Beyond the Stars: Unmasking the Truth Behind Rating Inflation in Atlanta's Restaurants

Team 003: Honglai Peng, Qinyang Song, Xiaoran Zhu, Xin He, Duanduan Zhou

Abstract

This study explores the phenomenon of "Rating Inflation," characterized by an increase in average ratings and a decrease in rating variance, which threatens the informativeness of ratings. Our primary objective is to uncover the underlying causes, particularly focusing on fake reviews and lack of emphasis on recency, etc. By targeting these issues, we aim to reduce Rating Inflation, leading to a recalibration and refinement of restaurant ratings. Ultimately, we want to ensure customers can make well-informed choices and the restaurants receive the recognition they truly deserve.

1 Introduction

Online ratings and reviews are crucial to consumer decisions, with over 80% of U.S. adults referencing them when choosing dining options [29]. These ratings facilitate quality discovery and help consumers benefit from others' experiences. However, 'Rating Inflation,' characterized by an increase in average ratings and a decrease in rating variance, which undermines their decision-making utility [4]. Recent studies, like Raval's, indicate platforms like Google and Facebook exhibit rating inflation, even for poorly-rated businesses by entities like the BBB [25]. Our project explores the driving factors behind rating inflation in online platforms and its effects on consumer behavior.

Currently, lots of research have been conducted to study the reasons why rating changes over the time. In [19], they pointed out that self-selection bias in online product reviews can affect consumer purchasing decisions, the trends in online ratings over time, and overall consumer satisfaction. In [30], Wang proposed that the social influence also gradually are embedded in the online rating systems, and their research demonstrated that the online rating systems are influenced by the social connections when using social-networking features. In addition, [13] [10] [7] [24] have shed light on reciprocity, emphasizing how feedback is often a response to previous interactions potentially affecting rating system. In [27] [21] [3], they indicated that herding behavior in rating systems causes users to mimic prevailing feedback, skewing authenticity and giving undue weight to early reviews. In [12], the research showed that ratings are prone to inflation, with raters feeling pressure to leave low ratings, which in turn pushes the average higher. In [4], they pointed out when design and manage the rating system, we also need consider the potential consequences of rating inflation. Chen et al.[8] demonstrated that multidimensional rating methods boosts the clarity and relevance of ratings, offering insights for creating reliable online rating systems. Kokkodis [18] proposed a rating deflation method to counteract the loss of informativeness because of rating inflation. In [9][22], the author indicated that aggregation of rating into a single metric will be another optimization to improve the rating systems.

2 Problem Definition

Rating Inflation presents a significant challenge in assessing the credibility of online ratings. While foundational research has comprehensively studied the mechanisms and biases influencing online ratings, the granularity and depth of rating data itself play a pivotal role. The dynamics of Rating Inflation, especially, demand a closer examination of how this data is generated, aggregated, and interpreted over time.

The problem can be defined as exploring and understanding the underlying causes of Rating Inflation, particularly by examining the intrinsic characteristics of the data - How rating scores are produced? Whether using average rating without considering the recency of reviews is adequate? Whether the rating inflation will be impacted by fake reviews? How the sentiment analysis potentially mitigates the rating inflation?

By addressing these concerns, we intend to unravel the factors contributing to Rating Inflation. This will not only broaden our understanding of the phenomenon but guide us towards recalibrating a more reliable and informative rating system for online platforms.

3 Research Method Survey

To effectively tackle Rating Inflation as outlined in our problem definition, our research methodology survey will be divided into three parts as following.

3.1 Recency of Online Rating

Based on a comprehensive survey of algorithms used in online restaurant ratings from Google Local Data[31], and an analysis of our dataset, we identified that Google's online rating system employs an average calculation, without factoring in the recency of reviews.

The significance of review recency has been underscored in prior research. Xie et al. pinpointed review recency as a primary factor that is strategically significant to business performance[17]. Overlooking the recency of online reviews can lead to the dissemination of outdated information, and may potentially diminish the incentives for restaurants to enhance their services. However, prior research has largely focused on recency modeling for recommendation systems[23], while few studies explore recency models to enhance online rating system. Therefore, we aim to bridge this gap by proposing an innovative recency model specifically tailored for online ratings.

3.2 Fake Review Detection Models

Though research on mitigating rating inflation by eliminating fake reviews is sparse, our study has identified several machine learning models acclaimed for their proficiency in detecting fraudulent reviews. Notably, these encompass character-level convolutional-LSTM, convolutional-LSTM, HAN, convolutional-HAN, BERT, DistilBERT, and RoBERTa[20]. We aim to delve deeper into the following specific models.

CNN + LSTM Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) are effective NN models for sorting text. CNNs are now adopted to NLP and can identify patterns using filters. LSTMs, a kind of Recurrent Neural Network, are good at remembering information for a longer time. Combining CNNs and LSTMs has been shown to increase accuracy in categorizing text[5].

BERT Bidirectional Encoder Representations (BERT) from Transformers, showcases outstanding results for its ability to understand text by looking at the words before and after each word in the text at the same time in a bidirectional manner. This helps it get a better understanding of the language and has led to great results in various language processing tasks including fake review detection[26].

3.3 Visualization

Traditional visualization tools have reached to their limits when encountered with vast, continuously evolving datasets [2]. To navigate through these challenges, we assessed several advanced visualization tools such as *Tableau* [14], *Gephi* [6], and *React D3*[11]. Tableau simplifies complex visualizations, Gephi specializes in network graphs, and React D3 merges D3's data prowess with React's DOM efficiency for dynamic, reusable visuals in React apps.

For visualization strategies, researchers found that users prefer a hierarchical way to present information, like varying icon sizes based on star ratings[1]. Other studies have shown that restaurants with review keywords aids users in quickly understanding general feedback and making decisions [16].

4 Proposed Method

4.1 Intuition and Innovations

To solidify our approach against Rating Inflation and to enhance the reliability of online restaurant ratings in Atlanta, we present a summary of our proposed methods as follows:

- Recency Adjustment: Refine the existing average rating approach by integrating a recency bias. This adjustment will ensure that recent reviews exert a more significant influence on a restaurant's overall rating, which incentivizes the restaurants to improve their services. Our approach to combat Rating Inflation involves a novel recency adjustment technique, combining time-decay modeling and Bayesian averaging to integrate recency into restaurant rating calculations. This time-decay model ensures that reviews lose their weighting influence as they age, thereby highlighting the importance of newer reviews[4]. Complementing the time-decay model, Bayesian averaging stabilizes the ratings for restaurants with fewer reviews, counterbalances the potential distortions from limited data, and helps in providing a more

robust rating by factoring in the number of reviews alongside their content.[15]. By combining these two methods, our model aims to provide a more accurate and timely reflection of a restaurant's performance and quality, which in turn, helps in mitigating Rating Inflation.

- Removal Fake Reviews : Our strategy employs a dual-layered detection system. Initially, we utilize BERT for initial fake review identification, followed by a CNN-LSTM model for secondary verification, extracting a reliable subset of genuine reviews.
- Data Visualization and Analysis: To demonstrate the efficacy of our system adjustments, we will recalibrate the rating system and create an interactive map to visualize the spatial distribution of rating, and illustrate the distribution of ratings before and after recalibration using pie charts, bar charts, etc. to depict the percentage changes, represented by markers of varying colors and sizes on the map to indicate rating differences, and display the proportion of detected fake reviews to highlight the impact of our detection process on the overall rating system, etc.

4.2 Data Preparation

4.2.1 Data Collection

For the purpose of our study, we have sourced extensive data from Google Local Data [31] across various business entities in the United States. The dataset encompasses a massive volume of 666,324,103 reviews, contributed by 113,643,107 users, and spanning across 4,963,111 businesses. To maintain specificity in our analysis, we extracted data of the state of Georgia, which includes 24,060,125 reviews associated with 166,381 businesses. This subset serves as our foundational dataset, ensuring that our project is grounded in Georgia's restaurant landscape, which is crucial for understanding Rating Inflation.

4.2.2 Data Pre-processing

- Filter the dataset to include only restaurant-related data, meticulously excluding those located within grocery stores or shopping malls, as they do not align with our research focus.
- Narrow down our scope to specifically target restaurants situated in the Atlanta area by by employing a geographic method, which defines the Atlanta region using specific latitude and longitude coordinates. This allows to enhance the precision of our location-based filtering but also facilitates an intuitive visualization of all the restaurants on a map.
- Merge the reviews dataset with the business data using the *gmap_id* as a unique identifier, creating a comprehensive overview of each restaurant.
- Address data redundancy by removing duplicate rows, identifying them based on a combination of *use_id*, *time*, and *gmap_id*.
- Eliminate entries corresponding to *permanently closed* restaurants, as these do not contribute to our analysis of the active restaurants in Atlanta.

It is important to note that during our preprocessing, we encounter entries with none values in the *text* column, which represents the user's reviews. Instead of discarding these, we choose to retain them in our dataset which are treated as legitimate reviews, acknowledging that a user's decision to rate without providing textual feedback is a form of expression in itself. However, for the purposes of our analysis, particularly when it involves content-based assessments, we focus only on the entries that do contain text.

Finally, our refined dataset comprises 1,327 restaurants in the Atlanta area. Of the 760,525 reviews associated with these restaurants, 442,964 reviews contain textual feedback, while 317,561 are devoid of any text, being represented as none values.

4.3 Experiments Setup

In this section, we describe the experimental setup for our study of Rating Inflation.

4.3.1 Recency Adjustment Experiments

Experimental Setup To validate our recency adjustment model, we conducted experiments using the dataset from Google Local Data. The dataset was preprocessed to filter out non-restaurant related data and focus exclusively on the Atlanta area.

Time-Decay The experiment involved applying a time-decay function to the reviews, reducing their influence over time. As per our implementation, the decay factor was calculated using the time-decay model with a life cycle of 365 days and a decay rate of 0.90. The formula

for the decay factor is given by:

$$\text{Decay Factor} = \text{Decay Rate}^{\frac{\text{Days Since Review}}{\text{Life Cycle}}}$$

Furthermore, a boost factor is applied to enhance the influence of the most recent reviews. Specifically, for reviews within a 90-day window, a boost factor of 1.2 is assigned, reflecting their increased relevance.

Bayesian Average Subsequently, we used Bayesian averaging to compute the final adjusted ratings. The Bayesian average is calculated using the formula:

$$\text{Bayesian Average} = \frac{v}{v+C} \times R + \frac{C}{v+C} \times m$$

where R is the mean rating for the restaurant, v is the number of reviews, C is the average number of reviews per restaurant, and m is the global weighted average rating calculated from the dataset, and the Bayesian average was applied to each restaurant's reviews. However, in our actual implementation, we choose the median of reviews which is smaller than the mean value as our C.

Finally, we performed min-max normalization within each group to scale the ratings between 1 and 5. The normalized weighted ratings were then rounded to one decimal place for consistency. These adjusted ratings were analyzed to assess the impact of our recency model on mitigating Rating Inflation.

4.3.2 *Fake Reviews Removal Experiments*

To effectively combat rating inflation, our approach involves a strategic focus on identifying and removing fake reviews from the dataset.

Our study applies CNN-LSTM and BERT models, recognized for detecting fake reviews, through transfer learning on Google restaurant reviews, followed by cross-validation to ensure accuracy and minimize false positives in identifying fake content.

- **Training Data Collection** Despite extensive research on fake reviews detection, there is a notable gap in the availability of pre-trained models and datasets. To bridge this gap, we've aggregated a balanced training dataset from two primary sources: [28] and the Kaggle Fake Reviews Detection challenge. Our compiled dataset consists of over 56,000 entries, evenly split between 28,000+ authentic reviews and 28,000+ deceptive reviews.
- **Model Construction and Training** Utilizing *bert-base* as our pre-trained model, we've appended custom layers to tailor a BERT-based architecture for fake reviews detection. Trained on our balanced dataset, this model achieved an accuracy of 0.85. Similarly, we developed a CNN-LSTM model, which after training on the same data, reached an accuracy of 0.82.
- **Application and Intersection Analysis** When deployed on the Google Reviews Dataset, our BERT-based model flagged over 110,000 potential fake reviews, while the CNN-LSTM model identified upwards of 120,000. By intersecting these findings, we isolated a subset of 80,000+ reviews with high confidence in their inauthenticity. The refined dataset now comprises 760,525 reviews, segregated into 672,176 true reviews and 88,349 fake reviews.

4.3.3 *Experiment Results Evaluation*

For the evaluation of our experimental results, we employed three critical indicators of rating inflation: *the predominance of high ratings*, *limited variance in review scores*, and *the high frequency of extreme ratings*. These metrics are pivotal in determining the effectiveness of our proposed methods in addressing the rating inflation issue.

Our analytical strategy encompassed both a comprehensive evaluation of overall ratings and focused case studies, *The Varsity* restaurant. This dual approach was instrumental in providing a holistic assessment of the effectiveness of our recency adjustment techniques and fake reviews removal methods to tackle the rating inflation problem.

Recency Adjustment Visualization We incorporated both the *Time Decay* and *Bayesian Average* methods to address rating inflation.

The impact of the *Time Decay* approach is particularly evident in the bar chart comparing the original dataset with the adjusted data, as shown in Fig. 1. In the original dataset, where recency was not considered, the number of 5-star ratings was approximately 4000. However, upon applying the recency adjustment, which gives less weight to older reviews, the number of 5-star ratings substantially decreased to around 400. This dramatic reduction suggests that a significant portion of the 5-star ratings in the original dataset originated from

older reviews, which may not accurately reflect the current state or quality of the restaurants.

Notably, the adjusted data exhibits a new category of *0* ratings, absent in the original dataset. This category represents reviews that, due to their age, no longer significantly influence the overall rating. The introduction of this category aligns with intuitive and realistic expectations, as older reviews typically become less relevant over time.

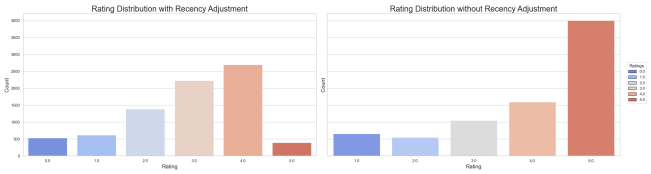

Figure 1: Rating Distribution for *The Varsity*

After that, the *Bayesian Average* method was incorporated to refine our rating analysis. The effectiveness of our method is evident in Fig. 2. Post-application, the distribution aligns more closely with a Gaussian distribution, as seen in the left chart. This contrasts with the pre-application distribution on the right, which appears more skewed.

The Gaussian-like distribution achieved with the Bayesian average suggests a more realistic and balanced rating pattern, countering the biases of extreme ratings and small sample sizes. This method ensures a fairer representation of restaurants' performance, enhancing the credibility and integrity of the rating system.

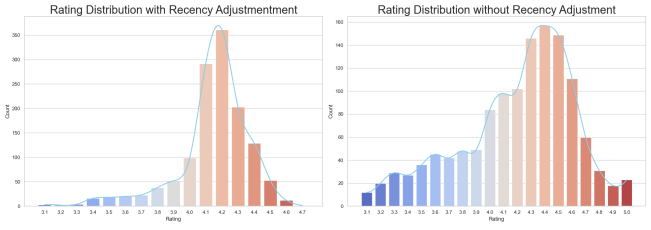

Figure 2: Rating Distribution in Atlanta

Fake Reviews Removal Visualization As depicted in Fig. 3, the overall rating distribution of Atlanta restaurants indicates a significant drop in 5-star ratings by approximately 80,000 after fake review removal. This notable decline, especially in the highest rating category, underscores the efficacy of our methods in addressing extreme ratings, which are often associated with rating inflation.

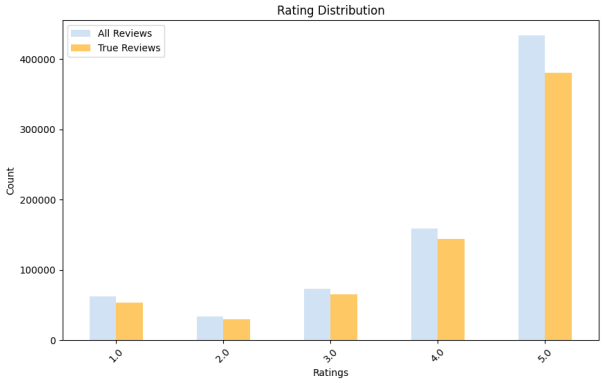

Figure 3: Rating Distribution for *The Varsity*

Fig. 4 visually represents the impacts of our strategy through pie chart using a detailed case study, *The Varsity*. This chart demonstrates a notable shift in the rating distribution following our applied measures. Initially, 5-star ratings formed a significant 53.4% of the total for this case. Following the removal of fake reviews, this percentage dropped to 51.0%, indicating a significant decrease in rating inflation. Meanwhile, the variance for Ratings of *The Varsity* has increased from 1.63 to 1.68, which also indicates a deflation for the rating distribution.

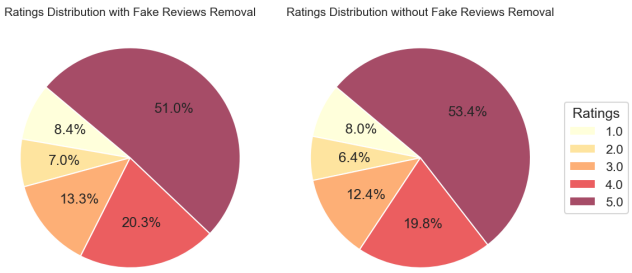

Figure 4: Rating Distribution for *The Varsity*

Additionally, we analyzed the monthly trends of 5-star ratings of this case, as depicted in Fig. 5. Prior to applying our review analysis techniques, the data exhibited pronounced peaks, indicative of potential rating inflation. After our fake reviews removal was implemented, a considerable smoothing of these peaks was evident, signaling a more stable and genuine rating pattern. This decrease in rating variability is a strong

indicator of the rating inflation improvement, consequently, enhancing the reliability and accuracy of the rating system.

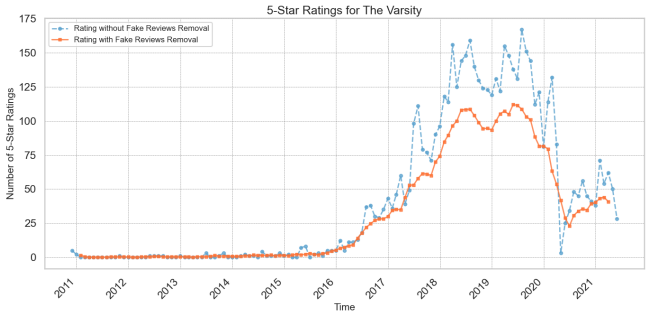

Figure 5: 5-Star Rating for *The Varsity*

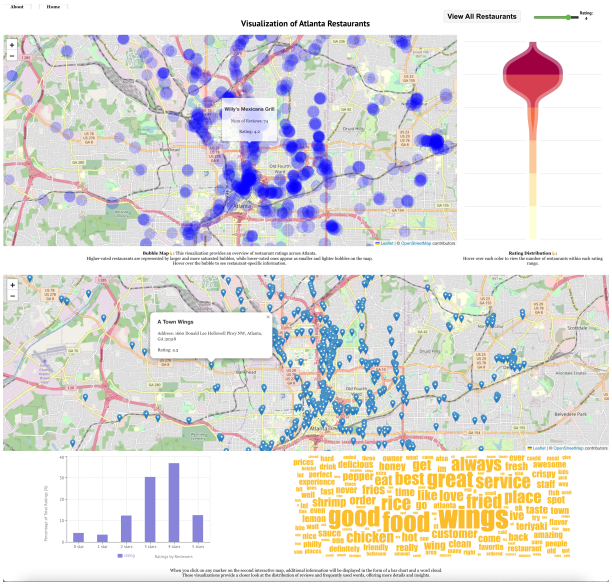

Figure 6: Interactive Map Demo

Interactive Maps Our website, built with React and D3, features interactive maps on the 'Home' page and a user guide on the 'About' page. Utilizing libraries like Leaflet, Nivo, and Visx, we offer a seamless visual exploration of restaurant ratings. As illustrated in the top part of Figure 6, a bubble map filters establishments by ratings, ranging from 4.0 to 5.0 in this case, with options to adjust the filter or display all restaurants. Adjacent funnel charts present the distribution across rating categories. The button part highlights an interactive feature where selecting a restaurant marker on the map displays its rating breakdown in a bar chart and aggregates common terms from reviews into a word cloud.

5 Conclusion and Discussion

Our investigation has significantly advanced the understanding of rating inflation in Atlanta's restaurant landscape, providing substantial methodological contributions and valuable data resources. Additionally, our detailed analysis, employing recency adjustments and the removal of fake reviews, has led to an ideal recalibration of rating distributions, resulting in a more genuine reflection of customer feedback.

By leveraging the metrics mentioned in 4.3.3, we have quantified the extent of rating inflation and, subsequently, the effectiveness of our methodologies in ameliorating its impact. The substantial reduction in 5-star ratings, notably those of a significant age, and the emergence of a zero-rating category, have both contributed to a rating distribution that we believe is more representative of the current consumer experience.

Despite the progress made, we acknowledge the potential for advancement in the detection models used for identifying fake reviews. Our reliance on self-trained models, due to the lack of accessible advanced resources, points to a gap in the availability of cutting-edge tools for broader application.

Furthermore, the integration of sentiment analysis into our methodology represents a forward-looking direction for subsequent studies. Sentiment analysis could provide a richer, qualitative layer to the predominantly quantitative methods currently in use, offering a more granular perspective on consumer feedback.

We also contribute to the research community by providing a labeled dataset of over 56,000 entries, a resource we hope will facilitate further advancements in review analysis.

In summary, this study underscores the widespread issue of rating inflation and actively contributes to the methodologies for mitigating it, thereby enhancing the integrity of online review systems. With the labeled dataset now available, and a refined approach to analyzing ratings, this work paves the way for a more reliable reflection of consumer opinions and experiences.

Contribution Statement: All members contributed similar efforts to the final project.

References

[1] Yang Jun Ah, Kim Hyun Jeong, Sung Jae Woo, and Lee Suk Ho. 2011. Visualization of restaurant information on web maps. In *The 5th International Conference on New Trends in Information Science and Service Science*, Vol. 2. IEEE, 270–272.

[2] Syed Mohd Ali, Noopur Gupta, Gopal Krishna Nayak, and Rakesh Kumar Lenka. 2016. Big data visualization: Tools and challenges. In *2016 2nd International conference on contemporary computing and informatics (IC3I)*. IEEE, 656–660.

[3] Sinan Aral. 2013. The problem with online ratings. *MIT Sloan Management Review* (2013).

[4] Arslan Aziz, Hui Li, and Rahul Telang. 2023. The consequences of rating inflation on platforms: Evidence from a quasi-experiment. *Information Systems Research* 34, 2 (2023), 590–608.

[5] Gourav Bathla, Pardeep Singh, Rahul Kumar Singh, Erik Cambria, and Rajeev Tiwari. 2022. Intelligent fake reviews detection based on aspect extraction and analysis using deep learning. *Neural Computing and Applications* 34, 22 (2022), 20213–20229.

[6] Anuja Bokhare and PS Metkewar. 2020. Visualization and interpretation of Gephi and Tableau: a comparative study. In *International Conference on Advances in Electrical and Computer Technologies*. Springer, 11–23.

[7] Gary Bolton, Ben Greiner, and Axel Ockenfels. 2013. Engineering trust: reciprocity in the production of reputation information. *Management science* 59, 2 (2013), 265–285.

[8] Pei-Yu Chen, Yili Hong, and Ying Liu. 2018. The value of multidimensional rating systems: Evidence from a natural experiment and randomized experiments. *Management Science* 64, 10 (2018), 4629–4647.

[9] Weijia Dai, Ginger Jin, Jungmin Lee, and Michael Luca. 2018. Aggregation of consumer ratings: an application to Yelp. com. *Quantitative Marketing and Economics* 16 (2018), 289–339.

[10] Chrysanthos Dellarocas and Charles A Wood. 2008. The sound of silence in online feedback: Estimating trading risks in the presence of reporting bias. *Management science* 54, 3 (2008), 460–476.

[11] Elad Elrom. 2021. *Integrating D3. js with React*. Springer.

[12] Apostolos Filippas, John Joseph Horton, and Joseph Golden. 2018. Reputation inflation. In *Proceedings of the 2018 ACM Conference on Economics and Computation*. 483–484.

[13] Andrey Fradkin, Elena Grewal, and David Holtz. 2021. Reciprocity and unveiling in two-sided reputation systems: Evidence from an experiment on Airbnb. *Marketing Science* 40, 6 (2021), 1013–1029.

[14] Jamie Hoelscher and Amanda Mortimer. 2018. Using Tableau to visualize data and drive decision-making. *Journal of Accounting Education* 44 (2018), 49–59.

[15] Adrian Raftery Jennifer Hoeting, David Madigan and Chris Volinsky. 1999. Bayesian Model Averaging: A Tutorial. *Statist. Sci.* 14 (1999), 382–417.

[16] Wei Jin, Hung Hay Ho, and Rohini K Srihari. 2009. OpinionMiner: a novel machine learning system for web opinion mining and extraction. (2009), 1195–1204.

[17] Shinyi Wu Karen Xie, Chihchien Chen. 2016. Online Consumer Review Factors Affecting Offline Hotel Popularity: Evidence from Tripadvisor. *Journal of Travel & Tourism Marketing* 33, 2 (2016), 211–223.

[18] Marios Kokkodis. 2019. Reputation deflation through dynamic expertise assessment in online labor markets. In *The World Wide Web Conference*. 896–905.

[19] Xinxin Li and Lorin M Hitt. 2008. Self-selection and information role of online product reviews. *Information Systems Research* 19, 4 (2008), 456–474.

[20] Rami Mohawesh, Shuxiang Xu, Son N Tran, Robert Ollington, Matthew Springer, Yaser Jararweh, and Sumbal Maqsood. 2021. Fake reviews detection: A survey. *IEEE Access* 9 (2021), 65771–65802.

[21] Lev Muchnik, Sinan Aral, and Sean J Taylor. 2013. Social influence bias: A randomized experiment. *Science* 341, 6146 (2013), 647–651.

[22] Chris Nosko and Steven Tadelis. 2015. *The limits of reputation in platform markets: An empirical analysis and field experiment*. Technical Report. National Bureau of Economic Research.

[23] Praveen Madiraju Paromita Nitu, Joseph Coelho. 2021. Improvising Personalized Travel Recommendation System with Recency Effects. *Big Data Mining and Analytics* 4, 3 (2021), 139–154.

[24] Davide Proserpio, Wendy Xu, and Georgios Zervas. 2018. You get what you give: theory and evidence of reciprocity in the sharing economy. *Quantitative Marketing and Economics* 16 (2018), 371–407.

[25] Devesh Raval. 2023. Do Gatekeepers Develop Worse Products? Evidence from Online Review Platforms. (2023).

[26] David Refaeli and Petr Hajek. 2021. Detecting fake online reviews using fine-tuned BERT. In *Proceedings of the 2021 5th International Conference on E-Business and Internet*. 76–80.

[27] Matthew J Salganik, Peter Sheridan Dodds, and Duncan J Watts. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *science* 311, 5762 (2006), 854–856.

[28] Joni Salminen, Chandrashekhar Kandpal, Ahmed Mohamed Kamel, Soon-gyo Jung, and Bernard J Jansen. 2022. Creating and detecting fake reviews of online products. *Journal of Retailing and Consumer Services* 64 (2022), 102771.

[29] Aaron Smith and Monica Anderson. 2016. Online shopping and e-commerce. (2016).

[30] Chong Wang, Xiaoquan Zhang, and Il-Horn Hann. 2018. Socially nudged: A quasi-experimental study of friends' social influence in online product ratings. *Information Systems Research* 29, 3 (2018), 641–655.

[31] An Yan, Zhankui He, Jiacheng Li, Tianyang Zhang, and Julian McAuley. 2023. Personalized Showcases: Generating multimodal explanations for recommendations. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2251–2255.