Insights into Customer Perception

Topic Modeling and Sentiment Analysis of Yelp Reviews

Team Members: Linh Chu, Linh Pham, Teresa Dang

Text Analytics and Natural Language Processing

Lorena Martin

Abstract

This study aimed to provide insights into customer perceptions of operating restaurants in California through topic modeling and sentiment analysis of Yelp reviews. The analysis consisted of multiple steps, including text preprocessing to break down individual reviews, Latent Dirichlet Allocation (LDA) to label review topics, and linear regression to predict restaurant ratings based on sentiment scores. The analysis confirmed a correlation between user star ratings and the overall sentiment expressed in reviews, with more positive sentiments corresponding to higher ratings. Among the four identified topics in customer reviews, sentiment toward service exhibited the strongest influence on ratings compared to other topics.

Based on these findings, we propose launching a "Yelp Rating Breakdown" feature that provides a rating breakdown and reviews keywords for businesses. This feature would enable restaurants to understand specific aspects contributing to higher or lower ratings, allowing them to strategically boost marketing campaigns highlighting their strengths and addressing common complaints. Additionally, it would offer customers a convenient way to gauge restaurant performance across various aspects, facilitating more informed decision-making based on personal preferences and priorities.

Keywords: sentiment analysis, topic labeling, Latent Dirichlet Allocation (LDA), linear regression, customer perceptions, Yelp reviews

Introduction

The rise of information technology has significantly impacted the restaurant industry. Consumer decision-making now frequently involves consulting online review platforms, such as Yelp, which have established user reviews as a standard practice. These platforms typically incorporate two key elements within their review systems: numerical ratings (e.g., one-to-five star scale) and textual reviews providing detailed commentary. Noh et al. (2023) found that star ratings had more influence on consumers who were comparing different products, while comments affected their choices of individual products. Due to such nuances, our paper aimed to explore the correlation between the numerical ratings and sentiment of commentary reviews to see if they provide consistent perception. Another challenge in understanding reviews is that a reviewer's overall rating might reflect service criteria in which the search user is not interested. Users often have no choice but to browse through massive amounts of text to find a particular piece of interesting information relevant to their preferences or priorities. Therefore we further looked into topic extraction from text reviews to investigate whether any specific topics had more influence on star ratings than others. Summarizing the granularity in text reviews is expected to uncover insights for businesses on how to improve their operations, as well as to provide a quick and efficient way to navigate decision-making processes for consumers.

Literature review

A. Impact of reviews on decision-making

Online consumer reviews, defined as a platform for consumers to access peer evaluations of products, services, or companies (Chan et al., 2017), have become a prominent feature of the online landscape. These reviews serve a dual purpose: information acquisition and decision-making (Adnyani & Pitanatri, 2023), enabling consumers to leverage them for seeking details and insights that influence their purchasing choices. Furthermore, online reviews not only function as decision-making tools but also as a feedback mechanism for consumers and a recommendation system embedded within online platforms (Ek Styvén & Foster, 2018). Interestingly, these reviews can also be viewed as a form of electronic word-of-mouth communication, influencing purchasing decisions like traditional word-of-mouth marketing (Ek Styvén & Foster, 2018).

B. Two types of reviews

Commercial online reviews incorporate both numerical star ratings and textual elements. The widely used one-to-five star rating system signifies the overall sentiment, with higher ratings indicating more positive evaluations. However, these ratings can be subjective, with individual interpretations varying slightly (Zhao, Xu & Wang, 2019). Although the distribution of star ratings typically approximates a normal curve, allowing for averaging, text reviews offer richer qualitative information beyond a simple numerical rating. According to Hu & Kim (2018), the subjective nature of text means reviewers assigning the same rating might express vastly differing opinions through their written comments. For instance, reviewers awarding four stars could praise a product as "very good" or find it "somewhat disappointing." This inherent subjectivity poses a challenge in summarizing or averaging large text review volumes. Consequently, most companies display text reviews sorted by factors like "likes" or recency, potentially introducing randomness or bias.

C. Sentiment Analysis

Text-mining techniques such as sentiment analysis are potent analytical tools that enable extracting social and business meanings from extensive and voluntary consumer data, thereby facilitating a response to the rapidly evolving tastes and fluid identities of postmodern consumers. Past research has successfully confirmed the influence of customer sentiment in reviews on restaurant overall ratings (Gan et tal., 2016). However, it was agreed that sentiment had to be taken into consideration along with context to be translated into actionable insights for businesses.

D. Topic Extraction

The method of topic extraction was inspired by the paper "Extracting Latent Topics from User Reviews Using Online LDA". Based on training the LDA model, the paper obtained 50 latent topics and lexical distributions of the corresponding topics. Wang (2018) suggested that this method could help restaurants "quickly and easily find the needs of users" and " improve and enhance the quality of service in all aspects to improve the rating of restaurant businesses and enhance their performance".

E. Study Gap and Research Focus

Online customer reviews have become a pivotal source of information influencing consumer decision-making processes. While there is a growing body of research investigating the relationship between numerical star ratings and textual review content, this study aims to address two notable gaps. Firstly, there is a lack of direct comparison of predictive model performance before and after incorporating topic extraction techniques from textual data. Secondly, limited efforts have been made to identify the specific topic dimensions exhibiting the greatest impact on rating outcomes. By conducting a comprehensive analysis integrating topic modeling and predictive modeling approaches, this study quantifies potential improvements in prediction accuracy through topic-based textual features and elucidates the relative importance of various topic dimensions in predicting customer ratings. The findings contribute to a nuanced understanding of the interplay between review text and numerical ratings, as well as the benefits of leveraging topic modeling, ultimately informing strategic decision-making and enhancing the customer experience.

Methodology

A. Data Source

For this study, the publicly available Yelp Academic Dataset served as the subject for our analysis. The dataset contains several subsets; however, we primarily focus on the business and review datasets. The business dataset provides comprehensive details about each business entity, including location, category, and additional services. On the other hand, the review dataset focuses on granular information regarding individual reviews, such as the actual review text, written date, and rating for the restaurant.

The Yelp Academic Dataset covers more than 150,000 businesses and almost 7,000,000 reviews. To maintain a focused scope of the analysis, we reduced our scope to restaurants in California that remain operational. By joining the review data with the corresponding business details and eliminating extraneous columns, we curated a smaller dataset containing 588 restaurants and approximately 168,000 reviews.

B. Text Preprocessing

Before conducting an in-depth analysis, it was crucial to preprocess available textual data. To streamline this process, we developed a function to systematically preprocess the dataset passed into it. The preprocessing pipeline consisted of the five following steps:

1. Word Tokenization: Each review was tokenized using NLTK.

2. Case Normalization: All tokens were converted to lowercase to ensure consistent treatment of words regardless of their capitalization in the original text.
3. Non-alphanumeric Character Removal: Non-alphanumeric characters, such as punctuation marks and digits, were removed from the tokens, as they typically do not contribute to the sentiment analysis process.
4. Stop Word Removal: Using an updated stop word list, common words that carry little semantic value (e.g., "he", "she", "they") were filtered out to enhance the efficiency.
5. Stemming: The remaining tokens were stemmed to reduce words back to their base or root form, thereby increasing consistency and reducing redundancy in the data.

An example of the text preprocessing stage was visualized through the word cloud for Luke's Sports Shack Bar & Grill. From the word cloud, it was clear that the restaurant was highly rated by customers for its service and food.

Figure 1: Word Cloud

The preprocessed text data provided a foundation for further analysis, enabling us to derive meaningful insights from the Yelp reviews while mitigating the impact of noise and irrelevant information.

C. Exploratory Data Analysis (EDA) and Sentiment Analysis

With stemmed tokens, we entered the exploratory data analysis (EDA) phase to gain insights into the dataset's characteristics and identify potential patterns. The Sentiment Intensity

Analyzer from the NLTK library was utilized to calculate sentiment scores for each review. Rather than relying on individual sentiment scores, we opted for the compound score, which aggregated negative, neutral, and positive sentiment into a single value ranging from -1 to 1.

Subsequently, we computed the mean sentiment score for each restaurant, enabling us to visualize the data through histograms and scatterplots. The distribution of review sentiment scores (Figure 2) and user ratings (Figure 3) were left-skewed, indicating that both reviews and ratings were overwhelmingly positive. Notably, the mean sentiment score was approximately 0.7, while the mean star rating hovered around 4.0.

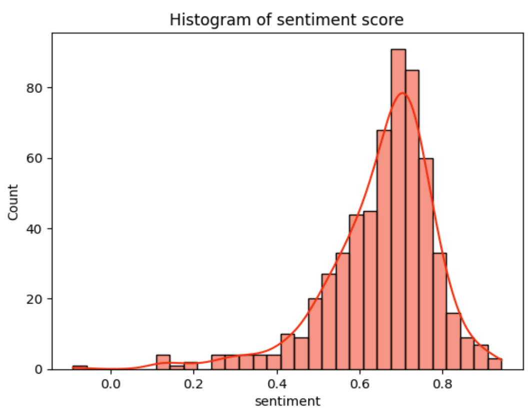


Figure 2: Histogram of Sentiment Score

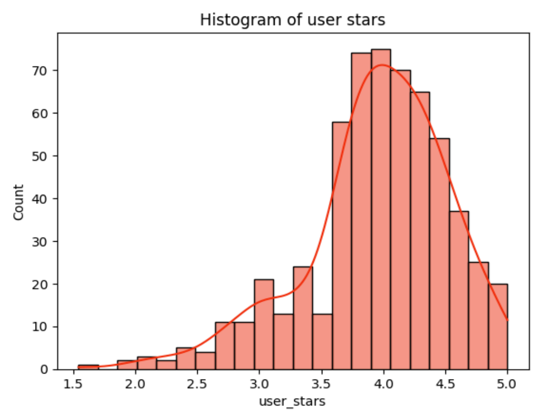


Figure 3: Histogram of User Ratings

To visualize the relationship between sentiment and user ratings, we plotted the review sentiment scores against the corresponding user ratings (Figure 4). The scatterplot unveiled a somewhat linear correlation between the two variables, suggesting that higher sentiment scores were associated with more favorable user ratings.

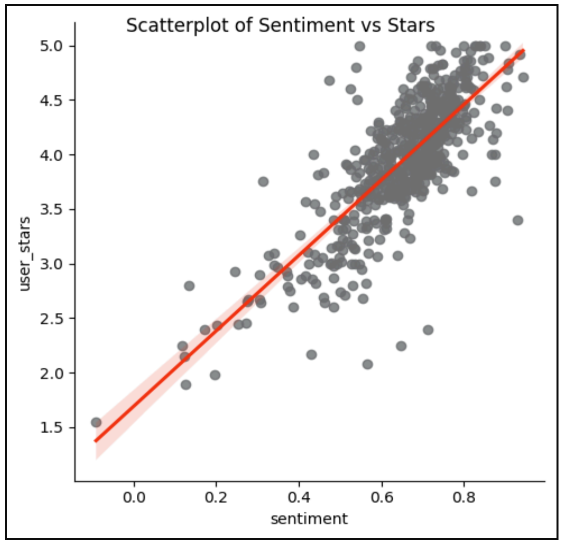


Figure 4: Scatterplot of Sentiment vs User Ratings

Through this exploratory phase, we gained valuable insights into the dataset's characteristics, including the overall positive sentiment exhibited in the reviews, as well as the connection between sentiment and user ratings.

D. Topic Modeling using Latent Dirichlet Allocation (LDA)

Topic modeling is a popular text-mining technique used to uncover the underlying semantic structure or topics in a collection of documents. This process involves identifying groups of words (topics) that frequently co-occur across different reviews, thereby revealing common themes that characterize customer opinions and experiences.

The Yelp review dataset consisted of unlabeled text data, making it unsuitable for applying supervised learning methods that rely on pre-existing labeled datasets. Therefore, we extracted insights directly from the text by identifying recurring patterns and similarities across the documents. The underlying assumption in topic modeling was that documents with similar word distributions or vector representations likely correspond to common topics, such as food quality, customer service, location, or pricing – themes that frequently appear in Yelp reviews.

In this analysis, we aim to gain insights into which factor has a greater impact on overall sentiment than others. Using topic modeling on the Yelp Review dataset can help businesses understand what matters most to their customers, identify areas for improvement, and monitor public sentiment toward their offerings.

Latent Dirichlet Allocation (LDA) is a probabilistic generative model that represents documents as mixtures of topics, where a distribution over words characterizes each topic. LDA discovers these latent topics from the observed text corpus without any prior knowledge of the topics themselves. LDA tokenizes the documents into bags-of-words (BoW) to effectively model topics in text data by converting raw textual information into a structured format. BoW enables LDA to uncover latent topics, estimate topic distributions within documents, and derive meaningful insights from text corpora based on word occurrence patterns.

Once the LDA model converged, it produced the document-topic distribution (indicating the proportion of topics in each document) and the topic-word distribution (representing the likelihood of words given each topic). By examining the top words associated with each topic in the topic-word distribution, we interpreted and labeled the discovered topics based on the underlying themes captured by the model. We then generated an elbow plot showing the relationship between the number of topics and the coherence score to help determine the optimal number of topics for the given corpus.

Results

A. Result of Topic Modeling

The maximum coherence score was achieved at 18 topics, then we hand-labeled them into four meaningful categories, including service, food, location, and time. Each review was assigned to one topic with the highest probability from the LDA model. The most popular topic is Service, accounting for 42%. The popularity of the topic of Food is 29%, followed by Location (16%), then Time (13%).

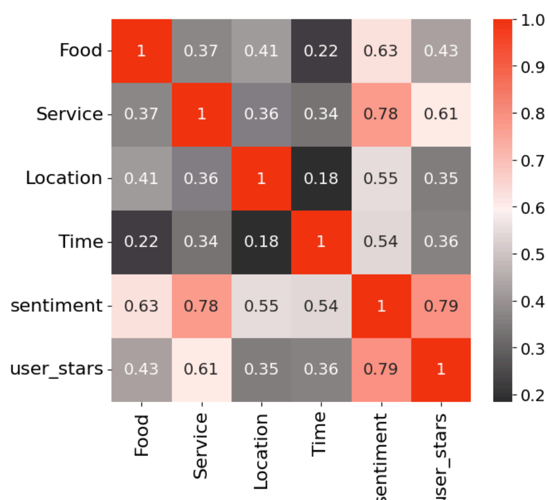
Table 1 showcases some of our LDA topics output:

Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**	**Topic 6**	**Topic 7**	**Topic 8**
burrito	first	night	fish	burger	wait	love	pork
mexican	disappoint	experi	seafood	chees	ask	friendli	tapa
bean	good	hotel	shrimp	fri	minut	recommend	belli
best	time	view	salad	onion	came	super	rib
authent	order	beauti	chicken	patti	never	amaz	deli
joe	back	parti	lobster	beet	server	love	prosciutto
Food	**Service**	**Location**	**Food**	**Food**	**Time**	**Service**	**Food**

Table 1: LDA Topic Output

From our correlation matrix, topics of 'Service' and 'Food' have the highest correlation with the users' stars rating at 0.61 and 0.43 respectively

Figure 5: Correlation Matrix



B. Result of Linear Regression between User Ratings and Overall Sentiment Score

We first leveraged Linear Regression using cross-validation R^2 to evaluate the relationship between user ratings (target variable) and the overall sentiment score of reviews. An R-squared value of 0.59 indicates a moderately strong linear relationship. The result suggests that the sentiment score can predict the majority of the star ratings from users, confirming our first hypothesis

C. Result of Linear Regression between User Ratings and Sentiment Score of Each Topic

However, the correlations became weaker when we examined the sentiment score of all topics including service, food, location, and time against the users' ratings. The R^2 unexpectedly dropped significantly to 36%. This problem happened due to the unclean topic extraction, followed by the imbalanced data in mean aggregation.

To validate our second hypothesis, we changed the approach to run a simple linear regression for each topic's sentiment score against star ratings. Service, food, location, and time had an R^2 of 0.31, 0.15, 0.10 and 0.08 respectively.

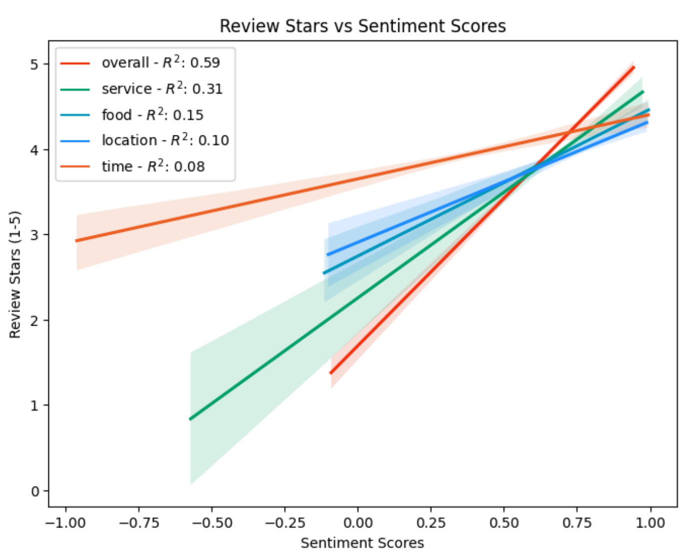


Figure 6: Linear Regression Results

This suggests that customers' perceptions of service quality carry more weight in determining their overall rating compared to other factors such as food, location, or timeliness, which confirms our second hypothesis.

Limitations and Potential Extensions of the Study

The analysis faces several key limitations that we acknowledged. Firstly, the dataset is geographically skewed, with around 76% of reviews coming from restaurants in Santa Barbara. This lack of diversity in the data can impair the model's ability to generalize effectively to other locations, potentially leading to poor performance and misleading sentiment scores.

Secondly, sentiment analysis algorithms often struggle to account for contextual nuances, as words or phrases can convey varying sentiments based on their surrounding context, which can significantly impact the accuracy of sentiment scores. For example, one negative comment about the service, the food, and the price receives a positive sentiment score of 0.77, whereas a positive review complimenting the price and the food receives a negative sentiment score of 0.59. The sentiment analysis likely recognizes the exclamation phrase “To die for” as a negative phrase.

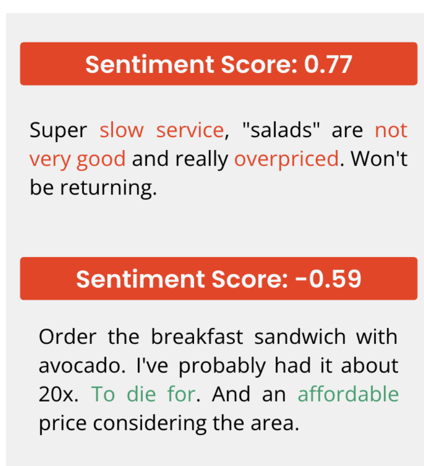


Figure 7: Sentiment Score Discrepancy

Thirdly, the topic modeling approach using LDA has its drawbacks. The iterative process of topic assignment is computationally expensive and time-consuming. Moreover, the topic labeling process, which involves manually examining dictionaries of frequently used words, is subjective and prone to human bias. These limitations underscore the importance of interpreting the findings with caution and recognizing the potential for inaccuracies or biases in the analysis.

In this project, we assigned each review to a single topic, despite the possibility that some reviews may convey multiple topics. For future studies, we would like to apply sentence-level topic modeling. Instead of tokenizing reviews into individual words, tokenizing them into sentences can provide a better basis for assigning multiple topics to a single review. By treating each sentence as a separate unit, topic modeling algorithms like LDA can be applied at the sentence level, allowing different sentences within a review to be assigned to different topics.

Conclusion and Business Recommendations

In conclusion, by analyzing the Yelp dataset, our study confirms a correlation between the user star ratings and the overall sentiment score of reviews. As the overall sentiment expressed in reviews becomes more positive, the assigned star ratings tend to be higher. In addition, among the four topics in customer reviews, sentiment toward Service has more influence than other topics.

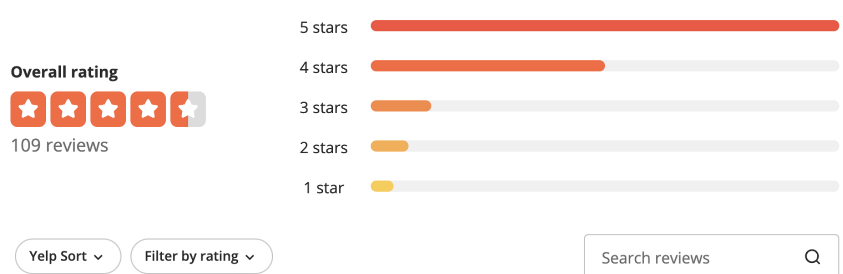


Figure 8: Business application (launch Yelp breakdown)

Our business recommendation is to launch Yelp Rating Breakdown, which provides "Rating Breakdown" and "Review Keywords" features on Yelp. For businesses, this feature can be invaluable as it allows them to understand the specific aspects that contribute to higher or lower ratings. To improve their overall ratings and customer satisfaction, restaurants can pinpoint the areas that resonate well with customers and boost marketing campaigns based on their strengths, as well as make informed decisions to address common complaints by analyzing dominant topics across customers' reviews

For customers, the proposed feature offers a convenient way to gauge a restaurant's performance across various aspects without having to read through numerous individual reviews. They can quickly identify the restaurant's strengths and weaknesses based on the rating breakdown and review keywords, enabling them to make more informed decisions based on their personal preferences and priorities.

References

Andayana, I & Pitanatri, Putu Diah. (2023). Understanding the Effect of Online Reviews on Consumer Decision-Making: Evidence from Alaya Resort Ubud. 9. 1-10. 10.18510/ijmier.2023.921.

Ek Styvén, M. and Foster, T. (2018). "Who am I if you can’t see me? The “self” of young travellers as driver of eWOM in social media". *Journal of Tourism Futures, Vol. 4 No. 1, pp. 80-92.* https://doi.org/10.1108/JTF-12-2017-0057

Gan, Q., Ferns, B. H., Yu, Y., & Jin, L. (2016). A Text Mining and Multidimensional Sentiment Analysis of Online Restaurant Reviews. *Journal of Quality Assurance in Hospitality & Tourism*, *18*(4), 465–492. https://doi.org/10.1080/1528008x.2016.1250243

Ganu, G., Elhadad, N., & Marian, A. (2009). Beyond the Stars: Improving Rating Predictions using Review Text Content. *International Workshop on the Web and Databases.*

Hu, Y., & Kim, H. J. (2018). Positive and negative ewom motivations and hotel customers’ ewom behavior: Does personality matter? *International Journal of Hospitality Management*, 75, 27–37. https://doi.org/10.1016/j.ijhm.2018.03.004

Lei, X., Qian, X., & Zhao, G. (2016). Rating Prediction Based on Social Sentiment From Textual Reviews. *IEEE Transactions on Multimedia, 18, 1910-1921.*

Noh, Y.-G., Jeon, J., & Hong, J.-H. (2023, March 20). Understanding of customer decision-making behaviors depending on online reviews. MDPI. https://www.mdpi.com/2076-3417/13/6/3949#metrics

Pykes, K. (2023, October 19). *What is topic modeling? an introduction with examples.* DataCamp. https://www.datacamp.com/tutorial/what-is-topic-modeling?fbclid=IwAR12IeGbSZiWh0

ABOR9z5BmEXVCdEsFLY8_wEMGkKiSnafleorFGwRewcm0

Wang, W., Feng, Y., & Dai, W. (2018). Topic analysis of online reviews for two competitive products using latent Dirichlet allocation. *Electronic Commerce Research and Applications*, *29*, 142–156. https://doi.org/10.1016/j.elerap.2018.04.003

Zhang, Sonya., Ly, Linda., Mach, Norman., Amaya, Christian. (2022). Topic Modeling and Sentiment Analysis of Yelp Restaurant Reviews. *International Journal of Information Systems in the Service Sector.*

Zhao, Y., Xu, X., & Wang, M. (2019). Predicting overall customer satisfaction: Big Data Evidence from hotel online textual reviews. *International Journal of Hospitality Management*, 76, 111–121. https://doi.org/10.1016/j.ijhm.2018.03.017