Unveiling Success: Sentiment Analysis and Restaurant Performance

Mustafa Elahi

October 5, 2024

1 Introduction

In the evolving landscape of the restaurant industry, online reviews have emerged as pivotal determinants of success. Platforms like Yelp have not only democratized customer feedback but also significantly influenced consumer choices and perceptions of quality. Previous studies, including those conducted by Luca (2016), have quantified the impact of numerical ratings on restaurant revenue, underscoring the tangible effects of digital word-of-mouth. However, less explored is the nuanced interplay between the sentiments expressed in these reviews and the indicators of restaurant success, such as ratings and business continuity. The sentiments—whether positive or negative—embedded within the text of customer reviews offer a rich layer of consumer insight, potentially providing a deeper understanding of what drives restaurant success beyond mere numerical ratings.

This research project seeks to examine how sentiments expressed in online restaurant reviews correlate with indicators of economic success, such as ratings and business continuity. Unlike previous studies that primarily focused on the quantitative aspects of consumer feedback (e.g., star ratings and review counts), this project delves into the qualitative dimensions of online reviews. By employing sentiment analysis alongside traditional econometric and machine learning techniques, including Ordinary Least Squares (OLS) regression and various clustering methods, this study aims to decode the linguistic patterns within reviews and their association with a restaurant's ability to thrive in a competitive market.

This project is timely and relevant, given the growing dependence of consumers on online reviews and the increasing pressure on restaurants to navigate their digital reputations effectively. As we delve into the nuanced effects of review sentiments on both restaurant ratings and operational longevity, our findings reveal that the impacts are not always as predictable as one might assume.

2 Data Sources and Descriptions

This study uses the Yelp Open Dataset, a publicly accessible dataset designed for academic research. It includes millions of reviews, business profiles, and user interactions from various global

cities, providing detailed insights into consumer behavior and business performance, especially in the restaurant industry. The research employs the `yelp_academic_dataset_business.json` and `yelp_academic_dataset_review.json` files, which include detailed business information and over 6 million reviews, respectively. Both datasets have been filtered during preprocessing to focus solely on restaurants with complete data.

To analyze the impact of review sentiments on restaurant ratings and longevity, variables were extracted from both data files and used to compile a single data file that was used for this research. Here are the variables contained in the new data file:

business id: a 22 character ID used to identify and distinguish different businesses from each other; this is used to match a restaurant's overall star-rating to each of its text reviews
is open: Whether the restaurant is open (1) or not(0).
stars: The average rating a restaurant receives on Yelp, ranging from 1.0 to 5.0.
review stars: The numerical rating a reviewer (who leaves a text review) gives a restaurant; ratings range from 1 to 5.

clean 'text': Text reviews of restaurant-goers experiences. This was originally extracted from the 'review' dataset as 'text', but became 'clean text' after being preprocessed.
sentiment: Sentiment of the reviews, expressed as 'POSITIVE' or 'NEGATIVE.' This is not extracted from any of the Yelp Open Dataset files; this was calculated using an algorithm after the preprocessing stage.

To provide a quantitative overview of the dataset utilized in this study, refer to Table 1 in the "Tables and Graphs" section.
Access to this dataset was obtained directly through the Yelp Dataset website $(https://www.yelp.com/dataset)$, where Yelp periodically releases updated versions of the dataset for research purposes.

3 Method

This study employs a blend of sentiment analysis, Ordinary Least Squares (OLS) regression, K-Means clustering, and Agglomerative clustering techniques to dissect the influence of online review sentiments on restaurant ratings and longevity. The inherent challenge of this analysis lies in the qualitative nature of review texts and the complex interplay between consumer sentiments and restaurant success indicators.

3.1 Sentiment Analysis

Firstly, we perform sentiment analysis on the review texts to categorize each review as either 'POSITIVE' or 'NEGATIVE'. This binary classification facilitates the quantitative assessment of review sentiments, setting the stage for further econometric analysis. The sentiment analysis is executed through a pre-trained model from the 'transformers' library developed by Hugging Face.

3.2 Ordinary Least Squares Regression

Following the sentiment classification, we apply OLS regression to model the relationship between review sentiments and overall restaurant ratings. The regression model is formalized as:

$$y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i \tag{1}$$

where y_i represents the average star rating of the i^{th} restaurant, X_{i1} is the sentiment of the review (1 for POSITIVE, 0 for NEGATIVE), and ε_i is the error term. The assumption here is that $E[\varepsilon_i|X_{i1}] = 0$, ensuring that the OLS estimates are unbiased. The coefficient β_1 illustrates the impact of review sentiment on the restaurant's average star rating, which addresses our research question on the qualitative influence of consumer feedback.

3.3 Clustering Methods

In order to identify groups within our restaurant data, we have applied two clustering techniques: K-Means and Agglomerative clustering. These methods sort restaurants into groups based on common features like whether they are still in business, their overall ratings, and the sentiments expressed in reviews.

K-Means clustering organizes restaurants into a pre-defined number of groups by minimizing the distance between the data points and the center of their assigned group. This helps us to see how restaurants are grouped based on review sentiments and their ratings.

On the other hand, Agglomerative clustering starts by treating each restaurant as a separate group and then progressively merges them into larger clusters. This is based on which restaurants are closest to each other, which allows us to see a hierarchy of groupings and understand the relationship between different variables in our data.

4 Results

4.1 OLS Regression Analysis

The Ordinary Least Squares (OLS) regression analysis provides quantitative evidence of the relationship between sentiment in reviews and the overall star rating of restaurants. The model's R-squared value of 0.101 indicates that approximately 10.1% of the variance in restaurant star ratings can be explained by the sentiment of reviews alone, which is notable considering the multifaceted nature of factors affecting restaurant ratings. The positive coefficient for *sentiment_num* (Table 3) confirms that positive sentiments are associated with higher star ratings. Given the high significance of the *sentiment_num* variable, we can infer that consumer sentiment is a substantial factor in the perceived quality of a restaurant.

4.2 Clustering Analysis

The clustering analysis offers a visual exploration of the data, revealing patterns that are not immediately apparent in the OLS regression. Through K-Means and Agglomerative clustering, we

investigate the relationship between sentiment and the key business metrics: operational status (is_open) and overall star rating (stars).

Upon examining Figures 3 and 4, we observe a discernible pattern where restaurants with positive reviews tend to have a higher probability of remaining open as compared to those with negative reviews. Intriguingly, the data also indicates that restaurants receiving predominantly positive reviews are not immune to closure and can, in fact, be more susceptible to shutting down than those with negative feedback. This paradox suggests that while positive sentiments from reviews are certainly beneficial, they are not definitive predictors of a restaurant's continued operation. The relationship between review sentiments and a restaurant's ability to stay in business is evidently more complex than a direct cause-and-effect scenario. Consequently, these findings imply that other factors beyond customer reviews may exert a more significant influence on the longevity of a restaurant.

The clustering analysis depicted in Figures 5 and 6 serves to reinforce the insights gained from the OLS regression, providing a nuanced view of the association between sentiment and overall star ratings of restaurants. These overall star ratings represent the aggregated judgment of a restaurant's quality by all its reviewers, as opposed to individual review stars. Our analysis demonstrates a clear trend: restaurants with predominantly positive reviews consistently exhibit higher aggregated star ratings. Conversely, those establishments that have amassed a greater proportion of negative feedback are reflected through lower overall star ratings. This correlation underscores the impact of collective customer satisfaction on the perceived excellence of a restaurant, as encapsulated by its overall star rating.

These findings answer our research question by confirming that while positive sentiments are associated with higher ratings, they do not singularly predict a restaurant's success or failure. The complexity of the hospitality industry demands a multifaceted approach to understanding and predicting business outcomes.

5 Conclusion

This research embarked on an exploration to understand the influence of online review sentiments on the success indicators of restaurants, specifically their ratings and longevity. The analysis conducted through the utilization of Ordinary Least Squares (OLS) regression and clustering techniques, supported by a substantial dataset from Yelp, enabled us to address the research question affirmatively. We discovered that positive review sentiments are indeed correlated with higher ratings, yet they do not solely determine a restaurant's survival, suggesting a more intricate relationship between customer feedback and restaurant success.

The strengths of this study lie in its methodological approach, combining econometric and machine learning tools to unravel the complexities of consumer sentiment and its impact. The extensive dataset provided a robust foundation for our analysis, allowing for a comprehensive examination of the nuanced ways in which sentiments contribute to the perceived quality and sustainability of restaurants. By employing both regression and clustering methods, the study offered a multifaceted

view of the data, enriching our understanding of the underlying patterns.

Despite the insights gained, this research faced limitations, chiefly the necessity to condense the vast Yelp dataset to manage the analysis feasibly. This condensation may have constrained the depth of our findings, suggesting that with more resources, a more granular analysis encompassing the full dataset could unveil further nuances. Future research directions could involve a deeper dive into the text of the reviews themselves, identifying specific keywords or phrases that most strongly correlate with positive or negative sentiments. Such an inquiry would address a new question: why do restaurants receive positive or negative reviews? Exploring this would not only extend our current understanding but also offer practical guidance for restaurants aiming to improve their online presence and operational success.

References

Luca, Michael. 2016. "Reviews, Reputation, and Revenue: The Case of Yelp.com." Harvard Business School Working Paper, No. 12-016.

Tables and Graphs

Table 1: Summary Statistics of Dataset

Description	Total Count	Mean	Min	Max	Positive Sentiments	Negative Sentiments
Total Rows	1,166,500	-	-	-	-	-
Stars (Average Rating)	-	3.806025	-	-	-	-
Review Stars (Average)	-	3.952815	-	-	-	-
Is Open (Status)	-	-	0	1	-	-
Sentiments	-	-	-	-	856,494	310,006

Table 3: OLS Regression Results

Variable	**Coef.**	**Std.Err.**	**t-Value**	**P$>\|t\|$**	**95% Conf. Interval**
Intercept	3.4728	0.001	3236.083	0.000	[3.471, 3.475]
Sentiment_num	0.4539	0.001	362.433	0.000	[0.451, 0.456]

Note: The dependent variable is the average star rating of a restaurant. The sentiment_num is coded as 1 for positive and 0 for negative sentiments. The R-squared value is 0.101, indicating that sentiment accounts for 10.1% of the variability in star ratings.

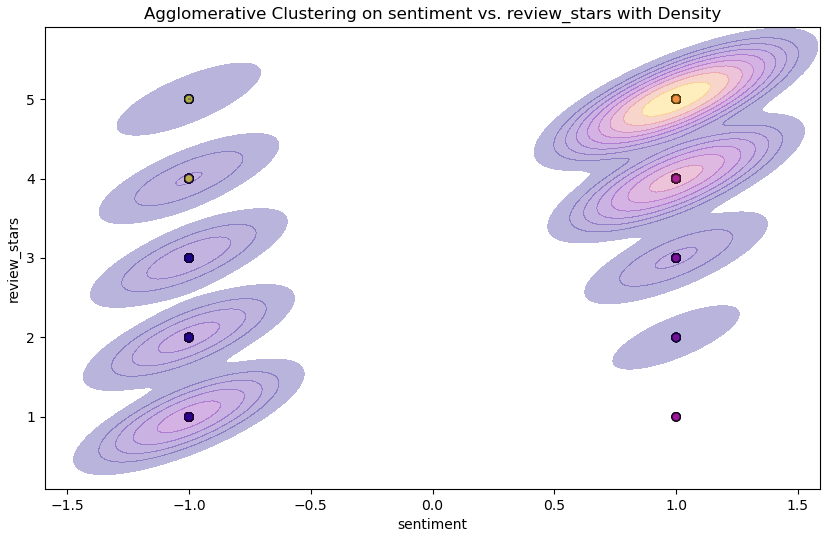

Figure 1: Agglomerative Clustering on sentiment vs. review_stars with Density

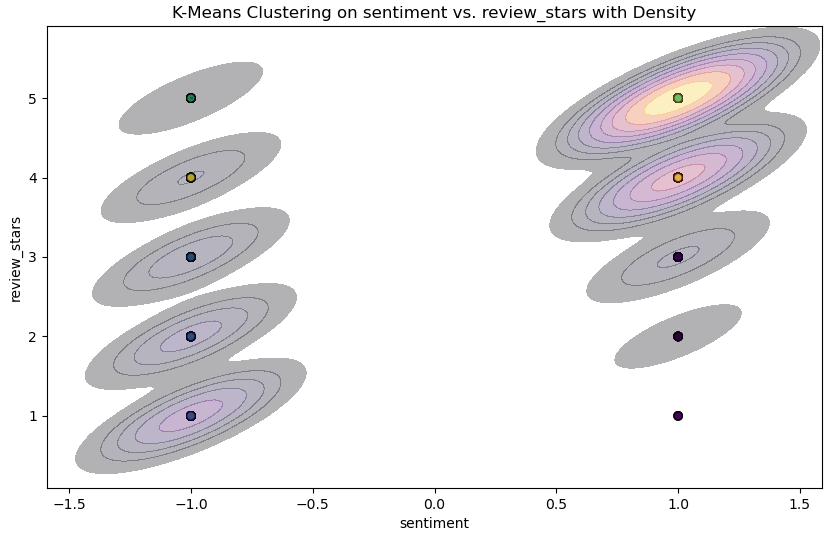

Figure 2: K-Means Clustering on sentiment vs. review_stars with Density

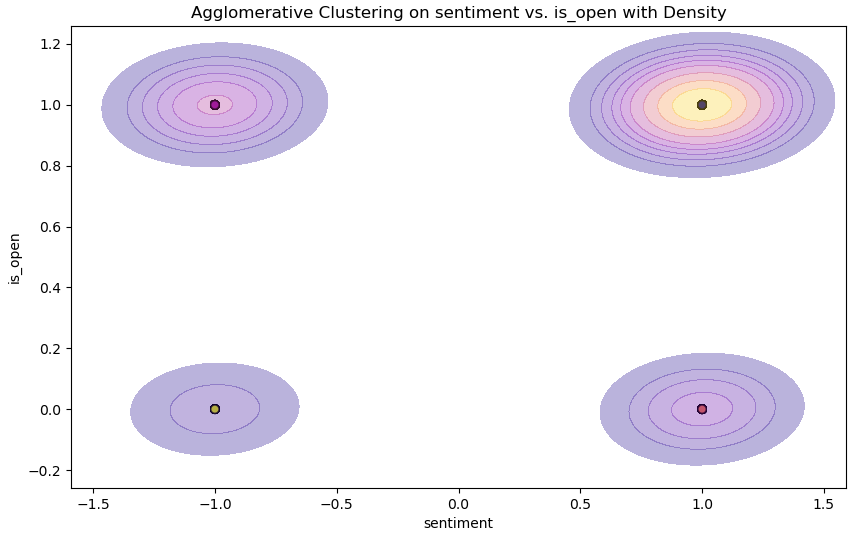

Figure 3: Agglomerative Clustering on sentiment vs. is_open with Density

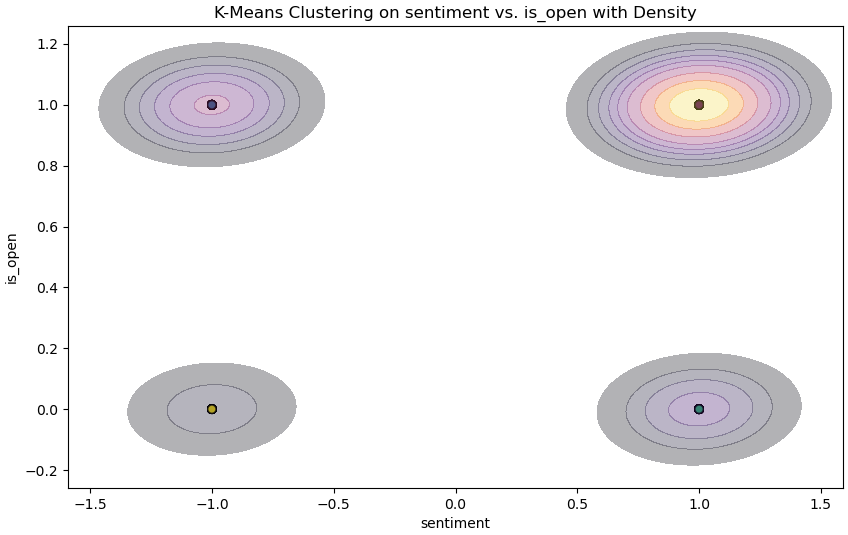

Figure 4: K-Means Clustering on sentiment vs. is_open with Density

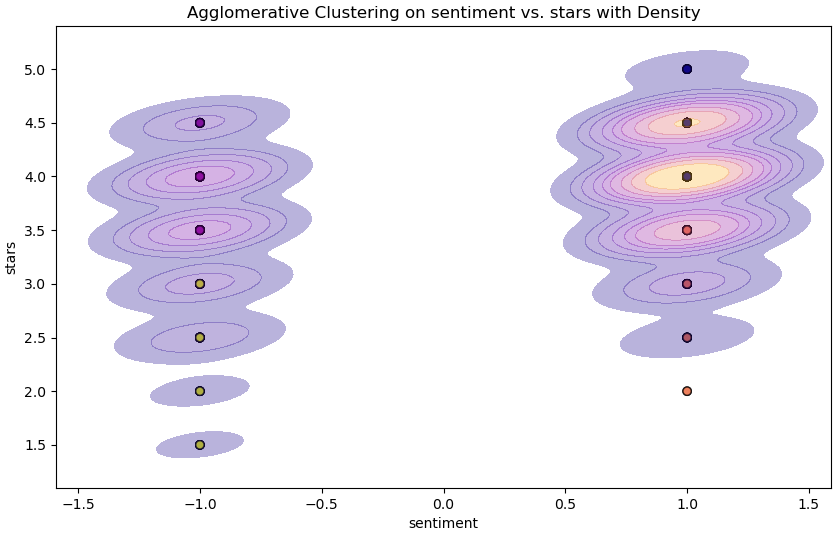

Figure 5: Agglomerative Clustering on sentiment vs. stars with Density

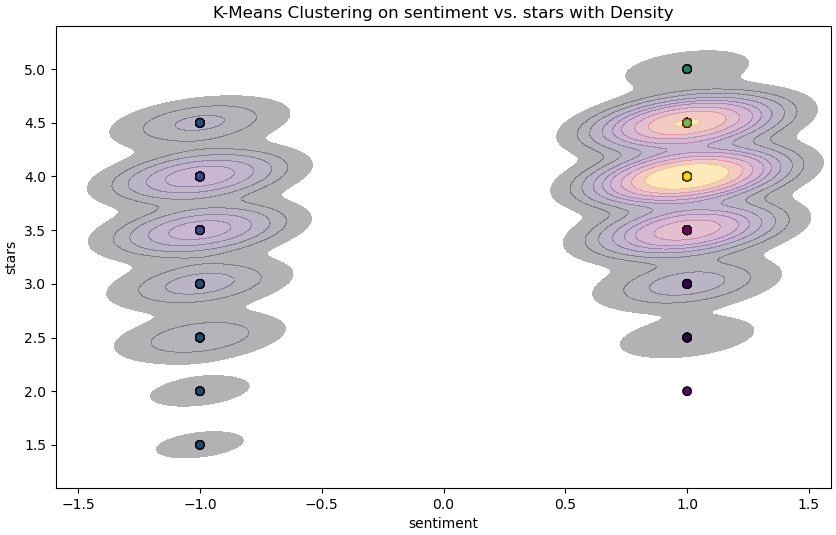

Figure 6: K-Means Clustering on sentiment vs. stars with Density