Solving the Hyderabadi Word Soup

Karolina Rączkowska 20241250 | Michał Wójcik 20241255 | Hubert Kołomański 20241253 |

Marek Rewoliński 20241452 | Christian Deluca 20241264

1. Introduction

Nowadays, online reviews are a crucial factor in choosing a restaurant. People want to ensure they pick a great spot to eat out. Most people are familiar with Google Maps which has the option of both navigating, viewing and rating everything from restaurants to tennis courts. Similarly, Zomato, tailored to the Indian dining industry, enables users to discover restaurants, read reviews, and provide feedback. This system benefits both customers and restaurants by facilitating informed choices and improvement based on feedback.

Our goal is to analyze a dataset of 105 restaurants in Hyderabad (state of Telangana, India) and a dataset of 10 000 of their Zomato reviews to uncover meaningful patterns. Using approaches like Multilabel Classification and Sentiment Analysis, we will be able to answer two major questions:

- How well can we classify a restaurant's cuisine type using the content of their reviews as input? (Multilabel Classification)
- How well can we predict a restaurant's Zomato score using the polarity of their reviews as input? (Sentiment Analysis).

In addition, we will conduct co-occurrence analysis and topic modeling to gain a deeper understanding of the key themes, word associations, and discover what meals are mentioned together in reviews.

2. Literature Review

The content of *Speech and Language Processing* book [1] provides a comprehensive introduction to natural language processing (NLP), including techniques relevant to Sentiment Analysis. Here we can find both lexicon-based approaches and machine learning methods for analyzing the emotional polarity of text. The chapters on text categorization and Sentiment Analysis were very helpful since they provided theoretical foundations and practical techniques for implementing Sentiment Analysis.

Research concluded by *Wankhade* [2] outlines challenges in sentiment analysis, such as detecting sarcasm, informal language, grammatical errors, and the impact of adverbs and intensifiers. They also highlight issues like limited labeled data and high computational costs that hinder model accuracy and progress.

A study by *Egger and Yu* [3] assessed LDA, NMF, Top2Vec, and BERTopic on Twitter data. While LDA excelled with large datasets, it lacked semantic depth. NMF was better for small datasets, Top2Vec excelled at clustering, and BERTopic, using pre-trained embeddings and HDBSCAN, produced the most coherent topics.

Methodology presented in the study described by a group of Stanford's researchers [4] trying to design a Machine Learning model for assigning multiple classes to various articles was a helpful resource for our tasks. However, some steps were augmented, changed and adapted to better fit our problem. One difference in data preprocessing was that we decided to use more than just TF-IDF vectorization. For the algorithm however, Naive Bayes Classifier used in the study turned out to be suboptimal for our problem generating the lowest F1_scores. As a result it was not chosen as one of the main algorithms in the wrapper during our testing.

3. Data Understanding

In this report, we analyze two datasets: one containing 10.000 restaurant reviews and another with details about **105** restaurants. The data is collected from Zomato in the region of Hyderabad, India. However, only 100 of these restaurants have reviews. The reviews span over the period **from May 2016 to May 2019**. Only 2.5% of the reviews are from 2016 and 2017, while the rest are evenly distributed between 2018 and 2019.

Posted reviews are, on average, **50 characters long**. Neutral and slightly positive reviews (3 and 4 stars) usually contain more characters. Restaurant ratings (ranging from 1 to 5 in 0.5 increments) are diverse - **around 40% are 5-star ratings**, and 25% are 4-star ratings. Negative reviews with 1 star also form a significant category, accounting for 17% of the total.

The restaurants with reviews are divided into **42 distinct cuisines**, with each restaurant, on average, **offering dishes from three cuisines**. The restaurants are also categorized into different pricing tiers, ranging from 150 to 2800. There is a moderate positive correlation between price and average ratings.

4. Data Preparation

The data preparation step is critical for ensuring the dataset is clean, standardized and ready for analysis. Initially, missing reviews and null values were removed to ensure data completeness, and reviews deemed uninformative, such as those consisting of fewer than three characters or containing excessive gibberish, were discarded.

The preprocessing pipeline was designed with a structured, multi-stage approach. During the first stage, non-essential elements like **emojis, hashtags, newlines, URLs, punctuation, contractions, numbers, and stop words** were removed as necessary, depending on the downstream task. Common c**ontractions** were standardized (e.g., m → am, n't → not), and **diacritics** were normalized (e.g., café → cafe) to handle special characters effectively. Optional steps, such **tokenization, lemmatization, part-of-speech tagging, and detokenization**, were implemented when required for specific analysis purposes and will be described in a further part of the report.

Further refinements were made to enhance text quality and uniformity. **Uppercase words** were split into separate components, and variations of frequently used expressions, such as **"good" (gud, goo, gd)** and **"ok" (kk, Oke, k)**, were standardized to their common forms. **Gibberish words,** such as excessively long tokens or words with repeated characters (e.g., "aaaaa"), were identified and removed. Additionally, **spaces around punctuation** were corrected to ensure clean formatting and proper spacing.

To proceed with our analysis on reviews, we have utilized pre-trained **NER RoBERTa** model to retrieve food entities from our dataset. The selected model was fine-tuned on 400 Instagram posts related to food and on test dataset had f1-score equal to 0.91. [5]

Our final output after changing retrieved meals to lowercase and removing rows with less than 3 characters resulted in final **7197 distinct meals/ food entities.**

We decided to pick Robustly Optimzed BERT model, because it holds some significant advantages in pre-training over base BERT. It enhances BERT by using **more training data (160GB**), dynamic masking (static in BERT), and eliminating the Next Sentence Prediction task. It optimizes hyperparameters, adopts byte-level Byte-Pair Encoding, and fine-tunes dropout rates, resulting in improved performance on benchmarks like GLUE and SQuAD. [6]

5. Data Modelling

5.1. Sentiment Analysis

To perform Sentiment Analysis, we employed two distinct approaches: the VADER model, that is well suited for Social Media and a Transformer-based method. For the Transformer-based approach, we used a fine-tuned on Twitter posts RoBERTa model. []. For both we used the same preprocessed dataset:

- Emojis, punctuation marks, and other special characters were kept to provide accurate data and preserve context and sentimental nuances.
- No lemmatization was applied during preprocessing.
- Stopwords were also kept.
- Remove reviews without ratings
- Remove review with rating = “Like”

In order to evaluate the outputs, we assumed that ratings under 2.5 stars are negative, under 4 stars are neutral and others are positive.

5.2. Multilabel Classification

To prepare the data for the classification task, we carried out the following steps:

- Removed missing reviews to ensure data completeness.
- Corrected spelling and cleaned gibberish in the messages using regular expressions.
- Converted diacritics and generated Part-of-Speech (POS) tags.
- Extracted food names using a pretrained NER model based on the RoBERTa transformer.
- Removed stop words, converted cuisines into One-Hot vectors, and applied lemmatization and tokenization as the final text preprocessing steps.
- Finally, in order to reduce the noise within the data, only the reviews containing food Named Entities were kept for the next parts of the analysis. Consequently, dropping almost half of all the observations, leaving us with a little over 5 thousand rows to work on.

Before vectorization, we prepared two versions of the data:

- One data frame containing all individual reviews.

- Another data frame with reviews aggregated by restaurants.

For text encoding, we utilized three vectorization methods: **Word2Vec, Bag of Words (BOW), and TF-IDF**.

To prevent data leakage, each prepared data frame was split into training and test sets. We evaluated three classification algorithms: **Logistic Regression**, **Random Forest**, and **Support Vector Machine**. These algorithms were paired with various wrappers designed for multilabel classification, including **Classifier Chain**, **OneVsRest**, and **MultiOutput**.

5.3. Meal Co-occurrence Analysis

To ensure high-quality input data for co-occurrence analysis, we performed extensive preprocessing steps.

- Extracted meal data from reviews with fine-tuned NER RoBERTa model.
- Removed rows with empty meal lists.
- Standardized dish names by converting them to lowercase and stripping extra spaces.
- Corrected spelling and merged synonyms (e.g., “biriyani” → “biryani”)
- Removed non-dishes and ingredients, for instance “chicken” or “vegetables”.

We chose 70 dishes to ensure clarity and ease of interpretation of the graph and keeping as frequent dishes from the reviews as possible. While including more dishes, the graph nodes were clogged together and irrelevant nodes, for instance ingredients were connected with dishes. Additionally, very infrequent co-occurrences were displayed, which hindered the observation and implemented unnecessary noise that did not offer any additional insights.

Clustering

To further analyze the underlying relationships between meals, we applied clustering methods to segment the extracted meals into distinct groups. Two approaches were implemented:

- **HDBSCAN clustering** based on a precomputed co-occurrence matrix.
- **BERTopic clustering** using **Transformer-based word embeddings** derived from the vocabulary vector of the co-occurrence matrix.

In the first method, the co-occurrence matrix was converted into distance matrices using **Euclidean distance**, **Cosine similarity**, and **Pairwise metrics**.

BERTopic method incorporated **word embeddings**, **UMAP** for dimensionality reduction, and **HDBSCAN** for clustering. The final step was customized by calculating **topic labels** using cosine similarity between the word embeddings of the **Cuisine Vector** and the embeddings of the **top 5 most frequent words** within each topic.

5.4. Topic Modelling

Two independent approaches were applied to discover the topics describing restaurants reviews: Latent Dirichlet Allocation (LDA) and BERTopic.

After the general preprocessing emojis, newlines, hashtages, urls, contractions and diacritics are already removed. The preprocessing steps for both Latent Dirichlet Allocation (LDA) and BERTopic methods begin by removing reviews with **fewer than 3 characters**. They then diverge: in LDA we immediately remove **stopwords and punctuation,** while in BERTopic retain these elements initially for contextual understanding. Both methods proceed to remove contractions, emojis, and numbers from the text. However, they differ again in their final processing steps - **LDA applies lemmatization** to reduce words to their base forms and uses word tokenization at a later stage, whereas BERTopic specifically avoids lemmatization and **employs sentence tokenization instead.**

LDA

The **LDA from Gensim** method involved analysing reviews using a bag-of-words (BoW). A dictionary was created with corpora.Dictionary, and reviews were converted to a corpus (doc2bow). Fine-tuning LDA parameters on the number of topics, alpha, eta to boost convergence consistently resulted in the lowest number of topics (e.g., 2) with minimal improvement from further adjustments.

BERTopic

BERTopic leveraged embeddings for better context and stability. Dimensionality reduction was done with UMAP (n_neighbors=15, n_components=5), and clustering with HDBSCAN (min_cluster_size=30). cTf-Idf and **MMR (diversity=0.3)** were used to extract descriptive topic names.

The initial model generated **53 topics**, which were then refined through **Hierarchical Visualization and Representations.** Through manual merging and analysis, we consolidated these into **20 more generalized topic** groups. The 20 refined topics encompass specific domains such as: **place and ambience**, **service and staff**, **delivery experience**, Zomato platform-specific feedback, food quantity and quality and cuisine types (Chinese, Indian, buffet, vegetarian, Italian, American, desserts).

6. Evaluation

6.1. Sentiment Analysis evaluation

To compare the performance of VADER and the transformer model [6] - RoBERTa, we analyzed confusion matrices (Figure 1 and 2) and its metrics (Table 1 and 2). Due to the imbalance in class distribution, we utilize weighted F1 score as the primary metric in our study. (Relying only on accuracy can be misleading because it favors the most common class while ignoring how well the model handles the less frequent ones). Thus, as we can see on Table 1 and 2, the weighted f1-score is significantly higher for RoBERTa-base model (**0.80**) than VADER (**0.73**). Furthermore, the Transformer-based model also outperforms VADER in terms of correlation with the initial rating. In order to compare the models, a custom Polarity Score was calculated for RoBERTa model using the formula of **(positive - negative)/ (positive + negative)**. The Spearman correlation for the Transformer is **0.73**, while for the VADER method, it is **0.57**.

Figure 1: Confusion Matrix – RoBERTa-base model.

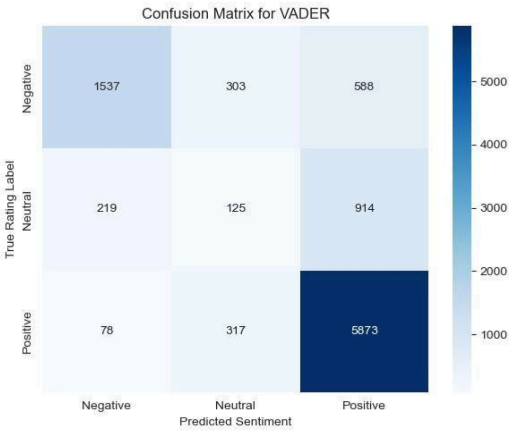


Figure 2: Confusion Matrix – VADER model.

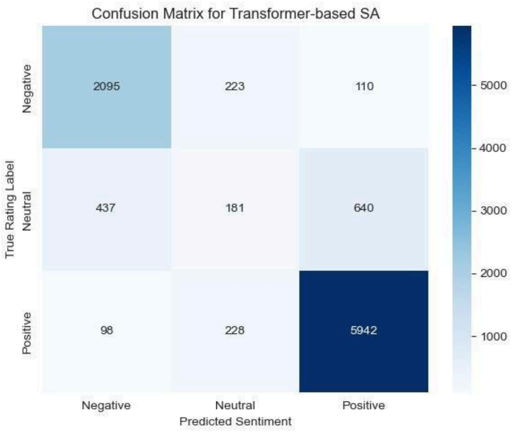
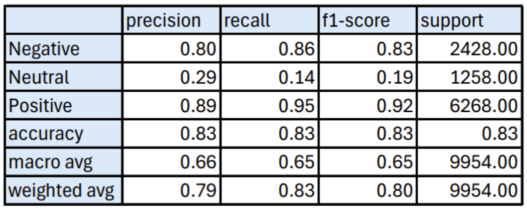
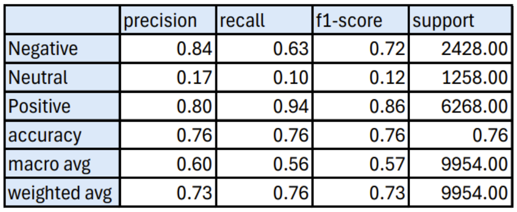


Table 1: Classification report – RoBERTa-base model.



	precision	recall	f1-score	support
Negative	0.80	0.86	0.83	2428.00
Neutral	0.29	0.14	0.19	1258.00
Positive	0.89	0.95	0.92	6268.00
accuracy	0.83	0.83	0.83	0.83
macro avg	0.66	0.65	0.65	9954.00
weighted avg	0.79	0.83	0.80	9954.00

Table 2: Classification report – VADER model.



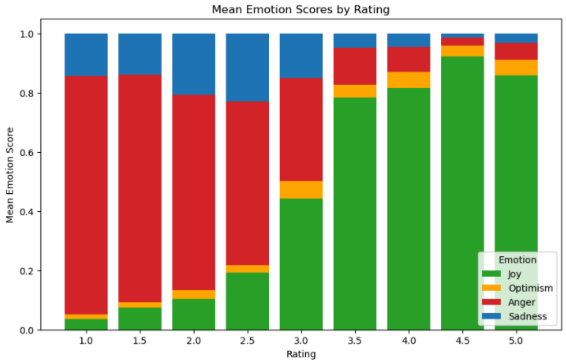
	precision	recall	f1-score	support
Negative	0.84	0.63	0.72	2428.00
Neutral	0.17	0.10	0.12	1258.00
Positive	0.80	0.94	0.86	6268.00
accuracy	0.76	0.76	0.76	0.76
macro avg	0.60	0.56	0.57	9954.00
weighted avg	0.73	0.76	0.73	9954.00

To address why the correlation is not very high, we calculated the Mean Squared Error (MSE) between VADER’s sentiment score and the normalized (0-1) rating. The highest MSE (3.94) was observed in the following review:

*"Fisherman's Wharf used to be our **fav joint** when we lived in Goa. Excellent food and bevs. Then we moved to Hyderabad, FMW followed. We were so happy, to have a Goan restaurant just 5 min drive from home. In the beginning, it was **great place to dine**. Then it started getting local. **Fishes started getting stale**, fewer varieties... **Not so great.** Now, after about 3 years, it's more a typical curry point. We ordered Mutton Xacuti and Kingfish Richado. Fish was so **firm and overcooked**, Richado overloaded with sweetness. Mutton was out of the world, with heaps of Curry leaves (Kadi patta/Karia paku), super high on heat... Reminded me of a Rayala curry in Dindi (Godavari distt). Meat **was cooked with extra affection** that it took me 5 min to chew it like bubblegum. I really thank myself for not ordering more. Fisherman's Wharf—such a disappointment."*

Despite the negative sentiment and a low **rating of 1,** both VADER **(0.986)** and the Transformer model **(0.869)** assigned relatively positive scores. This discrepancy arises from the review’s **mix of positive remarks** (e.g., initial excitement and fondness for the restaurant) and **negative experiences** (e.g., stale fish, fewer varieties, and disappointing meals), which can **mislead the algorithms.**

Figure 3: Emotions in reviews by ratings



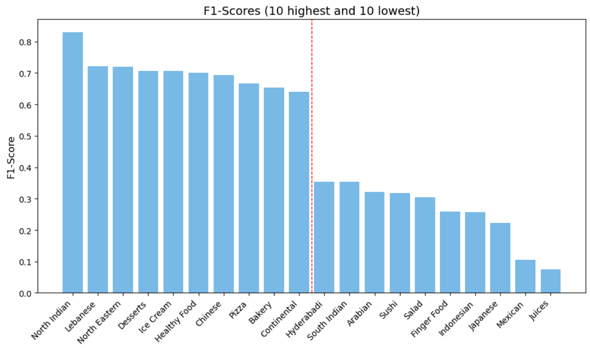
Using a transformer-based method for emotion scoring, the following results were obtained: **Joy: 0.38, Optimism: 0.05, Anger: 0.34, Sadness: 0.23.** The higher joy score compared to anger and sadness suggests that positive aspects of reviews outweighed negative comments. This mix of emotions likely affects the model's interpretation, making the sentiment score appear more positive (or negative) than expected as both aspects are considered.

Figure 3 illustrates the emotional associations with each rating score. Negative reviews are predominantly linked to anger, while positive reviews align more with joy.

6.2. Multilabel classification evaluation

Among all the combinations, **Logistic Regression** with the **MultiOutput** wrapper demonstrated the best performance, achieving a **weighted F1-score of 61%**. This score was weighted by the number of reviews for each cuisine in the dataset. Notably, the dataset contained **42 cuisines**, with some cuisines performing significantly better than others.

Figure 4: Multilabel classification scores



For the **Hyderabadi Council**, the most notable outcome of our model is its ability to accurately predict reviews related to **North Indian**, **Lebanese**, and **North Eastern** cuisines, achieving **F1-scores above 70%** (*as seen in the Figure X below*), with **North Indian cuisine reaching an impressive 83%**. However, the model performed poorly when predicting non-Indian cuisines such as **Japanese**, **Mexican**, or **Indonesian**. This underperformance can be attributed to the limited number of samples representing these cuisines in our dataset.

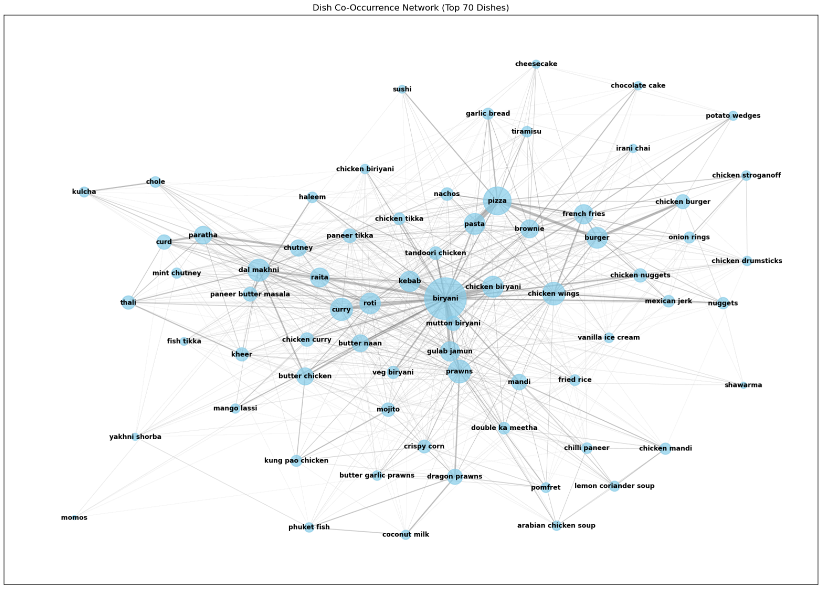
6.3. Meals Co-occurance and clustering evaluation

As we see on the graph (Figure 1), the biggest node, right in the center is "biryani" which also has the most connections to another dishes. On the other hand, it seems that for instance "momos" and "shawarma" nodes are the least connected dish to anything else and the most "abandoned" ones, meaning they may not belong to any group of meals in the restaurant. The biggest section of food include chicken. From the graph we can conclude some cuisines and food types based on the thickness of lines that are connecting the nodes and how the nodes are scattered. For instance:

- Thick connection between pasta and pizza and tiramisu may indicate the Italian cuisine.
- Biryani, mutton biryani, veg biryani and chicken biryani are strongly connected which makes sense since all those meals are the version of one dish. Moreover, connections with roti, curry, gulab jamun, tandoori chicken etc. suggest that this is Indian cuisine, also the most ordered type of food.

- Connections between phuket fish, dragon prawns and pomfret indicate sea food.
- Burger, pizza, French fries, chicken nuggets, chicken wings indicate fast-food.
- There are only two soups which suggest us that soups are not common order

Figure 5: Dishes Co-occurance network graph



Clustering

To evaluate the performance of HDBSCAN approaches, we used the **Silhouette score** to identify the best-performing metric. After fine-tuning the parameters, the **Pairwise distance** method achieved the highest Silhouette score (**0.32**). However, this approach failed to produce meaningful clusters, and the resulting plots were too dense for proper visualization. Despite numerous tuning attempts, the method was ultimately deemed unsuccessful.

Figure 6: Extracted dishes clusters

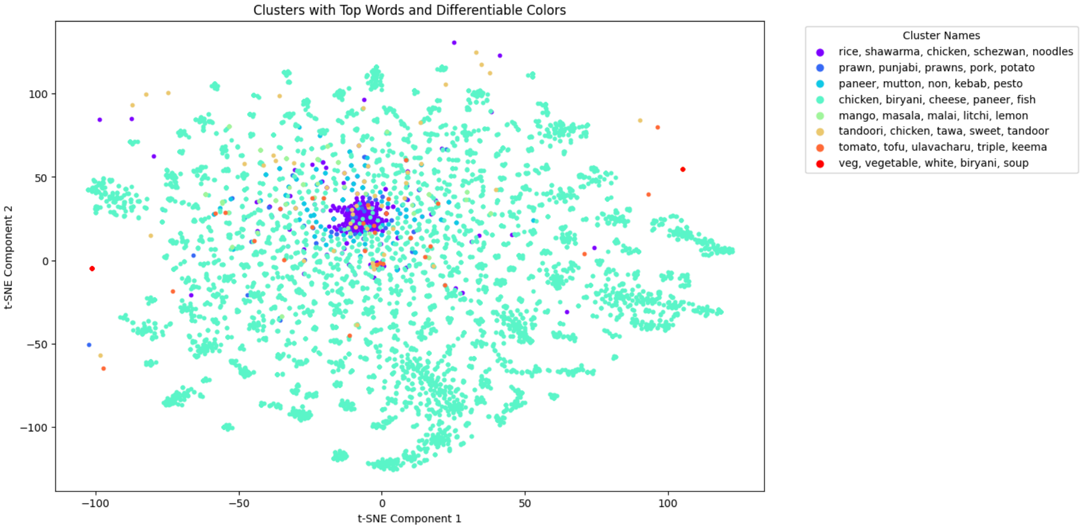
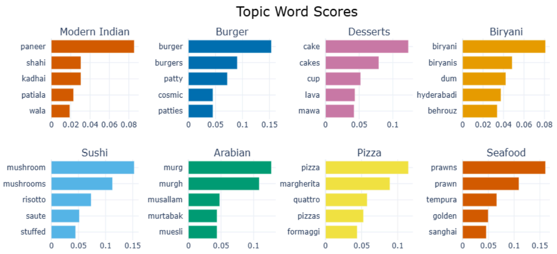


Figure 7: Cluster representations



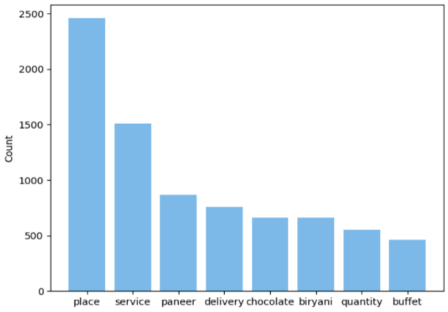
The **BERTopic method**, on the other hand, delivered more promising results compared to its predecessor. Although BERTopic achieved a lower Silhouette score (**0.10**), it significantly outperformed the first method in terms of **interpretability** and the differentiation between cuisines. The results provided clearer insights and more distinguishable clusters, as illustrated on Figure 7.

6.4. Topic Modelling evaluation

The LDA model was tested with 10, 12, 14, 16, 18, and 20 topics, with the **14-topic model** showing the highest coherence and a log_perplexity of **-936** on **10-fold cross-validation**. This suggests the model can assign topics to unseen documents effectively. However, coherence was around **40%**, indicating that the topics are weakly interpretable and consistent. The topics are overly generic (e.g., "good, food, service" in one topic), and top words repeat across topics. LDA's main drawback is its **inability to consider context**.

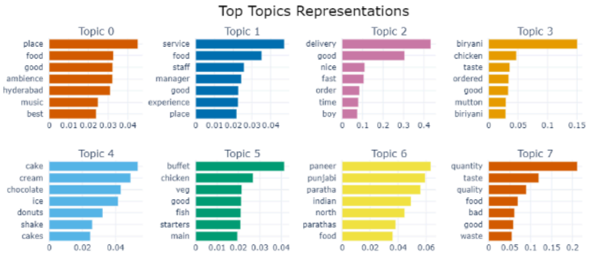
To evaluate BERTopic, human analysis based on visualizations, topics representations and review-topic assignment were applied.

Figure 8: Top topics extracted



Around **2600 reviews (27%)** focused on the restaurant's general **place, ambience, and location** in Hyderabad, making it the most discussed topic. **Service and delivery** were also prominent, with 1600 (17%) and 800 reviews respectively. The service topic captured feedback on the manager, staff, and overall experience, while delivery highlighted the efficiency of order handling. A significant number of reviews (1700) **emphasized Biryani or Indian cuisine,** reflecting their popularity and aligning with the restaurant's location.

Figure 9: Top topics word representations



Topics were diverse, with **top words holding less than 5% share**, especially for cuisine descriptions. However, in topics like delivery or quality/quantity, dominant words were evident. Reviews and assigned **topics matched well**, though merging into 20 generic categories led to broad terms like "cake," "cream," or "chocolate" representing all dessert reviews.

7. Conclusion

The aim of this report was to equip Hyderabad's city council with insights into the restaurant sector, enabling them to make informed and tailored decisions that effectively address the city's needs. Various approaches were employed to achieve this goal.

The sentiment analysis highlighted the strong connection between customer emotions and restaurant ratings, illustrating how dining experiences are reflected in the reviews customers leave.

Through multilabel classification, the range of cuisines available was explored, showing how customer reviews can reveal a restaurant's culinary identity. This information can assist businesses in customizing their offerings to meet market demands and help them differentiate themselves in a competitive environment.

The meal co-occurrence analysis offered a new perspective on popular dish pairings, giving restaurants the chance to create menus that align with customer preferences. Additionally, clustering and topic modeling identified key themes in reviews, such as ambiance, service, and food quality, indicating which aspects—beyond just food quality—are crucial in the dining industry.

Overall, this study provides the Hyderabad Council and local businesses with a fresh analytical perspective on the restaurant sector, offering tools to better understand customer needs, enhance dining experiences, and bolster the city's reputation as a vibrant culinary hub.

8. References

[1] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 3rd ed. Upper Saddle River, NJ: Pearson Prentice Hall, 2023.

[2] R. Egger and J. Yu, "A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts," 2022.

[3] M. Wankhade, A. Chandra, C.Kulkarni, *A survey on sentiment analysis methods, applications, and challenges*, 2022

[3] Dizex, *InstaFoodRoBERTa*, Hugging Face. [Online]. Available: https://huggingface.co/Dizex/InstaFoodRoBERTa-NER. [Accessed: Dec. 06, 2024].

[4] Zach CHASE Nicolas GENAIN Orren KARNIOL-TAMBOUR: Learning Multi-Label Topic Classification of News Articles: [link]

[5] Y. Liu and M. Ott, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," 2019.

[6] Cardiffnlp, twitter-roberta-base-sentiment-latest, Model for Sentiment Analysis. [Online]. Available: https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest. [Accessed: Dec. 06, 2024].