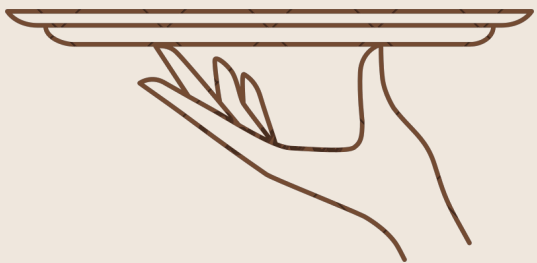
Nova Information Management School
BSc in Data Science

Solving the Hyderabadi Word Soup

TEXT MINING 2024-2025
PROJECT REPORT

Laura Matias (20221836)
Marta Aliende (20241453)
Marta Almendra (20221878)
Matilde Casimiro (20221940)
Teresa Simão (20221873)

TABLE OF CONTENTS



INTRODUCTION

This report focuses on analyzing restaurant reviews for the Hyderabad Tourism Board using text mining techniques. While scores and ratings are widely used to measure customer satisfaction, they often lose their meaning because of the tendency to cluster near the maximum value. On the other hand, text reviews give more detailed feedback about customer experiences, which can be used to gain useful insights.
The main goals of this project were:

- **Cuisine Classification:** Classifying restaurants by their cuisine type using the content of their reviews (multilabel classification).
- **Score Prediction:** Predicting a restaurant's Zomato score based on the sentiment of its reviews (sentiment analysis).

Additionally, we explored dishes that are frequently mentioned together, understanding their relationship to cuisine types, and discovering main themes in the reviews, such as service, ambiance, or food quality, to better explain customer feedback.
We used the CRISP-DM framework, a standard process for data mining projects, which includes six phases: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment. This step-by-step approach helped organize our analysis and ensure we met the project goals.

LITERATURE REVIEW

Several studies have explored advanced methods for analysing customer reviews across various domains, including tourism and food services, using techniques such as multilabel classification, sentiment analysis, topic modelling, and clustering. By combining ideas from the following papers, we wanted to build a solid basis for this field of study.
Laily et al. [1] used multilabel classification and Latent Dirichlet Allocation (LDA) to evaluate tourism reviews in Indonesia. This approach, which allows the classification of multiple topics or categories per review, inspired our use of Classifier Chains for multilabel classification of food reviews by cuisine type.
Islam et al. [2] employed sentiment analysis on food reviews using machine learning techniques to predict restaurant ratings. This work influenced our use of VADER and TextBlob, two sentiment analysis tools that allowed us to capture the overall polarity of the reviews, to predict restaurant ratings based on review sentiment and providing useful insights into customer preferences.
Eletter et al. [3] explored topic modelling with BERTopic for extracting themes from customer reviews. We applied a similar approach with BERTopic to extract topics related to cuisine, dishes, and customer experiences, which helped identify trends in the data and reveal common themes in restaurant reviews.
Jin and Bai [4] proposed a text clustering algorithm based on semantic word co-occurrence, which facilitated the grouping of similar documents. In our project, we applied K-Means clustering to group food reviews by cuisine, and understand the co-occurrence patterns in customer feedback, mainly the relationships between different dishes and cuisines,
Full references to these studies can be found at the end of this report.

DATA UNDERSTANDING

Data understanding involves collecting, exploring, and assessing the quality of data to ensure alignment with the project's objectives. This step is critical for identifying potential issues such as missing values, inconsistencies, and duplicates, which guide the subsequent data preparation and modelling strategies. For the purpose of this project, two datasets were provided to us.

Restaurants Dataset

This dataset contains 105 observations, each regarding a restaurant in Hyderabad, and six columns. We found that there were two variables with missing values: "Timings," which had only one missing value, and "Collections," which was missing approximately 50% of the observations. We also noticed some characteristics in our data that would require further preprocessing, mainly that the "Collections" and "Cuisine" columns contained nouns separated by commas, the "Cost" column included commas to denote numbers in the thousands, and the "Timings" column had inconsistent formats across observations. It is also important to note that the “Cuisine” labels were highly imbalanced (Annexes Fig. 1)
In conclusion, we handled the missing values, applied parsing to the "Collections" and "Cuisine" columns, converted "Cost" to numeric values, and standardized the "Timings" format.

Reviews Dataset

The Reviews dataset consists of 10,000 observations and seven columns. We found that approximately 0.45% of rows had missing values in them, and that "Restaurant" and "Pictures" were the only columns without any null entry. Regarding the need for any further treatment to the data, we found one non-standard entry (“like”) in the "Rating" column, which is otherwise a numeric feature. We also found that the column “Metadata” was comprised of observations with the format “[x] Reviews, [y] Followers”, so this information could be separated.
In conclusion, the Reviews dataset required treatment for the missing values, correcting the non-standard "Rating" entry and separating the “Metadata” column into two columns: “Number_Reviews” and “Followers”. After performing this changes, we made some visualizations to help us understand in more depth the data we were dealing with, and found that the it was in accordance with our initial preconception that ratings tend cluster near the maximum value (Fig. A).

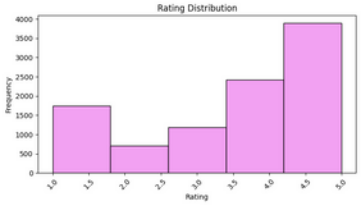


Fig. A - Ratings Distribution

DATA PREPARATION

Multilabel Classification [1]

The next step was to prepare the reviews data for Multilabel Classification. This included an initial data cleaning phase to remove noise, extraneous information, and formatting anomalies that might interfere with our research. Punctuation, URLs, special characters (emojis, @ , # , etc), HTML tags, and unnecessary whitespaces were all removed throughout the cleaning process. Moreover, sections of reviews with numerical information, such as 3.5/5 or 3-4 were converted to a more sensible representation ("three point five out of five" and “three out of four”). Repeated letters and isolated characters, excluding punctuation, were also adjusted to help remove spam words. A pipeline was created to combine the basic cleaning tasks with further preprocessing stages. These included case normalization (converting text to lowercase), tokenization, stopword removal (excluding negation terms), lemmatization achieved with spacy. To execute our pre-processing pipeline, we made a copy of the dataset called clean_reviews. After running the pipeline, it included new columns for the cleaned text and tokenized words. Additional changes were made, such as incorporating the cuisine types of each restaurant in the clean_reviews dataset and determining the lengths and sentence counts of each review.
Finally, we discovered an inconsistency between the restaurants and reviews datasets: two restaurants appeared in the restaurant dataset but had no corresponding entries in the review dataset's restaurant column. These two restaurants included cuisines types only mentioned once, Mithai and Malaysian. Because their information was not available, the list of cuisine types was transformed to 42 for the remainder of the investigation.

The next phase focused on preparing a numerical representation of the data for modeling. Before experimenting with different vectorization methods, we investigated if it was possible to remove irrelevant tokens based on their frequency—eliminating infrequent tokens (likely typos) and excessively frequent ones that would not help distinguish between cuisine types. We examined the most and least common words to determine their relevance and ultimately removed a proportion of uncommon tokens as well as some of the most common ones. However, some of the high-frequency words that were meaningful to different cuisine styles, were maintained despite being among the most common ones, as they proved to be relevant during the modeling stage. Finally, different vectorizers like term frequency-inverse document frequency (TFIDF) and bag of words (BOW), were employed on the text without the irrelevant words.

Additional visualizations were used to gain a greater understanding of the transformed data, allowing us to confirm the reasoning behind our earlier decisions and final conclusions needed before modelling. We investigated the review frequency by cuisine types, determining that we had imbalanced data, with large differences in size between the most common cuisines (North Indian and Chinese) and the least common (North Eastern and Pizza). Using donut charts, tree maps, and word clouds, we proved that our previous feature selection was successful and useful. We compared the word clouds by cuisine type before and after removing unnecessary tokens, and the results were substantially clearer and more distinguishable. As a result, the transformed vocabulary appeared to be more useful in classifying cuisine types based on the content of their reviews. Still, we discovered that cuisines like Bakery, Cafe and HealthyFood are probably very easy to distinguish, while others like Chinese and Andhra are more difficult due to their similar vocabulary in the reviews (even after it being reduced).

Fig.B - Easy/Challenging to distinguish Cusine WordClouds

Sentiment Analysis [2]

In this section, a slightly adjusted data cleaning and preprocessing was conducted, with the same logic regarding sections with numerical information, whitespace, repeated characters, and unwanted patterns (HTML tags, website links, social media tags email adresses) corrections. However, unlike earlier preprocessing steps, lemmatization and tokenization were not performed, and the original text case, emojis, punctuation, and stopwords were maintained because they could provide useful information for the sentiment analysis task.

Co-occurrence Analysis [4]

In this section of our project we searched for dishes mentioned together in reviews, and if they formed clusters that could possibly identify a cuisine type. This analysis was conducted using the previously prepared data for multilabel classification (cleaned reviews with feature selection - no exceptionally common and uncommon words).

Topic Modelling [3]

In this part of the analysis, we wanted to investigate if reviews could be classified according to emergent topics and their meaning, and if relevant insights could be extracted from the them. To do so, we used the initial resulting tokens from the multilabel pipeline (without any feature selection), as to not loose potential information helpful in understanding review topics, and the same preprocessed text but without tokenization for the final model tested (BerTopic). Additionally, reviews with less than two words were excluded, to ensure the model received meaningful input.

MODELLING

Multilabel Classification [1]

As referred before, the goal in this part of the project was to accurately classify a restaurant's cuisine type or types, using their reviews as input. Since each review could be associated with multiple cuisine types, we required a classification and encoding approach that could handle multiple labels for each input. For this, we used a MultiLabelBinarizer to encode the labels into a binary format and converted the reviews into numerical representations using Bag-of-Words (BoW) and TF-IDF.

We trained multiple models using different base classifiers, including Decision Tree, Logistic Regression, and Naive Bayes, and applied two multi-label classification strategies: One-vs-Rest (OvR) and Classifier Chain (CC). The One-vs-Rest strategy treats each label as an independent binary classification problem, making it a more straightforward approach, but unable to model label dependencies. The Classifier Chain, on the other hand, is a sequential approach capable of modelling label dependencies, since each label's prediction becomes an input feature for the next.

Then, to optimize the model's performance, we performed hyperparameter tuning for the different classifiers through a Grid Search, leveraging a predefined split of the data into training, validation, and test sets to ensure a robust evaluation on the validation set.

To automate the process of transforming text into numerical features and fitting the classifiers in a streamlined manner, we created pipelines for each model (Decision Tree, Logistic Regression, and Naive Bayes) with both the One vs Rest and Classifier Chain approaches. For each pipeline, we defined a parameter grid that included various hyperparameters for the vectorizer and classifier. This exhaustive search focused on optimizing the weighted F1-score, which is particularly relevant in multi-label classification tasks, and allowed us to identify the best combination of parameters for each model.

To further assess the performance and measure the effectiveness in predicting multiple labels of each hyper tuned model, we calculated the key metrics Accuracy, weighted Precision, Recall, and F1-Score, which helped us identify the most suitable approaches for this task.

Sentiment Analysis [2]

The objective in this section of the project was to predict a restaurant's Zomato score accurately by using the polarity of its reviews as input. We used both VADER and TextBlob sentiment analysis tools to assess the polarity of restaurant reviews and evaluate their ability to predict restaurant ratings.

When reviews consisted of multiple sentences, we also calculated the average of the scores of each sentence to obtain a more detailed understanding of the overall polarity of the review. We then analysed the correlation between VADER sentiment scores and the Zomato ratings, and found a reasonably strong, although not perfect relationship. We normalized both the restaurant ratings and the polarity scores to see if the correlation improved, but the difference was minimal. To assess the accuracy of VADER's sentiment scores in predicting restaurant ratings, we computed performance metrics, including RMSE (Root Mean Squared Error) and MAPE (Mean Absolute Percentage Error).

After analysing the VADER sentiment scores, we extended our approach by incorporating TextBlob, another popular sentiment analysis tool. Following the same approach as before, we calculated the TextBlob polarity score based on the overall sentiment of each review, and for reviews containing multiple sentences, we also computed the average of the polarity of each sentence. By analysing the correlation between the polarity scores and the ratings, we found that is was almost identical to the VADER correlation, which implies implies that both tools perform similarly in modelling the relationship between review sentiment and restaurant ratings. Finally, we obtained the RMSE and MAPE to evaluate the accuracy of TextBlob's polarity in predicting the restaurant ratings and compared the results from both tools to assess which provided the most reliable sentiment scores for predicting restaurant ratings.

Additionally, we performed regression analysis to predict restaurant ratings based on sentiment scores from VADER and TextBlob. We applied several regression models, including Linear Regression, Random Forest Regressor, Gradient Boosting Regressor, and Support Vector Machine (SVM). For each model, we split the dataset into training and testing sets (80-20 split), trained the models on the training set, and then evaluated the models' performance on the test set using two metrics: Mean Squared Error (MSE) and R^2 score, which indicates the proportion of variance in the target variable explained by the model. We compared the results to assess which model performed best in predicting the restaurant ratings based on review sentiment.

Co-occurrence Analysis [4]

In order to find dishes mentioned together in reviews, two pre-trained name entity recogniction (NER) models based on BERT were employed, alongside functions to extract the dishes from each review. The first model was FoodBaseBERT, which has been trained to recognize one entity: food, and that was fine-tuned on the FoodBase NER dataset. This model didn't show great results, and so we moved onto RoBERTa (Robustly Optimized BERT Pretraining Approach). It consists of a transformer-based language model developed as an improvement over BERT, and uses pretraining on big amounts of data to understand language context and semantics better. In fact, it showed better results, which made it our chosen model for the remainder of the analysis, but still struggled to recognize certain common dishes. To address this, we developed a function to adjust them into their correct textual form. Finally, we filtered out reviews where a dish wasn't mentioned.
The next step was to examine the overall and cuisine type-specific dishes co-occurrences by calculating their co-occurrence matrices and visualizing the results. Following that, we implemented clustering techniques to identify possible clusters of dishes mentioned together. To account for discrepancies in the dataset size, we padded the dishes' co-occurrence matrix to match the total number of reviews, ensuring it aligned with the dataset dimensions. This approach preserved the statistical properties of the original matrix while allowing it to integrate seamlessly with the rest of the data. We used the resulting padded co-occurrence matrix with a variety of clustering methods. Starting with K-means, we examined inertia plots to determine the optimal number of clusters (k), having the maximum value equal to the number of cuisines, and identifying the "elbow point" as the ideal k. In addition, we incorporated K-means and Latent Semantic Analysis (LSA), to transform the data into a lower-dimensional space while retaining the most meaningful features, and finding the best k by the same logic.
Furthermore, we experimented with HDBSCAN and OPTICS, once again using the dishes co-occurrence matrix, and adjusting the number of minimum samples for both models to select the values that yielded the best results. Finally, we evaluated and plotted the results of the approaches, using a 3D visualization that showed the resulting clusters and identified them with a label containing the top 5 most frequent dishes in them, which allowed for an easier understanding of the results.

Topic Modelling [3]

The goal of topic modelling in the context of this project is to identify underlying patterns within the reviews. We applied Latent Dirichlet Allocation (LDA) and Latent Semantic Indexing (LSI) by leveraging Gensim implementations, as well as BERTopic to assess if the reviews could be classified according to emergent topics, and what relevant insight could be extracted from these. We experimented LDA and LSI models with both a bag-of-words and tf-idf representation of the data, ultimately producing several model variations to be compared. To evaluate and guide the refining of our models, we calculated metrics such as perplexity for LDA, which helped us measure how well the model fit the data, and coherence for both LDA and LSI variations, which provided insights into how interpretable the results were. Since the coherence values showed poor results, we moved onto a different model.

Finally, after implementing BerTopic, to better understand the results obtained, we explored the topic information and visualized the relationships between topics as well as the distribution of documents across these topics. These visualizations offered an intuitive way to assess the model's output and gain a clearer perspective on how the topics were structured within the data.

EVALUATION

Multilabel Classification [1]

Given the imbalance in our data, the metrics used in our grid searches included accuracy, weighted F1 score, weighted precision, and weighted recall. We focused on the three latter ones, examinating the parameters that provided the best results on the grid searches, and their metrics performances on the training and test sets. We concluded the Naive Bayes with the Classifier Chain (with parameters alpha = 0.1, and order = random) model outperformed the rest, showing the highest weighted f1 score (61%), and a good recall (64%) and precision (59%) scores. This model also had a smaller likelihood of overfitting, shown by more comparable performance values in the training and test sets. However, there was still a considerable difference [Annexes Fig. 3].

In fact, the model demonstrates a clear ability to differentiate certain cuisine types effectively while struggling with others. For example, it performs very well for cuisines like North Indian (F1-Score: 71%) likely due to having a higher representation on the dataset compared to other cuisines, and North Eastern (F1-score: 77%) possibly due to having words easier to distinguish, as observed in its word cloud [Annexes Fig. 2]. On the other hand, cuisines like Mexican (12%) an Indonisian (20%) perform very poorly, likely because they are very underrepresented.

Moreover, as previously assumed by their wordclouds, cuisine types like Cafe (F1-score: 58%) and Bakery (F1-score: 69%) perform very well and cuisines such as Andhra perform poorly (F1-score: 39%). However, opposite to our previous assumptions, the model performs very well with the Chinese type (F1-score: 71%), likely to its higher data representiveness, and, on the other hand doesn't perform as well as expected for the Healthy Food cuisine type (F1-score: 45%).

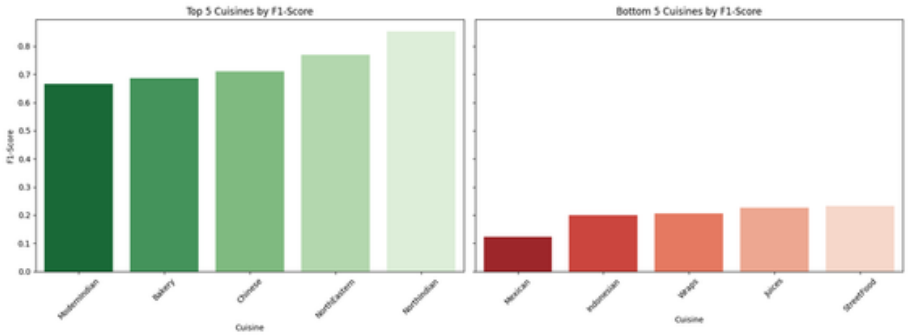


Fig. C - Top and Worst Cuisine Types F1 Scores

Sentiment Analysis [2]

While the RMSE of 0.284 obtained by the VADER tool indicates reasonable predictive performance, the error is not negligible, meaning the VADER sentiment scores capture only part of the information that determines restaurant ratings.

Since the ratings were normalized to a range between 0 and 1, this RMSE values indicates that, on average, the predicted ratings differ from the actual ratings by approximately 1.42 points on the original 0–5 scale. However, the MAPE of 0.152 is a relatively low percentage error, showing that the predictions are fairly accurate in terms of proportional error.

The TextBlob model obtained a RMSE of 0.278, which means that, on average, the predicted ratings deviate from the actual ratings by 0.278 on the normalized scale, equivalent to 1.39 points on the original 0–5 scale.

As for the MAPE metric, the value obtained was of 0.159, meaning the predicted ratings are, on average, 15.9% off from the actual ratings, so TextBlob also performs well in capturing the relationship between sentiment and ratings, achieving a relatively low prediction errors.
While TextBlob and Vader perform similarly, TextBlob's percentage error is slightly higher than VADER's, and its RMSE is only marginally better, suggesting that TextBlob provides slightly better predictions, but performs slightly worse in relative error terms. However, the difference is minimal and both tools capture the sentiment-ratings relationship to a similar extent. Together, the metrics obtained indicate that sentiment scores provide a reasonably good, but not perfect, prediction of restaurant ratings, indicating that sentiment alone cannot fully explain the variance in ratings in this case. Additionaly, visualizations were deployed to understand the distributions of the results better. [Annexes Fig. 4 and Fig.5]
As for the regression models, we found that Gradient Boosting was the most effective model for predicting restaurant ratings based on review sentiment, as it obtained both the lowest MSE (0.0540) and the highest R^2 score (0.6087). Linear Regression and Support Vector Machine (SVM) also perform reasonably well, with R^2 scores of 0.5896 and 0.5838, respectively. SVM, however, had a slightly higher MSE (0.0575) than Linear Regression (0.0567). [Annexes Fig. 6]

Co-occurrence Analysis [4]

A visual inspection of the overall dishes co-occurences was developed, showing that dishes such as 'chicken', 'biryani', 'soup', and others [Annexes Fig. 7], are frequently mentioned together. Also, by studying co-occurrences by cuisine type, we discovered that cuisines such as Pizza, Bakery, Burger, and Dessert are easily distinguished thanks to their unique dishes mentioned together, yet cuisines like Mexican and Street Food are more challenging to recognize. This is consistent with our prior findings from the cusine word clouds. We also noticed that certain dish combinations, such as chicken biryani with rice, veg, and curry, are common in multiple cuisines (e.g. Biryani.) Significant connections were discovered in Chinese dishes, connecting items such as soup, chicken, veg, fish, and biryani. Similarly, North Indian cuisine had dense connections between dishes such as chicken paneer, biryani, and soup.
Different methods were used to evaluate the different clustering techniques, and all models were visually evaluated, with an emphasis on cluster labels, checking also their position and density. With regards to K-Means, the models were compared in terms of inertia, silhouette, and Calinski-Harabasz scores. The goal was to find a model that balanced them in a way that minimized inertia (ensure points were closer to their cluster centers), and maximized silhouette (ensure well-separated and cohesive clusters) and Calinski-Harabasz scores (have a good ratio of between-cluster dispersion to within-cluster dispersion). The regular and LSA integrated models were compared, both showing poor results in terms of inertia, yet the regular K-Means with 6 clusters, with an inertia of 18 209 239, silhouette of 0.84 and Calinski-Harabasz score of 4541, had overall slightly better results.
Since OPTICS and HDBSCAN are density-based techniques, only the silhouette score was calculated for each model, both showing very poor results in comparison to K-Means. Nonetheless, HDSBCAN performed better with a score of -0.15, over the competing OPTICS silhouette of -0.30.

Topic Modelling [3]

Since model coherences for LSA and LSI were not great, we focused on our BerTopic model. The BerTopic results successfully demonstrated that reviews could be grouped according to emergent topics, identifying 75 of them. By visually examining the 2D reduced embeddings of all documents [Annexes Fig. 8] and the intertopic distance map, we decided that the topics could be grouped into a smaller number (around 10-15 based on the intertopic map). Yet, after a more through examination of each topic found, we decided to provide a summary of the topics sub-divided into two main categories:

Common Orders: Topics as "biryani_order_chicken", "cake_velvet_cupcake", 'chinese_noodle_rice', and much more, revealed common orders, allowing restaurants to identify which dishes are frequently discussed and potentially associated to client satisfaction or complaints.
Service Feedback: Topics like "excellent_service_bahadur" and "ambience_really_nice" provide great insights for restaurants to manage their service and staff based on overall customer opinions. Moreover, topics such as "rice_properly_not" and "salt_salty_balance_extremely_no" point toward specific issues, helping restaurants address recurrent problems they might be facing.

Fig. D - Intertopic Distance Mao obtained with BERTopic

CONCLUSIONS

This report successfully addressed the key information requirements by analyzing sentiment, cuisine classification, co-occurrence, and topic modeling from restaurant reviews. However, several challenges arose during the project, the most significant being the imbalanced dataset, which impacted the performance of most models, particularly in multiclass and multilabel classification. Despite attempts to mitigate this through model selection and parameter tuning, certain cuisines remained difficult to classify effectively. Additionally, while the sentiment analysis tools employed provided useful insights, they could not fully capture the complexity of restaurant ratings, highlighting the need for more sophisticated models.
Another issue we faced was concerning clustering analysis. While our co-occurence analysis produced great results, all clustering methods failed to meet our expectations, with K-Means performing averagely and other techniques such as OPTICS and HDBSCAN struggling due to low silhouette scores, implying that density-based clustering may not be the best fit for this type of data without further adjustments.
To better meet the information requirements, more focus could have been placed on addressing class imbalance, either by using advanced oversampling techniques or experimenting with ensemble methods. Overall, while the methods used answered the project's core questions, improvements in data preprocessing and model selection could lead to more accurate and insightful results in future works.

REFERENCES

[1] I. L. Laily, I. Budi, A. B. Santoso and P. K. Putra, "Mining Indonesia Tourism's Reviews to Evaluate the Services Through Multilabel Classification and LDA," 2020 International Conference on Electrical Engineering and Informatics (ICELTICs), Aceh, Indonesia, 2020, pp. 1-7, doi: 10.1109/ICELTICs50595.2020.9315392.

[2] N. Islam, N. Akter and A. Sattar, "Sentiment Analysis on Food Review using Machine Learning Approach," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, 2021, pp. 157-164, doi: 10.1109/ICAIS50930.2021.9395874.

[3] S. F. Eletter, K. I. AlQeisi and G. A. Elrefae, "The Use of Topic Modeling in Mining Customers' Reviews," 2021 22nd International Arab Conference on Information Technology (ACIT), Muscat, Oman, 2021, pp. 1-4, doi: 10.1109/ACIT53391.2021.9677049.

[4] C. -X. Jin and Q. -C. Bai, "Text Clustering Algorithm Based on the Graph Structures of Semantic Word Co-occurrence," 2016 International Conference on Information System and Artificial Intelligence (ISAI), Hong Kong, China, 2016, pp. 497-502, doi: 10.1109/ISAI.2016.0112.

ANNEXES

MultiLabel Classification

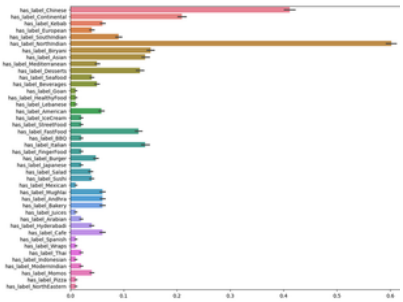


Fig. 1 - Cuisine types Distributions

Fig. 2 - North Eastern type Word Cloud

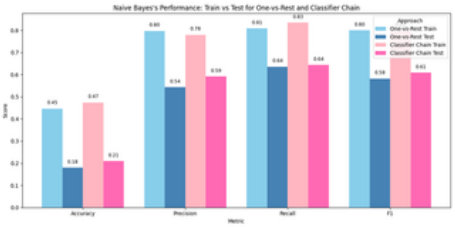


Fig. 3 - Naive Bayes Metrics

Sentiment Analysis

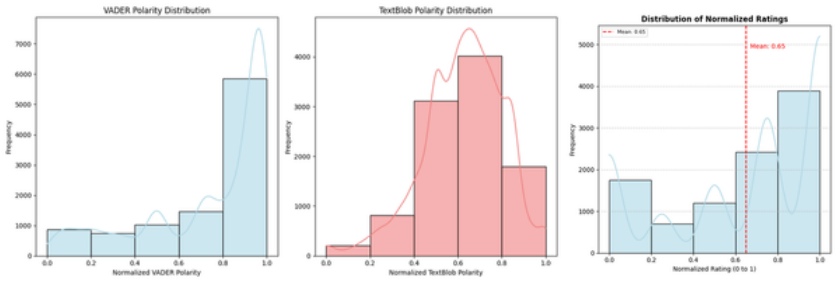


Fig. 4 - Ratings and Polarities Distributions

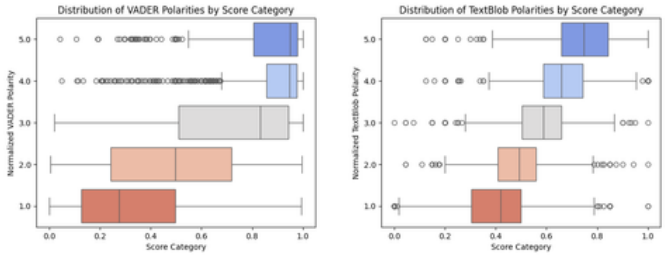


Fig. 5 - Polarities by Score Category

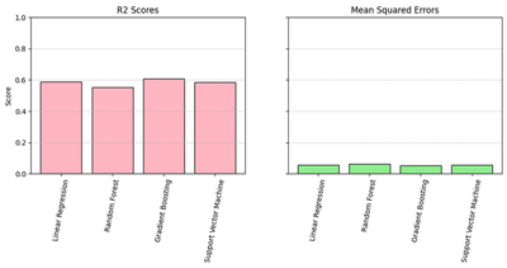


Fig. 6 - Regression Models Metrics

Co-occurence and Topic Modelling Analysis

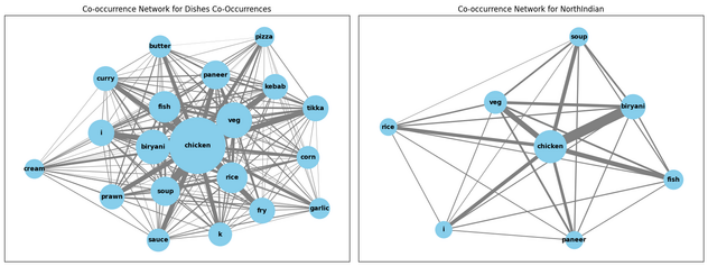


Fig. 7 - Co-occurrence Networks for dishes and NorthIndian cuisine type

Fig. 8 - BERTopic results