Sentiment Analysis of Yelp Reviews

Sonal Naveen Meda,

Owais Shaikh,

Sai Swetha Vadrevu

University of Maryland, Baltimore County DATA

690 Natural Language Processing

Dr. Antonio Diana

8th December 2023

Sentiment Analysis of Yelp Reviews

Abstract

The abundance of online reviews has significantly influenced customer decision-making, especially in the food service industry. This project delves into the intricate relationship between the ratings and the corresponding sentiment expressed in Yelp reviews. Through sentiment analysis, the study seeks to uncover how quantitative evaluations correlate with qualitative feedback, hypothesizing that positive sentiments are associated with higher ratings and vice versa. Employing machine learning algorithms such as the Generalized Linear Model (GLM), XGBoost Regressor, and Gradient Boost Regressor, the project evaluates this relationship using features consisting of review text, rating, and review length to predict the sentiment scores. The findings aim to provide deeper insights into consumer behavior and enhance understanding of the role of star ratings in reflecting customer experiences.

Keywords: Reviews, Ratings, XGBoost, VADER, Gradient Boost Sentiment Analysis, Yelp, Reviews, NLP.

Introduction

In the current digital era, online reviews, especially those from Yelp and related websites, significantly impact customer decisions, especially in the hospitality and service sectors. Star ratings have become a crucial tool for decision-making in food delivery applications (Xu et al., 2015). The problem is that these quantitative assessments frequently oversimplify the intricate feelings expressed in reviews, missing the overall complexity of the customer experience (Lak & Turetken, 2014).

This study aims to bridge the gap between quantitative evaluations and qualitative

feedback by analyzing the complex link between the feelings expressed in Yelp reviews and the star ratings that correlate with such reviews. The study is driven by the hypothesis that a significant relationship exists between the sentiment expressed in Yelp reviews and the star ratings assigned by users. Specifically, it is posited that reviews expressing positive sentiments will correlate with higher star ratings, while negative sentiments will correspond with lower star ratings (Sharma & Dutta, 2021; Kaviya et al., 2017). This hypothesis aims to explore the extent to which quantitative ratings reflect the qualitative sentiments of consumers.

The methodology involves cleaning and pre-processing the data to implement and test machine learning models like GLM, XGBoost, and gradient boosting regression on a sample of Yelp reviews to predict sentiment scores. The features are review length, rating, and review text. The predicted sentiment score was compared with the compound score obtained using VADER. Key metrics such as R-squared error, mean squared error (MSE) and mean absolute error (MAE) will be employed to assess model performance.

The findings of this project are expected to offer valuable insights into consumer behavior and decision-making processes in the context of online reviews. This can translate to enhanced customer engagement and service improvement strategies for businesses. For consumers, it provides a deeper understanding of what star ratings may signify regarding actual customer experience.

Literature Review

The interplay between star ratings and textual sentiments in consumer reviews has been a topic of significant interest in recent research. Al-Natour and Turetken (2020) comprehensively assessed how sentiment analysis tools compare with traditional star ratings in reflecting consumer opinions. Their study emphasized the effectiveness of sentiment

analysis in capturing the underlying tone of content, suggesting its potential as a complement or alternative to star ratings. This aligns with the current project's objectives, which seek to explore the relationship between textual sentiments in Yelp reviews and the corresponding star ratings. Similarly, Xu, Wu, and Wang (2015) specifically focused on Yelp's ratings, examining how text reviews can be used for sentiment analysis, offering direct relevance to the project's context.

Furthermore, the aspect-level sentiment analysis discussed by Qiu, Liu, Li, and Lin (2018) is particularly pertinent for extracting detailed sentiments from reviews. Their approach to predicting ratings for non-rated reviews using sentiment analysis can be adapted to understand and correlate sentiments in Yelp reviews with given star ratings. Lak and Turetken (2014) also contributed to this field by comparing explicit (star ratings) and implicit (sentiment analysis) measures of opinions, a comparison that is central to the project's hypothesis.

Wan (2022) provided a unique perspective by analyzing sentiment analysis of star ratings from a cross-cultural viewpoint, highlighting the importance of considering cultural factors in sentiment analysis. This factor might be significant in analyzing Yelp reviews. In addition, the work of Sharma and Dutta (2021) on developing a domain-specific sentiment lexicon using star ratings of reviews introduces a novel approach that can enhance the accuracy of sentiment prediction in this project. Kaviya, Roshini, Vaidhehi, and Sweetlin (2017) further complement this body of knowledge by demonstrating how sentiment analysis can be applied explicitly to restaurant ratings, closely mirroring the Yelp review context the project focuses on.

These studies provide a robust theoretical foundation for the project, offering insights into

various methodologies and approaches that can be leveraged to analyze and predict sentiments in Yelp reviews based on star ratings. Their findings and methods will be instrumental in guiding the project's approach toward understanding the intricate relationship between textual sentiment and numerical ratings in consumer reviews.

Methodology

Data Collection and Loading

df_final	**dfr**	**Selected Fields from the review.json file**	**Description**
		business_id	Unique identifier for each business
		stars	Rating given by a reviewer
		text	The text content of the review
		length	Length of the review
	dfb	**Selected Fields from the business.json file**	**Description**
		business_id	Unique identifier for each
		categories	Types of services or goods the business offers
		business name	Name of the business

Table 1: Overview of Selected Fields from Yelp's Dataset for Final Data Frame

We sourced our data from the Yelp Dataset, which is tailor-made for academic research. We used ‘business.json, ' which contained detailed information about various businesses, and ‘review.json,’ which included customer reviews for these businesses. The data from these files were loaded into two Data Frames. The dataset was streamlined by selecting specific columns

relevant to our study. Table 1 provides an overview of the different chosen fields and the streamlined process.

Data Cleaning and Preprocessing

This phase was critical to ensure data quality and usability. The null values were dropped, and the two data frames were merged on 'business_id.' The length of the reviews was computed and added to the data frame. The hospitality and service industries which had over ten reviews were filtered and retained for better modeling and analysis.

EDA

Basic statistics, such as mean, standard deviation, and quantile range were computed and analyzed. The distribution of the reviews across different categories and hospitality and service industries was analyzed and visualized. This analysis provided profound insights into the distribution and nature of the reviews and highlighted popular categories.

Correlation analysis was conducted to explore the relationships between star ratings and the lengths of the reviews.

Word clouds were generated for positive and negative reviews to visually represent the most common words in each category. This step provided insights into the prevalent themes and sentiments expressed in the reviews.

Review Text Processing

First, the review text was tokenized. Second, all the characters in the text were converted to lowercase to ensure uniformity, as text data is case-sensitive. Third, the punctuation and stopwords were removed. This was essential for focusing on significant words. Lastly, text contractions were expanded, and special characters were drawn to clean the data further. The list of words in the 'Reviews' column was converted to a continuous

string.

'Reviews,' 'Rating,' and 'Rev_len' were loaded into a data frame, 'final_df,' and saved as a CSV file for modeling and further analysis.

VADER Integration

Sentiment Analysis is used for detecting attitude or emotion. VADER (Valence Aware Dictionary for Sentiment Reasoning) will assign sentiment scores to the words in our Yelp reviews based on its existing dictionary. It maps the lexicon to the emotion intensities. For example, expressions like 'joy, smile, and 'happy' convey a positive sentiment. VADER is efficient in understanding the context of the statement as well. The subsequent result is a sentiment score assigned to each review in our data. The sentiment score, also called the compound score, reflects the intensity of the emotion associated with each review.

The range of the compound score is from -1 to 1; the interpretation of the range is as follows.

- Positive (0.05 to 1): Indicates a positive sentiment, with higher values signifying greater positivity.
- Neutral (-0.05 to 0.05): Suggests a neutral sentiment, neither positive nor negative.
- Negative (-1 to -0.05): Points to negative sentiment, with lower values indicating more negativity.

The compound score for our reviews was computed and included as a new column in the final data frame.

TF-IDF Vectorization

Each word in the review acts like a feature, but each review also has unnecessary features that might not be relevant to our model.TF-IDF (Term Frequency-Inverse Document Frequency) is a numerical statistic that reflects that a word is essential to a document in a

collection or corpus.TF-IDF gives higher weights to terms frequently used in a specific review but is not standard across all reviews. This helps identify words that are distinctive to a particular review and might carry sentiment or topic-specific information.

TF-IDF will return a sparse matrix with all the words in the review converted to vectors with a specific weight. Since we use ratings and the review length as a feature, we convert these to a sparse matrix and combine them to form one feature for modeling. Our dataset is now ready for modeling; we have our features, i.e. Reviews, ratings, review length combined, and the target variable (sentiment score).

Model Implementation

All our models follow the typical regression setup. The features are processed and as mentioned in the TF-IDF section, we have one combined input feature of Reviews, Ratings, and Review length, and our target variable is the Sentiment score.

The feature set is heterogeneous as it has ordinal values (ratings), integers (review length), and text (review). The inherent flexibility of the Generalized Linear Model (GLM) provides a robust foundation for regression analysis of heterogeneous features. This is also true for Gradient Boost and XGBoost.

Gradient boosting is a powerful ensemble technique that iteratively builds an additive model by minimizing the expected value of a given loss function, refining its predictions with a series of weak learners, typically decision trees, to achieve high accuracy and efficiency in various machine learning tasks (Bentéjac et al.,2020).

XG Boost is a novel algorithm tailored for sparse data and approximate tree learning, optimizes cache access patterns, data compression, and sharding, showcasing scalability beyond billions of examples while utilizing significantly fewer resources than existing

systems (Chen & Guestrin, 2016).

The dataset was split into training and test datasets; we implemented an 80-20 split. GLM, Gradient Boost, and XG Boost were trained and evaluated.

Model Evaluation

We evaluated the models and compared the results. The Mean Squared Error (MSE) measures the model's accuracy in predicting the target variable, quantifying the variance in prediction errors (James et al., 2013). R-squared (R^2) indicates how well the model's predictions approximate the actual data by explaining the variance in the target variable (Freedman, 2009). Mean Absolute Error (MAE) directly measures the magnitude of prediction errors (Hastie et al., 2009). Root mean square error (RMSE) and explained variance were also computed.

Hypothesis Testing

From our exploratory data analysis, it was evident that our data was not normally distributed. Therefore, a nonparametric test like Sprearman’s rank correlation was used to test the hypothesis.

Results

EDA

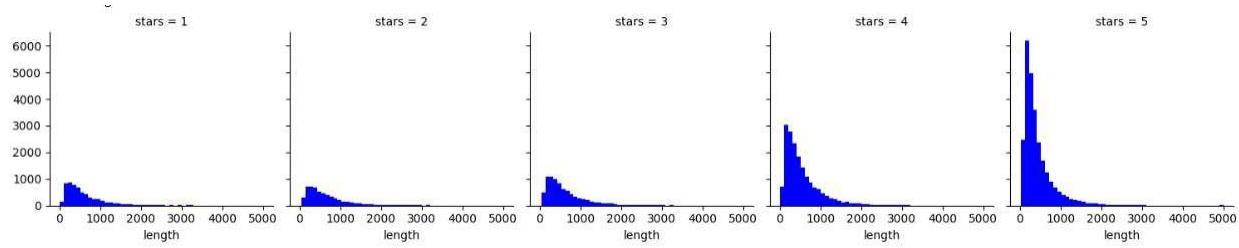

Figure 1: Count plot of stars vs length of the review

There was a relationship observed between the number of stars and the length of the

review. In Figure 1 we can see that reviews with higher star ratings were usually more verbose compared to the reviews with lower star ratings.

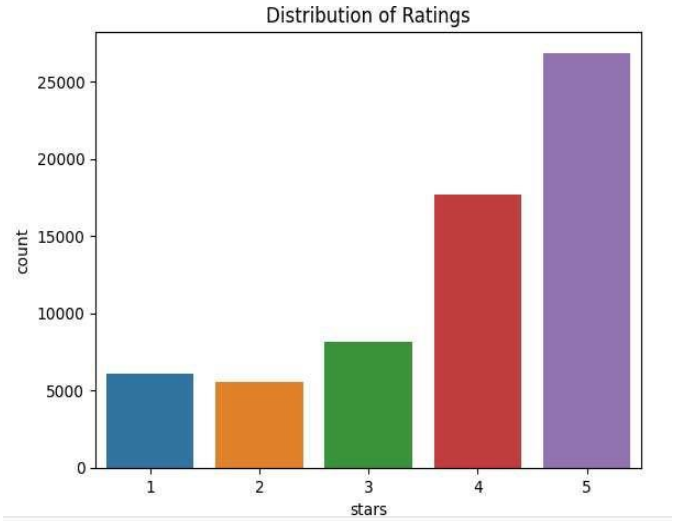

Figure 2: Plot for frequency of ratings

Figure 2 indicates that the dataset has significantly more reviews with 4-star and 5-star ratings than 1,2 and 3-star ratings.

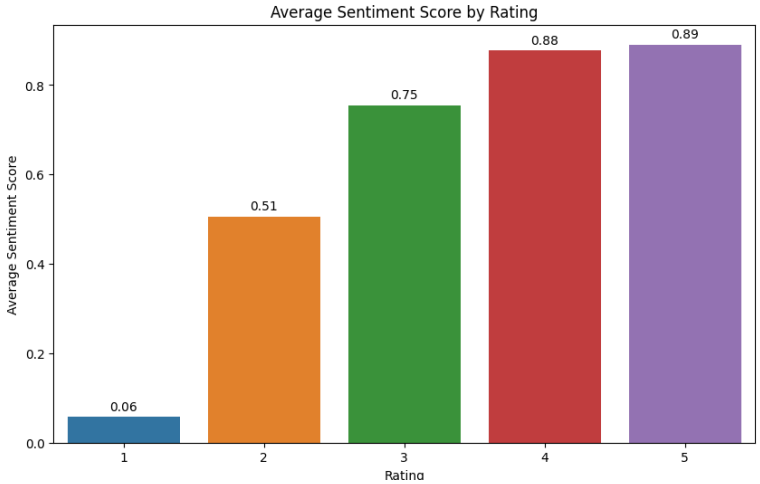

Figure 3: Average of the sentiment score for each rating

From Figure 3, we can infer that the average sentiment scores correspond to ordinal ratings from 1 to 5. The jump in sentiment scores is substantial between ratings 1 (0.06) and 3 (0.75). Whereas, the difference between ratings 4 (0.88) and 5 (0.89) is minimal. This pattern suggests that while there is a clear positive trend in sentiment as ratings increase, the distinction in perceived sentiment becomes less pronounced between the highest ratings.

Sample Analysis: Burger King

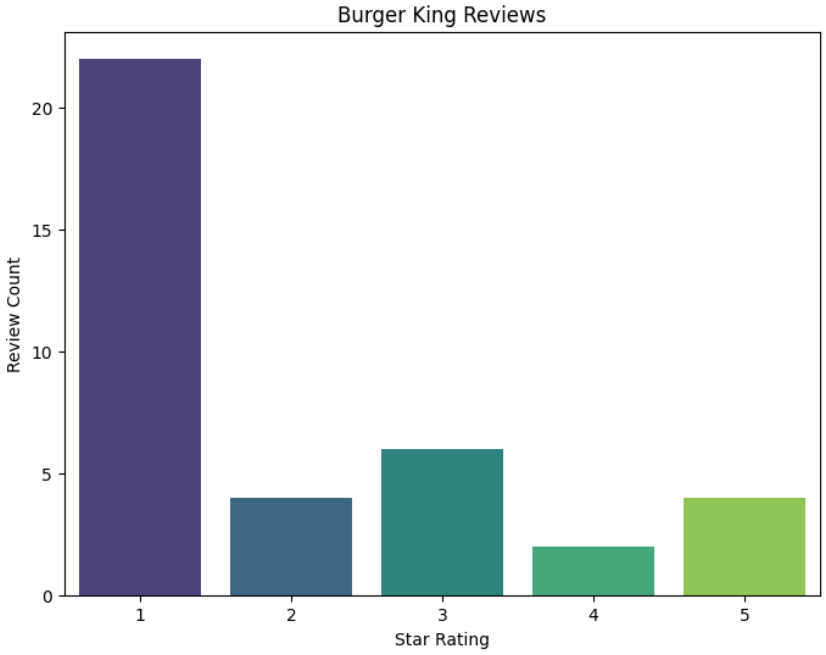

Figure 4: Ratings for Burger King

We analyzed a sample from our dataset, by filtering a specific business i.e. 'Burger King'. A chain restaurant like Burger King has generally bad ratings in the United States (Burger King, Consumer Affairs). The figure shows the review count of Burger King, we can see that it has more reviews with a lower star rating.

Figure 5: Word cloud for Positive Reviews

Figure 6: Word cloud for Negative Reviews

In Figure 5, the word cloud represents reviews the 5-star ratings and in Figure 6, the word cloud represents reviews the 1-star ratings. The sentiment analysis of the reviews can aid in recognizing the quantitative relationship between the ratings and the sentiment of the reviews. This can assist in improving business strategies, customer satisfaction, and engagement.

Model Evaluation

Lower values of MAE (Mean Absolute Error) and MSE (Mean Squared Error) indicate a

model's enhanced precision in forecasting sentiment scores, while a reduced RMSE (Root Mean Squared Error) signifies that the model's predictions are generally closer to the actual values. Conversely, a higher R-squared value implies that the model has a greater capability to encapsulate the variance in sentiment scores, effectively capturing the underlying patterns and trends. Additionally, a greater explained variance score denotes that the model has a strong ability to predict the sentiment scores with minimal deviation from the true data points.

Table 2: Model Evaluation results

Metric	**GLM**	**Gradient Boosting**	**X Gradient Boosting**
Mean Squared Error	0.067269	0.070790	0.059425
Root Mean Squared Error	0.259363	0.266063	0.243772
R-squared	0.611791	0.591474	0.657060
Mean Absolute Error	0.170324	0.152701	0.134581
Explained Variance	0.611798	0.591475	0.657066

Based on the results in Table 2, the X Gradient Boosting model consistently outperforms the other two models across all metrics, demonstrating the lowest error rates and the highest explanatory power. This suggests that it is the most effective model among the three for predicting the sentiment score based on review length, review text, and rating. The GLM performs moderately well, while the standard Gradient Boosting model lags slightly behind in this comparative analysis.

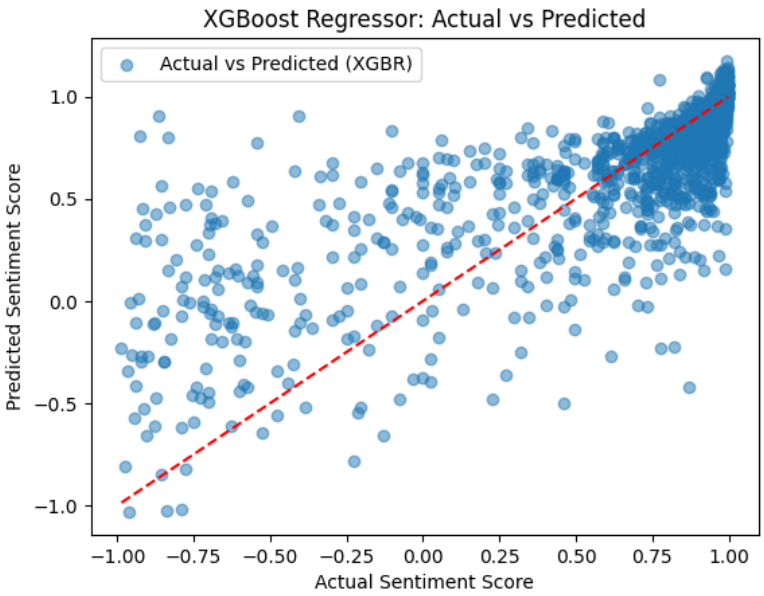

Figure 3: Scatter plot of actual vs predicted values for the XGB model

The scatter plot for the XGBoost Regressor indicates a strong correlation between actual and predicted sentiment scores, especially for higher sentiment values. The concentration of points along the line, representing the ideal match between predictions and reality, indicates a high degree of model accuracy. The model particularly excels at predicting positive sentiment scores, as evidenced by the dense clustering of points in this region.

Hypothesis Testing

Null Hypothesis (H0): No significant linear relationship exists between sentiment scores and star ratings.

Alternative Hypothesis (H1): A significant linear relationship exists between sentiment scores and star ratings.

Spearman's rank correlation coefficient was 0.47 with a p-value of 0.0 which negates our null hypothesis. The results suggest a significant relationship between the Rating and Compound sentiment score of the reviews.

Conclusion

The project successfully established a significant relationship between the sentiment expressed in Yelp reviews and star ratings, thus validating the hypothesis. The sentiment analysis, using VADER, revealed a nuanced understanding of customer feedback, while the machine learning models effectively predicted sentiment scores using user reviews and ratings as features. The models demonstrated varying degrees of performance. XG Boost outperformed Gradient Boost and GLM.

The quantitative relationship between sentiment score and ratings varies across regions, cultures, and sectors (Wan, 2022). The business strategies need to be tailored specifically by analyzing the sentiment of the reviews of the business that is in question. Our findings underscore the complexity of consumer opinions and the efficacy of combining qualitative and quantitative data for a comprehensive understanding. The study contributes to the broader understanding of customer sentiment in online reviews and can assist businesses in leveraging this information for improved customer engagement and service strategies.

References

Al-Natour, S., & Turetken, O. (2020).

A comparative assessment of sentiment analysis and star ratings for consumer reviews. International Journal of Information Management, 54, 102132. https://doi.org/10.1016/j.ijinfomgt.2020.102132

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2020).

A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, 54(3), 1937–1967. https://doi.org/10.1007/s10462-020-09896-5

Burger King. (n.d.). ConsumerAffairs.

Retrieved from https://www.consumeraffairs.com/food/burgerking.html

Chen, T., & Guestrin, C. (2016).

XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), August 13-17, 2016, San Francisco, CA, USA. ACM. https://doi.org/10.1145/2939672.2939785

Freedman, D. A. (2009). Statistical Models: Theory and Practice. Cambridge University Press.

Hastie, T., Tibshirani, R., & Friedman, J. (2009).

The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013).

An Introduction to Statistical Learning: with Applications in R. Springer.

Kaviya, K., Roshini, C., Vaidhehi, V., & Sweetlin, J. D. (2017).

Sentiment analysis for restaurant rating. In 2017, the IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy

and Materials (ICSTM) (pp. 140-145). Chennai, India: IEEE. https://doi.org/10.1109/ICSTM.2017.8089140

Lak, P., & Turetken, O. (2014). Star ratings versus sentiment analysis -- A comparison of explicit and implicit measures of opinions. In 2014, the 47th Hawaii International Conference on System Sciences (pp. 796-805). Waikoloa, HI, USA: IEEE. https://doi.org/10.1109/HICSS.2014.106

Qiu, J., Liu, C., Li, Y., & Lin, Z. (2018). Leveraging sentiment analysis at the aspect level to predict ratings of reviews. Information Sciences, 451–452, 295-309. https://doi.org/10.1016/j.ins.2018.04.009

Sharma, S. S., & Dutta, G. (2021). SentiDraw: Using star ratings of reviews to develop a domain-specific sentiment lexicon for polarity determination. Information Processing & Management, 58(1), 102412. https://doi.org/10.1016/j.ipm.2020.102412

Wan, Y. (2022). A sentiment analysis of star-rating: A cross-cultural perspective. In Proceedings of the 55th Hawaii International Conference on System Sciences.

Xu, Y., Wu, X., & Wang, Q. (2015). Sentiment analysis of Yelp's ratings based on text reviews. In 2015 17th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC) (Vol. 17, No. 1, pp. 117-120).

Yelp Inc. (2023). Yelp dataset. Retrieved from https://www.yelp.com/dataset