Shifting Standards or Changing Preferences? Unraveling Review Polarization via LLMs

Limin Fang [*] Chunhua Wu [†] Baohong Sun [‡§]

October 30, 2024

Abstract

This paper examines the drivers of online review polarization, an increasing trend of extreme (1-star and 5-star) ratings over time, and assesses its impact on review informativeness and business sales. Using Large Language Models (LLMs) to analyze Yelp reviews from 2005 to 2021, we distinguish between two main contributors to polarization: (1) the scale effect — shifts in rating standards, and (2) the content effect — changes in consumer experiences revealed in review texts. Our analysis shows that the surge in 5-star reviews is mainly due to the scale effect, reflecting more lenient rating standards, whereas the increase in 1-star reviews is largely driven by the content effect, indicating genuine declines in customer satisfaction. By analyzing votes on the usefulness of reviews and integrating Yelp data with Texas restaurant sales data, we demonstrate that the scale effect significantly reduces the informativeness of reviews and affects sales through displayed ratings. Our findings offer valuable insights for businesses and review platforms seeking to understand and navigate the changing dynamics of online review systems.

[*] UBC Sauder School of Business, University of British Columbia. Email: limin.fang@sauder.ubc.ca

[†] Corresponding author. UBC Sauder School of Business, University of British Columbia. Email: chunhua.wu@sauder.ubc.ca

[‡] Cheung Kong Graduate School of Business (CKGSB). Email: bhsun@ckgsb.edu.cn

[§] We thank Shin Oblander and Khai Chiong for invaluable feedbacks. We greatly appreciate the comments from seminar and conference participants at Hong Kong Polytechnic University, Temple University, and 2024 INFORMS Marketing Science. All errors are our own.

1 Introduction

Consumer reviews significantly influence consumer decisions and market outcomes. Extensive research has documented the substantial impact of online reviews on product sales (e.g., Chevalier and Mayzlin, 2006; Moe and Trusov, 2011; Anderson and Magruder, 2012). A notable feature of online reviews is the prevalence of polarity, characterized by an increasing prevalence of extreme (1-star and 5-star) ratings over time, leading to a bimodal or J-shaped distribution patterns (Schoenmueller et al., 2020; Hu et al., 2017). Notably, while reviews were initially moderate, they have become increasingly polarized over the years, resulting in a distinct trend of **review polarization**.

Yelp, a prominent online consumer review platform, exemplifies this trend. According to the Yelp Open Dataset (2023), its review distributions were predominantly unimodal and heavily skewed prior to 2012 (see Figure 1). In and after 2012, the share of 1-star reviews surpassed that of 2-star reviews, and the share of 5-star reviews exceeded that of 4-star reviews, creating a bimodal distribution. Since then, the share of middle-range ratings has declined from about 65% in 2005 to 27% in 2021, while the shares of 5-star and 1-star ratings have increased significantly. If this trend continues, by 2028 nearly all reviews could be either 5- or 1-star, effectively making Yelp's scale binary. This polarization raises critical questions about the informativeness of reviews and their impact on consumer decision-making.

Understanding the factors that drive review polarization is essential. Existing literature has explored self-selection (e.g., Li and Hitt (2008), Hu et al. (2017), Kramer (2007)) and the impact-effort trade-off (e.g., Wu and Huberman (2008)), primarily focusing on numerical ratings while often overlooking review content. A major challenge in analyzing this trend lies in distinguishing two key factors: changes in numerical rating standards over time (the "scale effect") and changes in consumer experiences revealed in review texts (the "content effect"). For example, a 4-star experience detailed in a review a decade ago might now receive a 5-star rating due to inflation in rating standards; conversely, a 2-star experience might now be rated 1 star. These shifts in rating standards are spurious and can introduce noise into ratings, as evidenced by rating inflation trends in credit ratings (Frenkel, 2015) and education (The Wall Street Journal, 2023). Zervas et al. (2021) specifically notes the prevalence of extreme positivity in online reviews. In contrast, changes in review texts are more likely to reflect genuine shifts in consumer experiences, as texts are where consumers articulate and justify their ratings. With Yelp's five-star rating scale potentially becoming a de facto binary system, key questions emerge: Is the scale effect or the content

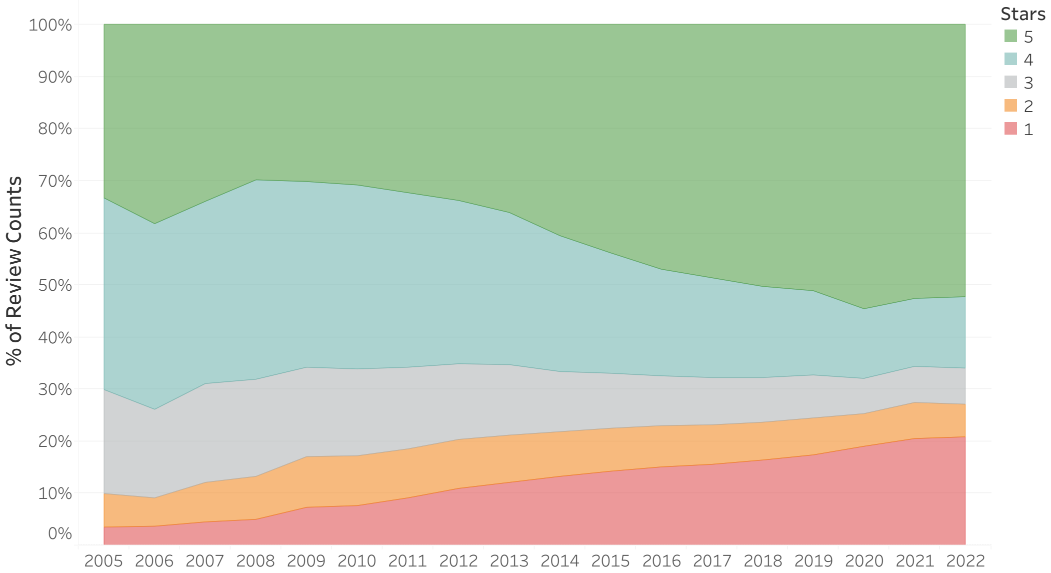

Figure 1: Distribution of Yelp Review Stars Over Years

effect driving this polarization? How does each effect impact review informativeness and consumer decision-making?

In this paper, we aim to achieve three research objectives: (1) to unravel the factors driving review polarization by distinguishing between shifts in rating scale and changes in revealed consumer experiences from content (the scale vs. content effects); (2) to explore the socioeconomic factors influencing both the scale and content effects; and (3) to investigate the impact of both effects on the informativeness of reviews and consumer purchase decisions. To achieve these objectives, we use the 2023 Yelp Open Dataset, which covers information on reviews, businesses, and users across 11 major metropolitan areas over 18 years, allowing us to track the long-term dynamics of reviews across diverse regions over time. We supplement this dataset with US demographic data and presidential election data to explore the relationship between review polarization and socioeconomic factors. Additionally, we collect restaurant sales data from the Texas restaurant market to assess the potential impact of review polarization on consumer purchase decisions.

To separate the scale effect from the content effect, we leverage recently developed large language models (LLMs) and develop prediction models mapping review texts to numerical ratings. We encode each review message into OpenAI's high-dimensional embedding space that captures the underlying semantics.[1] Using 2013 as our benchmark year, we train predictive models to es-

[1] Each message is transformed into a 1,536 dimensional vector. See `https://platform.openai.com/docs/guides/embeddings` for details.

tablish a consistent rating scale baseline. We then apply these benchmark-year model parameters to predict numerical ratings for reviews from other years, effectively holding the rating scale constant while allowing content to vary. We capture the content effect through the predicted ratings that reflect changes in consumer experiences on a fixed scale over time, and the scale effect by the difference between predicted and actual ratings. Our prediction models demonstrate strong performance, achieving an average accuracy rate of 70% — substantially higher than traditional sentiment analysis methods. This superior accuracy stems from LLMs' enhanced capability to understand subtle linguistic patterns and contextual relationships compared to conventional natural language processing approaches (Naveed et al., 2023; Asercion, 2024).

Our analysis reveals opposing patterns in the drivers of 5-star versus 1-star ratings across metros. The increase in 5-star reviews between 2013 and 2021 is predominantly driven by the scale effect (about 80%), suggesting that while the actual customer experiences described in review content have not significantly changed, users have become more inclined to award 5 stars instead of 4 stars. This result indicates significant scale inflation in 5-star ratings. Conversely, the increase in 1-star reviews is mainly due to the content effect (only 20% from the scale effect), implying a genuine decline in consumer experiences. We further account for business and user characteristics as well as internal review dynamics to capture mechanisms of self-selection and impact-effort trade-offs in the literature (Li and Hitt, 2008; Hu et al., 2017; Wu and Huberman, 2008). Even after controlling for these factors, we observe robust divergent patterns for 5-star and 1-star ratings.

We further examine how scale and content effects relate to socioeconomic factors, including demographic characteristics and political polarization at the county level. We find demographic factors, such as median age and the share of ethnic minorities, are positively correlated with the share of polarized reviews, with the correlation driven mostly by the scale effect. We also observe that higher political polarization correlates with fewer 1-star and more 5-star ratings, which aligns with prior research showing that consumers in politically polarized regions prefer products and establishments that match their ideologies, resulting in more positive reviews overall (Jost et al., 2017; Fernandes, 2020).

To assess the impact of scale and content effects on review informativeness, we analyze their relationship with the likelihood of a review being voted as useful. Results show that reviews with stronger scale effects are less likely to be deemed useful, suggesting that ratings more misaligned with their textual content are considered less informative. We further find a negative correlation between predicted ratings (based on content) and usefulness, indicating that reviews expressing negative experiences tend to be regarded as more helpful by users.

To further explore the impact of scale and content effects on business performance, we link Texas restaurant revenue data to Yelp business profiles. Using our LLM-based prediction model, we find both effects are significantly and positively associated with restaurant revenue, suggesting they influence consumer purchase decisions and that rating scale inflation likely drives up sales. Additionally, we find that higher shares of scale effects in displayed ratings reduce their impact on sales, indicating that consumers discount inflated ratings. Overall, these results demonstrate that while the scale effect reduces review informativeness, consumers adjust for rating inflation in their purchase decisions.

Our research makes two contributions to the literature. First, we document robust patterns of review polarization over time at both the business and reviewer levels. While existing studies either show a general decreasing trend in ratings (Chevalier and Mayzlin, 2006; Godes and Silva, 2012) or identify polarized reviews without considering temporal dynamics (Schoenmueller et al., 2020), our work captures the evolution of this phenomenon over time. Second, our paper is the first to distinctly separate and analyze the scale and content effects in online review dynamics. Identifying the contribution of the scale effect is especially important because inconsistency in rating scales can reduce the informativeness of reviews. We introduce a novel, practical, and efficient approach to measuring the scale effect in online reviews, demonstrating that scale inflation is a significant driver of review polarization and reduces review informativeness. Leveraging the advanced capabilities of Large Language Models (LLMs), our approach offers a more nuanced and sophisticated understanding of review content than traditional methods. While scale inconsistency has been shown to diminish signaling value in other contexts like education (Boleslavsky and Cotton, 2015; Rojstaczer and Healy, 2012), our work is the first to systematically analyze this phenomenon in online consumer reviews, made possible by recent advances in LLM technology.

Our study has important managerial implications: with recent advances in LLMs, platforms may consider adjusting ratings based on content to enhance credibility and consistency, given that consumers discount inflated ratings. Firms may need to find additional quality signals to stand out, as scale inflation blurs quality signals and can reduce incentives to maintain high-quality services.

The rest of the paper is organized as follows. Following the literature review, we describe our data and the LLM text encoding process in Section 2. Section 3 discusses the model and Section 4 presents the results. We conclude with policy and managerial implications in Section 5.

1.1 Literature Review

Our study contributes to several strands of literature. First, it relates closely to the extensive literature on the various ways that consumer reviews affect sales. For example, Chevalier and Mayzlin (2006) examine the impact of online reviews on book sales and find that positive reviews boost sales, but negative reviews have a greater impact in reducing sales, highlighting the asymmetric impacts. Similarly, Luca (2011) demonstrates that a one-star increase in Yelp ratings can lead to a substantial increase (5-9%) in revenue for independent restaurants. Liu (2006) and Dellarocas et al. (2007) show that review sentiment, volume, and valence all influence movie box office revenue. In addition to these static effects, a number of papers illustrate that these effects can change over time. For instance, Moe and Trusov (2011) find that early ratings can temporarily boost sales, but the effect diminishes as more reviews accumulate. Zhao et al. (2013) and Wu et al. (2015) build structural learning models to investigate how information acquisition from online reviews updates consumer product perceptions and affects product choices; Wu et al. (2015) further quantifies the economic value of online reviews to consumers and firms. Fang (2022) emphasizes how review platforms facilitate consumer learning about restaurant quality and show that tourists benefit from online reviews much more than local repeat consumers. Our work builds on these studies by measuring the impact of the scale and content factors on sales.

Second, our paper complements existing studies on the characteristics of online reviews. Review characteristics, such as polarity, extreme positivity, and the presence of fake reviews, significantly impact consumer decision-making and market outcomes. Dellarocas et al. (2007) show the bimodal pattern of review distributions, where reviews cluster at the extremes — very positive or very negative. Hu et al. (2009) find that acquisition bias and under-reporting bias contribute substantially to the J-shaped distribution of online reviews. Bayerl et al. (2023) explore factors influencing review distributions, such as the weekend effect, which affects the timing and nature of reviews posted by consumers. Chen et al. (2021) demonstrate how self-selection bias leads to inflated ratings, particularly for niche products, which could misguide firms' pricing and product assortment decisions. Moe and Schweidel (2012) further investigate how review polarity can influence consumer perceptions and the overall informativeness of reviews. Brandes et al. (2022) explore the role of reviewer attrition, where the loss of moderate reviewers over time leads to a disproportionate number of extreme reviews. This extremity bias affects the overall balance of the review distributions and the informativeness of reviews. Schoenmueller et al. (2020) expand on this extremity aspect by analyzing the polarity of reviews across multiple online platforms, and find that extreme evaluations from biased reviewers significantly impact the usefulness of reviews for decision-making. Similarly, De Langhe et al. (2016) stress that while attracting attention, extreme reviews are often perceived as less informative compared to more balanced reviews, suggesting that reviews need to strike a balance between detail and neutrality in order to be useful. To mitigate the negative impacts of review biases, especially those from extreme reviews, several studies have sought solutions. Karaman (2021) finds that soliciting reviews from moderate experience customers can lead to a more balanced distribution. Pocchiari et al. (2023) show that review-update solicitations help companies manage negative feedbacks by effectively encouraging positive updates. Li (2016) demonstrate that promotions can counterbalance initial negative ratings and low review volumes. Park et al. (2023) assess mandatory disclosure of incentivized reviews, finding the policy ineffective at reducing rating inflation. They suggest alternative approaches.

Other papers further investigate and stress the aspect of review inflation. For example, Aziz et al. (2023) explores the impact rating inflation in the restaurant industry, finding that it reduces user trails and makes sales more concentrated among popular restaurants. Zervas et al. (2021) study the prevalence of extreme positivity in online reviews, particularly on platforms like Airbnb, where users leave overwhelmingly positive feedbacks. This extreme positivity can distort consumers' perceptions of product or service quality, leading to higher expectation and potential dissatisfaction. Luca and Zervas (2016) address the issue of fake reviews, which are often characterized by extremely positive ratings and could mislead consumers and damage business reputations. Similarly, Filippas et al. (2020) discuss the deterioration of rating system effectiveness due to rating inflation, noting that the pressure to leave positive feedback erodes the informativeness of ratings. Our study uncovers an important driver of rating inflation, which is the overwhelming scale inflation in extreme positive reviews (i.e. 5-star ratings.)

Third, our study extends the literature on review dynamics. Among others, Godes and Silva (2012) find that initial reviews set a benchmark which influences subsequent feedbacks, leading to a decrease in ratings over review sequence. This decrease is likely due to a self-selection mechanism, where early positive reviewers are followed by less enthusiastic users, a theory also shared by Wu and Huberman (2008). Li and Hitt (2008) model this bias in Amazon book reviews and show that encouraging early positive reviews can boost new product sales, but this bias, if left unaddressed, may negatively impact consumer surplus. Moe and Schweidel (2012) demonstrate how previous reviews influence future review content. Recent studies by Park et al. (2021) and Karlinsky-Shichor and Schoenmueller (2023) further emphasize the influence of early reviews on sales and subsequent reviews, highlighting the significant effect of a product's first review and "oracle reviewers," whose early reviews reliably predict which products will become popular. Our study contributes to this literature by examining the drivers through the lens of the scale and content effects.

Fourth, our study relates to the literature on the interplay between social media, political polarization, and market outcomes. For example, Levy (2021) reveals that social media algorithms may limit exposure to opposing views, and thereby reinforce biases and contribute to ideological echo chambers. These effects influence not only political opinions but also consumer purchase behaviors. Jost et al. (2017) and Fernandes (2020) discuss how political ideologies shape consumer activism, noting that liberals are more likely to engage in boycotts and buycotts because they have a greater propensity to challenge institutions. Schoenmueller et al. (2023) explore how political polarization translates into preference polarization and show that political ideologies significantly impact brand preferences and purchase behaviors. Our study adds to this literature by examining the relationship between political polarization and review polarization over time. We find that politically polarized counties tend to have more 5-star reviews and fewer 1-star reviews. This finding is consistent with those in Jost et al. (2017) and Fernandes (2020), which state that in polarized markets, consumers only buy products that align with their ideology, leading to overall higher ratings.

Last but not least, our research adds to the growing body of literature that deploys the language-processing ability of LLMs. Among others, Yoganarasimhan and Yakovetskaya (2024) and Ye et al. (2024) pioneer research in this area. Yoganarasimhan and Yakovetskaya (2024) use LLMs to detect polarization in news content and find that these models are very effective. Ye et al. (2024) builds on this finding and explores how LLMs can be used to predict the appeal of various news content and potentially replace traditional experimentation methods to promote user engagement on news platforms. LLMs have also been used to simulate consumer preferences or responses to survey questions. For example, Li et al. (2024) show that LLM-generated survey responses are very similar to those from real humans with a high agreement rate of over 75%. Brand et al. (2023) demonstrates that GPT-3.5 accurately captures consumer behaviors and gives realistic estimates of consumer willingness-to-pay for products. Goli and Singh (2023) explores these aspects from the perspective of consumer intertemporal choices, finding that GPT models appear much less patient than humans and cautioning against using GPT to simulate consumer preferences. Our study adds to this strand of research by leveraging the efficiency in contextual understandings provided by the

LLM embeddings to analyze review content and its evolution over time.

2 Data and Polarization Patterns

We discuss our data sources in Section 2.1. The data on review show a consistent pattern of polarization over time, even after controlling for geography, business category, reviewer cohort, and reviewer tenure. We demonstrate these patterns in Section 2.2.

2.1 Datasets

The data used in this study come from multiple sources. To separate the scale effect from the content effect in polarization, we use the Yelp Open Dataset 2023, which includes nearly 7 million reviews written by almost 2 million users for over 150,000 businesses in 11 metropolitan areas from 2005 to 2021. This comprehensive dataset enables us to use LLMs to train predictive models that explore the connections between review content and ratings. To link review polarization to broader social changes, we collect demographic data from the decennial census and the American Community Survey for areas covered by the Yelp Dataset during the sample period. Additionally, we obtain Presidential election voting data at the county level from the MIT Election Lab, which includes the percentage of votes for Democrats, Republicans, and other parties in each county for the 2004, 2008, 2012, 2016, and 2020 elections. Using this dataset, we construct a measure of political polarization based on the method provided in Pan et al. (2024).[2]

To examine the potential impact of the scale and content effects on consumer purchase decisions, we construct a dataset linking Yelp reviews to business sales by combining multiple data sources. First, we obtain revenue data from the Texas Mixed Beverage dataset, collected by the Texas Comptroller Office of Public Accounts. The dataset includes monthly revenue from alcoholic drinks at the establishment level in Texas from January 1993 to September 2021. Although this dataset does not include food sales, changes in alcoholic drink sales are indicative of fluctuations in restaurant food sales, as consumers often have drinks with their meals when dining out.[3] Second, we collect Yelp reviews for these restaurants, including review text, ratings, and user IDs. Since

[2] The political distance between two counties A and B is calculated as the sum of the absolute differences between the fractions of Democratic, Republican, and independent voters: $Political\ Distance_{AB} = |d_A - d_B| + |r_A - r_B| + |o_A - o_B|$, where d, r and o represent the percentages of Demographic, Republican, and independent voters, respectively. A county's political polarization is measured by the average of this county's political distance to all other counties.

[3] Fang (2022) shows that a restaurant's alcoholic drink sales account for a roughly constant proportion of total sales.

metropolitan areas in Texas are not part of the Yelp Open Dataset, we gather this information using the Yelp API business search to identify restaurants listed on Yelp and then visit individual Yelp pages to collect review and user data. Compared to information in the Yelp Open Dataset, the Yelp review data we collected include only those restaurants that are part of the revenue dataset instead of all businesses ever listed on Yelp. Finally, we complement these data with market information, such as demographics and income, from the decennial census and the American Community Survey. In total, our dataset includes 11,096 restaurants with Yelp profiles in Texas.

2.2 Data Patterns for Review Polarization

The Yelp Dataset reveals a notable trend of review ratings becoming more polarized over time. We showcase this overall pattern and further analyze it from various dimensions, considering factors such as business turnover, reviewer turnover, and review content.

Overall Pattern As shown in Figure 1, there is a very clear and consistent pattern of review polarization on the Yelp platform over the years. Before 2010, the percentage of 5-star and 1-star reviews combined (referred to as the review polarity index) was less than 40% of the total reviews. However, by 2021, over 70% of the reviews were either 1 star or 5 stars. In the top panel of Figure 2, we show the increasing trend of this polarity index over time, with the projected values and confidence intervals beyond 2021. If this trend continues, almost all review ratings will be either 1 or 5 stars by the year 2028. Another discernible pattern in the review distribution is the concurrent trend in the ratio between the shares of 5-star and 1-star reviews, defined as “positive imbalance” illustrated in the lower panel of Figure 2. The ratio started at nearly 4 in 2010 with a decreasing trend over time; by 2021, this ratio had dropped to around 2.5. This trend indicates that the share of 1-star reviews has been increasing at a faster rate than that of 5-star reviews.

Patterns by Business Location and Type The polarization pattern may be driven by business turnover. Specifically, if more businesses with extremely low or high quality enter the market over time, we should expect to see a polarizing pattern of reviews. To examine this potential driver, we show the distribution of review ratings over time by business location and type. Figure 3 presents the pattern for each of the 11 metropolitan areas in the Yelp Dataset. Overall, there is a consistent and persistent trend of review polarization in each area, although the degree of polarization differs across metropolitan areas. Figure 4 displays the pattern by business category.

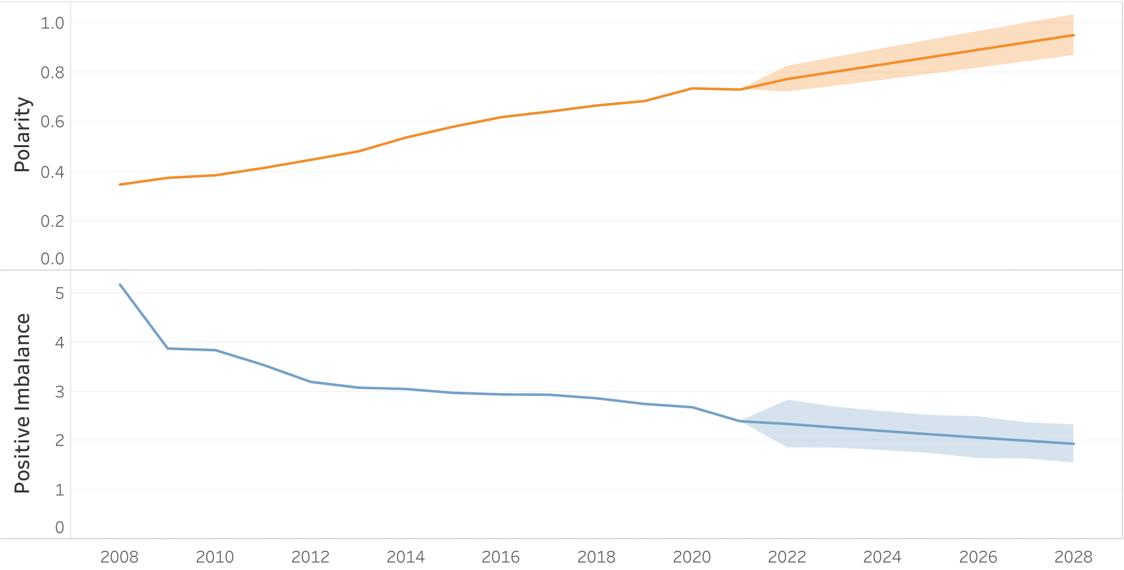

Figure 2: Yelp Review Polarity and Positive Imbalance

Similar trends of review polarization can be observed for each category as well. Interestingly, the Food and Restaurants categories are less polarized than other categories such as Automotive and Shopping. For example, in the Automotive category, less than 20% of the review ratings were in the middle range in 2013, whereas for Restaurants, this share was over 50%.

In addition to category, businesses of different quality levels may experience different trends in review polarization. For example, those businesses rated extremely high or extremely low might face boundary constraints in receiving 5 star or 1 star ratings. Figure 5 examines this aspect by graphing the distributions of ratings based on the overall average star ratings of businesses; the overall average ratings are calculated by averaging all reviews up to 2021 for each business. As shown in the figure, the overall polarization pattern persists across quality levels. For each rating range, there is a consistent pattern of review polarization, even for the highest and lowest rating categories. Nonetheless, as anticipated, businesses with lower overall average ratings receive higher shares of 1-star ratings, whereas businesses with higher overall average ratings receive greater shares of 5-star ratings.

Patterns by Reviewer Type In addition to business turnover, another significant contributor to review polarization could be the behavioral changes of reviewers over time or reviewer turnover. New cohorts of reviewers may be more extreme than earlier ones, or the same reviewers may become more extreme over time. Figure 6 shows the review distribution by reviewer cohort, defined by the

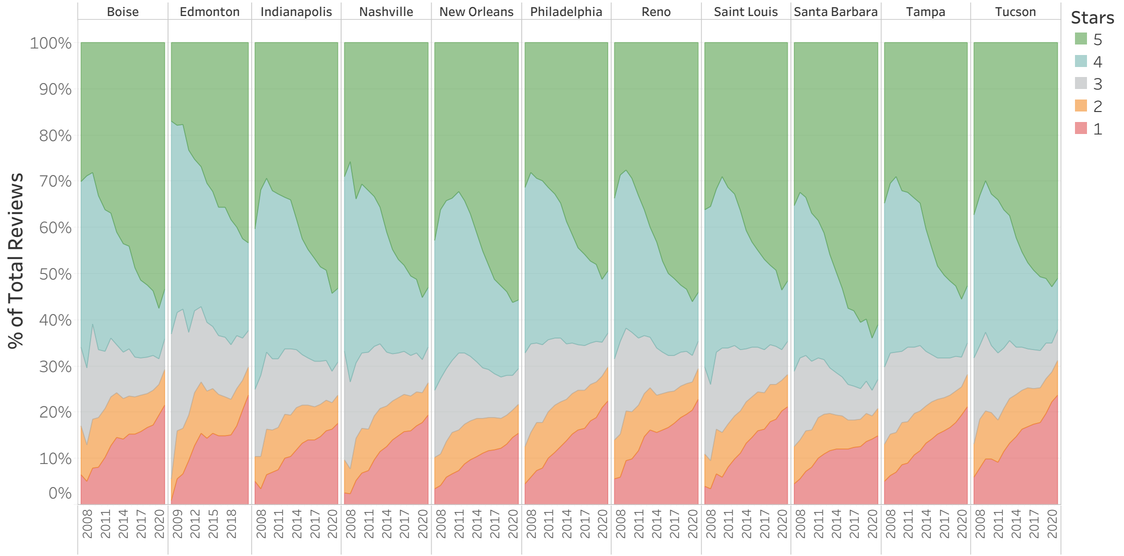

Figure 3: Yelp Review Polarization across Metropolitan Areas

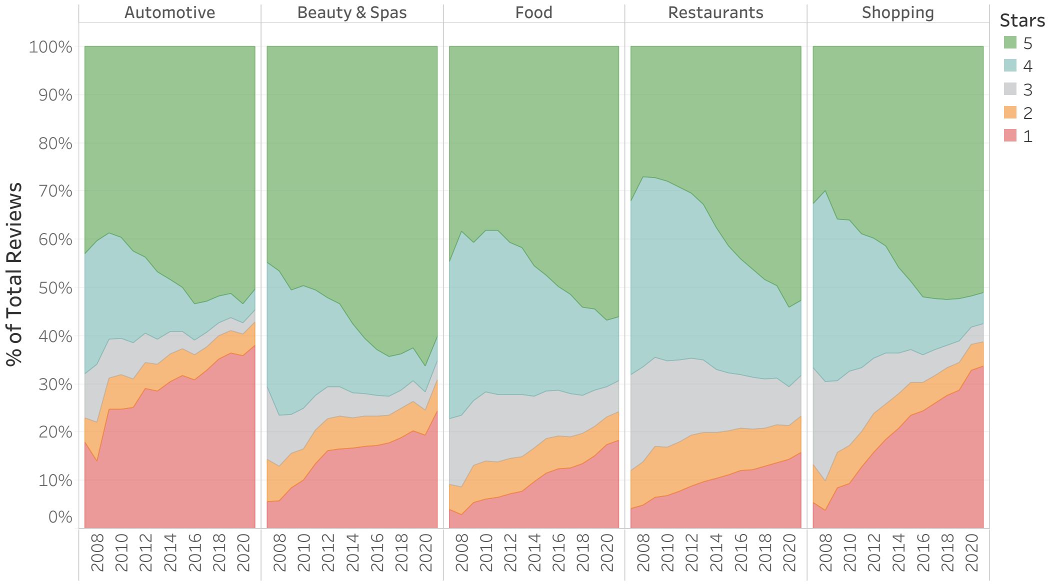

Figure 4: Yelp Review Polarization across Business Categories

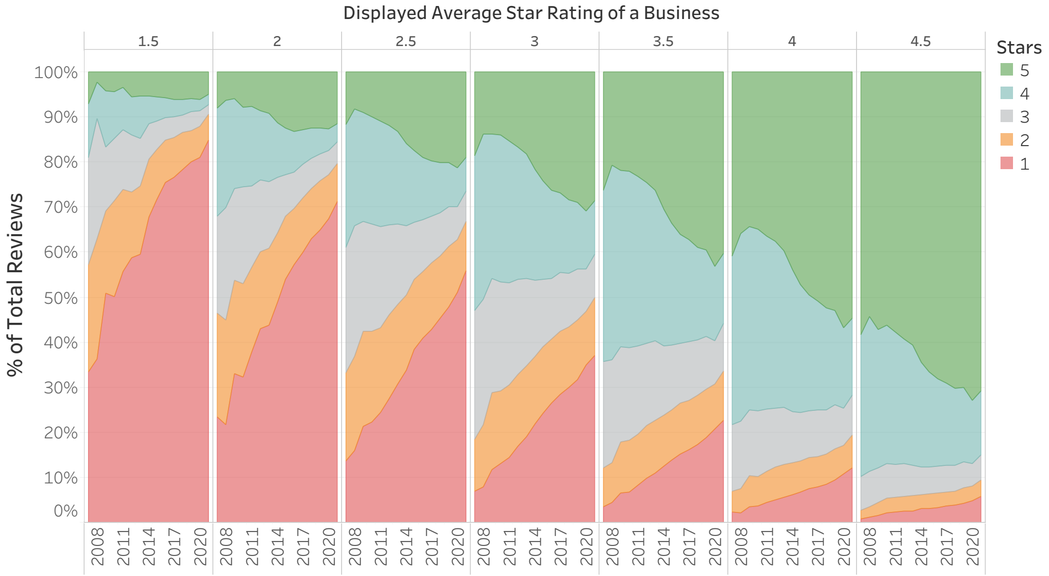

Figure 5: Yelp Review Polarization across Business Star Levels

year a reviewer joins the platform. There is a distinct pattern of cohort effect in polarization: the shares of 1- and 5-star ratings given by more recent users are much higher than those given by earlier users. For instance, users who joined after 2016 are over 60% more likely to give extreme ratings than those who joined before 2008. Figure 7 further decomposes the pattern by reviewer tenure, illustrating the long-term dynamics of reviews given by each user cohort. The figure shows clear evidence of review polarization over time, even within the same reviewer cohort. Specifically, reviewers who joined in 2008 are almost twice as likely to give extreme ratings in 2021 as they were when they first joined the platform.

Patterns by Review Content A key question underlying review polarization is its impact on the informativeness of reviews over time. We analyze the difference in the distribution of more informative versus less informative review messages. To classify the reviews into these categories, we use users' votes on whether a review is considered "useful" on the platform. Figure 8 shows the distribution of reviews with useful votes and those without. As illustrated, there is a consistent pattern of review polarization in both categories. The only exception is the decrease in the share of 1-star ratings for "useful" reviews and the decline in the share of 5-star ratings for "non-useful" reviews towards the end of the sample period (2020 to 2021). These changes in the last two years should be interpreted cautiously, as it takes time for reviews to receive "useful" votes, and the votes are sparse, particularly during the COVID pandemic.

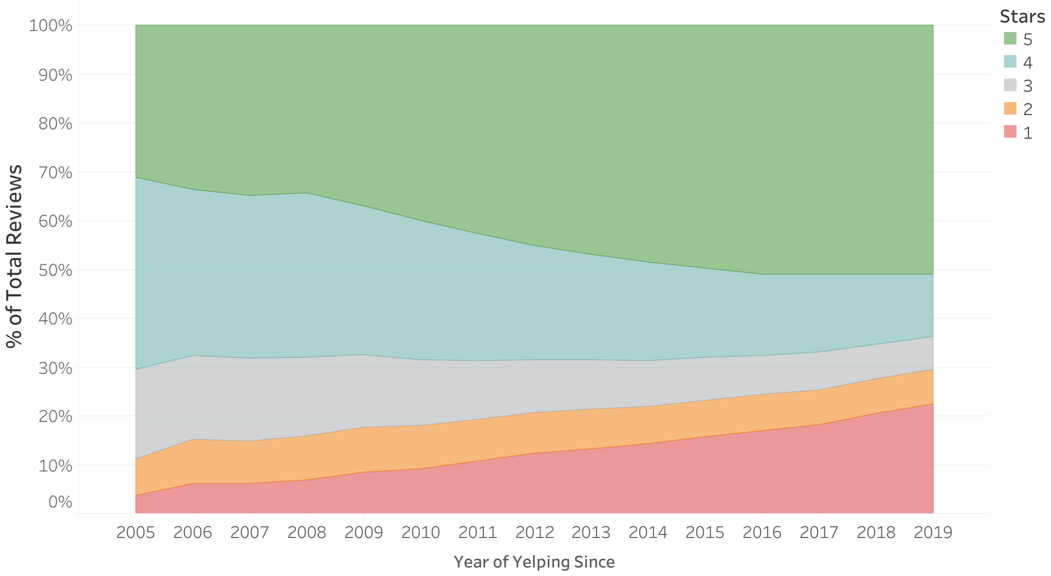

Figure 6: Yelp Review Polarization by Reviewer Cohort

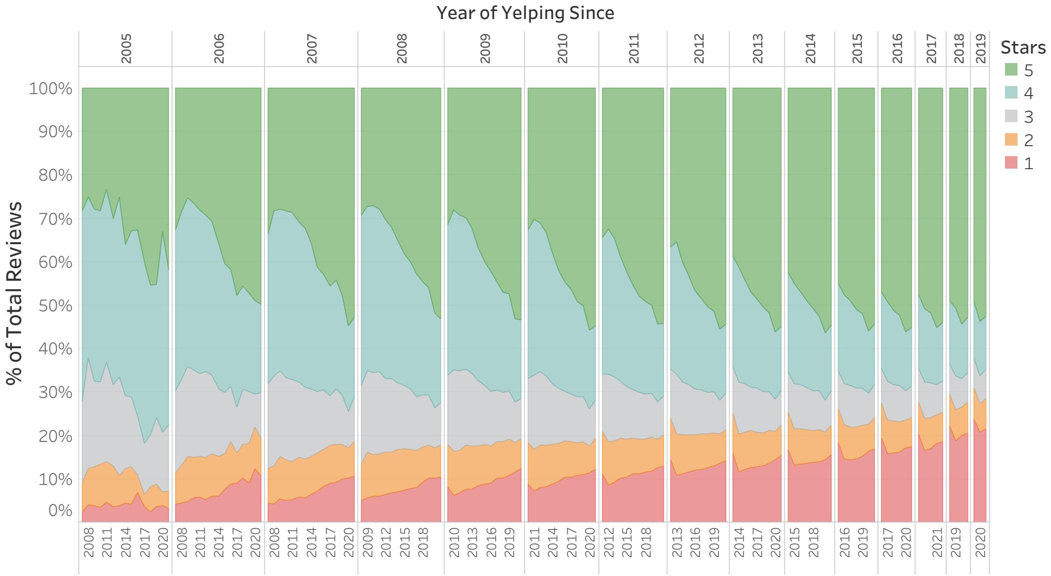

Figure 7: Yelp Review Polarization by Reviewer Tenure

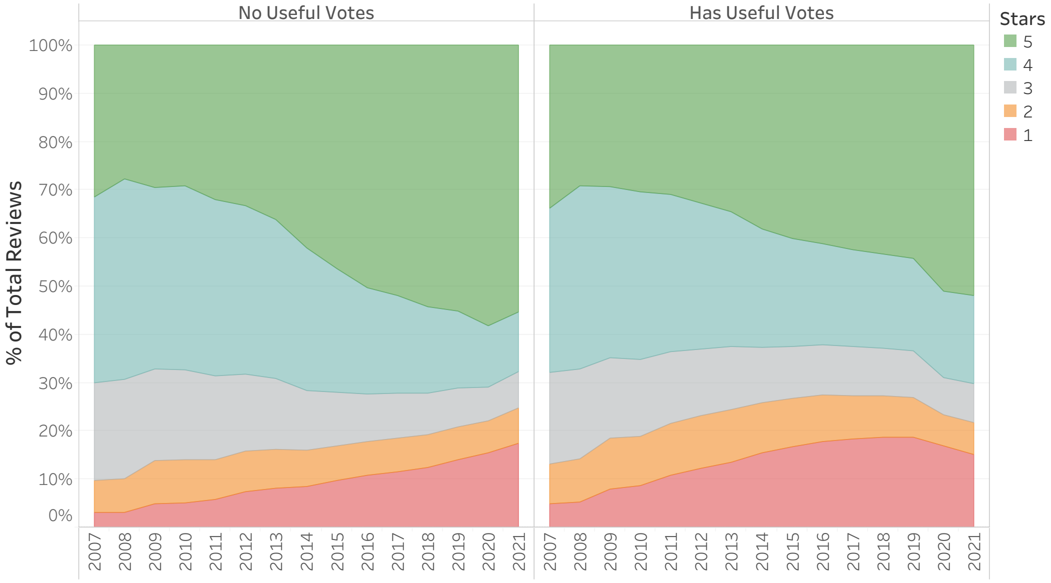

Figure 8: Yelp Review Polarization across Review Informativeness

3 Conceptual Framework and Empirical Approach

We first introduce our conceptual framework of the review generation process; in particular, what factors may contribute to the formation of review content and ratings. Our conceptual framework is summarized in Figure 9. Consumers write textual reviews to describe their experiences and share their evaluations through numeric ratings. Consumer experiences may change over time, and these changes can be captured in both the review text and the numeric ratings. The review text often provides detailed descriptions of experiences and justifications for the ratings given.

When examining review polarization, we focus on the separate evolutions of review content and ratings, particularly the evolution of the mapping between review text and ratings. This mapping, highlighted by the red arrow in the figure, is of special interest because aggregated numerical ratings are often the first information consumers see on a review platform. If the ratings are not consistently justified by the experiences shared in text over time, this inconsistency can undermine the credibility and informativeness of the ratings. The education literature has long established that inconsistency in grading scales reduces the signaling value of grades (Boleslavsky and Cotton, 2015; Rojstaczer and Healy, 2012). A similar outcome can easily occur in the consumer review setting. In this paper, the consumer experience shared in text is referred to as "content," and the mapping between text and rating is referred to as "scale."

Both review content and scale can evolve over time, influenced by various factors such as reviewer characteristics, business characteristics, timing and sequence of reviews, demographics, macroeconomic environment, and political polarization. Reviewer characteristics like tenure and review activity can affect the experiences shared in the review text; for example, more active reviewers may be less harsh than those who rarely post. Business characteristics, such as category and quality, are crucial in shaping consumer experiences. High-quality businesses are likely to give consumers better experiences and receive higher ratings.

Mechanisms like self-selection and impact-effort trade-off influence the types of consumers who write reviews throughout a business's lifecycle. As shown in the literature (Godes and Silva, 2012; Schoenmueller et al., 2020; Li and Hitt, 2008; Wu and Huberman, 2008), more enthusiastic reviewers are likely to try out a business early and write positive reviews. In contrast, less enthusiastic ones visit later and share more negative reviews or write reviews only if they substantially overturn the average of past ratings. This process can lead to a decrease in both ratings and consumer satisfactions shared in review text over time. Therefore, the timing, sequence, and average ratings of past reviews all contribute to the consumer experiences shared and the ratings given. In addition, broader socioeconomic trends such as demographics, macroeconomic conditions, and political polarization shape consumer experiences and rating standards over time.

Furthermore, reviewers' language use can influence both content and scale. Studies show that language use in computer-mediated communication varies based on participant characteristics, context, and geography (Squires et al., 2012; Vandekerckhove and Nobels, 2010). While individual language styles may remain stable, new reviewers and businesses can introduce different language patterns, affecting the mapping between review text and ratings. Thus, changes in the rating scale in our setting may result from evolving reviewer and business characteristics.

More importantly, societal trends like leniency in rating standards can cause overall scale inflation or deflation, leading to a distinct time trend of review polarization even after controlling for the aforementioned factors. Understanding the driving forces behind the evolution of review content and scale helps us decipher the trend of review polarization and assess its potential impact.

Building on this conceptual framework, we propose a model to disentangle the scale effect from the content effect in review dynamics and examine the factors influencing these effects. We first demonstrate our LLM encoding approach in Section 3.1 and then describe our model that deploys LLM embeddings to map review content to ratings in Section 3.2. We then decompose review dynamics into the scale and content effects in Section 3.3. In Section 3.4, we discuss the regression models to explore the socioeconomic drivers of the scale and content effects. Then in Section 3.5,

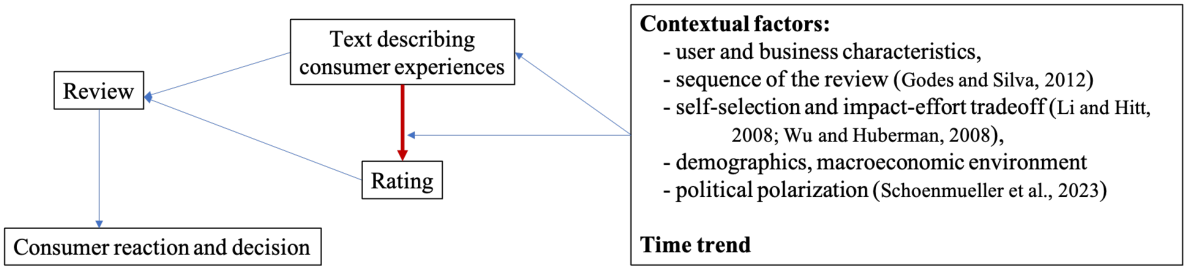
Text describing consumer experiences
Review
Rating
Consumer reaction and decision
Contextual factors:
- user and business characteristics,
- sequence of the review (Godes and Silva, 2012)
- self-selection and impact-effort tradeoff (Li and Hitt, 2008; Wu and Huberman, 2008),
- demographics, macroeconomic environment
- political polarization (Schoenmueller et al., 2023)
Time trend

Figure 9: Conceptual Framework of Review Generation Process

we describe the regression models to examine the impact of the scale effect on the informativeness of reviews and consumer purchase decisions. All results are presented in Section 4.

3.1 Encoding Yelp Review Text Using LLMs

To differentiate content from scale in reviews, we take advantage of the recently developed large language models (LLMs) to encode review texts. The output of the encoding process provides input to our prediction model to map review content to ratings. The existing literature that examines review content often relies on two main approaches: a) sentiment analysis, which focuses on the frequency of certain words (Fang, 2022; Hollenbeck, 2018; Cui et al., 2023); and b) bag-of-words analysis, which defines collections of specific keywords into content topics (Cummins et al., 2018). A commonly used tool for this approach is the Linguistic Inquiry and Word Count (LIWC) software[4], frequently employed in psychology and social science research (Tanana et al., 2021). Although these tools are informative, they have limitations in handling negation, sarcasm, multi-word expressions, and contrastive conjunctions, especially when word meanings are context-specific (Ravi and Ravi, 2015; Tanana et al., 2021).

The latest advances in Large Language Models (LLMs) present new opportunities to leverage state-of-the-art language processing tools to analyze online review content. The exceptional performance of LLM products, such as ChatGPT and Claude, in comprehending human language has been well-documented in research across various fields (Brown et al., 2020; Naveed et al., 2023; Asercion, 2024). These models typically encode texts into high-dimensional numeric values, known as the embedding space. A vector in the space encapsulates the semantic meanings of the underlying text and the contextual connections between text blocks efficiently. These vectors then serve as the foundation for further decoding and analysis. Consequently, embedding spaces are the

[4] See details in `https://liwc.app`.

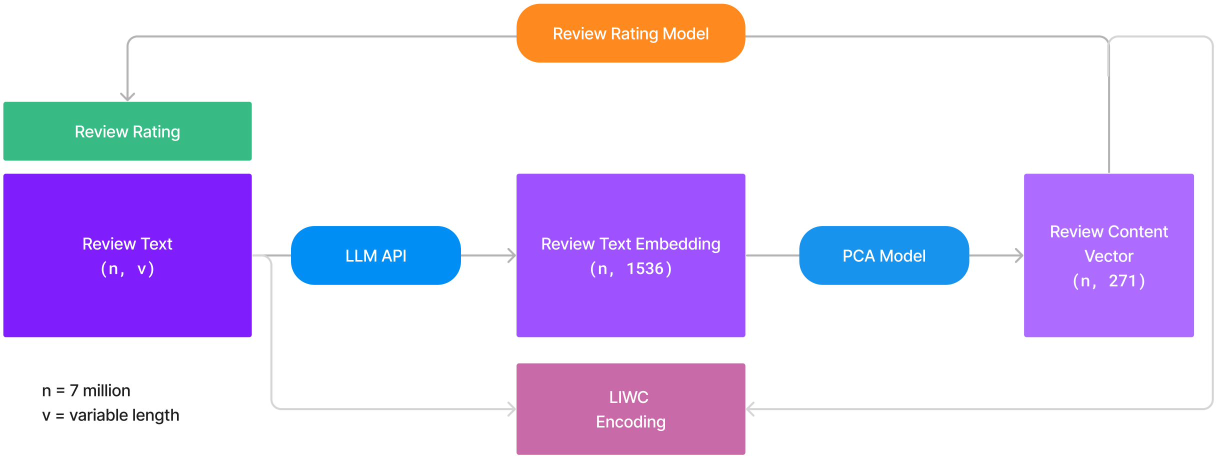

Figure 10: Review Content Encoding Process

fundamental building blocks for generative artificial intelligence in text processing.

We present a diagram of the content encoding process in Figure 10. We query OpenAI's API to obtain the embeddings for all the review messages in our dataset[5]. This process translates each review message into a $1,536$-dimensional vector. Obtaining embedding values directly from pretrained models without any task-specific training data or fine-tuning the model for a particular task is known as the zero-shot approach. This approach has been confirmed by the Yelp engineering team to deliver performance comparable to that of customized models, making it an efficient method for developing machine learning products.[6].

These embeddings convert the review text into numerical vectors, which can serve as input to machine learning models to predict ratings. However, the dimension of 1,536 is quite large. Training and validating prediction models with the full set of vectors and millions of observations would be computationally demanding. To overcome these challenges, we reduce the dimensionality of the embedding space by implementing a principal component analysis (PCA), which captures 80% of the variance in the embedding data with only 271 principal components. This represents a significant reduction in dimensionality compared to the original embedding space.

We experimented with various machine learning (ML) models to predict ratings using the full set of embeddings, including multinomial logistic regression (MNLogit) and deep neural networks (DNN). However, their predictive power did not surpass that of the models using the principal com-

[5] We use OpenAI's recommended 'text-embedding-ada-002' model as of October 2023. For details, see `https://platform.openai.com/docs/guides/embeddings`.

[6] For more information, see `https://engineeringblog.yelp.com/2023/04/yelp-content-as-embeddings.html`.

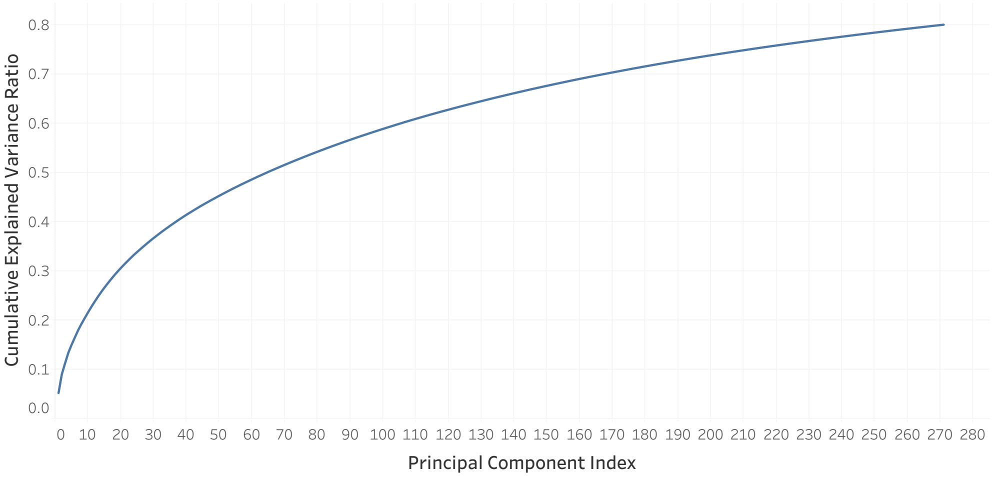

Figure 11: Scree Plot of Principal Component Analysis

ponents. In many cases, due to model convergence issues related to the large number of predictors, models with the original embeddings generated inferior predictions. The superior performance of principal-component-based models demonstrates that the principal components extracted the most relevant contextual information in online reviews. The missing information from the other 20% of the variance in the original embeddings was not very relevant for predicting ratings.

With the principal components in hand, we rank them based on their eigenvalue contributions to the original embedding space. We refer to them based on their rankings henceforth. For example, principal component 1 (PC1) contains the most information by capturing the largest variation across the review messages. We present the scree plot of the PCA analysis in Figure 11. The first 10 principal components capture over 20% of the total data variation, and the first 50 principal components capture nearly 50%. Once we get to the 271st principal component, the total data variation captured is 80%.

In addition to utilizing the OpenAI encoding process, we also encode the review text using the LIWC package. Specifically, for each review text, we use LIWC to create the following key psychometric measures[7]: WC (Word Count), WPS (Words per Sentence), Analytic, Clout, Authentic, Tone, BigWords, Linguistic, Drives, Cognition, Affect, Social, Culture, Lifestyle, Physical, Perception, and Conversation. In Section 4.2, we further link the principal components to these psychometric measures to illustrate more intuitively the semantic meanings of the most important principal components.

[7] For a full description of the procedure and list of available measurements, see `https://www.liwc.app/help/psychometrics-manuals`.

3.2 Predicting Review Ratings

Given review numeric ratings are in ordinal scales, we build an ordered logit model to predict the review ratings based on review content. Of all the ML prediction models (e.g. MNLogit or DNN) we experimented with, the ordered logit model with principal components perform the best. Let r_i denote the rating posted in a review indexed by i, $\boldsymbol{C}_i$ the review content, and $r_i^* = \boldsymbol{C}_i'\boldsymbol{\beta}$ the content-based score for the review rating.[8] The model is expressed as follows:

$$P(r_i = k) = \frac{1}{1 + \exp\big(-(\alpha_k - \boldsymbol{C}_i'\boldsymbol{\beta})\big)} - \frac{1}{1 + \exp\big(-(\alpha_{k-1} - \boldsymbol{C}_i'\boldsymbol{\beta})\big)}, \tag{1}$$

where $k \in \{1, 2, 3, 4, 5\}$ is the rating level from 1 to 5. Because Yelp uses a five star rating system, $\alpha_0 = -\infty$ and $\alpha_5 = \infty$ by definition. We need to estimate the coefficients $\boldsymbol{\beta}$, and thresholds $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)$. We use the principal components derived from the raw embeddings as the values representing the content $\boldsymbol{C}_i$.

To account for regional heterogeneity, we treat each metropolitan area and year combination as one entity and estimate a separate model for each. Each model we estimate can be represented as $P(r_{i,t}) = f(r_{i,t}|\boldsymbol{C}i; \boldsymbol{\theta}_t)$, where $\boldsymbol{\theta}_t = (\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t)$ is metropolitan and time specific. We also train deep neural networks for the prediction of review ratings based on the content. Overall, we do not observe a meaningful performance improvement over the simpler ordered logit model.

A significant advantage of the ordered logit model is the interpretability of its parameters. In particular, $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ capture the mapping between the review content and the rating and represent the scale factor in reviews. By comparing $\boldsymbol{\alpha}$'s for the same area over the years, we can examine the aggregate trends in scale for giving extreme ratings. Similarly, comparing the magnitudes of $\boldsymbol{\beta}$'s over the years allows us to identify differences in the weights of the underlying semantic elements that determine ratings.

With the estimated parameters $\boldsymbol{\beta}$, we can calculate the marginal impact of each content element C^c (i.e., each principal component indexed by c) as the change in the rating scale due to a one-standard-deviation change in C^c. Both review content embeddings and the principal components are vectors that capture complex features of review texts, making them difficult to interpret at face value. To identify the most important semantic elements driving review ratings, we rank the content elements based on their marginal impacts in the model and investigate the semantic meanings of the top ten elements through regression models that link them to psychometric measures from the

[8] The apostrophe here means transpose.

LIWC library. This process helps us partially decode the meanings of the LLM embeddings. These linkages will be shown in Section 4.

3.3 Decomposing Review Polarization

Using the review rating prediction model, we decompose the evolution of reviews into content effects (changes in $\boldsymbol{C}$) and scale effects (shifts in $(\boldsymbol{\alpha}, \boldsymbol{\beta})$). For a specific market (county), we express the distributional changes in review ratings from year s to year t as:

$$
\begin{aligned}
\Delta_{st}(R^k) =& R_t^k - R_s^k \\
=& \hat{R}_t^k - \hat{R}_s^k + (R_t^k - \hat{R}_t^k) + (R_s^k - \hat{R}_s^k) \\
=& \int_{\boldsymbol{C_t}} f(r = k|\boldsymbol{C}_t; \boldsymbol{\theta}_t) g(\boldsymbol{C}_t) \mathrm{d}\boldsymbol{C}_t - \int_{\boldsymbol{C_s}} f(r = k|\boldsymbol{C}_s; \boldsymbol{\theta}_s) g(\boldsymbol{C}_s) \mathrm{d}\boldsymbol{C}_s + \varepsilon_t^k + \varepsilon_s^k \\
=& \left(\int_{\boldsymbol{C_t}} f(r = k|\boldsymbol{C}_t; \boldsymbol{\theta}_t) g(\boldsymbol{C}_t) \mathrm{d}\boldsymbol{C}_t - \int_{\boldsymbol{C_t}} f(r = k|\boldsymbol{C}_t; \boldsymbol{\theta}_s) g(\boldsymbol{C}_t) \mathrm{d}\boldsymbol{C}_t \right) + \\
& \left(\int_{\boldsymbol{C_t}} f(r = k|\boldsymbol{C}_t; \boldsymbol{\theta}_s) g(\boldsymbol{C}_t) \mathrm{d}\boldsymbol{C}_t - \int_{\boldsymbol{C_s}} f(r = k|\boldsymbol{C}_s; \boldsymbol{\theta}_s) g(\boldsymbol{C}_s) \mathrm{d}\boldsymbol{C}_s \right) + \varepsilon_t^k + \varepsilon_s^k \\
=& \Delta_{st}^S(R^k) + \Delta_{st}^C(R^k) + \varepsilon_{st}^k,
\end{aligned}
\tag{2}
$$

where R_t^k is the proportion of k-star reviews in year t, and $\hat{R}_t^k$ is the model-predicted proportion, with the prediction error ε_t^k. $\hat{R}_t^k$ is calculated by integrating individual predicted probabilities $f(r = k|\boldsymbol{C}_t; \boldsymbol{\theta}_t)$ based on review contents $\boldsymbol{C}_t$, where $g(\cdot)$ is the probability density function of $\boldsymbol{C}_t$. The decomposition yields two components: the scale effect $\Delta_{st}^S(R^k)$, measuring changes in rating standards (S), and the content effect $\Delta_{st}^C(R^k)$, capturing shifts in consumer experiences expressed in the content (C). [9].

This type of distribution-based decomposition can be done at the individual review level as well. Specifically, we can write the difference between the probabilities of two reviews r_i and r_j having a rating k as

$$
\begin{aligned}
\Delta_{st}(P_{ij}^k) =& f(r_i = k|\boldsymbol{C}_i; \boldsymbol{\theta}_t) - f(r_j = k|\boldsymbol{C}_j; \boldsymbol{\theta}_s) + \varepsilon_i^k + \varepsilon_j^k \\
=& (f(r_i = k|\boldsymbol{C}_i; \boldsymbol{\theta}_t) - f(r_i = k|\boldsymbol{C}_i; \boldsymbol{\theta}_s)) + (f(r_i = k|\boldsymbol{C}_i; \boldsymbol{\theta}_s) - f(r_j = k|\boldsymbol{C}_j; \boldsymbol{\theta}_s)) + \varepsilon_i^k + \varepsilon_j^k \\
=& \Delta_{st}^S(P_i^k) + \Delta_{st}^C(P_{ij}^k) + \varepsilon_{ij}^k,
\end{aligned}
\tag{3}
$$

[9] Our empirical model utilizing LLM embeddings fits the observed ratings at the distribution level very precisely, and thus, the error contribution to the decomposition is negligible, i.e., less than 1% in all levels of ratings.

where r_i was written in year t and r_j was written in year s. $\Delta^S_{st}(P^k_i)$ and $\Delta^C_{st}(P^k_{ij})$ represent the scale and content effects respectively.

3.4 Drivers of the Scale and Content Effects

To understand what drives shifts in scale and content over time, we investigate both market-level factors and individual review characteristics. Our market is defined at the county level. For market-level drivers, once we fix the base year s (2013 in our application), we can link the decomposed effects with the observed market attributes Z_{mt} in market m and year t through the following relationships:

$$\Delta^S_{mst}(R^k) = \boldsymbol{Z}^{'}_{mt}\boldsymbol{\gamma}^{Sk} + \xi^{Sk}_{mt}, \tag{4}$$

$$\Delta^C_{mst}(R^k) = \boldsymbol{Z}^{'}_{mt}\boldsymbol{\gamma}^{Ck} + \xi^{Ck}_{mt}, \tag{5}$$

$$\Delta_{mst}(R^k) = \boldsymbol{Z}^{'}_{mt}\boldsymbol{\gamma}^{k} + \xi^{k}_{mt}. \tag{6}$$

The first equation (4) is for the scale effect, and the second equation (5) is for the content effect. The last equation (6) is for the overall effect that combines both the scale and content effects. The variables we consider in $\boldsymbol{Z}_{mt}$ include (1) broader societal trends, such as political polarization; (2) market-level demographic characteristics, such as population, age, income, and education, and (3) business factors, such as the number of businesses and the average number of reviews each business receives. In addition, we include county fixed effects in the regressions to account for any time-invariant county characteristics as well as year fixed effects to account for the aggregate time trend in the scale and content effects. The parameters, $\boldsymbol{\gamma}$, $\boldsymbol{\gamma}^{Ck}$, and $\boldsymbol{\gamma}^{k}$ are coefficients that illustrate how the various factors contribute to the scale, content and overall effects, respectively.

For individual review-level drivers, we use the following regression models:

$$\Delta^S_{st}(P^k_i) = \boldsymbol{Z}^{'}_{i}\boldsymbol{\gamma}^{Sk} + \phi^{Sk}t + \xi^{Sk}_{it}, \tag{7}$$

$$\Delta^C_{st}(P^k_i) = \boldsymbol{Z}^{'}_{i}\boldsymbol{\gamma}^{Ck} + \phi^{Ck}t + \xi^{Ck}_{it}, \tag{8}$$

$$\Delta_{st}(P^k_i) = \boldsymbol{Z}^{'}_{i}\boldsymbol{\gamma}^{k} + \phi^{k}t + \xi^{k}_{it}. \tag{9}$$

where t is the year when review r_i was written. We omit j from the $\Delta_{st}(\cdot)$ notations because regardless of which base review r_j (written in year s) we choose, its distributional probabilities will be absorbed into the constants in the above regression models. For review characteristics in $\boldsymbol{Z}_{it}$, we consider the ratings of past reviews, days since the last review, reviewer characteristics, and business fixed effects. $\boldsymbol{\gamma}$ are the coefficients that demonstrate how review characteristics contribute to the scale, content and overall effects at the individual review level. In these models, we include a specific time trend t to detect if the polarization trend still remains after accounting for all the individual review characteristics, and to explore through which effect (scale or content) the polarization trend propagates. We discuss more details on $\boldsymbol{Z}_{mt}$ and $\boldsymbol{Z}_i$ in Section 4.

3.5 Potential Impact of Scale and Content Effects on Review Informativeness

The scale and content effects can influence the informativeness of reviews and in turn the economic outcomes of businesses because consumers rely on ratings to inform their purchase decisions. To examine the potential impact, we use two sets of analyses. One is where we examine the impact of the scale and content effects on the "usefulness" of reviews, as voted by Yelp users, and the other is where we look at the impact on business performance, using the Texas restaurant revenue data, as described in Section 2. For the usefulness of reviews, we use the following models:

$$\log(Useful_i) = \phi_1 \Delta Rating_i + \varepsilon_i, \tag{10}$$

$$\log(Useful_i) = \phi_2 \Delta Rating_i + \phi_3 \widehat{Rating}_i + \varepsilon_i, \tag{11}$$

where $\Delta Rating_i$ is the difference between the displayed and predicted ratings of an individual review i, and $\widehat{Rating}_i$ is the predicted rating based on content using the model estimated for our base year 2013. In model (10), we use the difference between displayed and predicted ratings to gauge the impact of the scale factor. In model (11), we include the predicted rating to control for content and see if the scale factor still has a significant impact on the usefulness of reviews.

For the impact on business performance, we use the following models:

$$\log(Rev_{jt}) = \phi_1 Rating_{jt-1} + \phi_2 \widehat{Rating}_{jt-1} + \boldsymbol{X}_{jt}\theta_x + \theta_{jmnth} + \theta_t + \varepsilon_{jt}, \tag{12}$$

$$\log(Rev_{jt}) = \phi_3 Rating_{jt-1} + \phi_4 \%\Delta Rating_{jt-1} \times Rating_{jt-1} + \boldsymbol{X}_{jt}\theta_x + \theta_{jmnth} + \theta_t + \varepsilon_{jt}, \tag{13}$$

where $\log(Rev_{jt})$ is the revenue of restaurant j at time t. $Rating_{jt-1}$ is the cumulative average rating displayed on Yelp for restaurant j at time $t-1$; it is rounded to the nearest 0.5. $\widehat{Rating}_{jt-1}$ is the predicted cumulative average rating based on content. $\%\Delta Rating_{jt-1} \equiv (Rating_{jt-1} - \widehat{Rating}_{jt-1})/Rating_{jt-1}$ is the share of the scale factor in displayed ratings. $\boldsymbol{X}_{jt}$ includes demographic characteristics of the zip code tabulation area where restaurant j is located. θ_{jmnth} is the restaurant-calendar-month fixed effect. The calendar month here corresponds to January, February, etc. This fixed effect is to capture restaurant-specific seasonality, as different types of restaurants may respond to seasonality differently. θ_t is a year-month FE; it captures the aggregate time trend in restaurant revenues.

In these revenue regressions, we use the natural logarithm of revenue from alcoholic beverage sales as the dependent variable, operating under the assumption that for each restaurant, alcoholic drink sales consistently represent a fixed proportion of total sales over time. Consequently, fluctuations in alcoholic drink sales are indicative of changes in overall sales. In addition to seasonality, the restaurant-calendar-month fixed effects in our models account for the proportion of alcoholic drink sales within each restaurant's total sales. By controlling for these fixed effects, we assess how variations in a restaurant's revenue are influenced by changes in both the rating scale and the content of Yelp reviews. We employ lagged ratings from the previous period to acknowledge that consumers first observe a restaurant's ratings before deciding whether to dine there.[10]

Through model (12), we explore the individual associations of displayed ratings and review content with revenue. By controlling for predicted ratings, we assess whether the scale factor in displayed ratings accounts for additional revenue variations beyond those explained by review content, suggesting its potential influence on consumer choices. Conversely, controlling for displayed ratings allows us to evaluate if review content is independently associated with consumer decisions. Model (13) investigates whether a greater share of the scale factor in displayed ratings attenuates the relationship between revenue and displayed ratings. Since displayed ratings are the first information consumers encounter on Yelp, discrepancies between review content and displayed ratings may lead consumers to discount the latter. All results of the regressions are shown in Section 4.

4 Results

We first show the superior predictive performance of our LLM-based model compared to traditional sentiment-based approaches in Section 4.1. Then, we illustrate the linguistic factors that are most relevant in determining ratings in Section 4.2. In Section 4.3, we demonstrate the decomposition of ratings over time into the scale effect and the content effect. We then explore the socioeconomic

[10] Our analysis does not aim to establish a causal relationship between the scale and content factors and revenue, as these factors may be correlated with unobserved true restaurant quality, which can fluctuate over time and influence revenue changes. Nevertheless, the regression results provide suggestive evidence regarding the potential impact of the scale and content factors on consumer choices and business performance.

drivers of the scale and content effects in Section 4.4 and investigate their potential impact on the informativeness of reviews and consumer purchase decisions in Section 4.5. Finally, we provide a discussion on the implications of our findings in Section 4.6.

4.1 Review Rating Prediction

Using 2013 as the base year, we estimate the ordered logit model shown in equation 1 for each metropolitan area in the Yelp dataset. We present, as an example, the accuracy rates for the Philadelphia market in 2013 in Table 1. Each row of the table indicates the percentage of each star level in the data that matches the star levels predicted by our model. The diagonal line details the hit rates for each star level, representing the percentage of predicted ratings that match the original data. As shown, the highest hit rates are for 1-star and 5-star reviews, followed by 4-star reviews. Given that 1-star and 5-star reviews account for a significant share of the total ratings, the overall hit rate of the model is 0.7.

It is also evident that for all star levels except 2 stars, the highest percentages lie on the diagonal line, indicating that our model predictions are correct most of the time. Even when the model predictions miss, the majority of the mismatches are within a one-star rating difference. This level of accuracy in rating predictions is very reasonable, given the inherent uncertainties in the exact rating a user gives. Constrained by the discrete rating scale on Yelp, reviewers often have to round their assessment to the nearest numeric star, which is stochastic at the boundaries. The relatively high concentration of predictions within one star off the diagonal line provides evidence supporting our approach's exceptional performance.

Compared to alternative approaches, our prediction method performs substantially better. For example, if we do not utilize any of the review content information, a naive prediction model would predict all ratings at the star level with the highest percentage in the data (i.e. the popularity bias). The baseline hit rate from this naive approach is about 0.46. If we use the LIWC attributes of the review texts instead of the LLM outputs, the hit rate improves to around 0.51 to 0.52, a 10% to 13% improvement. Table 2 presents these hit rates under different approaches. If we use the LIWC attributes related to only sentiment, such as tone and emotions, the hit rate is 0.51. If we include all LIWC main attributes, the hit rate improves slightly to 0.52, suggesting that language sentiment in the review text has the strongest predictive power for the star ratings among all language attributes. Once we adopt the LLM outputs as the predictors, the accuracy rate improves drastically by almost 45% from the baseline approach and by about 32% from the

LIWC sentiment analysis. These results indicate that the LLM embeddings capture many more nuances in the review language than traditional sentiment characteristics. According to Tanana et al. (2021), LLMs excel in contextual understanding compared to traditional sentiment analysis methods, which struggle with contrastive conjunctions. For example, in a sentence like "I usually appreciate the chef's specials, but tonight's dish didn't impress me," LIWC will likely overestimate the positive sentiment because of the word "appreciate."

Table 1: Model Accuracy Rate for Philadelphia Reviews

	Model 2013 Star Rating				
Actual Stars	1	2	3	4	5
1	0.782	0.164	0.049	0.005	0.000
2	0.421	0.334	0.215	0.028	0.001
3	0.081	0.169	0.418	0.310	0.023
4	0.002	0.009	0.089	0.598	0.302
5	0.000	0.001	0.009	0.220	0.769

Numbers represent the percentage of observations falling into each category. The diagonal values indicates the accurate predictions for each star level.

Table 2: Model Performance Comparison Across Metropolitan Areas

	Hit Rate			
Metro	Baseline	LIWC sentiment	LIWC	LLM
Boise	0.50	0.56	0.57	0.69
Edmonton	0.33	0.44	0.44	0.62
Indianapolis	0.46	0.47	0.49	0.66
Nashville	0.47	0.50	0.52	0.67
New Orleans	0.48	0.48	0.51	0.66
Philadelphia	0.43	0.48	0.49	0.66
Reno	0.49	0.57	0.58	0.70
Saint Louis	0.44	0.49	0.50	0.66
Santa Barbara	0.55	0.58	0.59	0.71
Tampa	0.49	0.51	0.53	0.68
Tucson	0.47	0.52	0.54	0.68
Avg Score	0.46	0.51	0.52	0.67
Score Improvement	0	9.6%	12.7%	44.6%

Note: data based on year 2013.

Our model estimates allow us to examine the distributional changes in the underlying review content score r^* in detail (recall that r^* represents the underlying score that will determine the

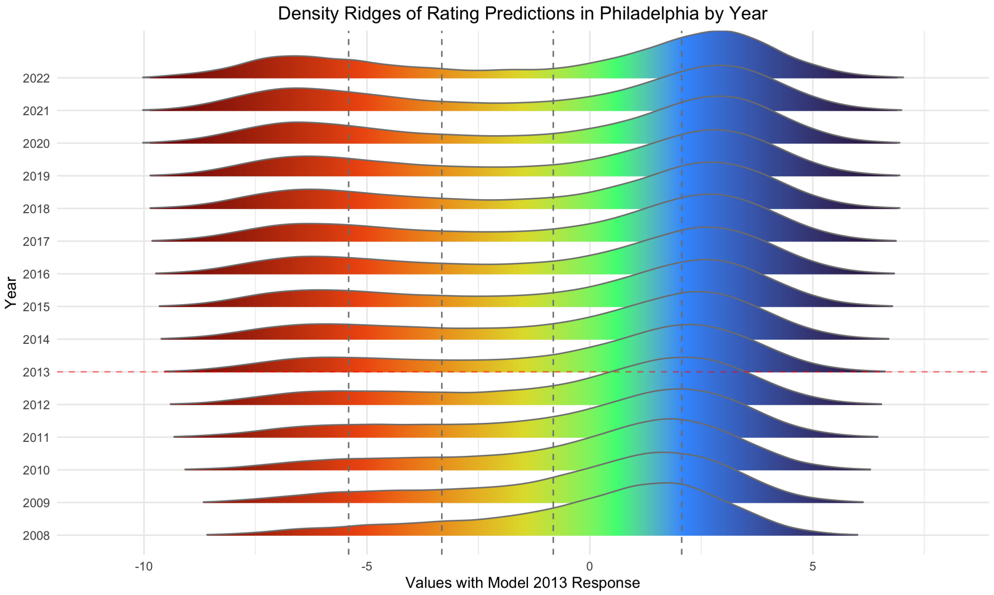

Figure 12: Review Content Distributions over Years with Cutoffs

likelihood of giving different ratings). Using the Philadelphia market as an example, we plot the empirical density of r^* over the years in Figure 12. The dotted vertical lines in the figure are the cut-off thresholds ($\boldsymbol{\alpha}$) of the benchmark ordered logit model with a base year of 2013. We compute the review content score by applying the model estimates of $\boldsymbol{\beta}$ to the content attributes for all years. The figure shows a clear pattern of changes in the distribution of the underlying review content score: the distribution is mostly unimodal in the early years of the platform's establishment before 2012, when the lower tail becomes thicker, and gradually, a second mode to the left emerges. In later years, we can see a clear bimodal distribution. These changes in the distribution signify the underlying content shifts that drive review polarization: a large proportion of the content written by the reviewers is becoming negative. On the other hand, we can also see an increase in the proportion of the extreme positive domain in the right tail of the distribution.

Applying the ordered logit model to the review data for each year, we can further examine the changes in the tendency to give a particular rating over time based on the estimated "cutoff" parameters $\boldsymbol{\alpha}$ for each year. These parameter estimates are essentially the thresholds for which a rating is more falls under a particular level. Figure 13 shows the changes in the "cutoff" parameters over the years for the Philadelphia market. Overall, the thresholds are decreasing for all levels of ratings, indicating that reviewers are more likely to give more positive ratings given the same experience expressed in the review text. However, the magnitudes of changes vary widely across

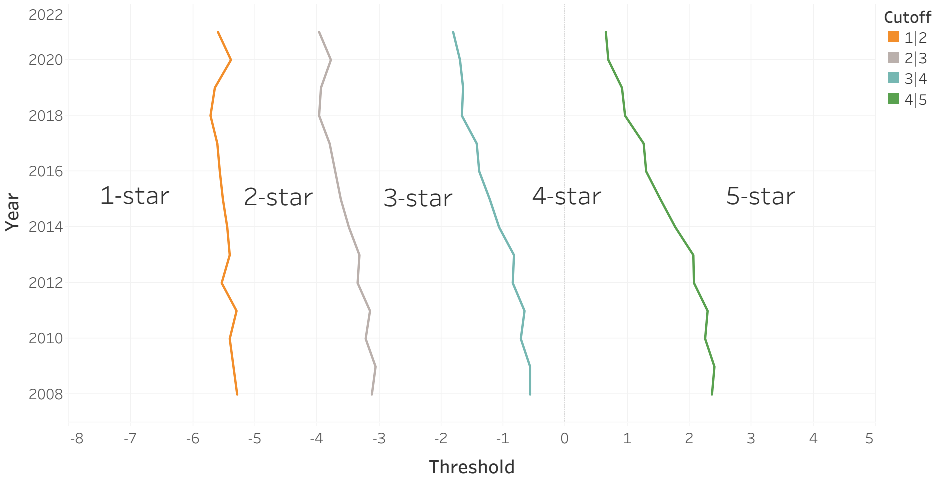

Figure 13: Cutoff Thresholds for Review Ratings over Years

rating levels: the threshold for 5-star ratings has the largest shifts, while the threshold for 1-star ratings remains relatively stable. As a result, the proportion of mid-level ratings (2, 3, 4 stars) is shrinking over time. This chart provides preliminary evidence for the impact of the scale effect in review polarization, suggesting that this effect is more likely to occur in the positive domain rather than in the negative domain.

4.2 Review Rating Determinants

Our model estimates allow us to examine which attributes of the review texts are most relevant for predicting the ratings. We rank the principal components by their marginal impacts, calculated based on the model estimates of $\boldsymbol{\beta}$ for the base year 2013 across all metropolitan markets. The top 10 elements that have the highest marginal impacts are presented in Table 3. Interestingly, all metropolitan markets share the same set of top 10 elements, suggesting that the most informative review attributes are consistent across all markets. For each metropolitan market, we report two numbers. The first is the impact weight, which represents the marginal impact of one standard deviation of each principal component on the underlying rating score r^*. The second number represents the ranking of the impact weight, with 1 being the most influential and 10 the least. The principal components are indexed by the rank of their eigenvalue contributions to the original embedding space. For example, as mentioned previously, PC1 has the highest eigenvalue contribution, PC2 the second highest, and so on.

Table 3 reveals a few notable patterns. First, the principal components with the highest eigenvalues also tend to have the highest impact on rating predictions; for example, PC1, PC2, PC3, and PC4 are the top four predictors of ratings. This pattern indicates that the principal components capturing the most data variation in review text (from the original LLM embeddings) are also the most indicative of ratings. Second, principal components with smaller eigenvalues, such as PC22, PC28, and PC30, are nonetheless among the top 10 predictors of ratings. This phenomenon implies that some subtleties in the review messages strongly predict review ratings. Although these less prominent principal components do not capture a large share of the variation in the textual data, they contain critical information for determining ratings. Third, across all markets, the results are very robust and consistent: the impact weight of each principal component is very similar across markets, and the impact rankings are also consistent. For example, 6 out of the 11 markets have exactly the same rankings of the top 5 principal components. The robustness of these estimates across markets demonstrates that the rating prediction models based on review content reliably and consistently reflect the underlying semantic elements in review ratings.

Table 3: Marginal Effects and Importance Rank by Market

		Principal Component									
		2	1	3	4	10	7	22	30	28	18
Boise	Impact Weight	3.433	1.434	0.645	0.582	0.502	0.414	0.237	0.242	0.178	0.194
	Rank	1	2	3	4	5	6	8	7	10	9
Edmonton	Impact Weight	3.138	1.520	0.595	0.581	0.610	0.325	0.215	0.202	0.235	0.267
	Rank	1	2	4	5	3	6	9	10	8	7
Indianapolis	Impact Weight	3.066	1.256	0.639	0.627	0.470	0.340	0.238	0.215	0.238	0.237
	Rank	1	2	3	4	5	6	8	10	7	9
Nashville	Impact Weight	3.163	1.245	0.642	0.608	0.388	0.387	0.251	0.231	0.210	0.205
	Rank	1	2	3	4	5	6	7	8	9	10
New Orleans	Impact Weight	2.971	1.094	0.600	0.584	0.421	0.331	0.223	0.214	0.190	0.187
	Rank	1	2	3	4	5	6	7	8	9	10
Philadelphia	Impact Weight	3.105	1.293	0.650	0.612	0.417	0.327	0.248	0.202	0.220	0.213
	Rank	1	2	3	4	5	6	7	10	8	9
Reno	Impact Weight	3.334	1.441	0.569	0.617	0.522	0.373	0.253	0.262	0.229	0.263
	Rank	1	2	4	3	5	6	9	8	10	7
Saint Louis	Impact Weight	3.094	1.262	0.629	0.597	0.469	0.344	0.263	0.225	0.255	0.215
	Rank	1	2	3	4	5	6	7	9	8	10
Santa Barbara	Impact Weight	3.33	1.28	0.68	0.60	0.45	0.40	0.25	0.25	0.18	0.21
	Rank	1	2	3	4	5	6	7	8	10	9
Tampa	Impact Weight	3.063	1.258	0.598	0.629	0.401	0.313	0.252	0.226	0.222	0.195
	Rank	1	2	4	3	5	6	7	8	9	10
Tucson	Impact Weight	3.155	1.283	0.594	0.597	0.469	0.345	0.230	0.206	0.207	0.181
	Rank	1	2	4	3	5	6	7	9	8	10

[a] Impact Weight indicates the marginal impact of one standard deviation of each principal component on the underlying rating score.

A challenge with the LLM-based analysis is understanding what contextual information LLMs capture. In our application, this challenge amounts to grasping the exact meanings of the most important drivers in the review prediction models. The principal components are a combination of LLM embeddings, which originate from complex and layered architectures and do not have intuitive meanings. It is very difficult to determine what semantic meaning a principal component encapsulates exactly, a challenge not unique to our study.

To understand the principal components more intuitively, we investigate the underlying semantic meanings of the top 10 principal components in Table 3 by relating these factors to the linguistic traits of review texts. We obtain the linguistic traits from the LIWC software, which analyzes the psychological state of content writers based on various characteristics in the text. Supported by decades of scientific research, LIWC is a widely used tool in social science, economics, and business research. It particularly aids in the comprehension, explanation, and quantification of psychological, social, and behavioral phenomena across various disciplines[11].

We first use the LIWC software to encode the review text into 119-dimensional vectors and focus on the most important 17 categories, including WC (Word Count), WPS (Words per Sentence), Analytic, Clout, Authentic, Tone, BigWords, Linguistic, Drives, Cognition, Affect, Social, Culture, Lifestyle, Physical, Perception, and Conversation. With the exception of WC and WPS, all the other dimensions have scores between 0 and 100. These scores are encoded based on the frequency of relevant words in each category. Due to this construction, a limitation of LIWC analysis (compared to more recent ML tools) is its restricted ability to understand the entire context of multiple blocks of text. Nonetheless, the semantic meanings in the LIWC library output are much more concrete and intuitive. For example, the Tone category includes positive tone and negative tone, which are direct measures of sentiment in a text. To relate the principal components to these more concrete meanings in the LIWC output, we regress each of the top 10 principal components onto the LIWC scores and compute their relative contributions in terms of the percentage of variance explained by each LIWC component.

We present the results of these analyses in Table 4. Consistent with Table 3, we rank the principal components based on their impact weights. The average impact weights across all metropolitan markets are shown in the top row of the table. We also report the R-squared results of each regression in the bottom row of the table. Across these principal components, we find an overall decreasing order of R-squared, implying that the variations explained by the LIWC scores diminish as the impact weights of the principal components decline. This pattern confirms that the information extracted from the LLM embedding is generally consistent with the LIWC output.

The most significant predictor of ratings, PC2, can be explained by the LIWC scores for 47.8%

[11] See details at https://www.liwc.app/help/howitworks.

of its data variation, with the dominant linguistic contributor being Tone, which represents the degree of positive/negative expressions in the text. Thus, to a large extent, PC2 represents the overall sentiment that closely predicts review ratings. The second largest predictor, PC1, can be explained by LIWC scores for over 50% of its data variation. The most significant linguistic contributors are Social and Physical, capturing important elements in consumption, such as the social environment and food and health considerations.

Similarly, the rest of the principal components mostly capture other factors in Yelp reviews, but to a lesser extent, as shown by the lower overall R-squared values. PC3 mainly represents the joint effect of text length and food and health considerations; PC4 captures more of the power or authority (e.g., own, order, allow, etc.) used in the language; PC10 is associated with text length and tone; PC7 represents logic; and PC22 represents authenticity and honesty. Overall, these LIWC regressions provide a partial explanation of the semantic meanings in the most important underlying predictors of review ratings. These analyses show that various aspects of the linguistic elements in review messages collectively influence the ratings users assign to their experiences. Prediction models with more straightforward textual encoding methods, such as sentiment analysis, may miss many of these linguistic nuances in review text.

4.3 Review Polarization Decomposition

Using the method developed in equation (2), we can decompose the overall changes in online review rating distributions into scale and content effects. We present the overall changes and the decomposition of these two effects in Table 5 for the Philadelphia market. Because 2013 is the base year, all the values are 0 for this year. The results are displayed separately for 1-star and 5-star ratings. The column "total effect" represents the overall change in the share of 1 and 5 stars in the distribution of reviews in various years. For example, the 1-star review share increased by 77.8% from 2013 to 2021. The next two columns decompose the change into the scale effect and the content effect.

As can be seen, there is a stark contrast between negative reviews and positive reviews in terms of the relative contributions of scale and content effects: for the 1-star reviews, the increase in share is mainly attributed to the content effect (80.7% from 2013 to 2021), whereas for the 5-star reviews, the dominant factor is the scale effect (80.9% from 2013 to 2021). Put simply, when more people give 1-star ratings, they actually mean it in the review message they write — overall, these messages reflect negative experiences in their semantics. However, when consumers give more 5-star

Table 4: Relationship Between Top Principal Components and LIWC Scores

	Principal Component									
	2	1	3	4	10	7	22	30	28	18
Avg Impact Weight[a]	-3.168	1.306	-0.622	-0.603	0.465	-0.354	0.242	0.225	0.215	-0.215
ANOVA[b]										
ln(WC)	**0.0624**	0.0195	**0.0268**	0.0253	**0.2662**	0.0007	0.0072	0.0000	0.0009	**0.0303**
ln(WPS)	0.0004	0.0320	0.0007	0.0049	0.0000	0.0000	0.0020	0.0008	0.0001	0.0005
Analytic	0.0051	0.0091	0.0187	0.0124	0.0001	**0.0215**	0.0019	0.0000	0.0002	**0.0114**
Clout	0.0187	0.0000	0.0152	**0.0313**	0.0023	**0.0306**	**0.0389**	0.0003	0.0068	0.0003
Authentic	0.0000	0.0283	0.0139	0.0160	0.0059	0.0003	**0.0137**	0.0013	0.0012	0.0001
Tone	**0.3173**	0.0870	0.0000	0.0081	**0.0535**	0.0004	0.0005	0.0002	0.0007	0.0053
BigWords	0.0043	0.0057	0.0006	0.0164	0.0001	0.0000	0.0028	0.0024	0.0001	0.0013
Linguistic	0.0002	0.0196	0.0005	0.0065	0.0026	0.0048	0.0093	0.0018	0.0031	0.0000
Drives	0.0051	0.0005	0.0000	**0.0570**	0.0000	0.0120	0.0001	0.0006	0.0002	0.0038
Cognition	0.0028	0.0006	0.0011	0.0170	0.0045	0.0015	0.0014	0.0028	0.0012	0.0017
Affect	0.0006	0.0001	0.0182	0.0005	0.0094	0.0082	0.0046	0.0019	0.0025	0.0046
Social	0.0079	**0.1405**	0.0012	0.0089	0.0011	0.0026	0.0131	0.0001	0.0001	0.0000
Culture	0.0000	0.0036	0.0132	0.0003	0.0000	0.0046	0.0007	0.0001	0.0074	0.0001
Lifestyle	0.0046	0.0559	0.0216	0.0033	0.0001	0.0128	0.0014	0.0013	0.0004	0.0018
Physical	0.0458	**0.1390**	**0.0361**	0.0000	0.0007	0.0010	0.0075	0.0091	0.0003	0.0001
Perception	0.0011	0.0004	0.0178	0.0140	0.0113	0.0025	0.0014	0.0008	0.0027	0.0022
Conversation	0.0007	0.0006	0.0000	0.0008	0.0010	0.0002	0.0002	0.0027	0.0000	0.0023
R-Squared[c]	0.478	0.542	0.186	0.223	0.359	0.104	0.107	0.026	0.028	0.066

[a] Avg Impact Weight is the average of the impact weights across all metropolitan markets in Table 3.

[b] Numbers in the table indicate the proportion of the variance of each principal component explained by the corresponding LIWC scores. Values are obtained through the Analysis of Variance (ANOVA) decomposition. The top two factors for each principal component's decomposition that has an R-Squared above 0.05 are bolded.

[c] R-Squared measures the proportion of the variance of each principal component explained by all the LIWC scores in linear regressions.

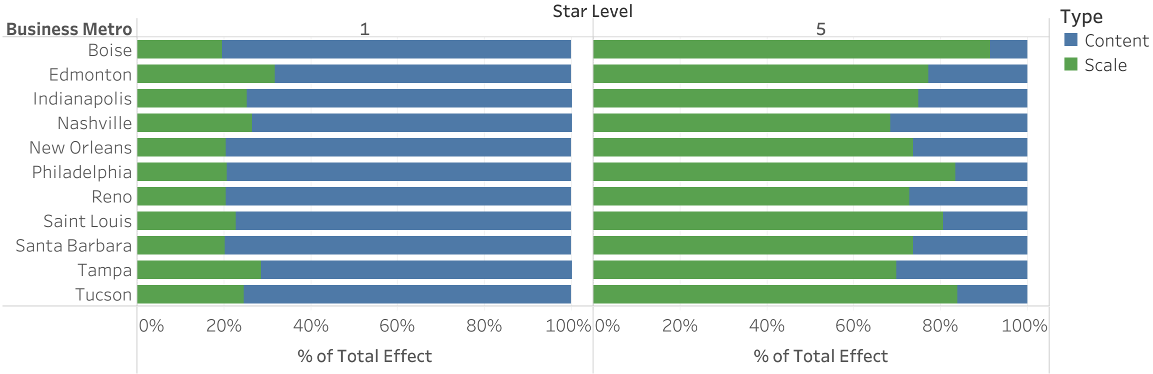

Figure 14: Scale and Content Effects 2013 to 2021 in Major Markets

reviews, they do not necessarily have a better experience; rather, they give a 5-star rating simply because they are more inclined to do so than before. For example, a "good" experience in 2013 might warrant a 4-star rating, whereas in more recent years, it deserves a 5-star rating, even if the underlying experience is the same. This table also shows the decomposition across multiple years. Overall, the contrast in the relative weights for the 1-star and 5-star reviews has been very consistent over the years.

Table 5: Scale and Content Effects over Years in Philadelphia

	1 Star Review			5 Star Review		
	Total Effect	Scale Effect	Content Effect	Total Effect	Scale Effect	Content Effect
2021	77.8%	19.3%	80.7%	42.5%	80.9%	19.1%
2017	31.8%	16.3%	83.7%	32.3%	70.0%	30.0%
2013	-	-	-	-	-	-

Note: 2013 is the base year.

We further present the decomposition for all the metropolitan areas in Figure 14 between 2013 and 2021. The effects are largely robust with the content effect being dominant for 1-star reviews while the scale effect dominant for 5-star reviews. The average splits for both sides seem to follow the 20-80 ratios.

4.4 Review Polarization and Societal and Business Factors

To investigate various socioeconomic factors that influence review polarization, we follow models (4) to (6) and (7) to (9) and regress the changes in the percentage of polarized reviews (i.e. 1- and 5-star ratings) onto a number of market-level and review-level characteristics. We also relate the scale and content effects to these socioeconomic factors. The results for the market-level factors are shown in Tables 6 and 7, and the results for the review-level characteristics are displayed in Tables 8 and 9.

Market-Level Factors For the market-level factors, we include the political polarization index, population, median age, median income, race, education, the number of new businesses each year, and the average number of reviews each business receives on Yelp at the county-year level. As shown in Table 6, for the combined percentage of 1- and 5-star ratings, political polarization does not have a significant correlation with review polarization once county and year fixed effects are controlled for. This is true for both the scale and content effects. For demographic characteristics, almost all factors are significantly correlated with either the scale or content effect, or both, except for population and the percentage of college-educated people. In particular, median age and the percentage of ethnic minorities exhibit positive correlations with review polarization, implying that in counties with older people and more ethnic minority groups, reviews tend to be more polarized. When decomposed, these correlations are significant only for the scale effect. Median income has a significantly negative correlation with the content effect but not the scale or overall effect, suggesting that higher-income people are less likely to use polarized language in reviews.

In terms of business-related factors, the number of new businesses each year has a significantly negative correlation with the scale effect, but not with the content effect, resulting in a significantly negative correlation with the overall effect. This pattern is consistent with the theory of self-selection bias proposed in the literature, i.e., reviews for the same business or product become more polarized over time (Li and Hitt, 2008; Wu and Huberman, 2008; Godes and Silva, 2012). The theory implies that a market with a higher number of new businesses is likely to see fewer polarized reviews. The finding is also consistent with the impact-effort trade-off conjecture by Wu and Huberman (2008), who posit that over time, only people with extreme views will post reviews because only extreme reviews justify the effort of posting. Our finding provides an additional insight that this impact-effort trade-off may incentivize reviewers to artificially exaggerate the rating scale, even if the underlying content is the same. As for the average number of reviews per business, the only marginally significant relationship is with respect to the scale effect.

Results in Table 6 focus on the combined percentage of extreme ratings but do not examine the heterogeneous effects of socioeconomic factors on either the share of 1-star ratings alone or that of the 5-star ratings alone. Table 7 explores these finer relationships. Here we present the

Table 6: Social, Political, and Business Factors for Review Polarization

	(1) Overall Effect	(2) Scale Effect	(3) Content Effect
	% 1- and 5-star	% 1- and 5-star	% 1- and 5-star
Political Polization Index	−0.005	0.019	−0.025
	(0.067)	(0.041)	(0.050)
Poppulation (log)	0.113	0.062	0.051
	(0.070)	(0.041)	(0.045)
Median Age (log)	0.302***	0.204**	0.098
	(0.096)	(0.079)	(0.067)
Median Income (log)	−0.064	−0.014	−0.050**
	(0.060)	(0.046)	(0.024)
% of Minority Population	0.270	0.216**	0.054
	(0.167)	(0.094)	(0.100)
% of with Colleague Degree and Above	0.219	0.213	0.006
	(0.265)	(0.169)	(0.138)
N New Business (log)	−0.031***	−0.025***	−0.006
	(0.010)	(0.009)	(0.004)
Avg N Reviews per Business (log)	0.068	0.070*	−0.002
	(0.050)	(0.036)	(0.025)
N	392	392	392
County FE	✓	✓	✓
Year FE	✓	✓	✓
Adj. R^2 (full model)	0.962	0.913	0.966
Adj. R^2 (proj model)	0.155	0.127	0.106

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.

results only for the political polarization factor. As shown, even though political polarization is not significantly correlated with the combined percentage of 1- and 5-star ratings, it is significantly associated with both of the separate percentages, with opposite signs. Political polarization has a negative correlation with the percentage of 1-star ratings and a positive one with the percentage of 5-star ratings, indicating that political polarization is associated with higher ratings in general. In areas with higher political polarization, we see fewer 1-star ratings but many more 5-star ratings. On the 1-star rating side, the negative relationship is significant for both the scale effect and the content effect. On the 5-star side, the positive relationship is significant only for the content effect, not the scale effect. These patterns can be explained by the findings in Jost et al. (2017) and Fernandes (2020), who show that consumers in politically polarized regions are more likely to boycott products that do not suit their ideologies and actively buy products that do, with liberal extremists tending to do so more than conservatives. This mechanism implies that consumers in politically polarized areas will consume/visit only products/businesses that they like and approve of, leading to a reduction in 1-star ratings and an increase in 5-star ratings. Interestingly, political extremists' enthusiasm shows up mostly on the content side instead of the scale side, as evidenced by the significant coefficients from the content effect regressions (3rd column) and the insignificant/only marginally significant coefficients in the scale effect regressions (2nd column).

Table 7: Review Polarization: Political Polarization Effect

	(1) Overall Effect	(2) Scale Effect	(3) Content Effect
% 1- and 5-star (Political Polarization Index)	-0.005	0.019	-0.025
	(0.067)	(0.041)	(0.050)
% 1-star (Political Polarization Index)	-0.129^{**}	-0.031^{*}	-0.099^{**}
	(0.060)	(0.016)	(0.047)
% 5-star (Political Polarization Index)	0.124^{*}	0.050	0.074^{*}
	(0.069)	(0.045)	(0.043)

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.

Review Characteristics For review-level factors, we focus more on elements related to the internal dynamics of reviews for a given business. Building on the literature (Moe and Trusov, 2011; Godes and Silva, 2012; Schoenmueller et al., 2020), we include the year the review was written, the number of days since the last review, the rating of the last review, the average rating and standard deviation of the last 10 reviews, the order of the review in the business's review sequence, the reviewer's tenure in months, and the reviewer's activity level, measured by the average number of reviews they write per year.[12] Additionally, we control for business-fixed effects to account for unobserved, business-specific factors.

Table 8 presents the results for the combined probability of a review receiving either a 1-star or 5-star rating. Even after accounting for both review and reviewer characteristics, a polarization trend persists over time. The coefficient for the year variable illustrates this trend in the internal dynamics of reviews for a given business. Positive and significant coefficients for both the scale and content effects suggest that polarization is increasing in both aspects.

Days since the last review and the rating of the previous review both negatively correlate with polarization. This means that longer intervals between reviews and higher previous ratings reduce the likelihood of the current review being polarized. Conversely, the average rating of the last 10 reviews shows a significantly positive correlation with review polarization. This is evident in both the scale and content effects, indicating that a higher average rating over recent reviews increases the probability of the current review being polarized. This relationship is reasonable, as consistently high average ratings likely reflect a high quality of service, making a 5-star rating more probable for subsequent reviews.

The standard deviation of the past 10 reviews has a negative correlation with the scale effect, but a positive one with the content effect, resulting in an overall positive correlation with polarization. A higher standard deviation suggests greater variability in previous reviews, leading to an increased chance of the current review being polarized. Additionally, the order of the review in the business's review sequence positively correlates with the scale effect but negatively with the content effect, resulting in an insignificant relationship with the overall polarization effect.

Regarding reviewer characteristics, reviewer tenure is significantly negatively associated with the scale effect but positively associated with the content effect. This combination leads to a notably positive correlation with the overall polarization effect, implying that more experienced reviewers are more likely to provide polarized reviews. In contrast, reviewer activity is significantly and negatively related to polarization in both the scale and content dimensions. This pattern aligns with the findings of Schoenmueller et al. (2020), which suggest that less selective reviewers tend to offer less polarized ratings. Therefore, highly active reviewers are more inclined to submit less polarized reviews.

Furthermore, Table 8 displays the results for the combined probability of a review receiving either a 1-star or 5-star rating. There may be heterogeneous effects of review characteristics on the

[12] We exclude reviewer-fixed effects because most reviewers submit only one review.

Table 8: Determinants for Review Polarization

	(1) Overall Effect Prob(1- and 5-star)	(2) Scale Effect Prob(1- and 5-star)	(3) Content Effect Prob(1- and 5-star)
Year	0.023^{***}	0.017^{***}	0.006^{***}
	(0.000)	(0.000)	(0.000)
Days Since Last Review	-0.000^{***}	-0.000^{***}	-0.000^{***}
	(0.000)	(0.000)	(0.000)
Rating of Last Review	-0.005^{***}	-0.002^{***}	-0.002^{***}
	(0.000)	(0.000)	(0.000)
Avg Rating of Last 10 Reviews	0.054^{***}	0.026^{***}	0.028^{***}
	(0.001)	(0.000)	(0.001)
Std Dev Rating of Last 10 Reviews	0.003^{***}	-0.001^{***}	0.004^{***}
	(0.001)	(0.000)	(0.001)
Reviewer Tenure (months)	0.001^{***}	-0.000^{***}	0.001^{***}
	(0.000)	(0.000)	(0.000)
Reviewer Activity (N/Year)	-0.000^{***}	-0.000^{***}	-0.000^{***}
	(0.000)	(0.000)	(0.000)
Order of Review (log)	0.000	0.003^{***}	-0.002^{***}
	(0.001)	(0.000)	(0.001)
N	1688609	1688609	1688609
Business FE	✓	✓	✓
Adj. R^2 (full model)	0.180	0.292	0.133
Adj. R^2 (proj model)	0.060	0.190	0.021

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.

separate probabilities of 1-star and 5-star ratings. We explore these effects in Table 9, focusing on the "year" variable to capture the time trend of polarization while controlling for all other factors. The table reveals that for 1-star ratings, both the scale and content effects exhibit significantly positive trends, with the content effect showing a stronger increase (0.006 vs. 0.002). This suggests that the primary driver behind the rising share of 1-star ratings over time is the content effect, meaning that consumers' experiences have genuinely declined.

In contrast, for 5-star ratings, the scale effect has a strong and significantly positive trend, while the content effect has a weak and negative trend (0.015 vs. −0.001). This indicates that the scale effect, rather than the content effect, is responsible for the growing share of 5-star ratings over time, leading to an inflation of the rating scale for positive reviews. These findings are consistent with earlier observations that the scale effect significantly contributes to changes in the proportion of 5-star ratings.

Table 9: Review Polarization: Effect Decomposition

	(1) Overall Effect	(2) Scale Effect	(3) Content Effect
Probability(1- and 5-star) (Year)	0.023^{***}	0.017^{***}	0.006^{***}
	(0.000)	(0.000)	(0.000)
Probability(1-star) (Year)	0.009^{***}	0.002^{***}	0.006^{***}
	(0.000)	(0.000)	(0.000)
Probability(5-star) (Year)	0.014^{***}	0.015^{***}	-0.001^{**}
	(0.000)	(0.000)	(0.000)

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.

4.5 Review Polarization and Informativeness of Reviews

Using the models shown in equations (10) to (13), we examine the potential effect of review polarization on the informativeness of reviews. In particular, we investigate the distinct impacts of the scale and content factors on the usefulness of reviews and restaurant revenue, which reflects consumer purchase decisions. For the usefulness analysis, our sample consists of all individual reviews in the Yelp Dataset. For the revenue regressions, our dataset includes information on restaurant revenue and Yelp reviews in Texas. We conducted the same rating prediction exercise on restaurant reviews in Texas as we did for the metropolitan regions in the Yelp Dataset, a process that produces the predicted ratings for the analysis.

The results are shown in Table 10. The first two columns show the results for the usefulness of reviews regressions and the last two columns show the results for revenue regressions. As shown in column 1, the coefficient for the scale effect in a review is negative and statistically significant, indicating that as the scale factor (i.e., the difference between actual and predicted stars) increases, the perceived usefulness of reviews diminishes. This suggests that reviews with ratings deviating more from predictions are considered less valuable by users. In other words, ratings without sufficient textual justifications are deemed as not helpful. Column 2 shows the results for including the predicted ratings (i.e., the content) in the model. The coefficient for the scale effect is still negative and statistically significant, although the magnitude is slightly smaller than that in Column 1. The coefficient for the predicted rating is negative and statistically signficant, suggesting that content that expresses negative experiences is regarded as more helpful. In other words, consumers put more weight on reviews with negative experiences when assessing the usefulness of reviews.[13]

Column 3 illustrates the results for the relationship of the scale and content factors with revenue.

[13] This result is the same even if we include indicator variables for the predicted ratings instead of the continuous predicted rating.

Table 10: Relationship Between Scale and Content Effects and Informativeness of Reviews

	(1) Useful	(2) Useful	(3) Log Revenue	(4) Log Revenue
Scale Effect in Individual Reviews	-0.092^{***}	-0.082^{***}		
	(0.008)	(0.008)		
Predicted Rating of Individual Reviews		-0.180^{***}		
		(0.009)		
Displayed Average Rating			0.061^{***}	0.125^{***}
			(0.014)	(0.009)
% Scale Effect × Displayed Average Rating				-0.065^{***}
				(0.012)
Predicted Average Rating			0.065^{***}	
			(0.012)	
Controls			✓	✓
N	6987305	6987305	666981	666981
Year×Month FE	204	204	196	196
Business FE	150281	150281		
Restaurant×Month FE			103867	103867

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. In the revenue regressions, all rating related independent variables are from the previous period. Controls variables include demographics information at the zip code tabulation area level, including shares of black, asian and hispanic populations and shares of populations aged 15 to 34, 35 to 65 and over 65.

As shown, both the coefficients for the displayed and predicted average ratings are positive and statistically significant, suggesting that businesses with better visible ratings tend to generate higher revenues. Additionally, conditional on the same displayed ratings, predicted average ratings still capture additional variations in revenue, suggesting that consumers take into the textual expression of consumer satisfication into consideration when making purchase decisions. It is worth noting that the coefficient for the displayed average rating also captures the relationship between the scale factor and revenue because the variation in displayed average ratings conditional on predicted ratings comes from the scale factor. The positive coefficient indicates that greater the scale factor is associated with higher revenue.

Column 4 shows how the relationship between displayed average ratings and revenue varies with the share of the scale factor in displayed ratings. The coefficient for the interaction term between the share of the scale factor and the displayed ratings is negative and significant, indicating that as the share of the scale factor increases, the positive correlation between displayed ratings and revenue weakens. This negative value implies that when the scale effect is large, consumers tend to discount the informativeness of displayed ratings, diminishing the impact of average ratings on revenue. Because the scale factor comes mostly from polarized ratings, especially 5 stars, the results also suggest that polarized reviews are less informative to consumers, a result that is consistent with the literature (Godes and Silva, 2012; Schoenmueller et al., 2020). Overall these results suggest that the scale effect reduces the informativeness of reviews.

4.6 Implications If Yelp Adopts a Content-Based Rating System

The findings in the previous section suggest that removing the scale effect from the rating system can make the ratings more informative, given that the adjusted ratings reflect consumer experiences on a more consistent basis. Removing the scale effect would alter the distribution of ratings, leading to several implications for businesses, platforms, and consumers.

Figure 15 illustrates the relationship between displayed cumulative average ratings and content-based cumulative average ratings for all businesses listed on Yelp in 2021. The content-based average ratings are calculated based on the 2013 scale. The vertical axis represents the displayed ratings, while the horizontal axis represents the content-based ratings, both rounded to the nearest 0.5. Each cell indicates the percentage of businesses with a specific displayed average rating corresponding to a particular content-based rating. Notably, the highest percentages cluster along the diagonal line, suggesting that the majority of businesses (over 70%) would not experience a change in their overall rating categories if the scale effect were removed. Among those that do see changes, nearly all adjustments are within 0.5 stars (approximately 96%). Most changes occur at the extremes, with 26% of 5-star ratings downgraded to 4.5 stars and 34% of 1-star ratings upgraded to 1.5 stars. Generally, ratings of 3.5 and above tend to decrease, those of 3 or 2.5 stars have an equal probability of increasing or decreasing, and ratings of 2 stars or below are more likely to increase. Nonetheless, the displayed ratings and content-based ratings share a strong correlation of 0.96, suggesting that adopting a content-based rating system would not distort the overall market signals while making them more consistent.

These patterns have several managerial implications. First, for businesses with displayed ratings of 3.5 and above, reduced content-based ratings are likely to disperse sales among a broader range of businesses, enabling consumers to discover more options. Second, for businesses with displayed ratings of 2 stars or below, increased content-based ratings can aid their survival by overcoming the initial "cold start" problem, where limited information about new businesses and early negative reviews may deter consumers from trying them, potentially causing these businesses to exit the market prematurely (Che and Hörner, 2018; Vellodi, 2018). Adopting a content-based rating system can help these initially negatively rated businesses endure longer, thereby offering consumers a wider variety of business options. Additionally, more dispersed sales and increased business longevity can enhance competition, resulting in greater benefits for consumers. For platforms like Yelp, a less polarized content-based rating system and a consistent scale between review content and numerical

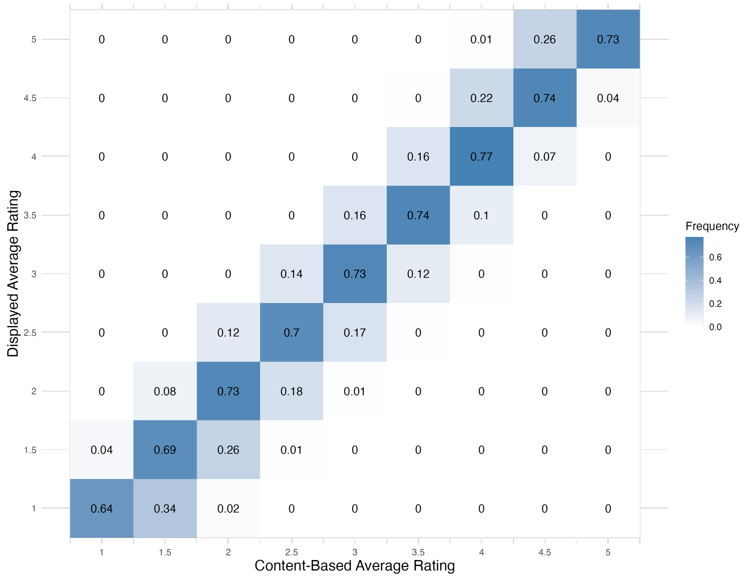

Figure 15: Correspondence Between Displayed and Content-Based Ratings on Yelp in 2021

ratings can enhance the informativeness of reviews, thereby increasing their signaling value to consumers. For consumers, the content-based rating system facilitates more accurate assessments of business performance, leading to more informed purchasing decisions.

5 Conclusion

This paper investigates the dynamics of online review polarization and its economic impact. By analyzing detailed Yelp reviews from 2005 to 2022, we differentiate between two key drivers of this phenomenon: the scale effect and the content effect. Utilizing advanced Large Language Models (LLMs), we decode review content and distinguish between shifts in rating standards and genuine changes in revealed consumer experiences. Our findings reveal that the surge in 5-star reviews is largely driven by a shift in rating standards (the scale effect), while the increase in 1-star reviews is primarily due to actual changes in the experiences (the content effect). Additionally, we explore the relationship between review polarization and broader societal trends, such as political polarization. We further assess the impact of the scale and content factors on the informativeness of reviews and business performance. The insights from this study offer important managerial implications for businesses and review platforms.

Despite the comprehensive analysis, this study has several limitations that suggest directions for future research. One limitation is the focus on a single platform — Yelp. While Yelp is a major player in the online review space, extending this analysis to other platforms like TripAdvisor, Google Reviews, and Amazon could provide a more holistic understanding of review polarization across different contexts and industries. Future research could explore whether similar trends in review polarization and their drivers are observed in geographic and cultural settings beyond the 11 regions covered by the Yelp Dataset. Furthermore, while LLMs provide powerful tools for content analysis, their interpretations are not always transparent. We link the model outcomes with the LIWC library to provide preliminary interpretations of the top semantic factors from LLMs. Future studies could combine LLMs with other methodologies to validate and extend the findings. Finally, exploring the long-term implications of review polarization on consumer trust and platform credibility would be valuable, particularly as businesses and consumers increasingly rely on online reviews for decision-making.

In summary, this study sheds light on the key drivers of online review polarization and provides managerial insights to both businesses and platforms. By distinguishing between scale and taste effects, we provide a nuanced understanding of the factors behind review polarization and its economic implications. Future research should continue to explore these dynamics across different platforms, regions, and methodologies to further our understanding of online consumer behaviors.

References

Anderson, M. and Magruder, J. (2012). Learning from the crowd: Regression discontinuity estimates of the effects of an online review database. *The Economic Journal*, 122(563):957–989.

Asercion, A. (2024). A comprehensive overview of large language models. *Epi Stories, University of Washington.*

Aziz, A., Li, H., and Telang, R. (2023). The consequences of rating inflation on platforms: Evidence from a quasi-experiment. *Information Systems Research*, 34(2):590–608.

Bayerl, A., Goldenberg, J., Schoenmueller, V. B., et al. (2023). The weekend effect in online reviews. In *45th ISMS Marketing Science Conference 2023.*

Boleslavsky, R. and Cotton, C. (2015). Grading standards and education quality. *American Economic Journal: Microeconomics*, 7(2):248–279.

Brand, J., Israeli, A., and Ngwe, D. (2023). Using gpt for market research. *Available at SSRN 4395751*.

Brandes, L., Godes, D., and Mayzlin, D. (2022). Extremity bias in online reviews: The role of attrition. *Journal of Marketing Research*, 59(4):675–695.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Che, Y.-K. and Hörner, J. (2018). Recommender systems as mechanisms for social learning. *The Quarterly Journal of Economics*, 133(2):871–925.

Chen, N., Li, A., and Talluri, K. (2021). Reviews and self-selection bias with operational implications. *Management Science*, 67(12):7472–7492.

Chevalier, J. A. and Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of marketing research*, 43(3):345–354.

Cui, J., Wang, Z., Ho, S.-B., and Cambria, E. (2023). Survey on sentiment analysis: evolution of research methods and topics. *Artificial Intelligence Review*, 56(8):8469–8510.

Cummins, N., Amiriparian, S., Ottl, S., Gerczuk, M., Schmitt, M., and Schuller, B. (2018). Multimodal bag-of-words for cross domains sentiment analysis. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4954–4958. IEEE.

De Langhe, B., Fernbach, P. M., and Lichtenstein, D. R. (2016). Navigating by the stars: Investigating the actual and perceived validity of online user ratings. *Journal of Consumer Research*, 42(6):817–833.

Dellarocas, C., Zhang, X., and Awad, N. F. (2007). Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *Journal of Interactive marketing*, 21(4):23–45.

Fang, L. (2022). The effects of online review platforms on restaurant revenue, consumer learning, and welfare. *Management Science*, 68(11):8116–8143.

Fernandes, D. (2020). Politics at the mall: The moral foundations of boycotts. *Journal of Public Policy & Marketing*, 39(4):494–513.

Filippas, A., Horton, J. J., and Zeckhauser, R. J. (2020). Owning, using, and renting: Some simple economics of the "sharing economy". *Management Science*, 66(9):4152–4172.

Frenkel, S. (2015). Repeated interaction and rating inflation: A model of double reputation. *American Economic Journal: Microeconomics*, 7(1):250–280.

Godes, D. and Silva, J. C. (2012). Sequential and temporal dynamics of online opinion. *Marketing Science*, 31(3):448–473.

Goli, A. and Singh, A. (2023). Language, time preferences, and consumer behavior: Evidence from large language models. *arXiv preprint arXiv:2305.02531.*

Hollenbeck, B. (2018). Online reputation mechanisms and the decreasing value of chain affiliation. *Journal of Marketing Research*, 55(5):636–654.

Hu, N., Pavlou, P. A., and Zhang, J. (2017). On self-selection biases in online product reviews. *MIS quarterly*, 41(2):449–475.

Hu, N., Zhang, J., and Pavlou, P. A. (2009). Overcoming the j-shaped distribution of product reviews. *Communications of the ACM*, 52(10):144–147.

Jost, J. T., Langer, M., and Singh, V. (2017). The politics of buying, boycotting, complaining, and disputing: An extension of the research program by jung, garbarino, briley, and wynhausen. *Journal of consumer research*, 44(3):503–510.

Karaman, H. (2021). Online review solicitations reduce extremity bias in online review distributions and increase their representativeness. *Management Science*, 67(7):4420–4445.

Karlinsky-Shichor, Y. and Schoenmueller, V. (2023). The oracles of online reviews. *Available at SSRN 4321683.*

Kramer, M. A. (2007). Self-selection bias in reputation systems. In *IFIP International Conference on Trust Management*, pages 255–268. Springer.

Levy, R. (2021). Social media, news consumption, and polarization: Evidence from a field experiment. *American economic review*, 111(3):831–870.

Li, P., Castelo, N., Katona, Z., and Sarvary, M. (2024). Frontiers: Determining the validity of large language models for automated perceptual analysis. *Marketing Science.*

Li, X. (2016). Could deal promotion improve merchants' online reputations? the moderating role of prior reviews. *Journal of Management Information Systems*, 33(1):171–201.

Li, X. and Hitt, L. M. (2008). Self-selection and information role of online product reviews. *Information Systems Research*, 19(4):456–474.

Liu, Y. (2006). Word of mouth for movies: Its dynamics and impact on box office revenue. *Journal of marketing*, 70(3):74–89.

Luca, M. (2011). Reviews, reputation, and revenue: The case of yelp. com. Technical report, Harvard Business School Working Paper.

Luca, M. and Zervas, G. (2016). Fake it till you make it: Reputation, competition, and yelp review fraud. *Management science*, 62(12):3412–3427.

Moe, W. W. and Schweidel, D. A. (2012). Online product opinions: Incidence, evaluation, and evolution. *Marketing Science*, 31(3):372–386.

Moe, W. W. and Trusov, M. (2011). The value of social dynamics in online product ratings forums. *Journal of Marketing Research*, 48(3):444–456.

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., and Mian, A. (2023). A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

Pan, Y., Pikulina, E., Siegel, S., and Wang, T. Y. (2024). Political divide and the composition of households' equity portfolios. *Available at SSRN 4381330*.

Park, S., Shin, W., and Xie, J. (2021). The fateful first consumer review. *Marketing Science*, 40(3):481–507.

Park, S., Shin, W., and Xie, J. (2023). Disclosure in incentivized reviews: Does it protect consumers? *Management Science*, 69(11):7009–7021.

Pocchiari, M., Schoenmueller, V., and Dover, Y. (2023). The dynamic potential of online reviews: Review updates and platform solicitations. *Available at SSRN*.

Ravi, K. and Ravi, V. (2015). A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-based systems*, 89:14–46.

Rojstaczer, S. and Healy, C. (2012). Where a is ordinary: The evolution of american college and university grading, 1940-2009. *Teachers College Record*, 114(7):1–23.

Schoenmueller, V., Netzer, O., and Stahl, F. (2020). The polarity of online reviews: Prevalence, drivers and implications. *Journal of Marketing Research*, 57(5):853–877.

Schoenmueller, V., Netzer, O., and Stahl, F. (2023). Frontiers: Polarized america: From political polarization to preference polarization. *Marketing Science*, 42(1):48–60.

Squires, L. et al. (2012). Whos punctuating what? sociolinguistic variation in instant messaging. *Orthography as social action: Scripts, spelling, identity and power*, 3:289.

Tanana, M. J., Soma, C. S., Kuo, P. B., Bertagnolli, N. M., Dembe, A., Pace, B. T., Srikumar, V., Atkins, D. C., and Imel, Z. E. (2021). How do you feel? using natural language processing to automatically rate emotion in psychotherapy. *Behavior research methods*, pages 1–14.

The Wall Street Journal (2023). Grade inflation makes a the new c. *The Wall Street Journal*.

Vandekerckhove, R. and Nobels, J. (2010). Code eclecticism: Linguistic variation and code alternation in the chat language of flemish teenagers 1. *Journal of sociolinguistics*, 14(5):657–677.

Vellodi, N. (2018). Ratings design and barriers to entry. *Available at SSRN 3267061*.

Wu, C., Che, H., Chan, T. Y., and Lu, X. (2015). The economic value of online reviews. *Marketing Science*, 34(5):739–754.

Wu, F. and Huberman, B. A. (2008). How public opinion forms. In *International Workshop on Internet and Network Economics*, pages 334–341. Springer.

Ye, Z., Yoganarasimhan, H., and Zheng, Y. (2024). Lola: Llm-assisted online learning algorithm for content experiments. *arXiv preprint arXiv:2406.02611*.

Yoganarasimhan, H. and Yakovetskaya, I. (2024). From feeds to inboxes: A comparative study of polarization in facebook and email news sharing. *Management Science, forthcoming*.

Zervas, G., Proserpio, D., and Byers, J. W. (2021). A first look at online reputation on airbnb, where every stay is above average. *Marketing Letters*, 32:1–16.

Zhao, Y., Yang, S., Narayan, V., and Zhao, Y. (2013). Modeling consumer learning from online product reviews. *Marketing science*, 32(1):153–169.