AAI-501-02-SU23 - Introduction to AI - Final Project

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 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 based on review text. We applied several machine learning models, including linear regression, and assessed their performance using suitable metrics. 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.

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, and predict star ratings 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.

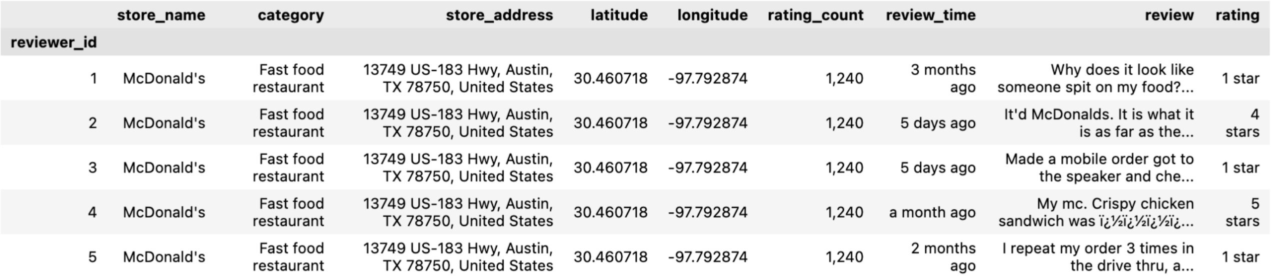
Dataset and Cleaning

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. 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



	store_name	category	store_address	latitude	longitude	rating_count	review_time	review	rating
reviewer_id									
1	McDonald's	Fast food restaurant	13749 US-183 Hwy, Austin, TX 78750, United States	30.460718	-97.792874	1,240	3 months ago	Why does it look like someone spit on my food?...	1 star
2	McDonald's	Fast food restaurant	13749 US-183 Hwy, Austin, TX 78750, United States	30.460718	-97.792874	1,240	5 days ago	It'd McDonalds. It is what it is as far as the...	4 stars
3	McDonald's	Fast food restaurant	13749 US-183 Hwy, Austin, TX 78750, United States	30.460718	-97.792874	1,240	5 days ago	Made a mobile order got to the speaker and che...	1 star
4	McDonald's	Fast food restaurant	13749 US-183 Hwy, Austin, TX 78750, United States	30.460718	-97.792874	1,240	a month ago	My mc. Crispy chicken sandwich was ï¿½ï¿½ï¿½ï¿...	5 stars
5	McDonald's	Fast food restaurant	13749 US-183 Hwy, Austin, TX 78750, United States	30.460718	-97.792874	1,240	2 months ago	I repeat my order 3 times in the drive thru, a...	1 star

The review lengths range from a minimum of 1 character to a maximum of 3115 characters. The most commonly used words are “I”, “food”, “order”, “service”, “The”, “McDonald”, “get”, “place”, “good”, and “time”.

Figure 2: McDonald's Review Lengths

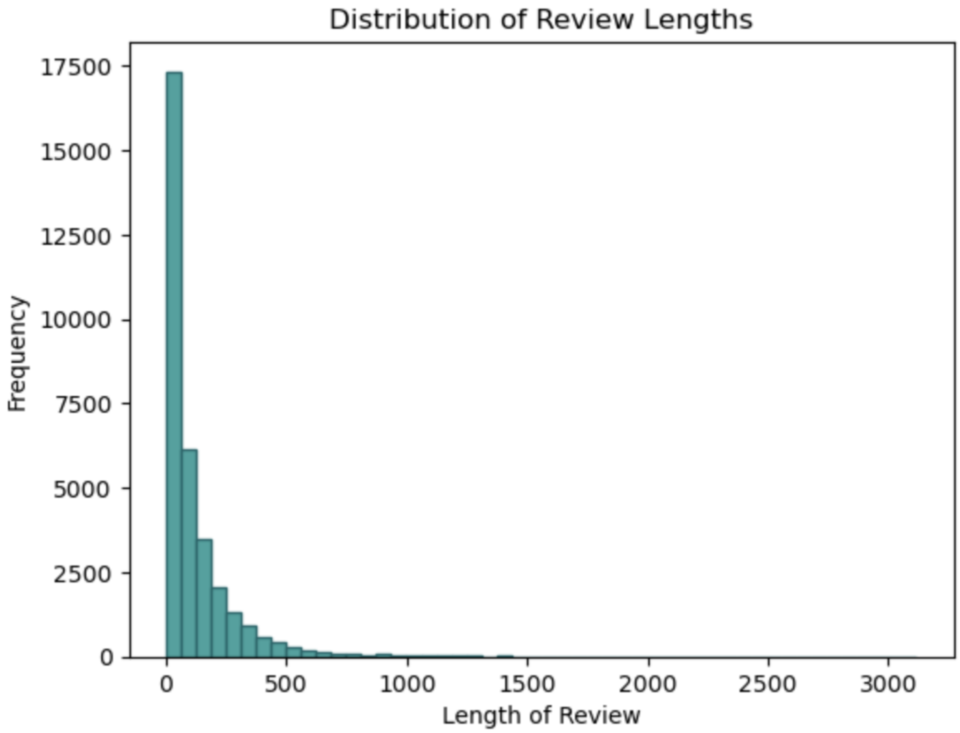
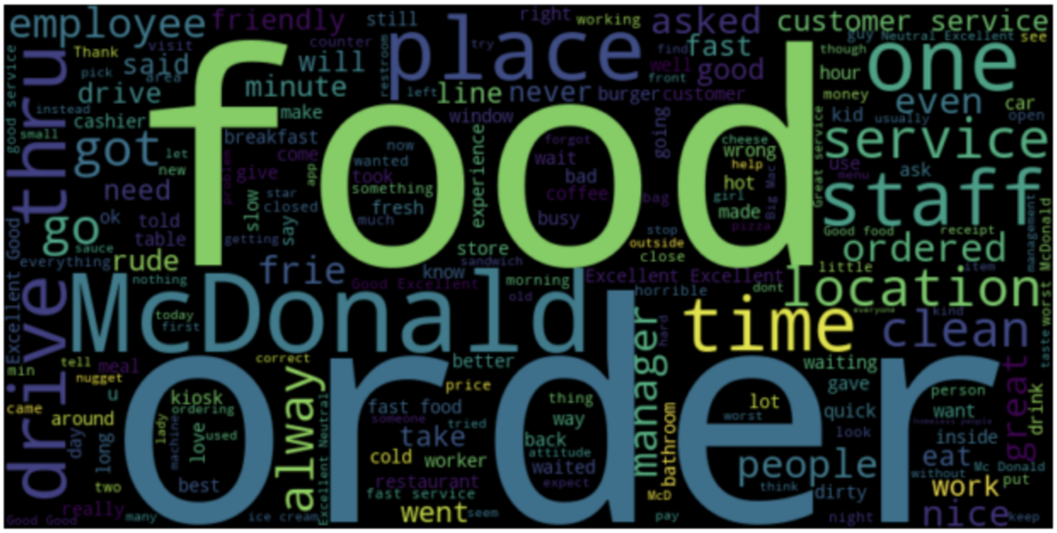
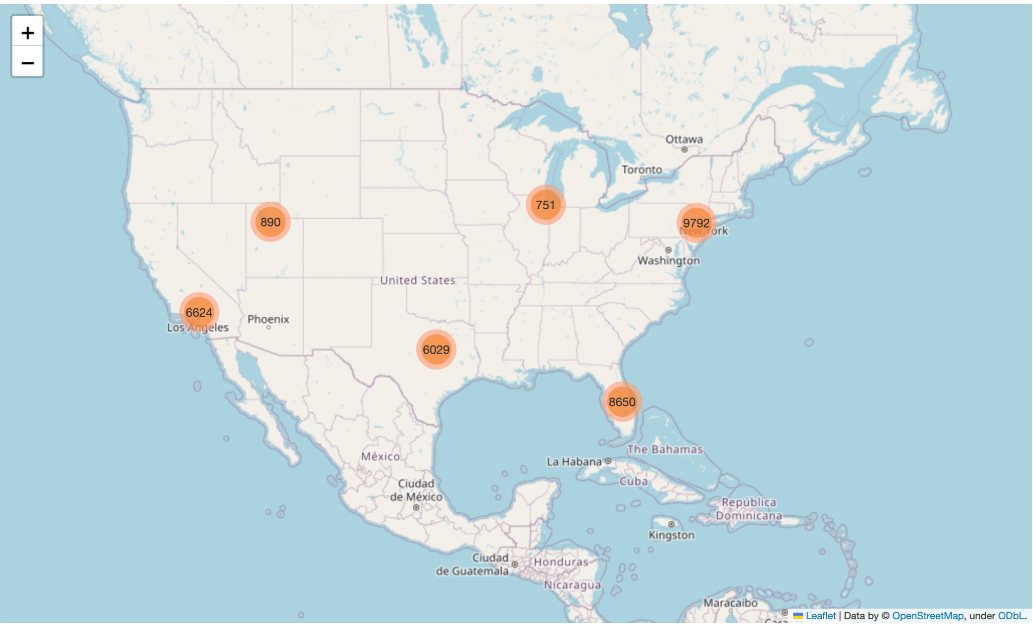


Figure 3: McDonald's Review Wordcloud



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

Figure 4: Map of Review Locations



Cleaning

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.

Several unwanted characters (e.g., '½ï', 'ï', '½', '¿') were identified in the 'review' and 'store_address' columns. These characters were removed from the dataset as they did not contribute meaningful information.

Finally, a new '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.

The cleaning process significantly enhanced the dataset's usability, preparing it for machine learning and analysis tasks.

Sentiment Analysis

Ridge Regression Model

Training and Testing Split

The dataset was divided into training and testing subsets, with 80% allocated for training and 20% for testing. The training set contained 26,716 samples, and the testing set included 6,680 samples.

Model Training

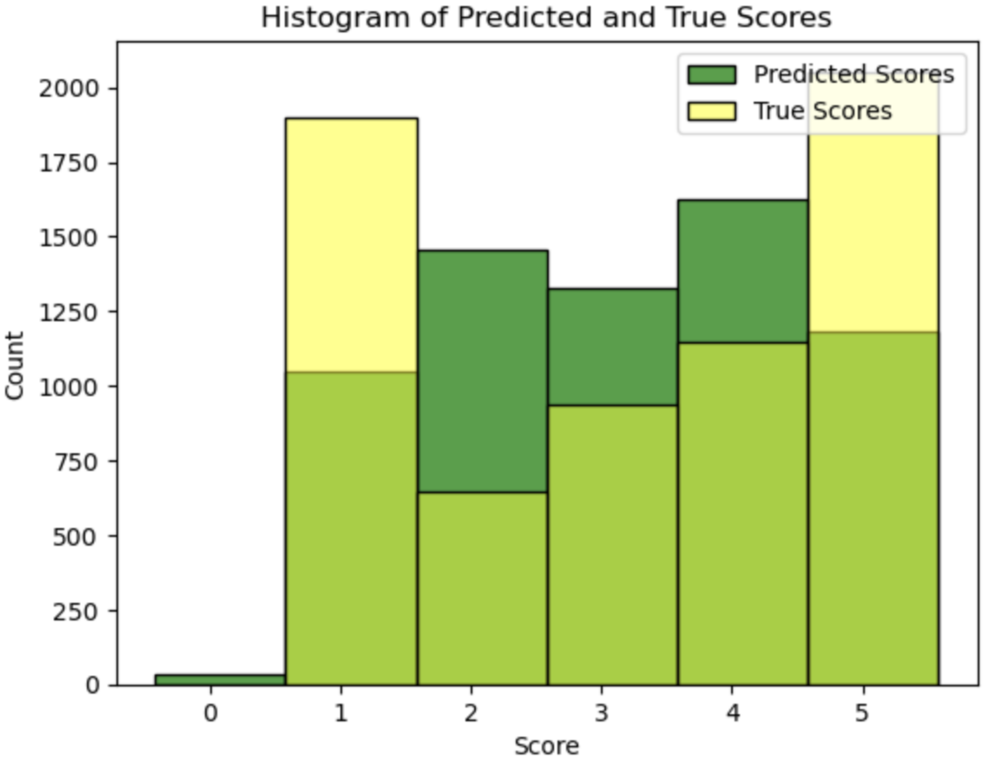
The Ridge Regression model was trained using TensorFlow and TensorFlow Hub to embed the reviews using the Google Universal Sentence Encoder. This encoder converts the text reviews into a numerical format that can be fed into the Ridge Regression model.

Performance

The coefficient of determination (R^2 score) for the Ridge Regression model trained on embeddings was 0.67705. This metric indicates the proportion of the variance in the dependent variable (ratings) that can be predicted from the independent variable (review text). The R^2 score of 0.67705 indicates a good fit since it is close to 1.

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 predicted ratings are higher in the 2-4 range while the actual ratings are higher in the 1 and 5 range.

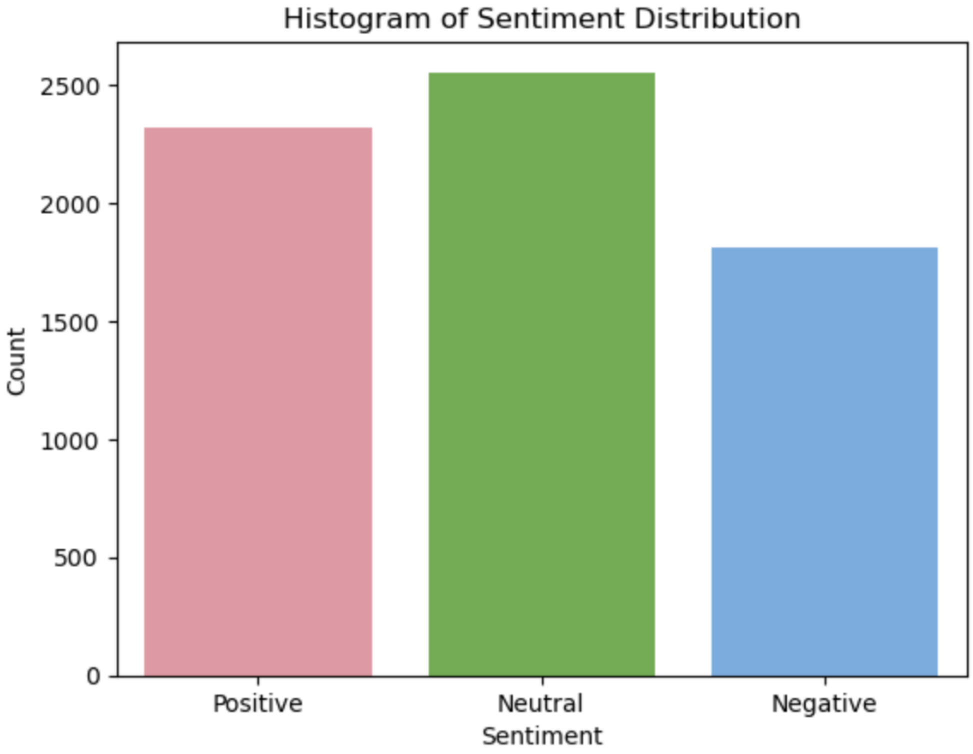
Figure 5: Histogram of Predicted and True Scores



Sentiment Analysis

The trained Ridge Regression model was used to calculate the sentiment of sample reviews. The sentiment was categorized into "Positive," "Neutral," or "Negative" based on the predicted rating. The sentiment was calculated by labeling a star rating greater than four as "Positive," a star rating between two and four as "Neutral," and a star rating lower than two as "Negative."

Figure 6: Histogram of Sentiment Distribution



Bayesian Model

Bayesian Ridge Regression was employed in this project to predict review ratings. 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.

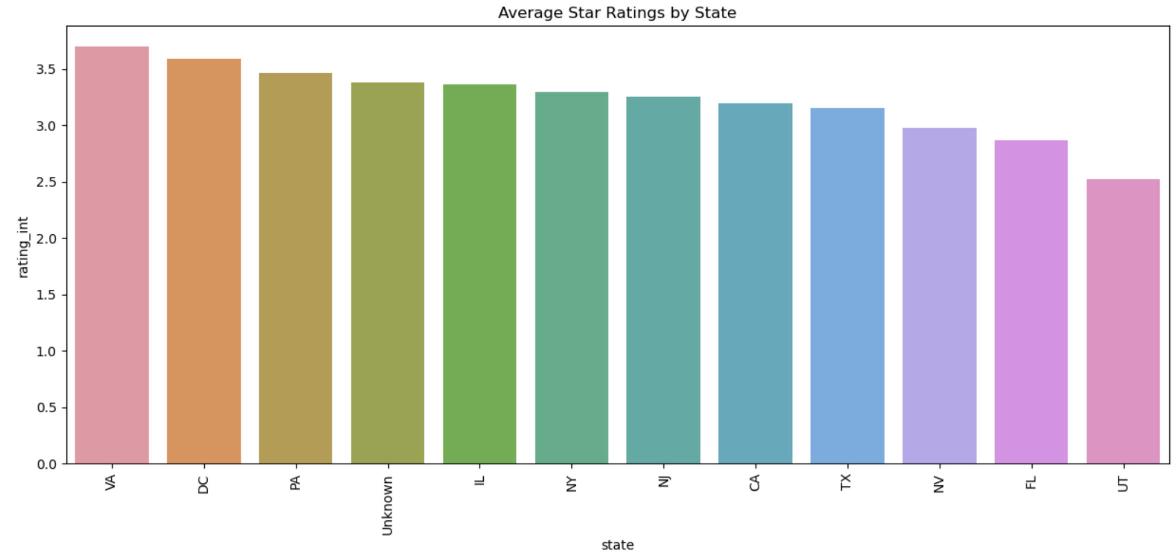
The coefficient of determination (R^2) for the Bayesian Ridge model on the test data was also 0.67711, indicating that the model could explain approximately 67.71% of the variance in the observed ratings. This score is very close to that of the Ridge model, suggesting that both models perform similarly in capturing the underlying pattern within the dataset.

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

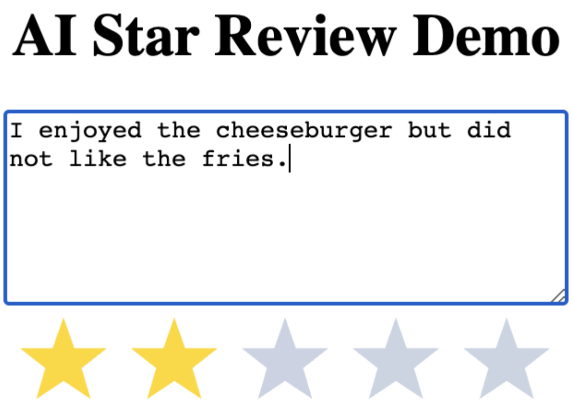
The implementation of the Ridge regression model aimed at predicting review scores using the Google Universal Sentence Encoder embeddings. This approach resulted in a coefficient of determination (R² score) of 0.67705, indicating a good fit of the model to the data.

The Bayesian Ridge model was also implemented, resulting in an R^2 score of 0.67711. The proximity of these scores suggests a similar level of performance between the Ridge and Bayesian Ridge models in predicting review ratings.

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



AI Star Review Demo

I enjoyed the cheeseburger but did not like the fries.

★★☆☆☆

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 promising results, though further tuning and additional features might improve their predictive accuracy. The exploration of geolocation data also highlighted potential areas for future research and practical application.

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 based on review text and has highlighted sentiment and geological data that future studies can utilize.

References

Russell, S. J., Norvig, P., &amp; Davis, E. (2022). Chapters 21, 23, 24. In Artificial Intelligence: A modern approach.Pearson Education.

Appendix

Presentation:

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

Github:

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