SDSC4008: Deep Learning

2024-2025 Semester B

香港城市大學
City University of Hong Kong

Topic: Sentiment & Geo-Spatial Analysis of Yelp Reviews using Deep Learning: Uncovering Business Insights

NAME	SID
MALHOTRA Akul	56698828
AIDARBEKOV Duman	57514520
KYDYR Meiirbek	57161931

Introduction

Background

The Yelp dataset is a full range of user reviews and ratings on the Yelp site, representing a rich data source for knowledge about consumer habits and tastes. It contains diverse elements such as reviews, users, and businesses, enabling detailed analysis of local companies. The interest here lies in the restaurant category, with reviews offering dense qualitative information and alongside structured fields such as location and business.

Motivation

As the restaurant market is more competitive, it is crucial to know what customers are thinking to succeed. While new restaurants emerge or existing chains consider expansion, insights from customer opinions can inform business strategies. Being able to determine what customers like or dislike regarding their dining experiences allows businesses to personalize their offerings and extract the maximum possible satisfaction from their customers. Besides, the Yelp dataset assists restaurants in identifying market trends, resolving recurring issues, and preserving their reputation.

Description

Our study aims to use sentiment analysis derived from the Yelp dataset to provide actionable suggestions for restaurants. By using reviews, we can see patterns that indicate customer preferences and problems. Our research will be able to advise new restaurant companies on what causes good meals, such as food quality and ambiance, and where they need to do better, such as service time and price.Moreover, our graph structure-based deep learning approach will enable geospatial analysis such that restaurants can determine the best locations to expand based on existing market performance. Through the analysis of the profiles of top-performing restaurants in various locations, we aim to equip businesses with information to create distinctive dining experiences that will stand out in the competitive market.

Literature Review

We have seen many papers focused on sentiment analysis of Yelp data, exploring various techniques to predict star ratings based on reviews and user interactions. However, there are relatively few studies that utilize graph structures in this context. One notable approach employs graphs for feature extraction, leveraging connected graph features to enhance traditional machine learning workflows (Shaikh, 2019). This method focuses on quantifying relationships within the data, such as user interactions and business influences, to improve classification accuracy. Another significant contribution involves creating a bipartite graph that connects users and businesses, which is used as part of the foundation for many different machine learning algorithms (Perez, 2017). This approach captures the interactions between users and the businesses they review, allowing for a structured representation of the data that facilitates better predictions. While these studies highlight the potential of graph-based methods, they primarily focus on feature extraction and classification.In contrast, our approach integrates graph neural networks directly into the prediction process, utilizing a Graph Attention Network (GAT) to dynamically learn from the graph's structure and node features. By utilizing business characteristics with edge attributes, such as distances, our model captures complex relationships in a more nuanced way. This direct incorporation of graph structures allows for a deeper understanding of the interactions within the Yelp dataset, enhancing the model's predictive capabilities beyond traditional feature extraction methods.

Methodology

We created two separate deep learning algorithms for sentiment analysis and geospatial analysis, since it was difficult to combine both of them in one neural network model. We have utilized 3 types of data out the requirements of the project, those being spatio-temporal, text and user ratings.

Sentiment Analysis:

1. Text Preprocessing:

In our examination, we focused particularly on reviews that were categorized under "Restaurants." The first thing was to divide the dataset to retain only the reviews. Second, we took care to retain only the English reviews, using a language detect library (langdetect) to exclude non-English reviews.Then we went for text normalization. All reviews were converted to lower case to ensure that they all have the same case. Punctuation and stopwords were then eliminated using NLTK's predefined list of English stopwords. This process greatly minimized noise in the text such that it was easier to analyze the content.In order to combat these issues, we also did stemming with the Porter Stemmer algorithm. This reduced the words to their root form, which helped to bring similar words together and also helped with the efficiency of our analysis.

2. Feature Extraction (Bag-of-Words)

After text preprocessing, we created a Document-Term Matrix (DTM) utilizing CountVectorizer. The matrix had the term frequency from the reviews. Then, we discarded sparse terms and retained words that appeared in over 1% of the documents. The filtering process was necessary to reduce the dimensionality and focus on key features.

3. Neural Network Model (Binary Classification)

For the binary classification task, we were looking to label the reviews as either "positive" or "negative." The labels were assigned based on predefined sentiment thresholds. We split the dataset into the training, validation and test sets in order to test the model's performance. We built Deep Neural Network (DNN) and Convolutional Neural Network (CNN) models to compare their performance metrics. Further, a better model will be used to analyze the results from Geospatial Analysis.

Model Architecture: Sequential model with:

DNN	CNN
Model Architecture Dense (512 units, ReLU, He normal) → BatchNorm → Dropout (0.5) Dense (256 units, ReLU, He normal) → BatchNorm → Dropout (0.4) Dense (128 units, ReLU, He normal) → BatchNorm → Dropout (0.3) Dense (64 units, ReLU, He normal) → BatchNorm → Dropout (0.2) Output (1 unit, sigmoid for binary classification)	Model Architecture Conv1D (128 filters, kernel=3, ReLU, He normal) → BatchNorm → MaxPooling (2) → Dropout (0.3) Conv1D (64 filters, kernel=3, ReLU, He normal) → BatchNorm → MaxPooling (2) → Dropout (0.3) Flatten → Dense (64 units, ReLU) → Dropout (0.5) Output (1 unit, sigmoid for binary classification) Optimizer:

Optimizer: AdamW (LR=0.001, β_1=0.9, β_2=0.999) Training Configuration: Loss: Binary cross-entropy Metrics: Accuracy, Precision, Recall Callbacks: -EarlyStopping (monitor="val_loss", patience=10, restore_best_weights) -ReduceLROnPlateau (factor=0.2, patience=5) Epochs: 30 (early stopping enforced) Batch Size: 128 Validation Data: (X_valid, y_valid) Key Features: Regularization: L2 weight decay (λ=0.001) + progressive dropout Stability: BatchNorm + adaptive LR scheduling Advanced Activations: LeakyReLU/PReLU in hidden layer	Adam (default parameters: LR=0.001, β_1=0.9, β_2=0.999) Training Configuration: Loss: Binary cross-entropy Metrics: Accuracy, Precision, Recall Callbacks: -EarlyStopping (monitor="val_loss", patience=5, restore_best_weights) -ReduceLROnPlateau (factor=0.2, patience=5, min_lr=1e-6) Epochs: 10 (early stopping enforced) Batch Size: 32 Validation Data: (X_valid_cnn, y_valid) Key Features: Regularization: Dropout (progressive: 0.3 → 0.5) + BatchNorm Stability: BatchNorm + adaptive LR scheduling

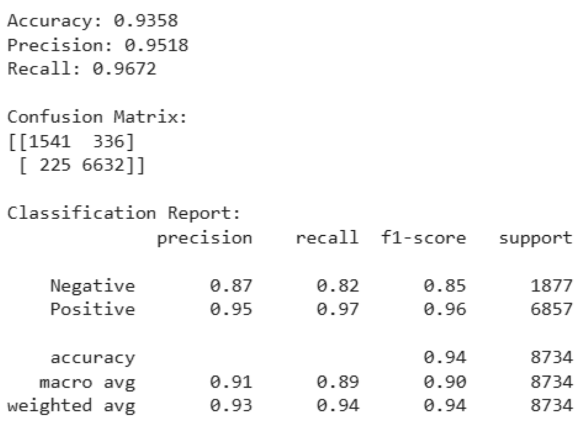
Geospatial Analysis:

1. **Data Preparation:**The analysis begins with the preparation of geospatial data, focusing on three key features: longitude, latitude, and review count. To address the skewness in the review count distribution, we applied log transformation. This transformation aims at reducing variance and making the data more normally distributed. Next, we standardized all features using the StandardScaler to ensure that they contribute equally to the model training process.
2. **Building the Spatial Graph:** In constructing the spatial graph, each node represents a restaurant. The graph is designed such that each restaurant is connected to its 15 closest neighbors based on geographic proximity. This spatial adjacency allows the model to leverage local context and relationships between restaurants, which is crucial for understanding rating patterns in the dataset.
3. **Neural Network Model:** The core of the methodology involves implementing a Graph Attention Network (GATv2). The architecture comprises two layers equipped with attention mechanisms that enable the model to weigh the importance of neighboring nodes dynamically. The input features for the GATv2 model consist of three attributes: longitude, latitude, and the log-transformed review count. The output of the model is a predicted rating between 0 and 1, reflecting the expected star rating for each restaurant. The detailed architecture can be examined in the code.

4. **Anomaly Detection:** To identify potential outliers in the predicted ratings, we incorporate anomaly detection mechanisms. This process involves calculating the residuals between the actual ratings and the predicted ratings generated by the GATv2 model. Then anomaly score is developed by combining the prediction error with the importance of the review count. We identify outliers in restaurants that specifically fall within the top 5% by this score. These outliers represent restaurants that significantly deviate from expected spatial patterns, warranting further investigation.

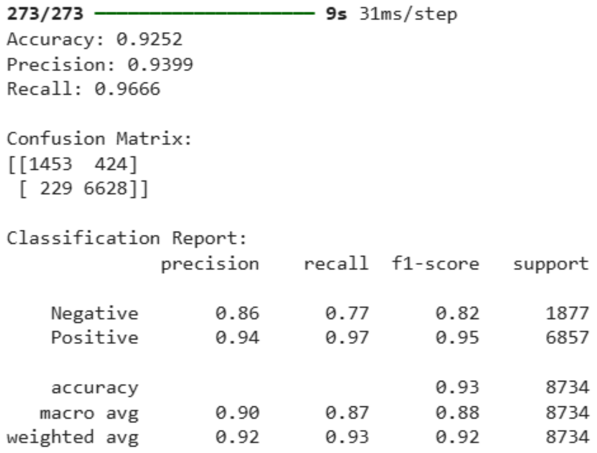
Analysis & Findings

DNN Performance metrics:



```
Accuracy: 0.9358
Precision: 0.9518
Recall: 0.9672

Confusion Matrix:
[[1541  336]
 [ 225 6632]]

Classification Report:
              precision    recall  f1-score   support

    Negative       0.87      0.82      0.85      1877
    Positive       0.95      0.97      0.96      6857

    accuracy                           0.94      8734
   macro avg       0.91      0.89      0.90      8734
weighted avg       0.93      0.94      0.94      8734
```

CNN Performance metrics:



```
273/273 ━━━━━━━━━━━━━━━━━━━━ 9s 31ms/step
Accuracy: 0.9252
Precision: 0.9399
Recall: 0.9666

Confusion Matrix:
[[1453  424]
 [ 229 6628]]

Classification Report:
              precision    recall  f1-score   support

    Negative       0.86      0.77      0.82      1877
    Positive       0.94      0.97      0.95      6857

    accuracy                           0.93      8734
   macro avg       0.90      0.87      0.88      8734
weighted avg       0.92      0.93      0.92      8734
```

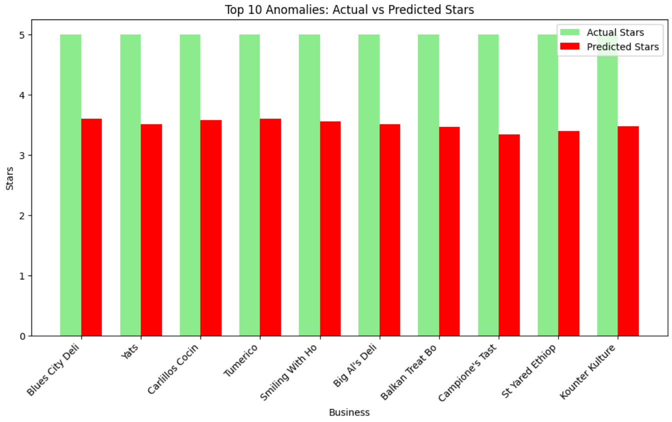
The DNN model performs better than the CNN model in accuracy (93.58% compared to 92.52%), precision (0.9518 compared to 0.9399), and F1-score of the negative class (0.85 compared to 0.82), reflecting improved overall performance and predictability.

Both models have very similar recall (0.9672 compared to 0.9666), reflecting similar efficacy in detection of true positives.

The DNN model produces fewer false positives (336 vs. 424) and is thus more accurate for the positive class, while both models have similar false negatives (225 vs. 229).

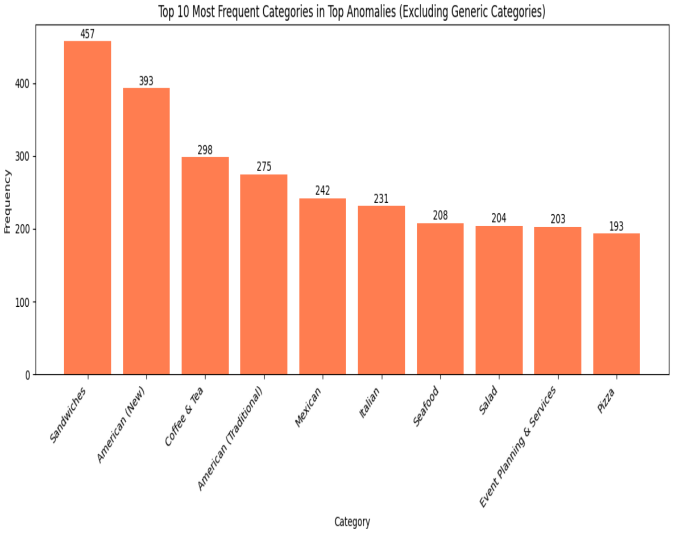
In imbalanced classification (more positive samples), the DNN's higher precision and balanced F1-scores mean that it generalizes better compared to the CNN model. Overall, DNN is the stronger model for this task, particularly in minimizing false alarms (FP) while maintaining high recall.

Post the obtainment of the anomalies by the procedure as described in the Methodology, an in-depth analysis was carried out for the characteristics possessed by the anomalies, also contrasting them to those possessed by general restaurants. This shall further assist in getting business insights, and be aligned with the research's focus on assisting someone who is planning to open a new startup restaurant or expand an existing chain, in understanding key characteristics which they can incorporate in their own business.

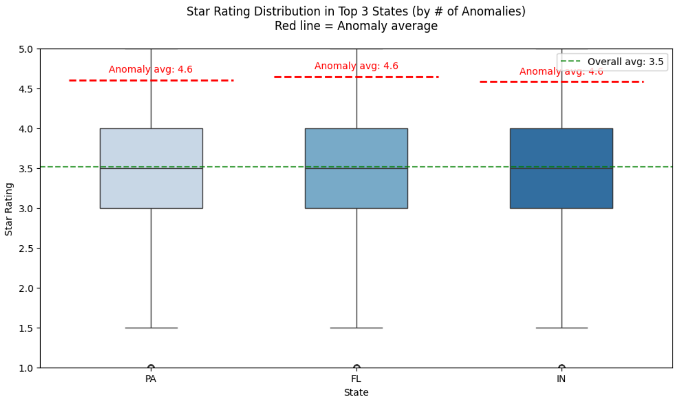


Firstly, it shall be useful to look at the top 10 anomalies which have been found by using the model.
All top 10 anomalies have a much higher average star rating than predicted by the model, indicating exceptional restaurant quality as well as service relative to other restaurants in the locality in which they are located. It is also interesting to see that All top 10 anomalies have an average rating close to the highest obtainable perfect 5 stars, thereby displaying extremely high quality and excellence. Though from the names, it is not evident the certain category these top anomalies belong to, and thus, it shall be useful to have a more broader look at the categories these anomalies come from by looking at top 10 most frequent categories in all of the top 5% anomalies, as predicted by the model.

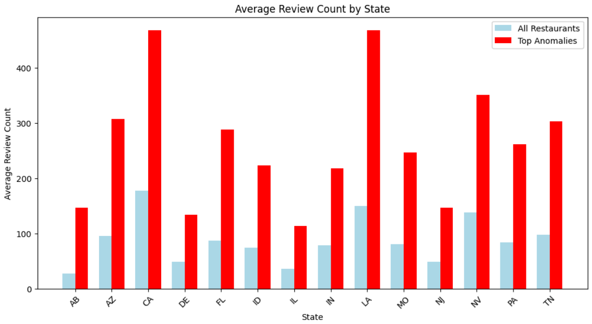
Looking at the top categories for the anomalies, it is notable to observe "Sandwiches" as the most common category among exceptional restaurants. American & Italian cuisine seems to be very prevalent in highly rated cuisine, with 2 variants of both appearing in top categories, as can be seen with the presence of American (Traditional) & American (New) for American and with Italian and Pizza specific restaurants as well. Country specific cuisines seem to be quite popular among anomalies as along with the presence of Italian and American cuisines, there is Mexican cuisine as well. It is also interesting to note that most top anomalies are fine dining or typical restaurant places, as seen from categories, and there is only one atypical cafe-like category 'Coffee and Tea' in the top most occurring anomalies.



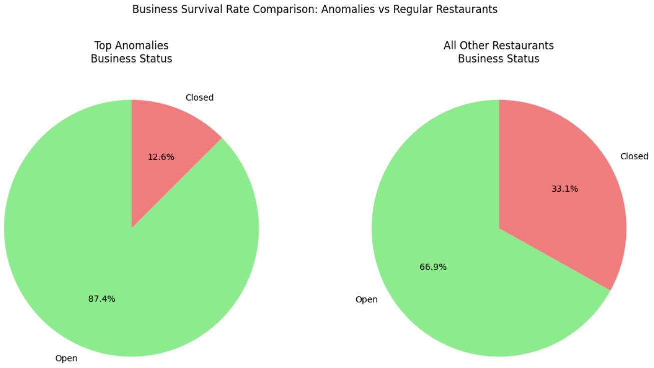
For a more geographic focused analysis, it shall also be useful to observe the top states with the most number of anomaly restaurants and also how the rating in such states is distributed.



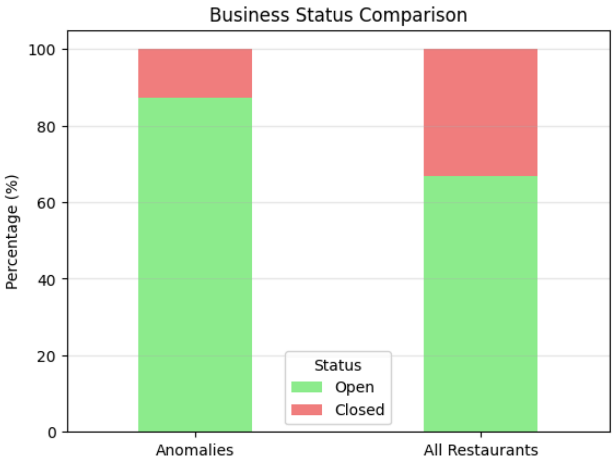
As it can be observed the states Pennsylavania, Florida and Indians have the most number of exceptional restaurants among all the states from the analysis. All the top 3 states, interestingly, have the same average stars for anomalies (4.6). The average stars for anomalies is about 31% higher than average stars for other restaurants (which have an overall average of 3.5 stars), thereby indicating that anomaly restaurants have outshined other usual restaurants when it comes to general rating as received from the customers, owing to their exceptional service and other characteristics that shall be further observed in the analysis henceforth. Another interesting piece of analysis would be to draw a contrast between the number of reviews as received by the anomaly restaurants and those by the other typical restaurants.



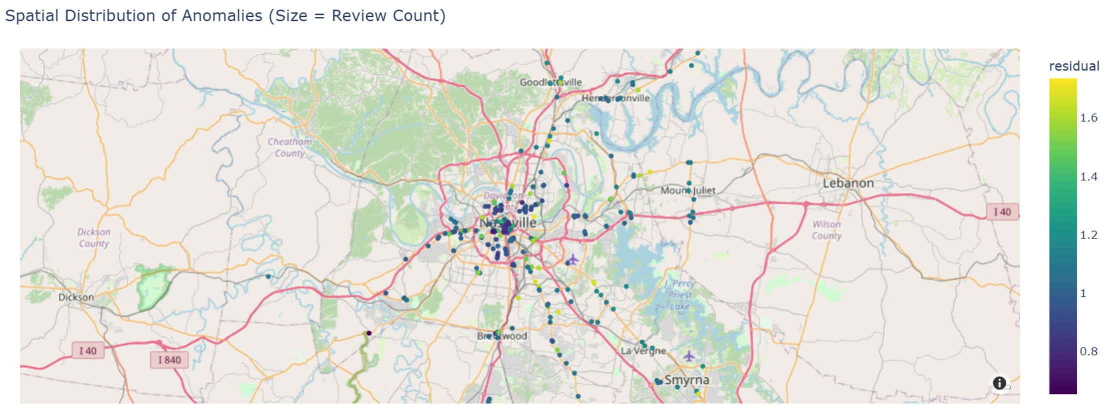
It is key to note that anomalies across all states have a much higher number of reviews than other typical restaurants. Although seen across all states, it is significantly noticeable in states like Arizona and Los Angeles. This shows that users are more likely to give more reviews to exceptional restaurants in comparison to lower rated ones. This is in line with the general logic, that customers often feel happy to speak out and reward places where they had a good experience in comparison to those places which offered a bad or an average experience to the customer. To further build upon anomaly analysis, and how exceptional restaurants display characteristics in contrast to other dining places, it shall be useful to observe how many exceptional anomaly restaurants have been closed in comparison to other restaurants.



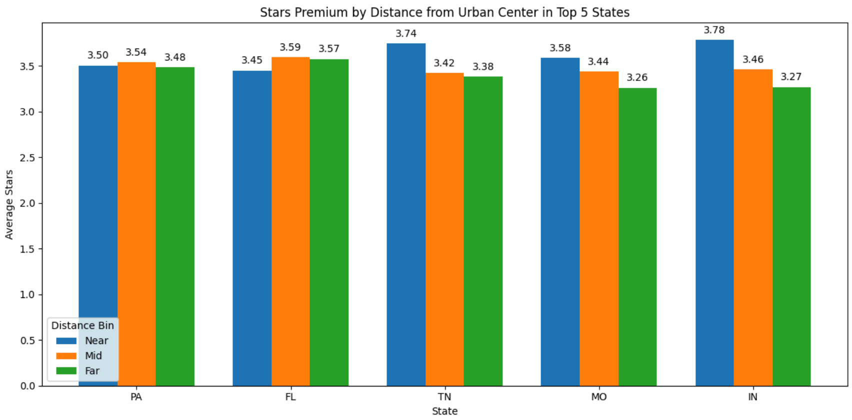
It is noticeable that 87.4% of exceptional restaurants are still open, whereas it’s only 66.9% for other dining places in the dataset chosen for evaluation.This shows that anomalies tend to show good, sustainable business performance and value in comparison to other lower rated dining places. They have a tendency of continued business as they have performed well relative to other restaurants in their geographical area, and thus are rewarded with more customers and higher rating for their exceptional service to visitors. We can further extend the analysis to some more geographical based analysis by looking at the top states and how many restaurants on average are open or closed.



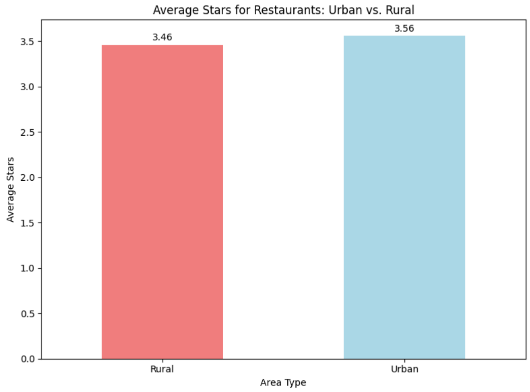
Looking at top states, Pennsylvania and Florida have much more places open in comparison to other top states. This finding is key as it reinforces our earlier finding about Pennsylvania and Florida being states with most anomaly exceptional restaurants. In other top states that have been presented, the difference is not major & a reasonable number of restaurants have been closed over time.It shall also be interesting to look at anomalies and other exceptional restaurants on a map of a particular region.

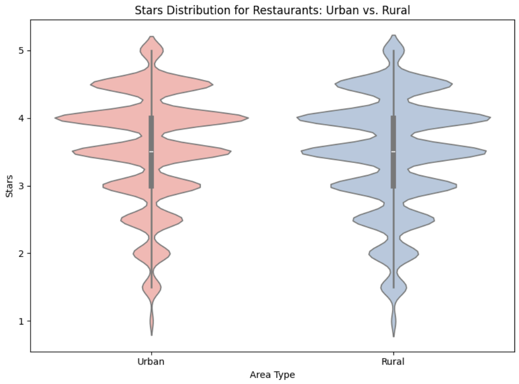


While the detailed map which shows anomalies across all regions of consideration is shown in the Python code provided, a more general map for Nashville City has been shown here, wherein the color of the restaurant has specifically been chosen to represent the residual value of the place. It is useful to observe the presence of anomalies across the City, but it is also observable that anomalies tend to decrease as we go further away from the city as shown in the map. A more in-depth analysis of this shall be useful to observe if this is a general trend as displayed by cities, and if the quality of restaurants goes down as we go further away from the city.

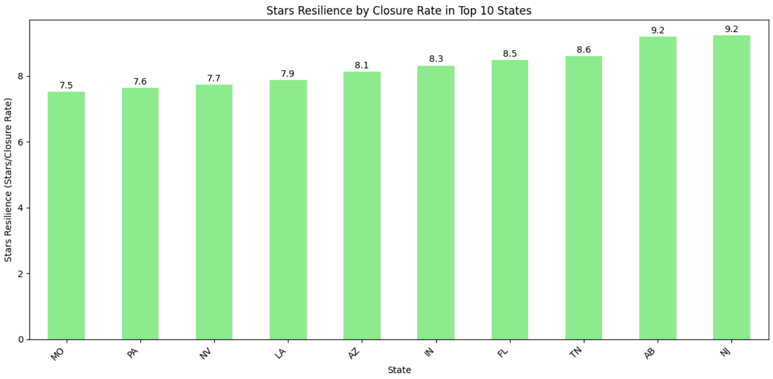


This figure helps assess how the quality of restaurants changes with distance as we go further away from the state. The distance has been divided into three separate bins for more extensive and accurate analysis. For the top states that have been analyzed, 3 of the 5 states which are Missouri, Tennessee and Indiana show a clear trend - as distance increases, the quality of dining places goes down. This is in line with the logical assumption that most extraordinary and great dining places lie closer to the main state and as we go further away, the quality drops. However, for the other 2 states in the analysis, they do not show this trend, and thus we can’t establish a clear relationship between the quality of restaurant and the distance from the state. However, a more useful analysis could be to analyse the difference in restaurant quality between the urban and rural places, as shown below.

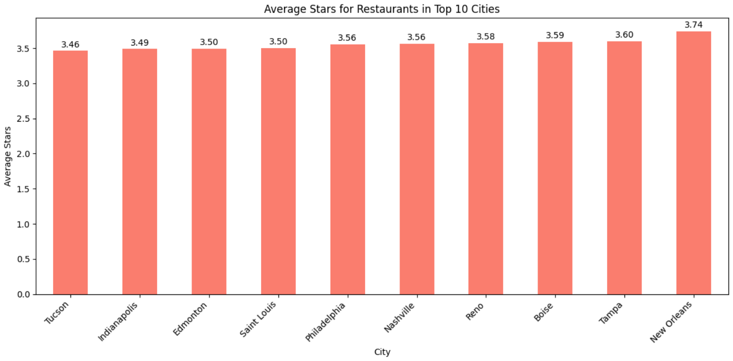




Urban places in this analysis are described as those with more than 200 restaurants in their geographic area, while rural places are those with lower than 200. It is observed that urban restaurants have shown slightly better performance in comparison to rural places, with urban places having an average rating score of 3.56 vs that of 3.46 displayed by rural places. Though the difference is not major, rural dining also has reasonably decent average ratings as received by their customers. This is an important insight into how restaurants in rural places do not lag behind their urban counterparts, and dining places across different geographic regions offer a good experience to the customers.



An interesting and key extension of the closure of restaurants analysis that was conducted previously would be to analyse the resilience rate of states. This would help us identify the states with the highest resilience rate.Stars resilience rate measures states which have high ratings despite restaurants getting closed, using the formula as Stars/Closure Rate. Alabama and New Jersey show very high rates indicating that dining places in these states have been of high quality despite restaurants being shut over time. This analysis is key, as it not only identifies states which have great restaurants but also which have shown sustainable business despite closure of restaurants with time. While sufficient analysis has been conducted at the state level, it shall also be useful to look at the cities with the highest average rating as received from the customers.



From the analysis of the top cities according to average stars, we can conclude that New Orleans is the City with the highest average star rating. All top cities with highest star ratings have decently high ratings, mostly around 3.5. It is evident that the difference in avg stars in top cities is not very significant, even though New Orleans shows a considerable margin over other cities with an average rating of 3.74.

In our last analysis, we reviewed the comments of the top 20 restaurants that we had ranked with the highest anomaly scores. Foretelling their sentiments, we found a staggering 97.5% of these comments to be marked as positive. This only goes to validate our hypothesis that the restaurants with the highest anomaly scores exhibit exceptional qualities that appeal well with customers.

In addition, we considered the most frequently occurring words used in the reviews. This frequency analysis provides valuable insights into the characteristics that make these exceptional restaurants stand out in a competitive market. Identifying these keyt characteristics aids our understanding as to what drives satisfaction and customer loyalty. For example, businesses can get the following insights:

- **'Sandwich'** - The most popular and outstanding menu category, which makes restaurants successful.
- **'Place'** - Prefer to open a restaurant in Illinois which has the lowest number of anomalies, so that the dining place can face less competition.
- **'Always'** - Suggests repeat customers ("always delicious"). Reward frequent visitors with discounts or free items.
- **'Order'** - Many top reviews indicate ease of ordering. Offer app-based preorders if not already available.
- **'Friendly'** - Staff behavior is a strength. Reinforce positive interactions to retain this reputation.

Discussion & Conclusion

In this study, we conducted a comprehensive sentiment and geospatial analysis of Yelp reviews to provide actionable insights for restaurant businesses. By employing a Graph Attention Network (GATv2), we integrated customer feedback with geographic data to predict star ratings and identify outlier restaurants. Our results revealed significant patterns in customer preferences, highlighting key factors that drive satisfaction and the importance of location in the restaurant industry.

Through our analysis, we successfully identified top-performing restaurants, characterized by their exceptional ratings compared to predictions. This not only validated our approach but also emphasized the critical role of specific menu items, such as sandwiches, in attracting positive customer experiences. Our findings indicate that understanding these characteristics can significantly enhance a restaurant's strategy, allowing new establishments to tailor their offerings to meet consumer demands effectively. However, there is also a limitation in our study. While our focus was on generating business insights, further research could enhance the granularity of our findings. For instance, incorporating additional Yelp data entities such as users in graph networks could provide a more comprehensive view in restaurant business development.In conclusion, our study offers a valuable framework for leveraging sentiment and geospatial analysis to derive business insights in the restaurant sector. By emphasizing actionable recommendations over predictive accuracy, we aim to empower restaurant owners with the knowledge needed to navigate a competitive landscape and enhance customer satisfaction. Future improvements could involve expanding the dataset and refining the analytical methods to further enrich the insights provided.

References

Perez, L. (2017). *Predicting Yelp star reviews based on network structure with deep learning*. arXiv. https://doi.org/10.48550/arXiv.1712.04350

Shaikh, A. H. (2019). *Yelp rating classification using connected graph feature extraction and feature importance in machine learning workflow* (Master's thesis, Dublin Business School).