful to remove certain words, such as food, place, order, get, go, restaurant, like, and try. Since these words are common in reviews of restaurants, they are likely to appear in any topic, thus offering little specific insight.

```
Topic 1: ['good', 'food', 'place', 'chicken', 'service', 'taste', 'one', 'ambience', 'ordered', 'really']
Topic 2: ['chicken', 'biryani', 'taste', 'fried', 'veg', 'ordered', 'fish', 'rice', 'spicy', 'dish']
Topic 3: ['good', 'taste', 'also', 'veg', 'nice', 'quantity', 'biryani', 'ok', 'biriyani', 'overall']
Topic 4: ['place', 'good', 'n', 'best', 'one', 'visit', 'friends', 'chocolate', 'hangout', 'amazing']
Topic 5: ['one', 'taste', 'like', 'n', 'restaurant', 'even', 'time', 'us', 'ordered', 'served']
Topic 6: ['n', 'prawns', 'great', 'food', 'sauce', 'best', 'garlic', 'u', 'topped', 'fish']
Topic 7: ['food', 'n', 'ordered', 'taste', 'place', 'biryani', 'good', 'quality', 'order', 'quantity']
Topic 8: ['service', 'n', 'biryani', 'chicken', 'time', 'order', 'ordered', 'great', 'worst', 'restaurant']
Topic 9: ['n', 'veg', 'us', 'even', 'chicken', 'order', 'place', 'restaurant', 'table', 'starters']
Topic 10: ['biryani', 'n', 'veg', 'starters', 'non', 'service', 'served', 'prawns', 'mutton', 'main']
Topic 11: ['one', 'best', 'try', 'chicken', 'must', 'also', 'good', 'hyderabad', 'food', 'biryani']
Topic 12: ['really', 'nice', 'taste', 'like', 'us', 'well', 'served', 'also', 'dish', 'staff']
Topic 13: ['great', 'time', 'us', 'good', 'served', 'order', 'experience', 'staff', 'biryani', 'hyderabadi']
Topic 14: ['really', 'ordered', 'biryani', 'nice', 'order', 'best', 'great', 'n', 'veg', 'cake']
Topic 15: ['one', 'like', 'really', 'great', 'veg', 'taste', 'order', 'chicken', 'food', 'service']
```

Figure 11: Top 10 Words for each Topic using Gensim LSI

Recalling the example review from Cream Stones mentioned above and employing LDA from Sklearn and Gensim, we can check the top 5 words on the topic that contribute the most to the review (Figure 12).

Furthermore, we can assess the similarity of the words assigned to each topic. From the Gensim LDA, we obtained an average coherence of 41.1%, with the highest coherence in Topic 6 at 50.8% and the lowest in Topic 9 at 33.1%, along with a perplexity of -7.69660. Based on the final topics, we can infer that Topic 0 refers to a restaurant specializing in sweet food and beverages, while Topic 6 describes establishments known for their fried chicken dishes (Figure 13).

These results show to be of easier interpretability, allowing for better conclusions to be drawn on the topics' meanings.

```
ice          335.370624
cream        335.072506
chocolate    270.628972
cake         240.633584
taste        189.155425
Name: 5, dtype: float64
```

Figure 12: Top 5 Words for Topic 5 using Sklearn LDA

```
Topic 0: ['chocolate', 'cake', 'cream', 'brownie', 'ice', 'shake', 'time', 'give', 'great', 'place']
Topic 1: ['place', 'food', 'good', 'taste', 'ambience', 'try', 'visit', 'coffee', 'awesome', 'best']
Topic 2: ['good', 'place', 'food', 'time', 'ambience', 'service', 'go', 'music', 'taste', 'decent']
Topic 3: ['delivery', 'order', 'food', 'even', 'serve', 'good', 'service', 'taste', 'bad', 'get']
Topic 4: ['good', 'food', 'order', 'taste', 'chicken', 'restaurant', 'paneer', 'place', 'biryani', 'service']
Topic 5: ['order', 'bad', 'chicken', 'food', 'taste', 'good', 'quality', 'get', 'time', 'quantity']
Topic 6: ['chicken', 'taste', 'place', 'food', 'try', 'dish', 'good', 'one', 'starter', 'fry']
Topic 7: ['place', 'food', 'service', 'one', 'order', 'good', 'go', 'time', 'great', 'visit']
Topic 8: ['good', 'food', 'service', 'place', 'nice', 'really', 'great', 'ambience', 'staff', 'also']
Topic 9: ['good', 'food', 'pork', 'quantity', 'less', 'taste', 'service', 'tasty', 'staff', 'biryani']
```

Figure 13: Top 10 Words for each Topic using Gensim LDA

7. Conclusions

The goal of this project was to extract the greatest number of insights possible from reviews of Hyderabad's restaurants. To achieve this, we employed several text mining techniques tailored to meet specific information requirements.

The first objective was to predict a restaurant's Zomato score based on the polarity of its reviews. By analyzing the sentiment of the reviews, we achieved a Mean Absolute Error (MAE) of 0.83. This means that our predicted ratings were off by an average of just 0.83 units, which is quite reasonable and shows that sentiment analysis can be a reliable predictor of restaurant ratings.

The second objective was to determine whether it's possible to predict a restaurant's cuisine based on the review. As observed from the results, classifier chain models were able to handle this task, although not perfectly.

The third objective was to identify dishes that are commonly mentioned together in the reviews. We did uncover some interesting insights here, but the process of extracting food-related words was somewhat manual. Initially, we considered using Spacy, but it lacks a prebuilt 'dishes' entity. We attempted to fine-tune a custom model, but due to the limited data, the results were not as effective as hoped. As a result, we resorted to a more manual approach to identify relevant terms.

Another key question was whether we could identify cuisine types based on clustering analysis. While we were able to create clusters, most of them were dominated by North Indian and Chinese cuisines. Nevertheless, we were still able to uncover some interesting clusters, demonstrating that clustering can provide valuable insights, though further refinement may be necessary.
Last, but not least, we verified that it is possible to classify reviews according to topics and extract the main subject of each topic. Although the insights taken from the topics were sometimes not independent from each other, for some words appeared repeatedly, it was somewhat effective in determining the most related topic to a specific review.

8. References

[1] N. V. Smirnov, A. S. Trifonov, "Classification of Incoming Messages of the University Admission Campaign", SmartIndustryCon, 2023.
Available: Classification of Incoming Messages of the University Admission Campaign" | IEEE Conference Publication | IEEE Xplore, SmartIndustryCon, 2023. (15/10/2024)

[2] S. Singh, T. Chauhan, V. Wahi, P. Meel, "Mining Tourists' Opinions on Popular Indian Tourism Hotspots using Sentiment Analysis and Topic Modeling", ICCMC, 2021.
Available: Mining Tourists' Opinions on Popular Indian Tourism Hotspots using Sentiment Analysis and Topic Modeling" | IEEE Conference Publication | IEEE Xplore, ICCMC, 2021. (14/10/2024)

[3] S. Liu, X. Fan, J. Chai, "A Clustering Analysis of News Text Based on Co-occurrence Matrix", IEEE, 2017.
Available: A Clustering Analysis of News Text Based on Co-occurrence Matrix | IEEE Conference Publication | IEEE Xplore", IEEE, 2017. (15/10/2024)