Nova Information Management School

BSc in Data Science

Text Mining 2024-2025

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

Project Report

Group 16

Miguel Matos – 20221925 Nuno Leando – 20221861 Patrícia Bezerra – 20221907
Rita Silva – 20221920 Vasco Capão – 20221906

Contents

1 Introduction

As digital content continues to increase at an unprecedented rate, text mining has emerged as an essential tool for data-driven decision-making, allowing organisations to extract valuable insights from textual data in several sectors.

The present report explores the use of text mining algorithms on two datasets based on restaurants and their reviews, focusing on sentiment analysis, topic modelling, clustering and multilabel classification. The analysis uses Natural Language Processing (NLP) and machine learning to detect patterns in consumer feedback, identify developing topics, and predict features such as cuisine types based on review content.

This project uses a comprehensive approach, beginning with data understanding and pre-processing, which are fundamental to ensure the quality and consistency of textual data. Tokenization, stemming, and stop word removal are all important techniques for preparing data for further analysis. Sentiment analysis, which uses tools like VADER and TextBlob, quantifies consumer sentiment and provides insights into satisfaction levels and emotional patterns. Similarly, topic modelling using algorithms, such as Latent Dirichlet Allocation (LDA), extracts theme structures that indicate important features of users' experiences.

The results of this report demonstrate the importance of text mining in extracting actionable insights from unstructured data, providing significant assistance for restaurants looking to improve customer satisfaction and operational efficiency. Text mining transforms raw text into an important tool by connecting qualitative narratives to quantitative data.

2 Literature Review

Considering that this report was mainly based on the content given during theoretical and practical classes, we only did a brief search to clarify some topics.

Data Processing

Several important steps are required for data pre-processing in text mining projects, such as those stated by Moleenar et al.[1] and Zhang et al.[2], especially in sentiment analysis tasks. The process usually starts with removing stop words using tools like the Natural Language Toolkit (NLTK), which helps keep only the most meaningful words in the document.

Another essential stage is the removal of punctuation. Eliminating punctuation simplifies the content, prevents inconsistencies, and improves the subsequent tokenization process.

Once tokenization is complete, Zhang et al.[2] highlight the importance of filtering tokens depending on their length. This includes eliminating overly short or long terms that could contribute noise to the analysis. Following this, stemming is done, that is lowering the words to their most basic or root form, allowing related word variants to be grouped. This approach ensures a more consistent representation of text data, which is crucial for increasing the performance of text mining algorithms.

Sentiment Analysis

As mentioned in the work of Moleenar et al.[1], Valence Aware Dictionary and Sentiment Reasoner (VADER) is a tool used to detect sentiment analysis in social media text. It is especially important when dealing with consumer reviews, as it assigns a polarity score to words, considering grammatical structures, degree modifiers and emojis. Then the scores are combined into a compound score, which is normalized and classified as very negative, negative, neutral, positive, or very positive.

Topic Modelling

Topic modelling is another key task in text-mining projects. As Moleenar et al.[1] discuss, Latent Dirichlet Allocation (LDA) is a widely used machine learning algorithm for this purpose. It is an algorithm that determines the optimal number of topics and employs coherence scores to assess the semantic similarity of the top words within each topic.

Additionally, Zhang et al.[2] describe LDA as a generative probabilistic model that calculates the likelihood of a document belonging to each topic. This algorithm is particularly effective for extracting keywords from short documents and identifying frequent word groups in a corpus where similar documents share common terms. Furthermore, their work highlights how adjusting parameters, such as the number of topics, helped eliminate irrelevant

French reviews in the output. The review also mentions Non-Negative Matrix Factorization (NMF), another algorithm for topic modelling. Unlike LDA, NMF is deterministic and assigns relative weights to each topic for a document.

Other Important Findings Using Text Mining

In their research, Li et al.[3] concluded that review length positively impacts review helpfulness, as longer reviews provide more detailed product information, which aids consumers in making judgments and decisions. However, they also found that the negative effect of positive emotional content on review helpfulness was weaker in longer reviews, while the positive effect of negative emotional content was stronger in reviews with more words.

3 Data Understanding

When performing a text mining project, it is essential to begin by understanding the dataset. In this case, the group first explored the restaurant dataset, focusing on the 'Cuisines' column as the target variable for the multilabel classification task. The analysis revealed missing values in the 'Collections' and 'Timing' columns, and identified 92 unique cuisine types in the restaurants, with 'North Indian, Chinese' being the most common. A visualization of the cost variable showed an average price of 861.43, though two restaurants had prices exceeding 2000, as shown in Figure 1.

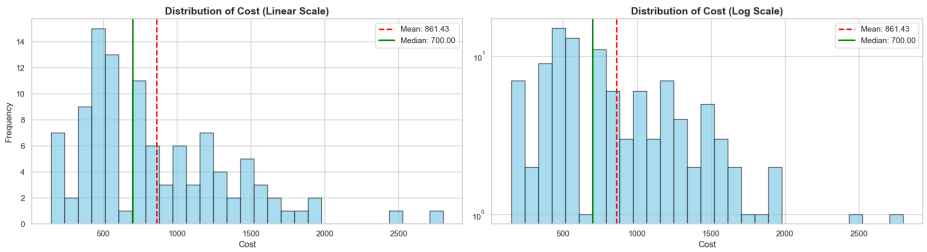

Figure 1: *Distribution of Cost*

Regarding the reviews dataset, the 'Rating' column was selected as the target variable for the sentiment analysis task. Observations with the value "Like" in the 'Rating' column were removed, as this variable only accepts values between 1 and 5. Two new variables, 'Nº_Reviews' and 'Nº_Followers', were derived from the 'Metadata' column, which was subsequently dropped. The analysis revealed missing values in the columns 'Reviewer', 'Review', 'Rating', 'Time', 'Nº_Reviews', and 'Nº_Followers'. Additionally, there are 7446 users, and the restaurant with the most reviews is 'Beyond Flavours'.

Several visualizations were created to analyze the data. It was observed that the most frequent ratings are 4.0 and 5.0, with 1.0 also being common, as shown in Figure 2. The mean rating is approximately 3.6, while the median is 4.0, suggesting users tend to express either strong dissatisfaction (1.0) or high satisfaction (4.0-5.0).

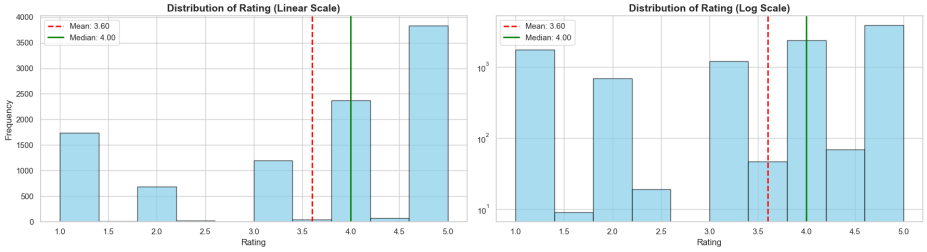

Figure 2: *Distribution of Ratings*

Additionally, most users contribute only a few reviews, with at least half of the users writing four or fewer. However, a small group has written a disproportionately high number of reviews, with some exceeding hundreds or even 1000.

After completing the visualizations, we proceeded with dataset integration by merging the restaurant and review datasets into a single one. The dataset information was then reviewed, and entries with missing values in the 'Review' column were removed to avoid introducing bias into future analyses. A main data frame was created, containing only the columns 'Name', 'Review', 'Rating' and 'Cuisines'.

Subsequently, the main pre-processing pipeline was applied to the 'Review' column, and the processed results were stored in a new variable named 'preproc_reviews'.

Concerning the word clouds, the corpus that will be used in the following steps was first defined. It was also created another three corpus that included only the positive ratings (all ratings above 3.0), the neutral ratings (rattings with a value of 3.0) and finally, the negative ratings (the remaining ones).

Word Count

To better understand the sentiment patterns in customer reviews, word count analysis was conducted using both the Bag of Words (BoW) and TF-IDF models. By examining the frequency and significance of uni-grams, bi-grams, and tri-grams, some key insights were derived about customer reviews.

Positive reviews commonly feature words like "good", "food", and "service", with phrases such as "good food" and "must visit" emphasizing customer satisfaction. Highly enthusiastic sentiments are reflected in tri-grams such as "must-visit place" and "best food ever". Negative reviews, on the other hand, are dominated by words like "worst", "bad", and "cold", with frequent phrases like "worst experience" and "food cold" highlighting dissatisfaction. Specific complaints appear in tri-grams such as "worst service ever" and "ordered chicken wings".

Neutral reviews demonstrate a more balanced language, with terms like "good" and "food" recurring but with less intensity. Bi-grams and tri-grams such as "food good" and "visited place" suggest mixed or moderate experiences. Figure 3 illustrates the uni-grams, bi-grams, and tri-grams by rating using BoW vectorization and demonstrates all the insights mentioned.

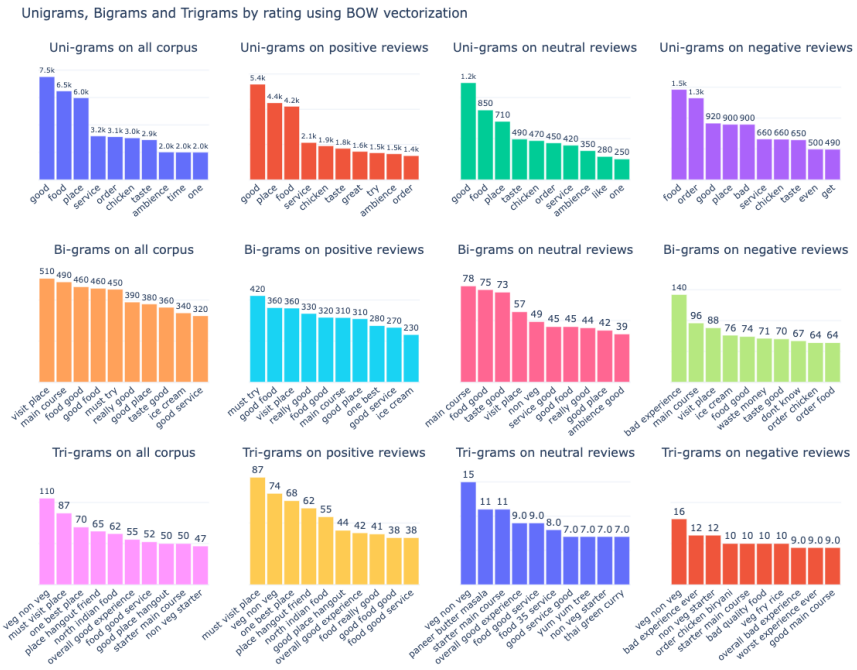

Figure 3: *Unigrams, Bigrams, Trigrams by Rating using BoW Vectorization*

Word Clouds

To have a deeper understanding of sentiment in customer reviews, it was also performed word cloud analysis using BoW and TF-IDF models. Regarding customer experience, several important insights were obtained by investigating the frequency and relevance of the most common words.

In the BoW Word Cloud, keywords such as “food”, “place”, “good”, and “service” emerged as central themes in the reviews, emphasizing customer focus on meal quality and overall experience. The use of colour in the visualisation, with words shaded based on their corresponding review ratings, provided further insight into sentiment patterns. Words in vivid green, such as “great”, “chicken”, and ”amazing”, indicate positive sentiments and are associated with higher ratings while the absence of dominant red words suggests an overall positive trend in the reviews.

In the TF-IDF Word Cloud, as shown in Figure 4, words like “good”, “place”, “food”, and “service” once more appear significantly, highlighting their high importance to customer feedback. Similarly, with the BoW analysis, the colour-coded words, ranging from red for lower ratings to green for higher ratings, reinforce the sentiment distribution. Terms such as “amazing”, “excellent”, and “taste”, coloured with green, reflect positive reviews and are associated with higher ratings. This analysis demonstrates that positive sentiments dominate the reviews, with customers generally reporting positive experiences.

Figure 4: *Word Cloud using TF-IDF Vectorization*

Co-occurrence Matrix

A co-occurrence matrix was used to identify correlations between words in customer reviews by examining how often pairs of terms appear together in the same sentence. The resulting heatmap showed darker shades for higher co-occurrence frequencies. Positive reviews commonly feature co-occurring terms like ”good”, ”food”, and ”place”, emphasizing positive meal experiences. Clusters of related terms such as ”taste”, ”chicken”, and ”biryani” appear frequently in discussions of food quality. Similarly, words like ”staff”, ”ambience”, and ”service” often group together, highlighting the importance of service in the overall experience. Negative reviews show distinct patterns, such as ”chicken” and ”wings”, indicating specific items that generate criticism. Figure 5 illustrates these correlations, emphasizing sentiment-driven word associations.

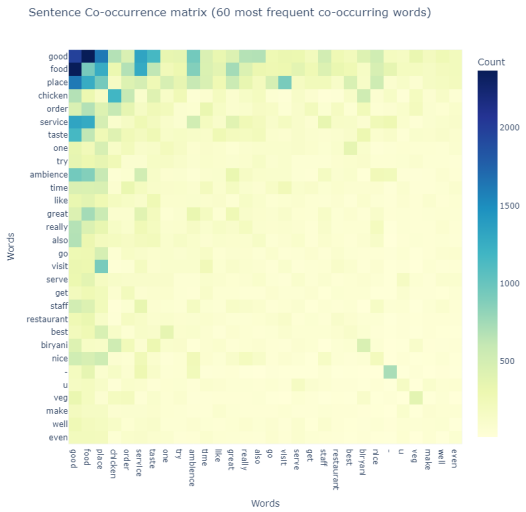

Figure 5: *Sentence Co-occurrence Matrix*

4 Data Preparation

4.1 General Data Preparation

Once the data understanding stage is complete, it is crucial to proceed to the general data preparation phase. In this regard, considering all the proposed tasks, a 'complete_df' will be created. This will be the primary data frame used for both clustering/co-occurrence analysis and multilabel classification, as it contains all the necessary reviews for both tasks. Since certain dish names are less frequent in clustering and co-occurrence analysis, it is important to include all the reviews to extract the most dish names possible. Additionally, in multilabel classification, some labels have very limited data representation, so removing reviews would aggravate this problem. Therefore, it will maintain all the reviews to ensure adequate data availability for both analyses.

Then, we proceeded with the creation of two new variables: the number of words and the number of sentences per review. By analyzing the summary statistics of the data frame, it is possible to conclude that reviews vary significantly in length, with an average of 182 words and 4.1 sentences, but the range is extremely wide. While most reviews are relatively brief, ranging from 100 to 200 words, others are particularly long, with some exceeding 3394 words. Similarly, the number of sentences per review ranges from 1 to 5, with a few reviews containing as many as 66 sentences. This implies that, while most evaluations are brief, there is significant variation in both length and structure, along with an overall tendency toward positive sentiments.

To gain a clearer understanding of the reviews' length, a plot represented by Figure 6 was created to analyze the distribution of the number of words per review. Most reviews have a relatively low word count, although a few contain significantly higher word counts. The mean number of words per review is 181.77, which is notably higher than the median of 126, influenced by outliers with long word counts. Based on this, the 95th percentile was calculated to be 535, and it was decided that reviews containing more than 600 words would be eliminated. Similarly, the 5th percentile was also calculated; however, since reviews with fewer words provided meaningful insights for the sentiment analysis model, they were maintained.

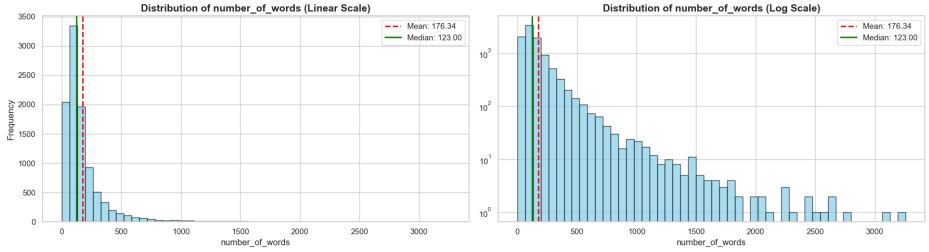

Figure 6: *Distribution of Number of Words*

Similarly, to better understand the sentence count distribution, a plot was created, revealing that most reviews contain a small number of sentences. The distribution showed that the majority of reviews are brief, typically consisting of three or fewer sentences, with fewer instances of longer, more detailed reviews. The 95th percentile was then calculated, yielding a value of 9.0. Since outliers can distort analysis and models, particularly in tasks like sentiment analysis and topic modelling, extremely long reviews may introduce noise and bias. As a result, reviews with an unusually high number of sentences were excluded from the analysis.

4.2 Multilabel Classification Data Preparation

Regarding the data preparation for multilabel classification, we started by changing the dataset to include only the 'Review' and 'Cuisines' columns, creating new "labels" columns with the cuisine types. Initially, it identified 42 unique cuisines and created binary columns for each, to check their presence in reviews (1 if present and 0 if not). A frequency analysis showed North Indian as the most frequent cuisine (5995 occurrences), followed by Chinese (4095), while Pizza was the least frequent (<100 occurrences).

Using a heatmap, we analyzed co-occurrences between rare ($\leq$100) and common cuisines. Labels with $\leq$100 occurrences were highly correlated and removed, and a similar approach was applied to labels with 199 or fewer occurrences, except for 'Lebanese_label', which always appeared alone, so we decided to keep it. A total of 10

invalid labels were identified and deleted.

After shuffling and ensuring a balanced dataset, we confirmed the distribution remained consistent post-filtering. We transformed some labels' names if they had 2 words, and joined them, like "Fast Food" were corrected to "FastFood", and no missing values were found. The dataset was updated to 32 labels, and word clouds were created for each cuisine after removing frequent generic words (e.g., "good", "food", "service").

Moving to data pre-processing, we created a column of pre-processed reviews in the main dataset to prepare for testing before performing a grid search. Following this, we performed the binarization of labels, converting each label into a binary list where each entry indicates the presence (1) or absence (0) of the corresponding label.

To further support testing, we generated columns representing TF-IDF vectors and Bag of Words (BoW) vectors based on the pre-processed reviews. These vector representations transform the text data into numerical formats, allowing us to test various approaches and evaluate model performance effectively.

4.3 Sentiment Analysis Data Preparation

Regarding sentiment analysis of user reviews, we started by creating a data frame named, 'main_sentiment', only containing the 'Name', 'Rating' and 'Review' columns from the original dataset. Using a text pre-processing pipeline, the 'Review' column is cleaned by handling punctuation, emojis, stopwords, etc. For this step, a new column called 'Preproc_text' is created, to store the cleaned version of each review. To further analyze sentiment at the sentence level, each review was tokenized into individual sentences, pre-processed, and stored as a list in also a new column named 'preproc_sentences'. Additionally, the number of words per review was calculated and this word count was stored in a new column, 'Nº Words', to provide an understanding of the length of the reviews.

4.4 Topic Modelling Data Preparation

The pre-processing for topic modelling involved applying tokenization to the pre-processed text data. The column 'preproc_reviews' already contained cleaned reviews where steps such as removing punctuation, stopwords, and other text cleaning operations had been performed. A dictionary was created using Gensim, which mapped each unique token to a unique ID. Following this, the reviews were converted into a Bag of Words (BoW) representation, encoding the frequency of each token in each document.

4.5 Co-occurrence Analysis/Clustering Data Preparation

As mentioned during the general data preparation, only a few reviews referred to specific dishes, limiting the co-occurrence analysis. To address this, web scraping was employed to create a list of 41353 dishes. The 'complete_df' dataset was imported, and a new data frame was created containing only the 'Review' and 'Cuisines' variables.

The pre-processing phase involved creating the 'preproc_content' column with cleaned text and the 'doc2vec_content' column, which retained words while removing punctuation for the Doc2Vec vectorizer. A regex operation was applied to identify dishes present in both the 'dishes_list' and the 'preproc_content' column. This process revealed 1791 missing entries, which were subsequently removed.

Two visualizations were then created to identify common words. Some irrelevant terms, resulting from web scraping issues, were excluded. Three new columns were added to the dataset to prepare for the modelling phase: one for Bag of Words vectors, one for TF-IDF vectors, and one for Doc2Vec vectors.

Subsequently, a co-occurrence matrix and a network graph were generated. 'chicken' exhibited the strongest correlations, with 'naan' and 'butter' appearing together 74 times and 'cake' and 'chocolate' 67 times. Due to its dominance, 'chicken' was excluded, and the co-occurrence matrix and network graph were revised. Figure 7 highlights significant associations such as 'rice' and 'biryani' (96 times), 'pasta' and 'sauce' (51 times), and 'pizza' and 'sauce' (34 times). Finally, dishes were vectorized without 'chicken' to enhance clustering results.

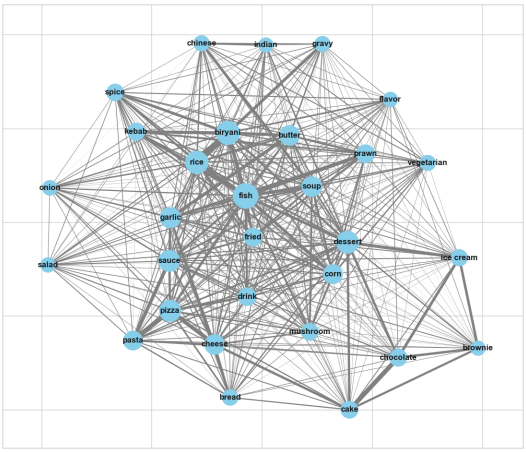

Figure 7: *Network Graph*

5 Modelling

5.1 Multilabel Classification

We evaluated two multilabel classification models, OneVsRestClassifier and ClassifierChain, both using Logistic Regression. The dataset was pre-processed with TF-IDF vectorization for features and binarization for labels and split into training (70%) and testing (30%) sets. The OneVsRestClassifier independently predicted each label, while the ClassifierChain modelled label dependencies.

The OneVsRestClassifier outperformed the ClassifierChain, with higher metrics across all measures, including an F1 score of 0.4980 compared to 0.4368 for the ClassifierChain. The OneVsRestClassifier demonstrated better Precision (0.5486 vs. 0.4491), indicating fewer false positives, and handled label predictions more effectively. The ClassifierChain struggled to model label dependencies, leading to lower overall performance. This suggests that independent label predictions are more suitable for this dataset.

Using the Bag of Words vectorizer, we tested the performance of the OneVsRestClassifier and ClassifierChain models. The OneVsRestClassifier achieved an accuracy of 0.1079, a precision of 0.5312, a recall of 0.4990, and an F1 score of 0.5000. The ClassifierChain model, in comparison, achieved a slightly higher accuracy of 0.1226 but lower precision at 0.4912, recall at 0.4834, and an F1 score of 0.4789. These results indicate that while the ClassifierChain was better at overall accuracy, the OneVsRestClassifier performed better in terms of precision and the balance between precision and recall, as reflected in its F1 score. The Bag of Words vectorizer appears to provide a slight improvement in model performance compared to previous tests with TF-IDF and will be selected as the feature representation for further analysis, including grid search optimization.

Following, we developed and evaluated the "Hermetic Classifier", a custom pipeline integrating pre-processing, vectorization, and classification. The classifier supports multiple vectorization methods, including TF-IDF, Bag of Words, and Doc2Vec, combined with robust classification models like Logistic Regression.

Training and evaluation were conducted on a binary classification task, comparing different vectorization techniques. Metrics such as accuracy, precision, recall, and F1 score revealed substantial differences in performance. TF-IDF and BoW consistently achieved good results, while Doc2Vec performed poorly, with metrics such as accuracy (0.01) and F1 score (0.03), indicating its embeddings failed to capture meaningful patterns for this task.

So, with this testing, we reach the conclusion that the Doc2Vec vector has a terrible performance compared with the BoW and the TF-IDF vector. For this reason, we will not use Doc2Vec in the grid search.

5.2 Sentiment Analysis

Regarding sentiment analysis of user reviews, we decided to use two natural language processing tools, VADER and TextBlob. Starting with VADER, Valence Aware Dictionary and sEntiment Reasoner, it was performed the calculation of two metrics resulting in the creation of 2 columns, 'vader_polarity', which is the overall sentiment

score for the entire pre-processed review text, and 'mean_sentence_polarity_vader', that is the average sentiment score across all sentences in the review. Regarding TextBlob Sentiment Analysis, also two columns were created, one storing the overall polarity score of the full review text and another calculating the average polarity of individual sentences within each review.

5.3 Topic Modeling

In topic modelling, we intentionally tested a smaller number of topics for LDA, LSA, and BERTopic to ensure manageable and interpretable results.

LDA

For the LDA model, we evaluated coherence scores for 3, 5, 10, and 15 topics to identify the optimal topic distribution. It was verified that increasing the number of topics led to lower scores. We opted for an early stopping condition to halt the tests when there was no improvement in the results after two consecutive iterations.

The best LDA model obtained, for each of the five topics, calculates a probability distribution over the 13342 unique words in the corpus. Each word is assigned a weight that signifies its relevance to a given topic. The extracted topics are meaningful and describe distinct themes:

- Topic 0: Negative Dining Experience, that includes terms such as "food", "order", "restaurant", "worst", and "don't".
- Topic 1: Positive Service and Buffet, include words like "service", "good", "food", "staff", and "buffet".
- Topic 2: Restaurant Atmosphere and Food Quality, is characterized by words such as "place", "food", "good", "service", and "ambience".
- Topic 3: Desserts and Sweets, have words like "chocolate", "cake", "cream", "ice", and "brownie".
- Topic 4: Positive Reviews on Food (Chicken & Biryani), includes terms like "good", "chicken", "ordered", "taste", and "biryani".

We made an interactive graph to demonstrate what we previously said, where it is possible to observe the most relevant words within each topic.

LSA

For the LSA model, we also tested different numbers of topics to identify the best configuration. LSA captured broad themes like food, service, and specific dishes but struggled to separate topics clearly, with overlapping terms such as "good", "food", and "place" appearing repeatedly. The mathematical nature of LSA led to less interpretable outputs compared to LDA, which better-captured topic diversity and specificity.

BERTopic

Lastly, as mentioned the BERTopic model was applied to the dataset and initially we handled the missing values in the text data and the model was configured to generate 20 topics.

The results topics with key themes reflecting food quality, service, ambience, delivery, and specific dishes such as chicken, biryani, and desserts. Topics were represented by words with strong relevance and combined with representative documents, providing clear examples of the text associated with each theme. While most topics were fairly coherent and aligned with common themes in the dataset, some topics appeared less relevant. This could be attributed to pre-processing issues, such as improper tokenization or the inclusion of irrelevant terms.

5.4 Co-occurrence Analysis/Clustering

During the modelling phase, the K-means algorithm was applied to each of the vectors created earlier. Initially, the algorithm was tested with the Bag of Words vectorization, using the elbow method to help to determine the optimal value for k. After analysing the inertia graph, a new column was added to the data frame containing 6 clusters. Although the "elbow_finder" function indicated that 9 clusters were the best k, we opted to choose 6 clusters as with this number the clusters would be easier to identify and we could conclude by the graph that 6 was also an "elbow" in the graph. The same approach was then applied to both TF-IDF and Doc2Vec vectorizations, resulting in 8 clusters for TF-IDF and 6 clusters for Doc2Vec.

6 Evaluation

6.1 Multilabel Classification

We used a grid search to test two vectorization techniques, TF-IDF and Bag of Words, and two classifiers, OneVsRestClassifier and ClassifierChain, with various hyperparameter configurations.

The best configuration used a TF-IDF vectorizer, and a OneVsRestClassifier with Logistic Regression, achieving a weighted F1 score of 0.535, precision of 0.496, recall of 0.592, and accuracy of 0.102.

Due to class imbalance, the model's F1 score was impacted in a negative way. We tried to fix this problem by removing 10 labels that strongly correlated with more frequent labels. Despite this adjustment, class imbalance persisted. To evaluate the impact of this decision, a comparative analysis was performed between the two models. The model trained with 42 labels achieved a weighted F1 score of 0.54, while the model trained with 32 labels achieved a slightly higher score of 0.55. However, the first model was considered the better since it can make predictions for all labels.

6.2 Sentiment Analysis

For the evaluation of Sentiment Analysis, we started to calculate the Pearson correlation coefficient between two sentiment analysis metrics, 'mean_sentence_polarity_vader' and 'vader_polarity'. The result obtained, 0.8436, shows a strong positive correlation, indicating that both measurements are well-aligned. This statistical measure was also calculated between the normalized ratings and the polarity scores, to evaluate the alignment between sentiment scores and user ratings. The data was normalized using MinMaxScaler, and the correlation between VADER polarity and ratings was 0.7178, indicating a strong positive relationship. For TextBlob polarity, the correlation was 0.7016 also showing a strong positive relationship. At the sentence level, the mean VADER polarity correlated with 0.704, while TextBlob's sentence-level polarity was 0.685. These results suggest that VADER slightly outperforms TextBlob in capturing sentiment that aligns with user ratings.

Two error metrics were calculated to assess the differences between normalized ratings and normalized VADER polarity scores. The Root Mean Squared Error (RMSE) was 0.272, and the Mean Absolute Percentage Error (MAPE) was 0.144.

The scatter plot represented by Figure 8 compares VADER polarity and TextBlob polarity, revealing a positive correlation between the two tools.

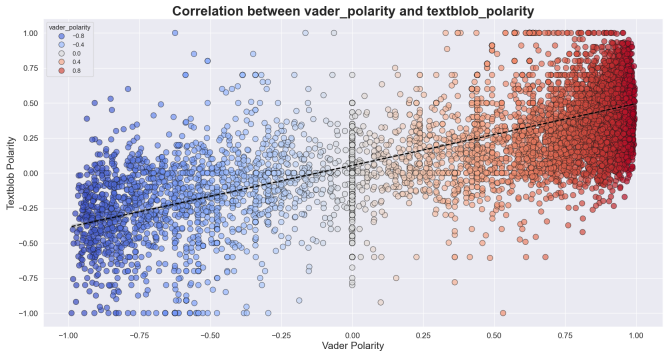

Figure 8: *Correlation Between Vader Polarity and TextBlob Polarity*

The table summarizing the dataset's statistics showed that the Rating column has an average value of 3.6 on a scale of 1 to 5. VADER polarity has a mean of 0.45, while TextBlob polarity has a lower mean of 0.25, suggesting that VADER assigns stronger positive sentiment scores overall. Normalized values for ratings and sentiment scores confirm this trend, with normalized VADER polarity averaging 0.727 compared to 0.627 for TextBlob. Lastly, we evaluated these two models used. Starting with VADER, through the normalized distributions made, we can notice that Vader tends to classify the reviews as 5-star rating way more than they actually are. It seems to classify pretty well the ratings of 2 and 3 stars, which distributions are very close to reality, although these are the ratings with

the least amount of reviews, so they are less significant. TextBlob, on the other hand, has a way more centred distribution, with a very low variation.

6.3 Topic Modelling

LDA was considered the most effective method, achieving the highest coherence score of 0.5529 and producing distinct topics. These topics captured important themes such as negative dining experiences, positive service and buffet, restaurant atmosphere and food quality, desserts and sweets, and positive reviews on food (chicken & biryani).

In contrast, LSA achieved a lower coherence score, 0.5072, and its topics were less distinct and harder to interpret. While it captured broad themes such as food and service, its outputs were less useful for extracting specific insights.

The BERTopic model produced a range of fairly coherent topics, reflecting common themes found in the dataset, such as food quality, service, ambience, and specific dishes. However, some topics might not seem entirely relevant or meaningful, and this could be due to issues in the pre-processing stage, such as incorrect tokenization or the inclusion of irrelevant terms.

6.4 Co-occurrence Analysis/Clustering

In the clusters evaluation, with Bag of Words vectorization, it was possible to identify, by the cluster names, that the first cluster corresponded to 'north indian, dessert' cuisines, the second to 'biryani, chinese', the third to 'italian, pizza', the fourth to 'desserts, fast food, bakery', the fifth to 'ice cream, desserts', and the seventh to 'seafood, chinese'. After determining which cuisines were represented in each cluster, two data frames were created to summarize the findings. The first included the top three most frequent cuisine types for each cluster along with their respective frequencies. The second, created to capture more specific cuisine types, excluded the six most frequent cuisine types.

By comparing the cuisines identified in each cluster with the actual most frequent cuisine types, it was concluded that it was possible to successfully identify cuisine types based on their respective cluster names.

The clusters produced using TF-IDF and Doc2Vec were then evaluated with the same methodology as those produced by Bag of Words, and the cuisine types for those clusters were also successfully identified. Finally, by calculating the silhouette score for each K-means algorithm, it was determined that Doc2Vec achieved the highest score, producing the most well-defined clusters.

7 Conclusion

With this project, we intended to demonstrate the utility of using text mining techniques in analysing restaurant reviews to classify cuisines, predict Zomato scores, identify topics and cluster related dishes and cuisines.

Regarding multilabel classification, the OneVsRestClassifier, using the TF-IDF vectorizer, achieved the best results with an F1 score of 0.5409, outperforming the ClassifierChain.

In predicting Zomato scores, VADER outperformed TextBlob, with a correlation of 0.7178 to normalized ratings and lower error metrics, RMSE with a value of 0.272 and MAPE with 0.144. While VADER tended to overestimate 5 star ratings, it dealt well with 2 and 3 star ratings.

Topic modelling using LDA achieved the highest coherence score, 0.5529, extracting themes like service quality, food satisfaction, and dining experiences.

Using clustering it was identified cuisine types, with the first cluster being 'north indian, dessert' cuisines, the second to 'biryani, chinese', the third to 'italian, pizza', the fourth to 'desserts, fast food, bakery', the fifth to 'ice cream, desserts', and the seventh to 'seafood, chinese'.

Despite the results obtained, we encountered some limitations while doing this project. First one, was class imbalance, that reduces the effectiveness of multilabel classification for rare cuisine types. When performing topic modelling, BERTopic has some topics affected due to pre-processing issues. Lastly, the fact that VADER overestimates 5 star ratings suggests that sentiment models need to be adjusted.

Despite these limitations, the steps applied demonstrated the usefulness of text mining in extracting actionable insights from unstructured data, aiding restaurants in understanding customer feedback and improving services.

References

[1] Annika Molenaar et al. "Using Natural Language Processing to Explore Social Media Opinions on Food Security: Sentiment Analysis and Topic Modeling Study". In: *J Med Internet Res* 26 (Mar. 2024), e47826. ISSN: 1438-8871. DOI: 10.2196/47826. URL: https://doi.org/10.2196/47826.

[2] Sonya Zhang et al. "Topic Modeling and Sentiment Analysis of Yelp Restaurant Reviews". In: *International Journal of Information Systems in the Service Sector* 14 (Jan. 2022), pp. 1–16. DOI: 10.4018/IJISSS.295872.

[3] Susan (Sixue) Jia. "Motivation and satisfaction of Chinese and U.S. tourists in restaurants: A cross-cultural text mining of online reviews". In: *Tourism Management* 78 (2020), p. 104071. DOI: 10.1016/j.tourman.2019.104071. URL: https://doi.org/10.1016/j.tourman.2019.104071.