**Rating and Sentiment Prediction of McDonalds Reviews Using NLP**

Paul Parks, Lishi Wang, Ivan Steklov

University of San Diego

AAI-501-02-SU23 - Introduction to AI

Ying Lin

August 02, 2023

# Abstract

This paper comprehensively analyzes a dataset of over 33,000 McDonald's Google reviews utilizing advanced Natural Language Processing (NLP) and machine learning techniques. Our study aimed to extract actionable insights by conducting sentiment analysis on customer reviews, investigating geographic trends in sentiments and ratings, and deploying models to predict star ratings, and sentiment ranges, based on review text. We applied several machine learning models, including linear regression, Bayesian regression, and Support Vector Machine. Performance between these models were compared. An emphasis was placed on data cleaning to handle non-standard characters, followed by strategic model selection. The results of this study show the potential of NLP and machine learning in the business domain, providing avenues for enhancing customer satisfaction, analyzing user reviews, guiding business growth strategies, optimizing resource management, and improving operational efficiency.

*Keywords*: natural language processing, linear regression, bayesian regression, support vector machine, sentiment analysis.

**Table of Contents**

[Abstract 2](#_Toc142948666)

[Introduction 4](#_Toc142948667)

[Dataset Cleaning, Preprocessing and Visualizations 5](#_Toc142948668)

[Loading the Dataset 5](#_Toc142948669)

[Cleaning and Preprocessing 8](#_Toc142948670)

[Score Prediction and Sentiment Analysis 9](#_Toc142948671)

[Text Encoding and Train/Test Split 9](#_Toc142948672)

[Ridge Regression Model 9](#_Toc142948673)

[Bayesian Model 11](#_Toc142948674)

[SVM (Support Vector Machine) Classifier. 12](#_Toc142948675)

[Figure 7: Histogram of Predicted and True Scores (SVM) 12](#_Toc142948676)

[Geolocation Analysis 13](#_Toc142948677)

[Results and Discussion 13](#_Toc142948678)

[Conclusion 15](#_Toc142948679)

[References 16](#_Toc142948680)

[Appendix 17](#_Toc142948681)

# Introduction

Customer feedback and online reviews significantly influence consumer behavior. Understanding and analyzing this feedback is essential for businesses aiming to thrive in competitive markets. McDonald's, one of the world's largest fast-food chains, receives many customer reviews that serve as valuable data. The analysis of over 33,000 Google reviews of McDonald's restaurants offers an opportunity to glean actionable insights that could reshape business strategies and operational efficiency.

The primary objectives of this research are to conduct sentiment analysis on McDonald's customer reviews, explore geographic trends in reviews and ratings, predict star ratings and sentiment based on review text. This exploration involves deploying Natural Language Processing (NLP) and machine learning techniques such as linear regression models.

This study emphasizes the importance of sentiment analysis in understanding customer preferences, dissatisfaction, and overall sentiment toward products and services. Examining geographical trends opens the door to recognizing regional variations in customer experience, allowing for localized strategies. Predicting star ratings based on review text provides an automated way to gauge customer satisfaction and identify areas for improvement.

This research contributes to the growing field of NLP and machine learning applications in business analysis by employing and assessing several machine learning models. The outcome can enhance customer satisfaction, user reviews, business growth, resource management, and operational efficiency.

The insights drawn from this research can serve as a model for McDonald's and other businesses seeking to leverage NLP and machine learning techniques for improved customer engagement and business decision-making.

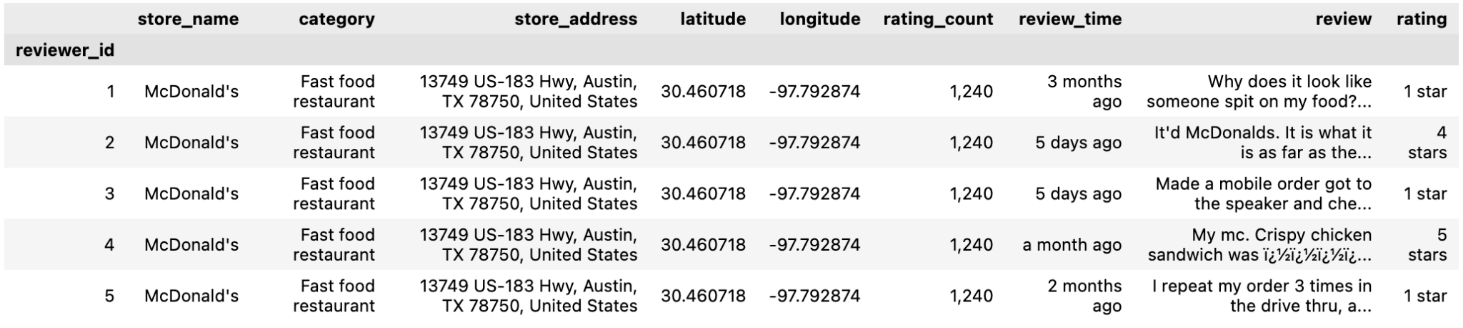
# Dataset Cleaning, Preprocessing and Visualizations

## Loading the Dataset

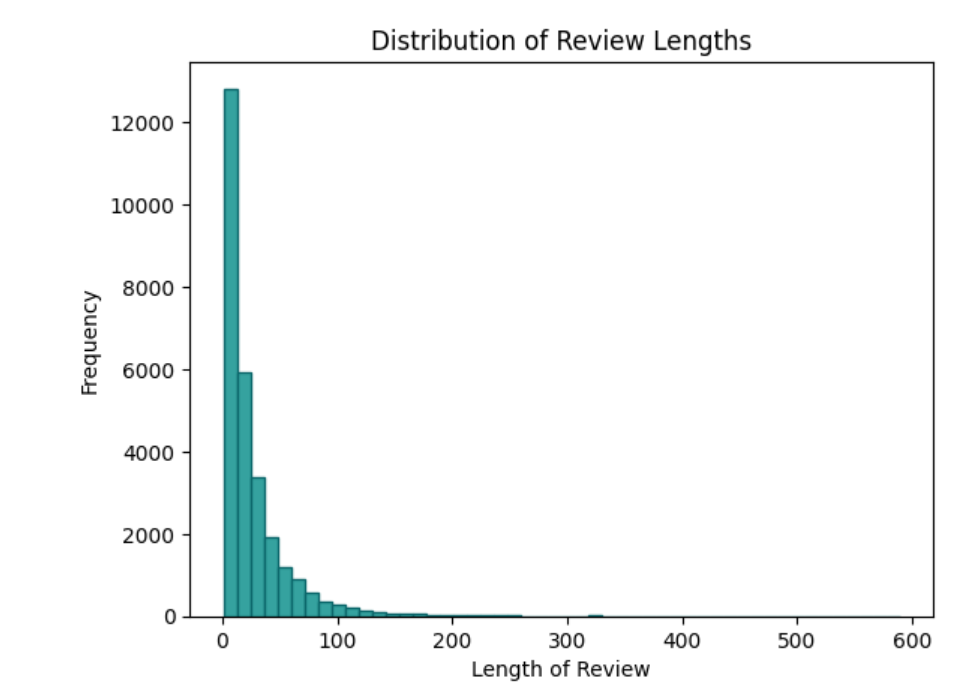
The data utilized in this research was gathered from a collection of over 33,000 anonymized Google reviews for various McDonald's locations within the United States (<https://www.kaggle.com/datasets/nelgiriyewithana/mcdonalds-store-reviews>). This dataset offers a glimpse into the diverse customer experiences at McDonald's restaurants and provides a window into broader consumer behavior and preferences. Key features of the dataset include:

* **reviewer\_id**: A unique identifier for each reviewer, ensuring anonymity.
* **store\_name**: The name of the specific McDonald's store under review.
* **category**: The category or type of store, providing context to the location.
* **store\_address**: The physical address of the store.
* **latitude**: The latitude coordinate of the store's geographical location.
* **longitude**: The longitude coordinate of the store's geographical location.
* **rating\_count**: The total number of ratings or reviews for the particular store.
* **review\_time**: The timestamp when the review was posted.
* **review**: The textual content of the customer's review.
* **rating**: The rating provided by the reviewer on a predetermined scale.

#### Figure 1: McDonald's Reviews Dataset

**** The review lengths range from a minimum of 1 word to a maximum of 589 characters. The most commonly used words are “I”, “food”, “order”, “service”, “The”, “McDonald”, “get”, “place”, “good”, and “time”.

#### Figure 2: McDonald's Review Lengths (Word Count)

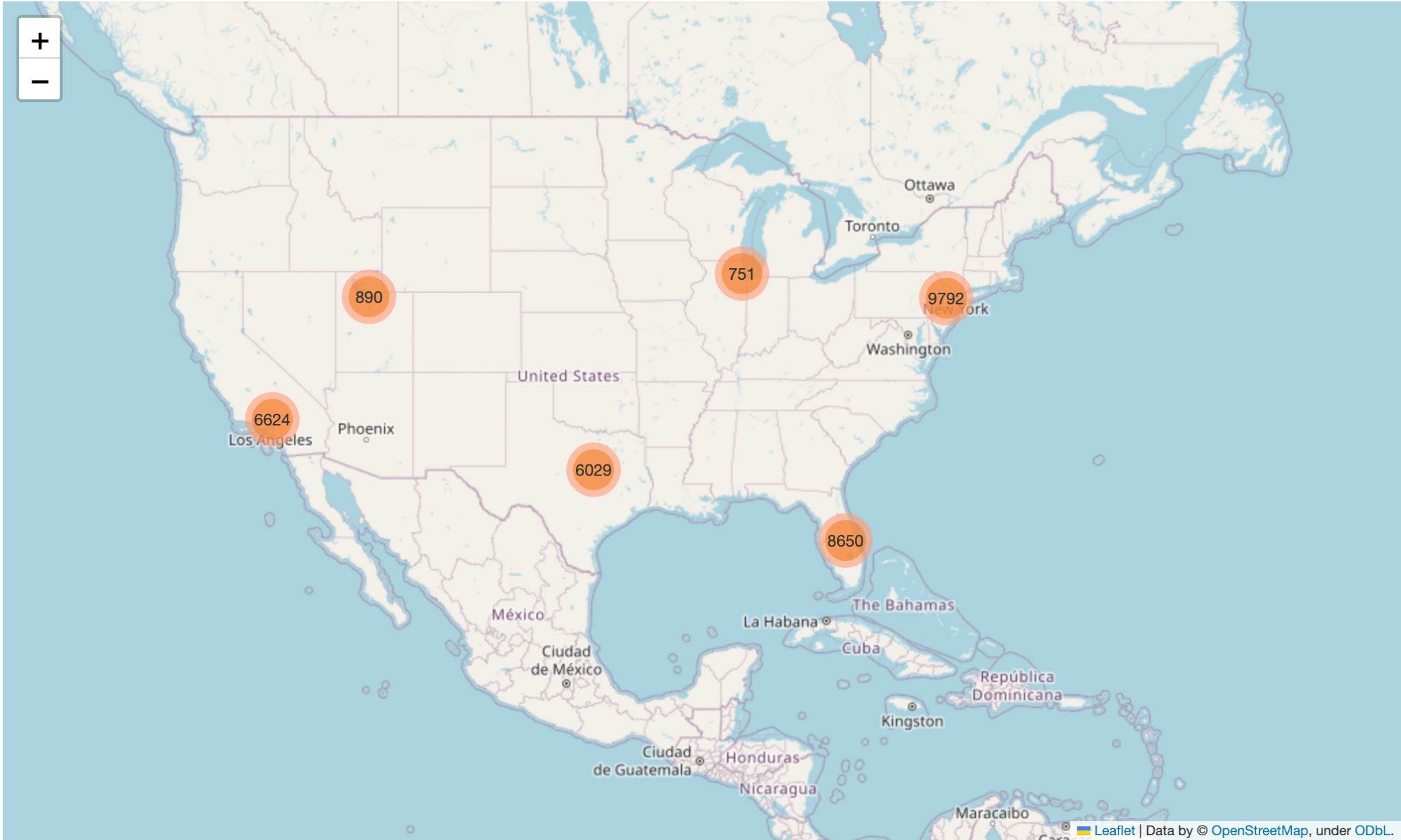


#### Figure 3: McDonald's Review Wordcloud



The dataset contains 33,000 anonymized Google reviews from 11 States.

#### Figure 4: Map of Review Locations



## Cleaning and Preprocessing

The cleaning process began with renaming the 'latitude ' column, which had an extraneous space, to 'latitude' to ensure consistency within the dataset. A new column, ‘rating\_int’, was created, converting the rating data from string to integer format. This transformation involved extracting the numeric portion of the rating and converting it into an integer data type, enabling more precise data manipulation and analysis. A ‘review\_string\_list’ column was added. This is the review text split into a list of its individual words, without punctuation. This is to facilitate the word counting. A ‘state’ column was created, extracting the two-letter state abbreviation from the 'store\_address' column. This extraction was performed using regular expression matching, allowing for the isolation of the two-letter state codes within the address text. Any address lacking a recognizable state code was assigned the value 'Unknown'. This new column enabled geographic data segmentation, enabling state-based analysis.

Several error characters, “mojibake” (e.g., '½ï', 'ï', '½', '¿') were present due to encoding errors from the dataset source. These characters were removed from the review texts of the dataset, as the data here is unrecoverable.

Present in the dataset were many reviews consisting of only one word. Upon examination, these reviews seem to be the default meaning of the star ratings. In case of reviews were only a star rating was given, with no accompanying text, the review defaulted to a one-word sentiment. The scale is 1 = “Bad” or “Terrible”, 2 = “Poor”, 3 = “Neutral”, 4 = “Good”, 5 = “Excellent”. Since we are attempting to conduct sentiment analysis and review prediction based on review text, these reviews were removed as they are directly correlative of sentiment.

One final step of data preprocessing that was implemented was adjusting the imbalance in the review ratings. The majority of star ratings were either 1 or 5, with 2, 3, and 4 individually comprising a low amount. For the SVM classifier model, we decided to combine 2, 3, 4 into one “Neutral” category, and created a sentiment scale for the model to class. This being 1 = “Terrible”, 2 = “Neutral”, 3 = “Excellent”.

After cleaning and preprocessing was done, the resultant dataset contained 28228 reviews.

# Score Prediction and Sentiment Analysis

### Text Encoding and Train/Test Split

The Google Universal Sentence Encoder was used to embed the review text. This encoder converts the text reviews into a numerical format that can be fed into the models used in this study.

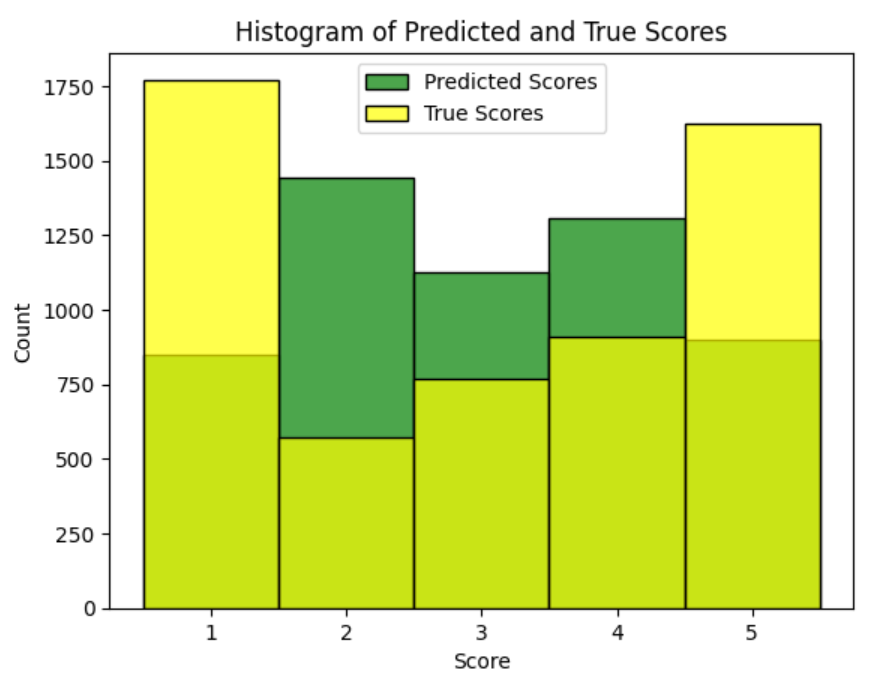
The dataset was divided into training and testing subsets, with 80% allocated for training and 20% for testing. The training set contained 22582 samples, and the testing set included 5646 samples.

### Ridge Regression Model

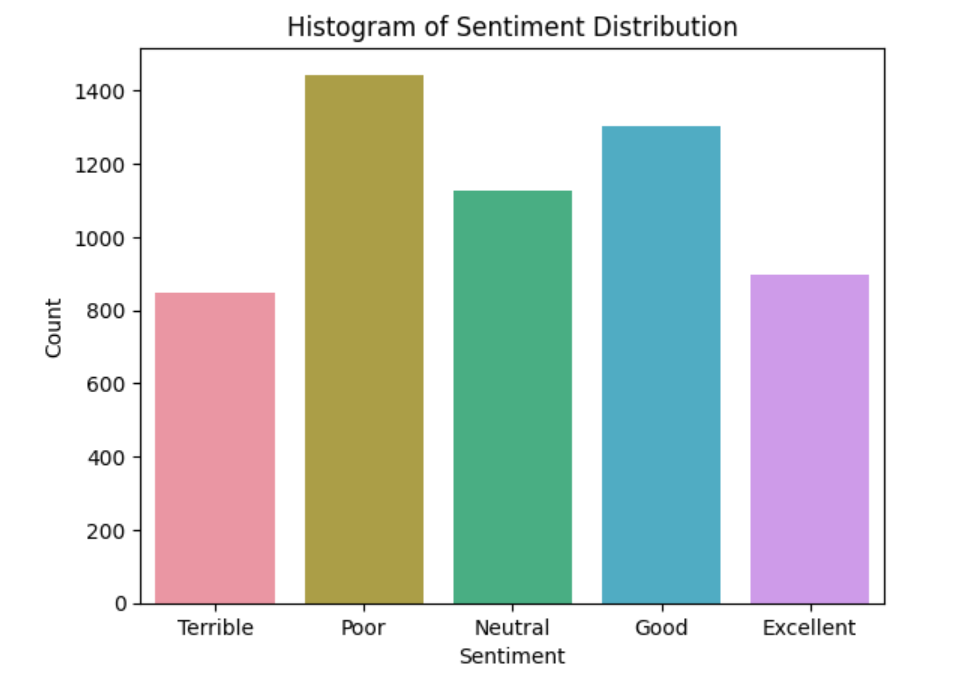
The Ridge Regression model was trained using the embedded review data. The coefficient of determination (R² score) for the Ridge Regression model was 0.634. This metric indicates the proportion of the variance in the dependent variable (ratings) that can be predicted from the independent variable (review text). The score indicates that about 63.4% of the variation in scores are accounted for.

Figure 5 presents a histogram comparing the predicted and actual scores. The green bars represent the predicted scores, while the yellow bars represent the actual scores. This visual representation provides insights into how closely the predictions align with the actual ratings. The histogram shows that the model overestimates the amount of ratings in the 2-4 range while underestimating the 1 and 5 reviews. This is likely due to the imbalance of 1 and 5 reviews versus 2, 3, and 4. As well as the fact that Ridge regression estimates an exact floating point score, as opposed to an integer, which has implications in rounding. There is some subjectivity in interpretation as to what exactly constitutes a score of 1. For instance, 1.4 would round down to 1 and 1.5 would round up to 2. While the Ridge model can be used to predict the star rating, the added complexity of a continuous scale poorly reflects the reality of user ratings.

#### Figure 5: Histogram of Predicted and True Scores (Ridge Model)



#### Figure 6: Histogram of Sentiment Distribution (Ridge Model)



#### Using the original sentiment scale, a visualization is present of the Ridge regression model prediction of sentiment.

## Bayesian Model

Bayesian Ridge Regression was employed in this project to predict review ratings, to compare to the simple Ridge Regression. This statistical model extends traditional ridge regression by utilizing a Bayesian model. The model was trained using the same training and test split used in the Regression model.

The coefficient of determination (R²) for the Bayesian Ridge model on the test data was 0.6341, indicating that the model could explain approximately 63.41% of the variance in the observed ratings. This score is very close to that of the Ridge model, suggesting that both models perform similarly in capturing the underlying pattern within the dataset.

## SVM (Support Vector Machine) Classifier.

As was noted before, the Ridge and Bayesian Ridge model performed poorly in accurately reflecting the sentiment of reviews, even if they could predict the review rating with some accuracy. A SVM model was implemented, with correcting the data imbalance and combining the scarcer middle scores into one category. The new review to sentiment scale is   
1 = “Terrible”, 2 = “Neutral” (combining 2, 3, 4 star scores), 3 = “Excellent”. After training the model using the normalized dataset, the model predicted the sentiment classification with 71.7% accuracy. The predicted sentiment more accurately reflected the ranges of the user reviews. The SVM did take longer to train and predict than the regression models.

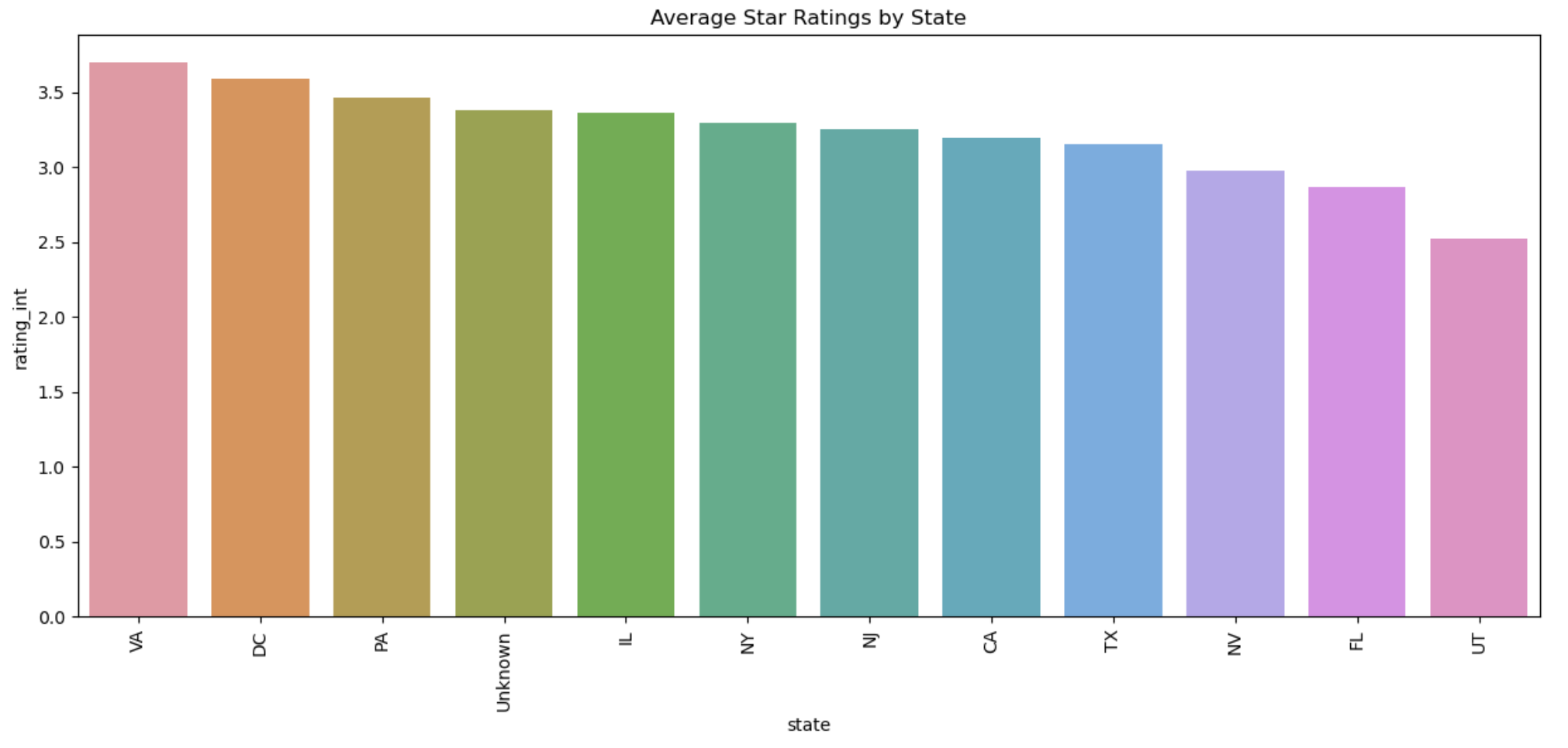
### Figure 7: Histogram of Predicted and True Scores (SVM)

# Geolocation Analysis

A geolocation analysis was conducted to understand the geographical patterns of customer satisfaction within the dataset, focusing on the average star ratings across different states in the United States.

By grouping the reviews according to the states and then calculating the mean rating for each state, a clear picture of regional variations in customer satisfaction emerged. A barplot was then used to visualize the results (see Figure 6).

#### Figure 7: Average Star Ratings by State



# Results and Discussion

The implementation of the Ridge regression model aimed at predicting review scores using the Google Universal Sentence Encoder embeddings. This approach resulted in a coefficient of determination (R² score) of 0.634. While the score prediction was less accurate due to the data imbalance, and rounding of precise score numbers that did not reflect the integer model of the reviews.

The Bayesian Ridge model was also implemented, resulting in an R² score of 0.6341. The proximity of these scores suggests a similar level of performance between the Ridge and Bayesian Ridge models in predicting review ratings.

The SVM model performed the best in terms of accuracy. After correcting the data imbalance, creating a new sentiment range and having the model be a classifier all contributed to the model more accurately reflecting the sentiment expressed by the reviews. SVM performance in terms of compute time was much longer than the regression models

Analysis of the average star ratings by state revealed specific geographical patterns in customer satisfaction. The visual representation provided insights into regional customer experience differences, potentially reflecting cultural preferences, management practices, and socio-economic conditions.

To further illustrate the uses of the rating model created in this project, an example web application has been created that uses our model to update a 5-star user interface component based on the review text entered into the review box.

***Figure 7: Star Prediction User Interface***



# Conclusion

This project successfully applied and evaluated various predictive models to a comprehensive dataset of over 33,000 McDonald's reviews. The study detected individual customer opinions and geographical trends by integrating text data and geographical insights.

The Ridge and Bayesian Ridge models demonstrated that predicting review scores using review text was possible with some accuracy. However, the predictions are continuous scores which do not map to integer reviews one to one, and may not accurately reflect the general sentiment of the review. The SVM model, with correcting data imbalance and implementing it as a classifier of sentiment, more accurately predicted sentiment expressed by the reviews.

Overall, the project illustrates the power of combining diverse data sources and analytical techniques to derive meaningful insights into customer experience. Our model can predict review ratings and sentiment based on review text, and has highlighted geological data that future studies can utilize.

Further areas of research and exploration include using models expressly built for sentiment analysis to improve accuracy and be able to extract more nuance from sentiment analysis. SentiWordNet and BERT are models that remember context of words. Tokenization techniques like only including important words may also improve accuracy (Hariharan et al., 2023).

# 

# References

Russell, S. & Norvig, P. (2021). Artificial Intelligence, A Modern Approach (4th). Pearson.

Hariharan, K. et al. (2023). Analyzation of sentiments of product reviews using Natural

Language Processing*. 2023 9th International Conference on Advanced Computing and*

*Communication Systems (ICACCS),* 1851–1854.

doi:10.1109/icaccs57279.2023.10112777

# Appendix

Presentation:

<https://vimeo.com/854560121?share=copy>

Github:

<https://github.com/p-parks/AAI-501-Final-Project>