Shifting Standards or Changing Experiences? Unraveling Review Polarization via LLMs

Limin Fang * Chunhua Wu † Baohong Sun ‡§

May 30, 2025

Abstract

This paper investigates the drivers of online review polarization, characterized by a growing prevalence of extreme ratings (1-star and 5-star). Leveraging nearly seven million Yelp reviews across 11 metropolitan markets and using Large Language Models (LLMs) to analyze review content, we separate genuine changes in consumer experiences (content effect) from shifts in numerical rating standards (scale effect). We find that rising 5-star reviews primarily reflect scale inflation (65%), whereas increased 1-star ratings largely capture declining consumer experiences (80%). Significant contributors to scale shifts include temporal trends, explaining about half of the increase in extreme ratings, and reviewer heterogeneity. The scale effect reduces review informativeness by making numerical ratings less consistent with review content, disproportionately benefiting newer businesses and distorting market competition. Our findings suggest that review platforms should consider user-experience-weighted ratings or content-based AI-generated scores to enhance informativeness, consistency, and credibility.

Keywords: Online Reviews; Review Polarization; Scale Effect; Content Effect; Large Language Models; Self-Selection Bias; Review Informativeness

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

†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, Khai Chiong, Tong Guo, Tai Lam, and the seminar and conference participants at Workshop on Platform Analytics, Shanghai University of Finance and Economics, Hong Kong Polytechnic University, Temple University, and 2024 INFORMS Marketing Science for invaluable feedback. All errors are our own.

1 Introduction

Online consumer reviews significantly influence consumer decisions, shape business reputations, and impact market outcomes. An extensive body of literature has demonstrated how small changes in online ratings can profoundly affect sales, revenue, and consumer perceptions (e.g., Chevalier and Mayzlin, 2006; Moe and Trusov, 2011; Anderson and Magruder, 2012; Luca, 2011). However, a notable and growing feature is increasing **review polarization**, characterized by a shift from moderate ratings towards extreme evaluations (1-star and 5-star). This polarization results in a bimodal or J-shaped distribution (Schoenmueller et al., 2020; Hu et al., 2009), raising critical questions about the informativeness of online ratings and their consequences for businesses, consumers, and review platforms alike.

Yelp, a leading online consumer review platform, exemplifies this trend (see Figure 1). Utilizing the Yelp Open Dataset (2023), we observe that prior to 2012, Yelp's review distribution was predominantly unimodal and moderately skewed toward positive ratings. However, starting in 2012, a distinct shift toward extreme ratings emerged, with a significant decline in moderate reviews. Specifically, the share of moderate ratings (2-4 stars) fell from approximately 65% in 2005 to only 27% in 2021, while the prevalence of extreme ratings substantially increased. If this trend continues, Yelp's five-star rating system could effectively become binary by 2028.

This ongoing polarization raises fundamental questions about its underlying drivers. Prior research has examined factors such as reviewer self-selection (Li and Hitt, 2008; Hu et al., 2017; Schoenmueller et al., 2020) and the impact-effort trade-off (Godes and Silva, 2012; Wu and Huberman, 2008). However, most existing studies focus primarily on numerical ratings, overlooking the rich textual content that reviewers provide. Review texts are critical, as they directly reflect consumers' expressed experiences and can offer deeper insights into whether observed polarization arises from genuine shifts in consumer experiences (the *content effect*) or from changes in how reviewers translate these experiences into numerical ratings (the *scale effect*). A concern, however, is that consumers' language use norms may change over time, potentially making textual content an unstable measure of genuine consumer experiences. We provide evidence that consumers' language use norms in Yelp reviews have changed very little during our study period of 2012 to 2021, making review texts a reliable measure of consumer experiences. Figure 2 vividly illustrates the temporal shifts in rating scales: words previously associated with moderate ratings (e.g., "good" or "bad") have become increasingly linked with extreme ratings over time, suggesting a shift in consumers'

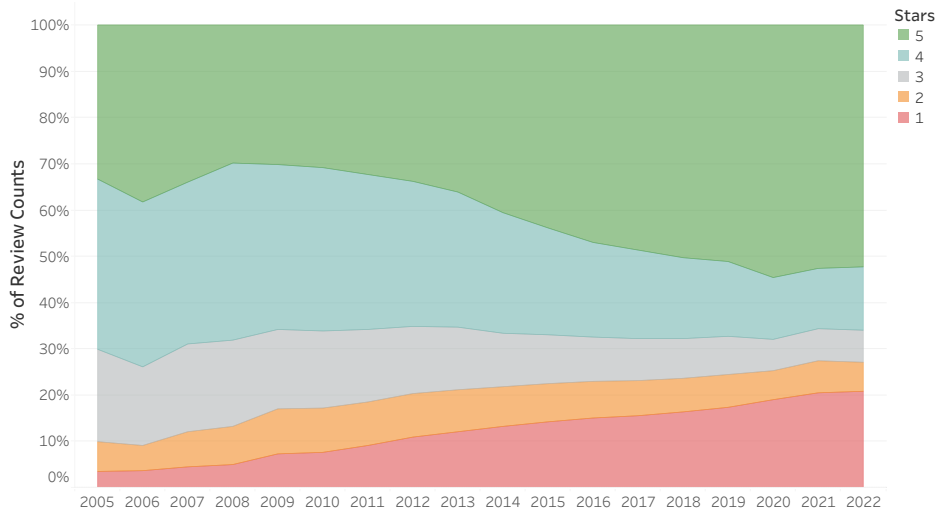


Figure 1: Distribution of Yelp Review Stars Over Years

rating standards rather than an actual change in experiences.

In this paper, we aim to unravel these dynamics by clearly distinguishing between the scale and content effects and exploring their implications. Our study addresses three primary research objectives: (1) identify the key factors driving review polarization by decomposing shifts in numerical rating scales from changes in consumer experiences as revealed through text; (2) explore how reviewer characteristics, business attributes, and review timing contribute to these scale and content effects; and (3) evaluate the impact of the scale effect on the informativeness of online reviews and explore methods to maintain review rating consistency.

To achieve these objectives, we leverage recent advancements in large language models (LLMs) to analyze nearly seven million restaurant reviews from the Yelp Open Dataset across 11 major North American metropolitan areas between 2012 and 2021. Each review is encoded into OpenAI's high-dimensional embedding space, capturing nuanced semantic meanings. Employing principal component analysis (PCA), we distill these embeddings into a reduced set of components, which we use to train an ordered logit model mapping review text to numerical ratings. Our model allows parameters to vary systematically by year, reviewer characteristics, and business attributes, enabling a clean decomposition of the scale and content effects.

Our findings demonstrate that the polarization of reviews is driven by both scale and content effects, but their relative importance differs significantly by rating extremes. Specifically, the increase in 5-star reviews is primarily attributable to scale inflation, with about 65% driven by changing rating standards rather than improved consumer experiences. Conversely, the growth in 1-star



Figure 2: Distribution of Negative- and Positive-Tone Words and Associated Stars OverYears

reviews largely reflects genuine declines in consumer satisfaction, with content changes accounting for over 80% of this increase. Of the various factors that drive the shifts in rating scale, the year trend contributes substantially to the increase in the shares of both 1-star and 5-star ratings, about 42% and 54%, respectively, on average for all metropolitan markets. We further identify significant contributions from reviewer and business heterogeneity, especially the rising share of one-time reviewers who disproportionately contribute to extreme ratings.

Crucially, we also find that the scale effect adversely affects review informativeness. Reviews with a stronger scale effect, where the numerical rating significantly diverges from the textual content, are substantially less likely to be marked as useful by consumers. Additionally, we show that scale-driven rating inflation systematically favors newer businesses over older ones in average ratings, creating competitive distortions.

This paper makes several contributions to the literature. First, we document robust, longitudinal patterns of review polarization at both reviewer and business levels, explicitly accounting for temporal dynamics overlooked in prior studies. Second, we distinctly quantify and separate the scale and content effects, highlighting the critical role of scale drift in compromising the reliability of online ratings. Third, we introduce an effective and scalable method for identifying and decomposing these effects using state-of-the-art language processing techniques.

Our findings offer actionable managerial implications for review platforms. Given the disproportionate impact of one-time reviewers on rating scales, platforms might consider weighting ratings by reviewer experience or utilizing AI-generated, content-based numerical scores to present more accurate and credible assessments of business quality. We provide preliminary evidence that such content-based scores better reflect consumer experiences, offering a practical approach to mitigate scale inflation.

The remainder of this paper is structured as follows. Following the literature review, Section 3 describes our data and documents the observed polarization patterns. Section 4 outlines our conceptual framework and empirical strategy. Section 5 presents our primary decomposition results and examines the factors contributing to review polarization. Section 6 discusses the implications of the scale effect on review informativeness and business competitiveness, and proposes actionable managerial recommendations to address the scale effect. Finally, Section 7 concludes with a summary of our findings and a discussion of future research directions.

2 Literature Review

Our study contributes to several strands of literature. First, it closely relates to research on how consumer reviews shape market outcomes. For example, Chevalier and Mayzlin (2006) show that while positive reviews boost book sales, negative reviews have an even stronger negative impact. Luca (2011) finds that a one-star increase in Yelp ratings can raise independent restaurant revenue significantly. Other studies demonstrate that review sentiment, volume, and valence influence movie box office revenue (Liu, 2006; Dellarocas et al., 2007), and that the effects of reviews can evolve over time (Moe and Trusov, 2011). Structural models further highlight how online reviews update consumer perceptions and choices, and quantify their economic value (Zhao et al., 2013; Wu et al., 2015). Recent work also shows that review platforms facilitate consumer learning, especially for tourists (Fang, 2022), and that easier access to reviews can incentivize firms to improve quality (Donati, 2025). Our study adds to this literature by showing that scale effects in consumer reviews can reduce informativeness and distort market competition, particularly disadvantaging older restaurants.

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 nega-

tive. Hu et al. (2017) find that acquisition bias and under-reporting bias contribute substantially to the J-shaped distribution of online reviews. Bayerl et al. (2023) discover that online reviews submitted during weekends are significantly more negative than those submitted during the week, primarily due to temporal self-selection, where users more likely to leave negative reviews are disproportionately active on weekends. Bondi et al. (2024) study the impact of consumer heterogeneity on online ratings and find that experienced users are more picky and give stringent ratings on high quality products, leading to a reversal of ratings, where poor-quality products receive higher ratings. They propose an algorithm to correct this bias. 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 moderately experienced customers can lead to a more balanced distribution. Pocchiari et al. (2023) show that review-update solicitations help companies manage negative feedback 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 and suggesting alternative approaches.

Other papers further investigate and emphasize the aspect of review inflation. For example, Aziz et al. (2023) explores the impact of 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 feedback. This extreme positivity can distort consumers' perceptions of product or service quality, leading to higher expectations 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. He et al. (2022) propose an algorithm to detect products that buy fake reviews by examining reviewer network structure. 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, namely 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) emphasize the importance of the sequence of reviews and find that initial reviews set a benchmark which influences subsequent feedback, leading to a decrease in ratings over the review sequence. This decrease is likely due to a self-selection mechanism, where early positive reviewers are followed by less enthusiastic users. Wu and Huberman (2008) find that exposure to previous reviews can lead to increasingly extreme reviews; however, if the expression of an opinion is costly and previous reviews are known, the impact-effort trade-off—where people will only write a review if it makes an impact on the ratings—can soften extreme reviews. 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 and how this effect varies across active and inactive reviewers. 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.

Last but not least, our research adds to the growing body of literature that deploys the language-processing ability of LLMs. Yoganarasimhan and Yakovetskaya (2024) use LLMs to detect polarization in news content and find that these models are highly effective. Ye et al. (2024) build on this finding and explore 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. Brucks and Toubia (2025) show that the way prompts are structured, such as the order and labels, can systematically bias LLM outputs, and that aggregating across multiple prompt variations can effectively eliminate the bias. Our study adds to this strand of research by leveraging the efficiency in contextual understanding provided by LLM embeddings to analyze review content and its evolution over time.

3 Data and Patterns

Our research utilizes the Yelp Open Dataset 2023, which contains nearly 7 million reviews written by almost 2 million users for over 150,000 businesses across 11 metropolitan areas from 2005 to 2021. This comprehensive dataset allows us to leverage LLMs to train predictive models that examine the relationship between review content and ratings[1]. The review data exhibit a consistent pattern of polarization over time, even after controlling for geography, business characteristics, reviewer cohort, reviewer tenure, and reviewer activity intensity. We illustrate these patterns below.

3.1 Overall Pattern

As shown in Figure 1, there is a clear and consistent pattern of review polarization on the Yelp platform over the years. Before 2010, the combined percentage of 5-star and 1-star reviews (referred to as the *polarity*) was less than 40% of total reviews. By 2021, however, over 70% of reviews were either 1 star or 5 stars. Figure 3 further illustrates this increasing trend of polarity and projects its value (with confidence intervals) beyond 2021. If this trend continues, almost all review ratings will be either 1 or 5 stars by 2028.

3.2 Patterns by Business Location and Type

One possible driver of the polarization pattern is business turnover. Specifically, if more businesses with extremely low or high quality enter the market over time, we might expect to see a polarizing pattern of reviews. To examine this potential factor, we show the distribution of review ratings over time by business location and type. Figure 4 presents the pattern for each of the 11

[1]In this paper, we treat all reviews in the dataset as genuine, since Yelp was among the first platforms to implement fake review detection algorithms, and the reviews included are all recommended by Yelp (Roberts, 2013; Singhal et al., 2023)

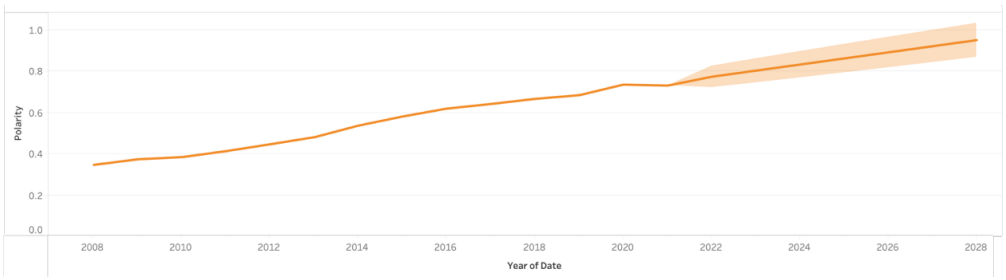


Figure 3: Yelp Review Polarity Over Time

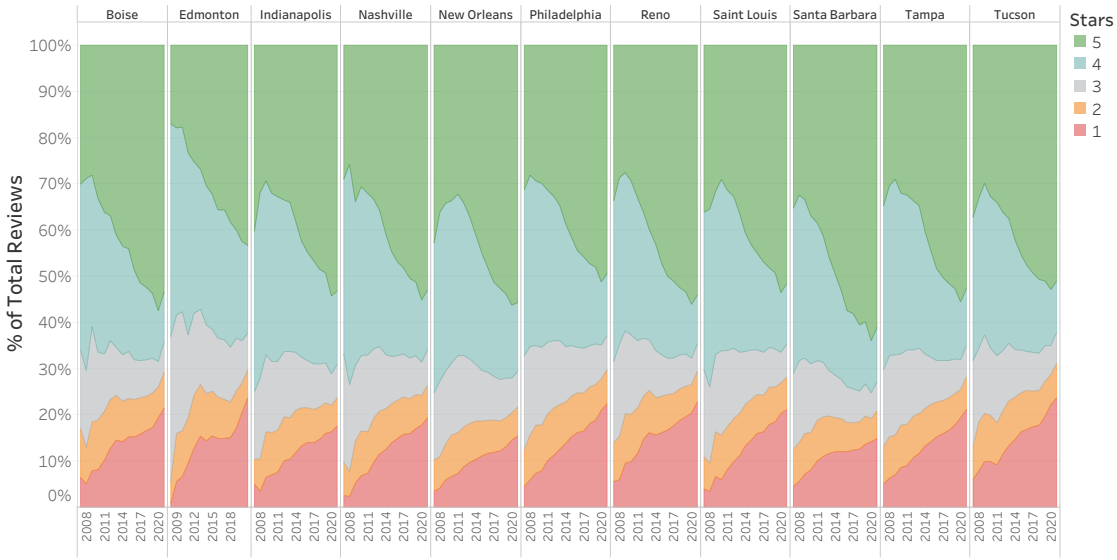


Figure 4: Yelp Review Polarization across Metropolitan Areas

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.

Businesses of different categories may also attract varying levels of polarized reviews. For example, restaurants may be more likely to receive moderate reviews than other business types. As more non-restaurant businesses entering the Yelp platform, the overall polarization trend can become more pronounced. Figure 5 shows the distribution of review ratings over time by business category. As can be seen, the polarization pattern is consistent across different business categories.

Businesses of different quality levels may also experience different trends in review polarization. Figure 6 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

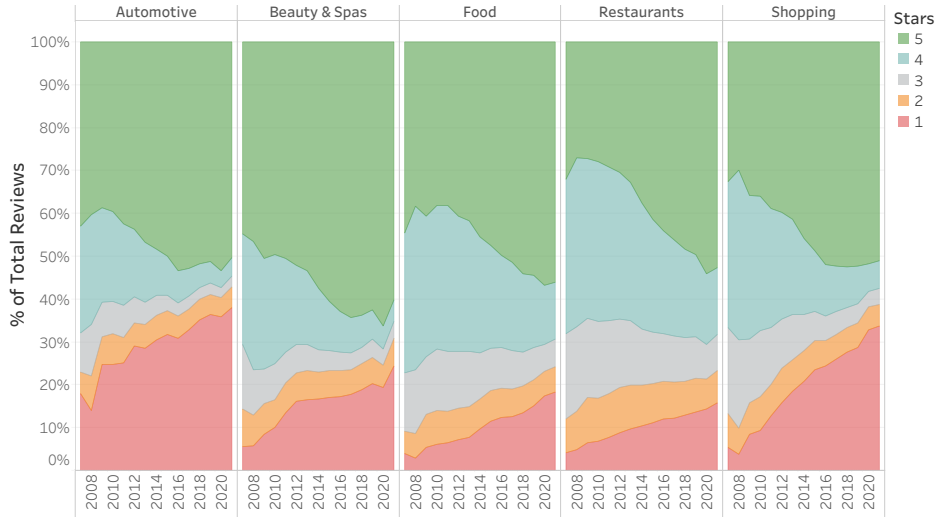


Figure 5: Yelp Review Polarization across Business Categories

the highest and lowest rating categories.

3.3 Patterns by Reviewer Tenure

In addition to business turnover, another significant contributor to review polarization could be 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 7 shows the review distribution by reviewer cohort, defined by the year a reviewer joins the platform. There is a distinct pattern of cohort effect: the shares of 1- and 5-star ratings given by more recent users are much higher than those given by earlier users. Figure 8 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.

3.4 Patterns by Reviewer Activity

The literature (e.g. Moe and Schweidel (2012)) suggests that frequent reviewers are less likely to write extreme ratings, implying that review polarization could be influenced by the turnover between frequent and infrequent reviewers. Figure 9 shows the distributions of star ratings over time for reviewers who wrote multiple reviews and for those who wrote only one review. As shown, although one-time reviewers are more likely to give polarized reviews, especially negative ones, compared to more frequent reviewers, the polarization trend is present for both groups.

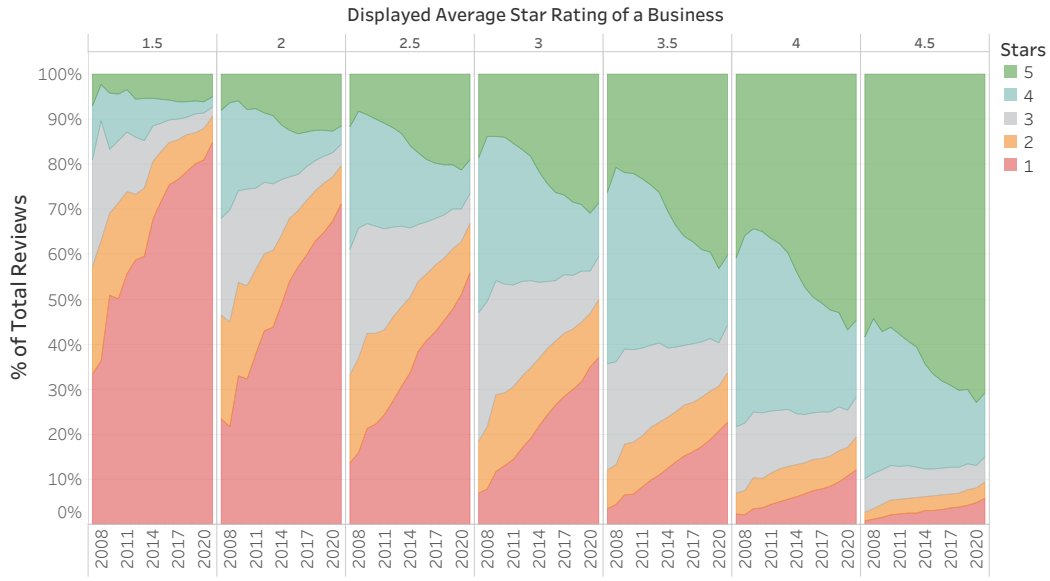


Figure 6: Yelp Review Polarization across Business Star Levels

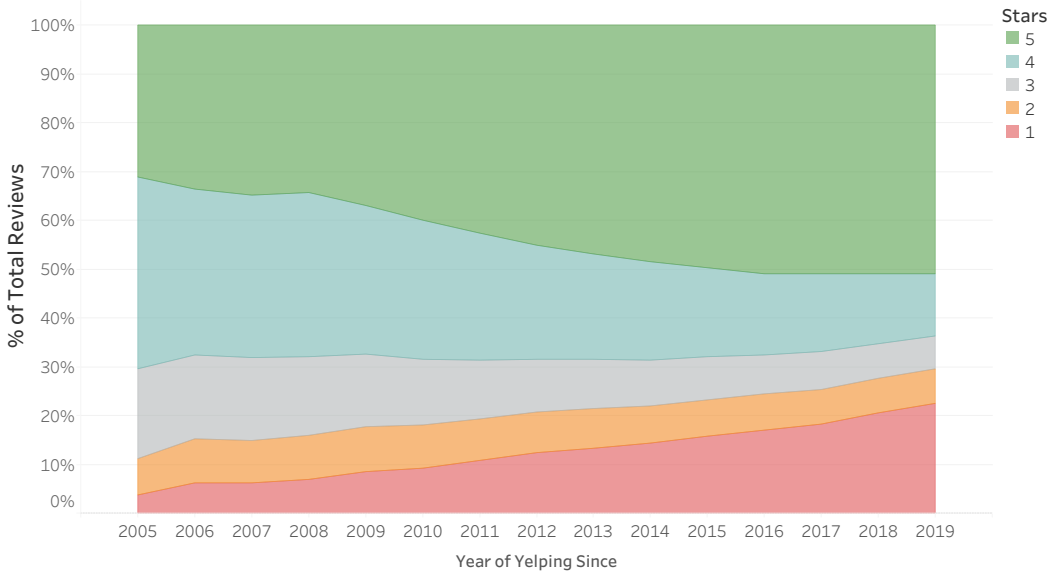


Figure 7: Yelp Review Polarization by Reviewer Cohort

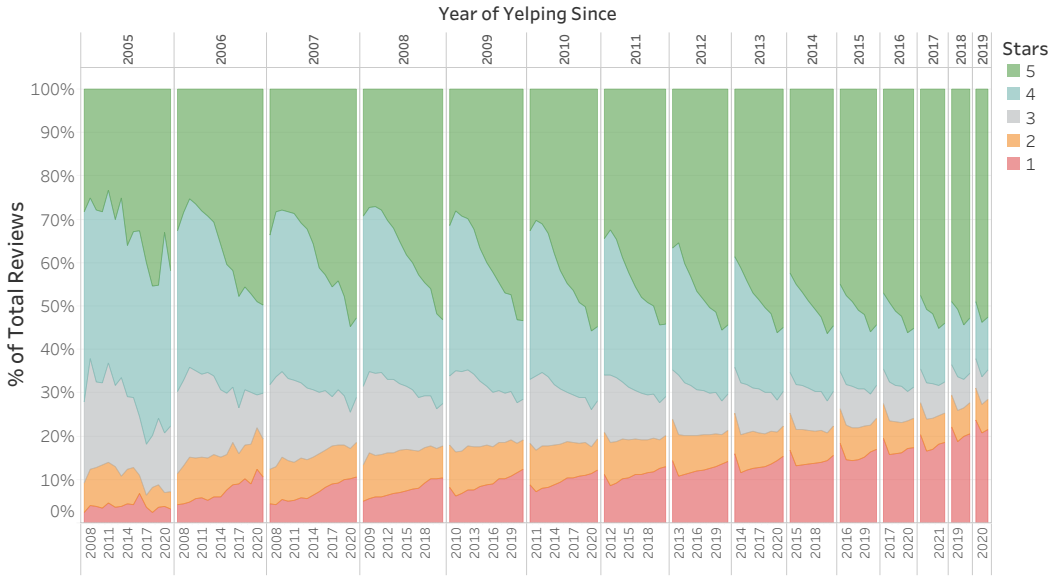


Figure 8: Yelp Review Polarization by Reviewer Tenure

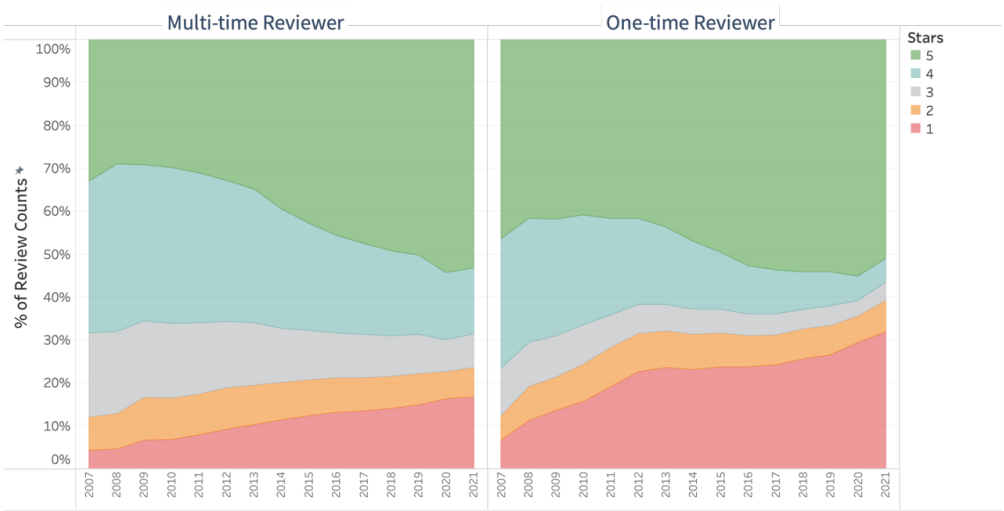


Figure 9: Yelp Review Polarization by Reviewer Activity

The patterns discussed above demonstrate that the polarization trend over time is robust and consistent across geographic markets and various aspects of reviewer and business characteristics.

4 Conceptual Framework and Model

We propose a method to uncover the underlying drivers of review polarization. Our conceptual framework, presented in Figure 10, starts with the observation that each review consists of both a numerical rating and accompanying text that describes the consumer's experience. To investigate potential shifts in rating standards, we use review text as our main proxy for consumer experience.[2] A potential concern in using review text as a measure is that the language norms reviewers use to describe similar experiences may change over time. To assess whether such changes occurred during our sample period (2012-2022), we analyze the distribution of language uses in reviews for businesses with stable product and service quality. If these distributions remain relatively consistent over time, we can infer that language norms have not shifted. In Section A of the Appendix, we show that for the coffee chain category, where offerings remained relatively unchanged during the sample period, the distribution of review embeddings was indeed stable throughout. To further examine possible shifts in word usage, we analyze the set of adjectives used in review texts across years. Our findings indicate that the language used in Yelp reviews remained largely consistent from 2012 to 2022, with the proportion of reviews containing different words between any two years typically varying by less than 1%. These findings confirm that review text is a reliable measure of consumer experience in our setting.

With text as our experience measure, we define the mapping from the review text to the numerical rating as the rating *scale*. Figure 10 illustrates our framework: the red arrow denotes this mapping between the text and the stars, i.e. the rating scale.

To identify potential drivers of the polarization trend, we draw on mechanisms discussed in the literature. First, reviewer characteristics can influence both review content and the rating scale (Moe and Schweidel, 2012; Simchon et al., 2022; Bondi et al., 2024). For example, Moe and Schweidel (2012) find that active reviewers tend to write more negative and more differentiated reviews, while less frequent reviewers are more likely to be influenced by previous opinions. Over time, different user cohorts may express experiences and sentiments in distinct ways, leading to more or fewer extreme ratings. Reviewer experience is also shown to affect reviews (e.g., Bondi

[2] An alternative approach is to infer consumer experience from revealed preferences, such as purchase behavior. However, such data are typically unavailable in online review settings. In addition, revealed preferences are relative and depend heavily on the set of available alternatives.

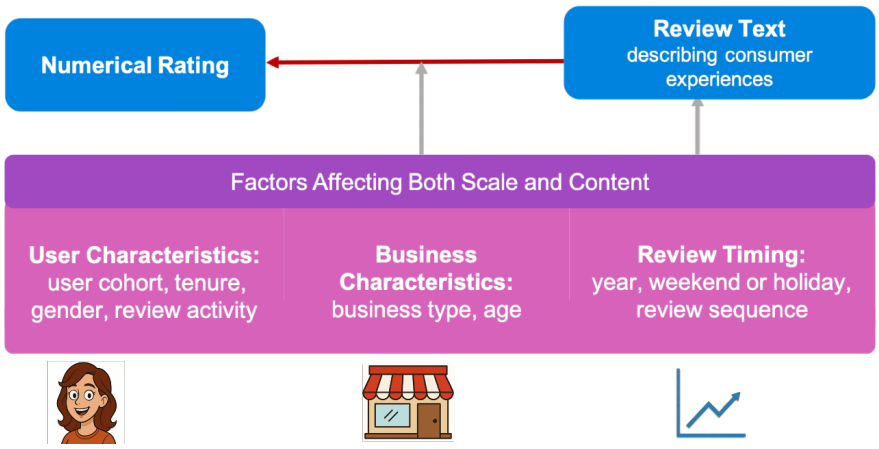


Figure 10: Conceptual Framework

et al. (2024)). Accordingly, we include all these reviewer characteristics in our model: user cohort, tenure, gender, and activity frequency.[3]

Second, business characteristics such as business type and age can also affect both content and scale. Reviewers may apply different standards to different types of businesses, and the language used may vary accordingly. For this reason, we restrict our analysis to restaurant reviews on Yelp, which account for the majority of reviews (about 68%).[4] Business age may influence content and scale due to a self-selection process: reviewers with different preferences arrive at different times. As shown by Li and Hitt (2008), early reviewers are likely to be early adopters with a higher preference for the product and thus more likely to rate positively, while later reviewers may have lower preferences and be more likely to write negative reviews. This self-selection process can affect both the content and the scale used for numerical ratings. For these reasons, we include business age in our model.

Finally, review timing — such as year, day of the week (e.g., weekend or holiday), and the sequence of the review in a business's review history — has also been shown to matter for reviews (Bayerl et al., 2023; Godes and Silva, 2012; Wu and Huberman, 2008). Both the self-selection biases discussed above and the impact-effort trade-off proposed by Wu and Huberman (2008) can shape the evolution of reviews and thus influence both content and scale. Bayerl et al. (2023)

[3]Reviewer demographics and political positions can also influence how experiences are rated and reviews are written. However, we do not observe reviewer demographics except for gender, which is inferred from names. We also cannot assume that reviewers share the average demographic characteristics of people in the areas surrounding the businesses they review. Therefore, only gender is included in our analysis.

[4]The small sample sizes for other categories, as shown in Table 15 of Web Appendix B, also make it difficult to train our prediction model accurately. Therefore, we limit our sample to restaurants only.

further illustrate that reviewers who post during weekends and on holidays are different from those who review at other times, with weekend reviewers tending to be more negative. Year, restaurant age, and review sequence have all been shown to impact ratings as well (Godes and Silva, 2012). Therefore, we also include these factors in our model.

Building on this conceptual framework, we develop a model to separate the scale effect from the content effect in review dynamics and to examine the factors that influence these effects. Details of the model are provided in Section 4.2. To capture the nuances of language, we employ an LLM encoding approach, described below.

4.1 Encoding Yelp Review Text Using LLMs

We use large language models (LLMs) to encode review texts, generating inputs for our model that maps content to ratings. The existing literature on review content analysis typically relies on two main approaches: (a) sentiment analysis, which focuses on the frequency of certain words (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 widely used tool for this purpose is the Linguistic Inquiry and Word Count (LIWC) software, which is frequently employed in psychology and social science research (Tanana et al., 2021). While 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). In this paper, we demonstrate that models using LLM embeddings achieve much higher prediction accuracy than those based on traditional approaches.

Recent advances in Large Language Models (LLMs) offer new opportunities to leverage state-of-the-art language processing tools for analyzing online review content.[5] These models typically encode texts into high-dimensional numeric values, known as the embedding space. A vector in this space efficiently encapsulates the semantic meanings of the underlying text and the contextual connections between text blocks. To encode the review text, we use OpenAI's API to obtain embeddings for all the review messages in our dataset[6]. This process translates each review message into a $3,072$-dimensional vector, which can then serve as input to machine learning models for predicting ratings.[7]

[5]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).

[6]We use OpenAI's most advanced 'text-embedding-3-large' embedding model as of January 2025. For details, see https://platform.openai.com/docs/guides/embeddings.

[7]Obtaining embedding values directly from pretrained models without any task-specific training data or fine-tuning

However, the 3,072-dimensional embedding is quite large. Training and validating prediction models with the full set of vectors and millions of observations would be computationally demanding. To address this, we reduce the dimensionality of the embedding space by implementing principal component analysis (PCA), which captures 80% of the variance in the embedding data with only 404 principal components. This represents a significant reduction in dimensionality compared to the original embedding space, without sacrificing much predictive power. As we show in Section B of the Appendix, for the class of ordered logit models (OLMs) — the main approach in our paper — the model with the full embeddings achieves a similar level of prediction performance as the model with principal components. Moreover, for other models we experimented with, such as the MNL and DNN models, the principal components actually outperform the full embeddings in prediction. This likely reflects the high dimensionality of the full embeddings, which increases the variance of model estimates. There is a variance-bias tradeoff: while full embedding models may capture more information, they can underperform due to increased noise and variance, especially when the extra dimensions do not add predictive value. These results suggest that the remaining 20% of the variance in the LLM embedding space is not essential for predicting ratings. This may be because the principal components allow us to capture the key language features of review texts for the restaurant industry specifically, while the LLM embeddings are based on the universe of all industries and contexts.

With the principal components in hand, we rank them based on their eigenvalue contributions to the original embedding space and refer to them by their rankings. 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 20 in Web Appendix B. The first 10 principal components capture over 20% of the total data variation, and the first 50 principal components capture nearly 44%. From PC184 onwards, the marginal contribution of each principal component to the total data variation is less than 0.1%. In addition to using 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: WC (Word Count), WPS (Words per Sentence), Analytic, Clout, Authentic, Tone, BigWords, Linguistic, Drives, Cognition, Affect, Social, Culture, Lifestyle, Physical, Perception, and Conversation. We provide the definition of each of these measures in Table 18 of Web Appendix B.[8] In Section 5, we further link the principal

is known as the zero-shot approach. Yelp engineering team discussed the performance of this approach is comparable to customized models. See https://engineeringblog.yelp.com/2023/04/yelp-content-as-embeddings.html.

[8]For a full description of the procedure and list of available measurements, see https://www.liwc.app/help/psyc

components to these psychometric measures to more intuitively illustrate the semantic meanings of the most important principal components.

4.2 Prediction Model

Among the various prediction models we considered, we chose the ordered logit model for its strong balance of predictive accuracy and interpretability. We also tested other machine learning models, such as multinomial logit (MNL) and deep neural networks (DNN). Although these alternatives achieved slightly better predictive performance, the ordered logit model is much easier to interpret. Further details can be found in Section B of the Appendix. Let r_i denote the rating given in review i, t represent the year in which review i was written, $\boldsymbol{C}_i$ denote the review content, and $r_i^* = \boldsymbol{C}_i^{'}\boldsymbol{\beta}_t$ denote the content score for the review. The ordered logit model is specified as:

$$P(r_i = k) = \frac{1}{1+\exp\big(-(\alpha_{it}^k - \boldsymbol{C}_i^{'}\boldsymbol{\beta}_t)/\sigma_{it}\big)} - \frac{1}{1+\exp\Big(-(\alpha_{it}^{k-1} - \boldsymbol{C}_i^{'}\boldsymbol{\beta}_t)/\sigma_{it}\Big)}, \tag{1}$$

where $k \in \{1,2,3,4,5\}$ represents the rating level from 1 to 5. Here, α_{it}^k are the threshold parameters, $\boldsymbol{\beta}_t$ are the coefficients for the content embeddings, and σ_{it} is the scale parameter for the logit error distribution. We use the principal components to represent the content $\boldsymbol{C}_i$.

For α_{it}^k, since Yelp uses a five-star rating system, we set $\alpha_{it}^0 = -\infty$ and $\alpha_{it}^5 = \infty$ by definition for any review. In our model, we allow the thresholds $\boldsymbol{\alpha_i} = (\alpha_{it}^1, \alpha_{it}^2, \alpha_{it}^3, \alpha_{it}^4)$ to vary at the individual review level to account for all factors that might influence the rating scale, such as reviewer and business characteristics. Specifically,

$$\begin{aligned}\alpha_{it}^k =& \sum_t \theta_t^k YEAR_{t(i)} + \sum_c \theta_c^k COHORT_{ci} + \sum_c \theta_r^k COHORT_{ci} \times TENURE_{ci} + \theta_g^k GENDER_i \\ &+ \theta_1^k ONCE_i + \theta_o^k SEQUENCE_i + \theta_a^k BIZ_AGE_i + \theta_w^k WKND_i \end{aligned} \tag{2}$$

where $YEAR_{t(i)}$ is an indicator for the year in which review i was written, and θ_t^k is the year fixed effect; $COHORT_{ci}$ is an indicator for the cohort of the user who wrote review i (i.e., the year the user joined Yelp), and θ_c^k is the cohort fixed effect; $TENURE_{ci}$ is the user's tenure, defined as the log of the number of days since the user joined Yelp up to review i. We interact $TENURE_{ci}$ with $COHORT_{ci}$ to capture cohort-specific trends in rating scale changes. $GENDER_i$ is an indicator for the inferred gender of the user, based on their name; names with unclear gender are labeled as unknown. $ONCE_i$ is an indicator for whether the user posted only one review during the sample period, reflecting user activity frequency. $SEQUENCE_i$ is the sequence number of review i in the

hometrics-manuals.

business's review history. BIZ_AGE_i is the age of the business in years at the time of review i, calculated as the difference between the date of the first review for the business and the date of review i. $WKND_i$ indicates whether review i was written on a weekend or holiday.

For the coefficients $\boldsymbol{\beta_t}$ associated with the content embeddings, we initially allow them to vary by year to capture potential changes in the importance of different embeddings in the mapping from review text to rating scale over time. However, our analysis shows that this added flexibility does not improve prediction performance, while greatly increasing computational complexity. Detailed results of this analysis are provided in Section C of the Appendix. Therefore, in our estimation, we hold $\boldsymbol{\beta_t}$ constant over time. For σ_{it}, we allow it to vary by all the factors included in the specification of $\boldsymbol{\alpha}$, i.e., equation (2), with details described in Web Appendix A. In doing so, we account for the fact that the variance of the underlying logit errors can change based on the various factors. More importantly, σ_{it} governs how sensitive the rating prediction is to the content score.

A key advantage of the ordered logit model is the interpretability of its parameters. By examining how $\boldsymbol{\alpha}_{it}$ and σ_{it} change with respect to all the factors in equation (2), we can explore how rating thresholds are affected by the various factors, and in turn decipher the contribution of each factor to the polarization trend. In particular, the year fixed effects θ_t^k capture how the thresholds evolve over the years after all observable reviewer and business characteristics are controlled for. θ_c^k (the cohort fixed effects), θ_g^k (gender fixed effects), and θ_1^k (one-time user fixed effect) illustrate how rating thresholds change due to the turnover of users with different characteristics. The coefficient of the interaction term between cohort and tenure, θ_r^k, captures the separate trend of threshold change within each cohort. θ_o^k, θ_a^k, and θ_w^k control for the effect of other review timing factors, especially those relating to the selection biases discussed in the literature (Bayerl et al., 2023; Godes and Silva, 2012; Wu and Huberman, 2008).

Additionally, examining $\boldsymbol{\beta}$ 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 for a given σ_{it}. The score $r^* = \boldsymbol{C}_i'\boldsymbol{\beta}$ summarizes the content in a review as a single number, making it straightforward to examine how content evolves over time.

4.3 Method to Decompose Review Polarization

Our prediction model enables us to disentangle the evolution of online reviews into two distinct components: content effects (changes in the review text itself) and scale effects (shifts in how reviewers translate their experiences into star ratings). Because the rating thresholds and variability (represented by $\boldsymbol{\alpha}$ and σ) differ across reviewers and businesses, isolating changes in content alone requires standardizing these scale parameters. To do this, we anchor $\boldsymbol{\alpha}$ and σ to a common baseline—specifically, the typical rating scale from our reference year, 2012.

Establishing the Baseline Rating Scale In practice, this standardization involves estimating an "average" rating scale for 2012. Here, "average" does not refer to a simple arithmetic mean, but rather to the combination of rating thresholds $\boldsymbol{\alpha}$ and σ that best explains (maximizes the likelihood of) all reviews in 2012, conditional on their content scores $\boldsymbol{C}_i'\boldsymbol{\beta}$. In other words, we re-estimate the ordered logit model using the original content scores from our main estimation, but constrain the rating scale parameters to be constant across all reviewers and businesses. The resulting parameters, $\bar{\boldsymbol{\alpha}}_b$ and $\bar{\sigma}_b$ (where the subscript b denotes the base year), thus represent the typical rating scale used in 2012. This approach sets 2012 as the reference point, allowing us to measure how rating scales have shifted in subsequent years relative to this baseline.

Decomposition Concept Conceptually, the decomposition proceeds as follows. The change in the share of 1- or 5-star ratings from the base year (2012) to year t can be attributed to:

- **Content Effect:** The change in the predicted share of 1- or 5-star ratings resulting from changes in content alone, holding the 2012 baseline $(\bar{\boldsymbol{\alpha}}_b, \bar{\sigma}_b)$ fixed.
- **Scale Effect:** The change in the predicted share of 1- or 5-star ratings resulting from changes in $(\boldsymbol{\alpha}_{it}, \sigma_{it})$ compared with the baseline, holding content fixed at year t values.

The scale effect can be further decomposed into components attributable to different sources:

- **Scale Effect due to Year Trend:** The portion of the change in the predicted share of 1- or 5-star ratings explained by changes in $(\boldsymbol{\alpha}_{it}, \sigma_{it})$ associated with year fixed effects. By setting all year fixed effects to their 2012 values, we obtain a counterfactual predicted share for year t that excludes the impact of year fixed effects. The difference between the actual predicted share for year t and this counterfactual gives the scale effect due to the year trend.
- **Scale Effect due to Heterogeneity:** The remaining portion, capturing the influence of changes in reviewer and business characteristics.

We formalize this decomposition algebraically below, beginning with the predicted probability that

a review receives rating k in year t.

Decomposing Individual Review Probabilities For any review i written in year t with content C_i, we can predict its probability of receiving a k-star rating using different scales:

1. **Actual Predicted Probability (P_{it}^k):** This uses the review's specific content ($\boldsymbol{C}_i$) and its estimated individual scale ($\boldsymbol{\alpha}_{it}, \sigma_{it}$), which reflects year t and all the non-year factors in Equation (2), denoted by X_i: $P_{it}^k = P(r_i = k|\boldsymbol{C}_i; \boldsymbol{\alpha}_{it}, \sigma_{it})$, or equivalently $P_{it}^k = P(r_i = k|\boldsymbol{C}_i; \boldsymbol{\alpha}(YEAR_t, X_i), \sigma(YEAR_t, X_i))$.
2. **Content-Based Probability ($P_{it}^{k,C}$):** This uses the review's content ($\boldsymbol{C}_i$) but applies the baseline 2012 scale ($\bar{\boldsymbol{\alpha}}_b, \bar{\sigma}_b$): $P_{it}^{k,C} = P(r_i = k|\boldsymbol{C}_i; \bar{\boldsymbol{\alpha}}_b, \bar{\sigma}_b)$. This value represents what the rating would likely be if the 2012 baseline rating standards were applied to year t's content.
3. **Heterogeneity-Adjusted Baseline Probability ($P_{it}^{k,H}$):** This uses the review's content ($\boldsymbol{C}_i$) and a scale that incorporates its specific heterogeneity factors X_i, but uses the baseline 2012 year effects: $P_{it}^{k,H} = P(r_i = k|\boldsymbol{C}_i; \boldsymbol{\alpha}(YEAR_b, X_i), \sigma(YEAR_b, X_i))$. This helps isolate the part of the scale effect due purely to time trends.

Using these probabilities, we can decompose the predicted probability for review i in year t as follows:

$$P_{it}^k = P_{it}^{k,C} + \Delta S_{it}^{k,Y} + \Delta S_{it}^{k,H}, \tag{3}$$

where:

- $\Delta S_{it}^{k,Y} = P_{it}^k - P_{it}^{k,H}$: This term represents the **scale effect due to year trend**. It captures the change in probability due to shifts in the year fixed effects from 2012 (year b) to year t, holding user and business heterogeneity (X_i) constant.
- $\Delta S_{it}^{k,H} = P_{it}^{k,H} - P_{it}^{k,C}$: This term represents the **scale effect due to heterogeneity**. It captures how much the predicted probability changes due to reviewer and business characteristics (X_i) when compared to the 2012 baseline average, holding the year effect at the 2012 (year b) level.

Equation (3) shows that the actual predicted probability is the sum of a baseline content-based probability plus two scale effect components.

Decomposing Changes in Annual Aggregate Rating Distributions We can aggregate this individual-level decomposition to understand the change in the proportion of k-star reviews (R^k) from the base year ($b = 2012$) to any given year t. Let R_t^k be the observed proportion in year t, and $\mathbb{E}[P_{it}^k]$ be the model-predicted proportion (the average P_{it}^k across all reviews in year t). The total

change $\Delta_{bt}(R^k) = R^k_t - R^k_b$ can be written as the change in predicted proportions plus a prediction error, which is very close to zero in our estimation:

$$\Delta_{bt}(R^k) = \mathbb{E}_t[P^k_{it}] - \mathbb{E}_b[P^k_{jb}] + \varepsilon^k_{bt}, \tag{4}$$

where ε^k_{bt} is the prediction error, and j indexes reviews written in year b. Since year b is our baseline, its scale effects are zero by definition, so $\mathbb{E}_b[P^k_{jb}] = \mathbb{E}_b[P^{k,C}_{jb}]$. Substituting our decomposition of P^k_{it}:

$$\Delta_{bt}(R^k) = \mathbb{E}_t[P^{k,C}_{it} + \Delta S^{k,H}_{it} + \Delta S^{k,Y}_{it}] - \mathbb{E}_b[P^{k,C}_{jb}] + \varepsilon^k_{bt}. \tag{5}$$

Rearranging gives our final decomposition (ignoring the small prediction error):

$$\Delta_{bt}(R^k) \approx \underbrace{(\mathbb{E}_t[P^{k,C}_{it}] - \mathbb{E}_b[P^{k,C}_{jb}])}_{\text{Content Effect}} + \underbrace{\mathbb{E}_t[\Delta S^{k,Y}_{it}]}_{\text{Scale Effect (Year Trend)}} + \underbrace{\mathbb{E}_t[\Delta S^{k,H}_{it}]}_{\text{Scale Effect (Heterogeneity)}}. \tag{6}$$

Equation (6) breaks down the total change in the share of k-star reviews into the three components: content effect, scale effect (year trend), and scale effect (heterogeneity).

5 Results

We begin by presenting an overview of our model's performance and highlighting key estimates. Next, we examine the estimated model parameters in detail, with particular attention to how various factors contribute to variation in the thresholds $\boldsymbol{\alpha}$. Because the parameter estimates for σ are very small — resulting in σ_i values close to 1 for most observations — we do not discuss these in detail here, but instead refer readers to Web Appendix A. We also identify the linguistic features that most strongly influence ratings, based on the estimates of $\boldsymbol{\beta}$, and interpret the most important principal components (PCs) by linking them to LIWC scores. Finally, we decompose changes in rating distributions over time into scale effects (arising from year trends and user/business heterogeneity) and content effects. Throughout this section, we primarily use the Philadelphia market as a detailed example, but also report results for all metropolitan areas where relevant.

5.1 Overview of Model Performance and Estimates

We estimate the ordered logit model separately for each metropolitan area in the Yelp dataset. Because the number of reviews before 2012 is limited,[9] we focus on a ten-year sample period from 2012 to 2021. As an illustration, Table 1 presents the confusion matrix for the Philadelphia market. Each row shows the distribution of model-predicted ratings for reviews with a given actual star rating, with rows summing to 100%. The diagonal entries indicate the hit rates for each star level,

[9] See Figure 19 of Web Appendix B.

i.e., the percentage of predictions that match the observed data. The highest hit rates are for 1-star and 5-star reviews (81% and 87%, respectively), followed by 4-star (56%), 3-star (52%), and 2-star (44%) reviews. The overall hit rate of the model is 0.71 for the Philadelphia market.

For all star levels, the largest percentages are on the diagonal, indicating that the model predicts the correct rating most of the time. Most misclassifications are within one star of the true rating. This level of accuracy is reasonable, given the inherent uncertainty in how users assign ratings: Yelp's rating scale is discrete, reviewers often round their assessments to the nearest star, introducing some randomness at the boundaries. The high concentration of predictions within one star of the true rating supports the strong performance of our approach.[10]

Table 1: Model Accuracy for Philadelphia Reviews (2012-2021)

	Predicted Star Rating				
Actual Star Rating	1	2	3	4	5
1	**0.806**	0.158	0.034	0.002	0.000
2	0.339	**0.436**	0.212	0.013	0.001
3	0.038	0.162	**0.523**	0.261	0.017
4	0.001	0.004	0.071	**0.556**	0.369
5	0.000	0.000	0.004	0.124	**0.872**

Note: Each row shows the distribution of predicted stars for a given actual star rating, with values in each row summing to 1. The diagonal entries (highlighted in bold) represent the proportion of correctly predicted reviews for each star category.

Model Comparison Our prediction method outperforms alternative approaches by a substantial margin. For instance, a naive model that ignores review content and always predicts the most common rating (popularity bias) achieves a baseline hit rate of 0.46 (see Table 2). Using LIWC attributes from the review texts, the hit rate rises to 0.51 with only sentiment-related attributes, and to 0.52 when all main LIWC attributes are included — a 10.9% to 12.7% improvement over the baseline. The small difference between sentiment-only and all LIWC attributes suggests that sentiment is the most predictive language feature for star ratings. When we use LLM embeddings as predictors, the accuracy jumps to 0.72, representing a 55.7% improvement over the baseline and a 38.5% improvement over LIWC analysis. This is equivalent to a 48.1% reduction in prediction errors compared to the baseline. These results indicate that LLM embeddings capture much more nuance in review language than traditional sentiment features.

[10]The model's hit rates are lower for moderate ratings (2, 3, and 4 stars) partly because these ratings are less common. To address this, we overweight moderate ratings in the estimation so that all star levels are equally represented. Table 17 of Web Appendix B shows the resulting accuracy rates. The hit rates for moderate ratings, especially 2 stars, improve, but at the expense of 1- and 5-star ratings, e.g., 5-star hit rate drops from 87% to 76%.

Table 2: Model Performance Comparison Across Metropolitan Areas

	Hit Rate			
Metro	Baseline	LIWC sentiment	LIWC	LLM
Boise	0.50	0.56	0.57	0.73
Edmonton	0.33	0.44	0.44	0.69
Indianapolis	0.46	0.47	0.49	0.71
Nashville	0.47	0.50	0.52	0.73
New Orleans	0.48	0.48	0.51	0.72
Philadelphia	0.43	0.48	0.49	0.71
Reno	0.49	0.57	0.58	0.73
Saint Louis	0.44	0.49	0.50	0.72
Santa Barbara	0.55	0.58	0.59	0.74
Tampa	0.49	0.51	0.53	0.73
Tucson	0.47	0.52	0.54	0.71
Avg Score	0.46	0.51	0.52	0.72
Score Improvement	0	10.9%	12.7%	55.7%
Error Reduction	0	9.3%	11.1%	48.1%

Note: Performance metrics are calculated using out-of-sample predictions on a 20% test set, with models trained on the remaining 80% of data.

Overview of the Threshold and Content Evolution To provide an overview of the model estimates, Figures 11 and 12 show the evolution of the average numerical values of α_{ik} by year and the content score by year, respectively, for the Philadelphia market. These visualizations offer a preliminary look at the trends in polarization and their sources. Figure 11 illustrates how the cutoff thresholds between different star ratings have shifted over time. These thresholds determine when a reviewer's underlying assessment translates into a specific star rating. The figure reveals a clear trend toward increasing polarization. The average cutoff for 1-star ratings has risen, meaning reviewers now require a higher content score to give more than 1 star. In contrast, the thresholds for higher star ratings (e.g., 4 to 5 stars) have declined, indicating that reviewers are more likely to award higher ratings over time. This divergence suggests that the rating distribution is becoming more polarized, with a growing tendency toward extreme ratings.

On the content side, Figure 12 displays the empirical density of the content score over time in Philadelphia. The dotted vertical lines represent the 2012 baseline average $\bar{\boldsymbol{\alpha}}_b$. The content score is computed by applying the estimated $\boldsymbol{\beta}$ to the content attributes for all years. The figure shows that the distribution of content scores was mostly unimodal in the early years, with a thickening lower tail before 2012 and the emergence of a second mode to the left. In later years, the distribution becomes bimodal. These changes indicate the shifts in review content: a larger share of reviews became negative, while the proportion of extremely positive content in the right tail also increased. This pattern suggests that the review content itself became more polarized over time.

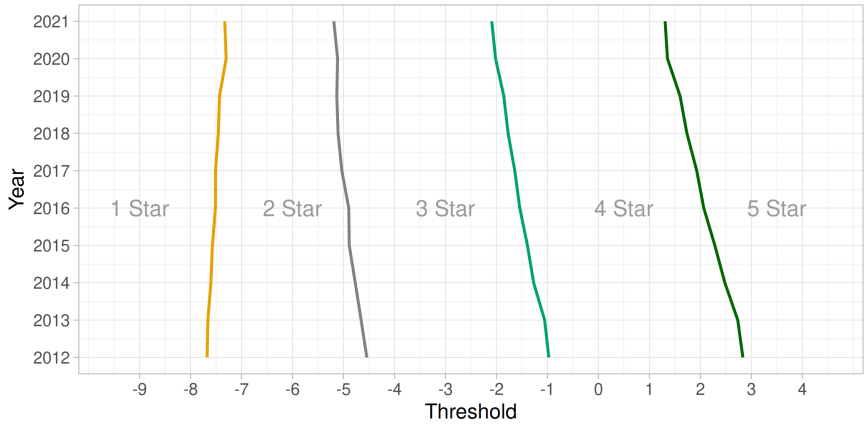


Figure 11: Average Cutoff Thresholds for Review Ratings by Year, Philadelphia

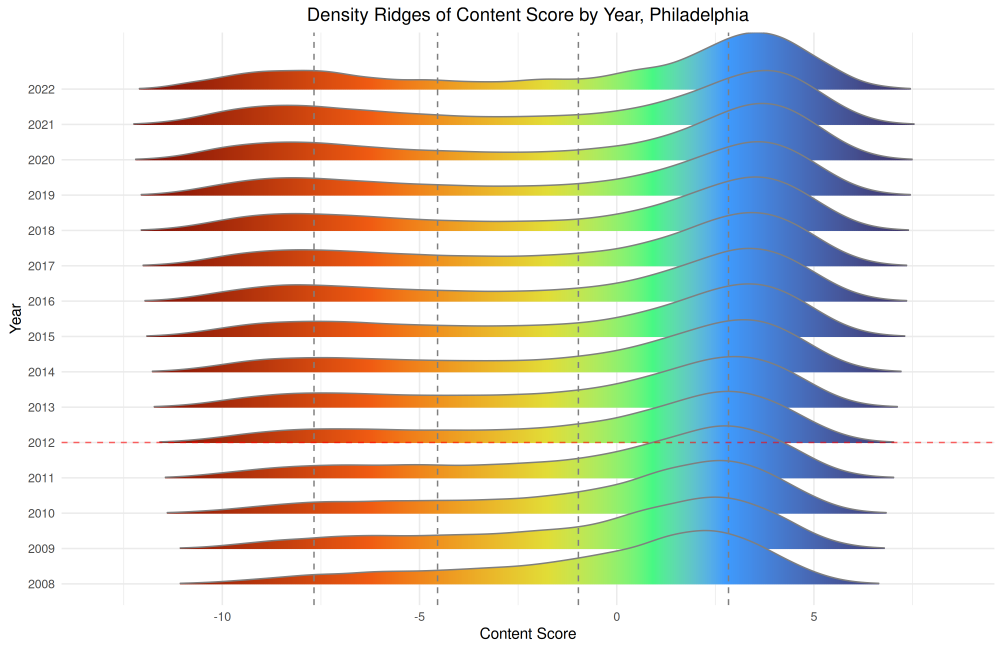


Figure 12: Review Content Distributions over Years with Cutoffs

5.2 Contributing Factors to Changes in Rating Thresholds and Content

Decomposing the Contribution of Various Factors to Threshold Variation By specifying $\boldsymbol{\alpha}_i$ as a function of the factors in equation (2), we can identify which factors drive variation in $\boldsymbol{\alpha}$, and the estimated coefficients quantify their contributions. Below, we summarize the effects of year, cohort, within-cohort trends, reviewer and business characteristics, and the contribution of each group of factors to threshold variation. Detailed regression results are provided in Section D of the Appendix. Table 3 presents the decomposition of each factor's contribution to variation in $\boldsymbol{\alpha}_i$ and highlights the key findings. Year effects account for the largest share (48.3%), underscoring the importance of temporal trends. This is followed by one-time reviewer effects (19.1%). User cohort and cohort tenure together contribute 11.5%, reviewer gender 5.3%, review sequence 8.7%, and business age 6.9%. The weekend/holiday effect is the smallest at 0.32%. We describe the effect of each factor in more detail below.

Table 3: Decomposition of Rating Threshold ($\boldsymbol{\alpha}$), Philadelphia

Factor	Contribution	Key Results
Year	48.27%	Largest driver; 1|2 threshold rises while 4|5 threshold declines over time, creating stronger polarization.
Cohort	3.91%	Newer cohorts show negative effects for 3|4 and 4|5 thresholds, making higher ratings more likely.
Tenure	7.59%	Experienced reviewers become more positive; all thresholds decline with tenure for 2010-2016 cohorts.
Gender	5.25%	Males have higher 1|2 thresholds but lower 4|5 thresholds than females, creating more polarized ratings.
One-time reviewer	19.10%	Second largest factor; coefficient much larger than other factors, strongly increasing extreme ratings.
Weekend/Holiday	0.32%	Smallest effect; positive coefficients for 1|2, 2|3 but negative for 3|4, 4|5, increasing polarization.
Review sequence	8.67%	Later reviews show de-polarization; thresholds for 1|2, 2|3, 3|4 decrease while 4|5 increases.
Business age	6.89%	Consistent positive effect across all thresholds; older businesses receive more negative ratings.

Notes: The decomposition illustrates the contributions of different factors to the overall variation in α_i. It considers the correlation between these factors. The decomposition is based on ANOVA.

Year Effects The year fixed effect estimates are shown by the colored solid lines in Figure 13. The gap between these lines and the overall average thresholds from Figure 11, shown by the gray dotted lines, represents the influence of non-year factors on the threshold changes. Two key patterns emerge. First, the year effects closely track the overall average threshold trend, highlighting the central role of year trends in driving threshold shifts. Second, the distance between the colored and gray lines varies by star rating, indicating that the influence of non-year factors is not uniform across thresholds. For the 1|2 threshold, the lines are mostly parallel, suggesting that non-year factors have a steady effect over time. In contrast, for higher thresholds, especially 4|5, the lines

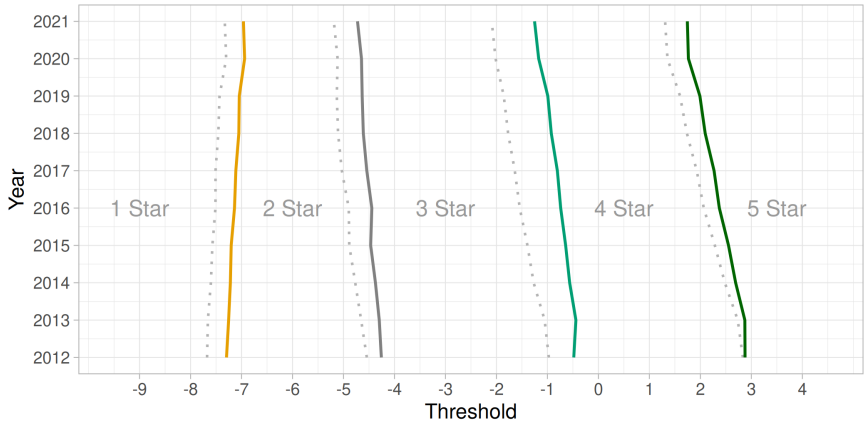


Figure 13: Estimates of Year Fixed Effects on Cutoff Thresholds, Philadelphia

diverge, indicating that non-year factors increasingly shape the variation in these thresholds. This pattern implies that while non-year factors have little impact on the rise in 1-star ratings, they play a much larger role in the increase in 5-star ratings, as we will show in the decomposition analysis in Section 5.3.

Cohort, Tenure, and Reviewer Characteristics Our analysis shows that newer cohorts are more likely to give higher ratings, with significant negative effects for higher thresholds (3|4 and 4|5), contributing to rating polarization. Experienced reviewers generally become more positive over time, with tenure effects primarily reducing the likelihood of negative reviews. Gender also influences rating behavior, with male reviewers exhibiting more polarized patterns than females. One-time reviewers display even stronger polarization tendencies, with substantially larger coefficients than gender effects, making them much more likely to give extreme ratings.

Business and Contextual Factors As more reviews accumulate for a business, ratings become less polarized, suggesting a de-polarization pattern in the review sequence. Business age consistently shows positive effects across all thresholds, indicating that older businesses receive more negative ratings. Reviews written on weekends or holidays demonstrate increased polarization in both directions, complementing findings by Bayerl et al. (2023) that weekend reviews tend to be more negative.

Determinants of Review Content We now turn to the content side of our model estimates, examining which features of review text are most predictive of ratings and how these content attributes relate to linguistic characteristics. Table 4 presents the top 10 principal components (PCs) with the highest marginal impacts on rating predictions across metropolitan areas. The

results are highly consistent: all markets share the same top 9 PCs, with only minor variation in the 10th. For each market, we report both the impact weight (the effect of a one-standard-deviation change in each PC on the content score r^*) and its ranking. Three key patterns emerge: 1) PCs with the largest eigenvalues (e.g., PC1 and PC5) also have the highest predictive impact, with PC1 substantially dominating; 2) some PCs with smaller eigenvalues (e.g., PC11, 18, 13, 17) are also strong predictors, indicating that subtle textual features matter; and 3) impact weights and rankings are highly consistent across markets.

Table 4: Marginal Effects and Importance Rank by Market

		Principal Component									
		1	5	2	11	18	13	17	8	3	4
Boise	Impact Weight	3.670	0.283	0.236	0.217	0.210	0.232	0.186	0.223	0.169	0.149
	Rank	1	2	3	6	7	4	8	5	9	11
Edmonton	Impact Weight	5.213	0.386	0.369	0.351	0.339	0.346	0.315	0.295	0.306	0.276
	Rank	1	2	3	4	6	5	7	9	8	10
Indianapolis	Impact Weight	4.197	0.372	0.290	0.267	0.248	0.273	0.224	0.226	0.232	0.187
	Rank	1	2	3	5	6	4	9	8	7	10
Nashville	Impact Weight	4.512	0.376	0.296	0.295	0.276	0.252	0.254	0.238	0.220	0.197
	Rank	1	2	3	4	5	7	6	8	9	10
New Orleans	Impact Weight	4.016	0.342	0.299	0.259	0.233	0.233	0.233	0.247	0.189	0.176
	Rank	1	2	3	4	8	6	7	5	9	10
Philadelphia	Impact Weight	4.476	0.375	0.307	0.305	0.298	0.291	0.273	0.239	0.230	0.203
	Rank	1	2	3	4	5	6	7	8	9	10
Reno	Impact Weight	4.134	0.283	0.276	0.258	0.258	0.287	0.236	0.250	0.198	0.198
	Rank	1	3	4	5	6	2	8	7	9	10
Saint Louis	Impact Weight	4.259	0.349	0.282	0.279	0.240	0.270	0.226	0.226	0.202	0.172
	Rank	1	2	3	4	6	5	8	7	9	10
Santa Barbara	Impact Weight	3.963	0.344	0.286	0.236	0.227	0.287	0.207	0.200	0.181	0.207
	Rank	1	2	4	5	6	3	8	9	11	7
Tampa	Impact Weight	4.614	0.369	0.303	0.281	0.270	0.316	0.258	0.291	0.223	0.195
	Rank	1	2	4	6	7	3	8	5	9	11
Tucson	Impact Weight	4.773	0.370	0.279	0.290	0.311	0.320	0.259	0.266	0.220	0.231
	Rank	1	2	6	5	4	3	8	7	10	9

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

To interpret the semantic meaning of these PCs, we regress the top 10 PCs on 17 key LIWC linguistic categories (see Section 4.1), including WC (Word Count), WPS (Words per Sentence), Clout, and Tone.[11] The regression results are summarized in Table 20 of Web Appendix B. For PC1, LIWC scores explain about 51% of the variance (as indicated by R-squared), mainly through Word Count and Tone (i.e., sentiment). PC5 is most associated with Word Count and Clout (i.e., words showing leadership and status). Across the top PCs, Word Count, Clout, Tone, and Physical (i.e., words associated with health and wellness) are the most important semantic predictors.

[11] Except for WC and WPS, LIWC scores range from 0 to 100 based on relevant word frequencies. Although LIWC captures less context than modern ML tools, its categories are intuitive and concrete (e.g., Tone directly measures sentiment).

Finally, we regress the content score on the same factors as in equation (2) to examine what drives changes in review content. The results show that review content increases with the year trend but decreases with cohorts' entry year. Female reviewers are more positive than their male counterparts, and weekday reviewers are more positive than weekend reviewers. Detailed findings for Philadelphia are discussed in Section E of the Appendix.

5.3 Polarization Decomposition Results

Using the decomposition method described previously, we break down changes in the distribution of review ratings into scale effects (arising from year trends and reviewer/business heterogeneity) and content effects. Table 5 presents the overall changes in rating proportions and their decomposition for Philadelphia. Since 2012 serves as the base year, all values are zero for that year. Results are reported separately for 1-star and 5-star ratings. The "total effect" column shows the overall change in the shares of 1- and 5-star reviews in each year. For example, the share of 1-star reviews increased by 82.4% from 2012 to 2021. The subsequent columns decompose this change into scale and content effects.

The contributions of each effect differ by rating: for 1-star reviews, the increase from 2012 to 2021 is driven predominantly by the content effect (80.4%), with the scale effect from year trends accounting for 18.1%, and reviewer/business heterogeneity contributing only 1.6%. In contrast, the share of 5-star reviews rose by 69.2%, with the scale effect from year trends as the largest contributor (48.6%), followed by the content effect (34.8%) and reviewer/business heterogeneity (16.6%). These results are consistent with the patterns in Figure 13, where reviewer/business heterogeneity (i.e., factors other than year) had little influence on the 1|2 threshold over time, but played a more significant role in the 4|5 threshold. The substantial impact of the scale effect from year trends, especially on the increase in 5-star reviews, may reflect evolving social norms around providing online feedback. A growing body of literature suggests that social-pressure mechanisms, such as reciprocity, platform design, and reputation inflation, have increasingly pushed ratings toward the top of the scale, often independent of actual changes in quality (Fradkin et al., 2021; Pera et al., 2019; Filippas et al., 2022).

Overall, these results indicate that increases in 1-star reviews mainly reflect real changes in user experiences as captured in the review content, whereas the rise in 5-star reviews is more strongly shaped by shifting rating standards.

Figure 14 shows the breakdown of scale and content effects for all metropolitan areas from 2012

Table 5: Scale and Content Effects over Years in Philadelphia

	1 Star Review				5 Star Review			
	Proportion	Scale Effect		Content	Proportion	Scale Effect		Content
	Change	Year Trend	Heterogeneity		Change	Year Trend	Heterogeneity	
2021	+82.4%	18.1%	1.6%	80.4%	+69.2%	48.6%	16.6%	34.8%
2017	+39.7%	17.6%	-2.0%	84.4%	+48.1%	43.6%	20.5%	35.8%
2012	0.0%	-	-	-	0.0%	-	-	-

Note: 2012 is the base year. The table shows the decomposition of changes in 1-star and 5-star review proportions into scale effects (driven by year trends and reviewer heterogeneity) and content effects. For 1-star reviews, content effects dominate (80.4% in 2021), while for 5-star reviews, scale effects are more prominent (65.2% combined in 2021).

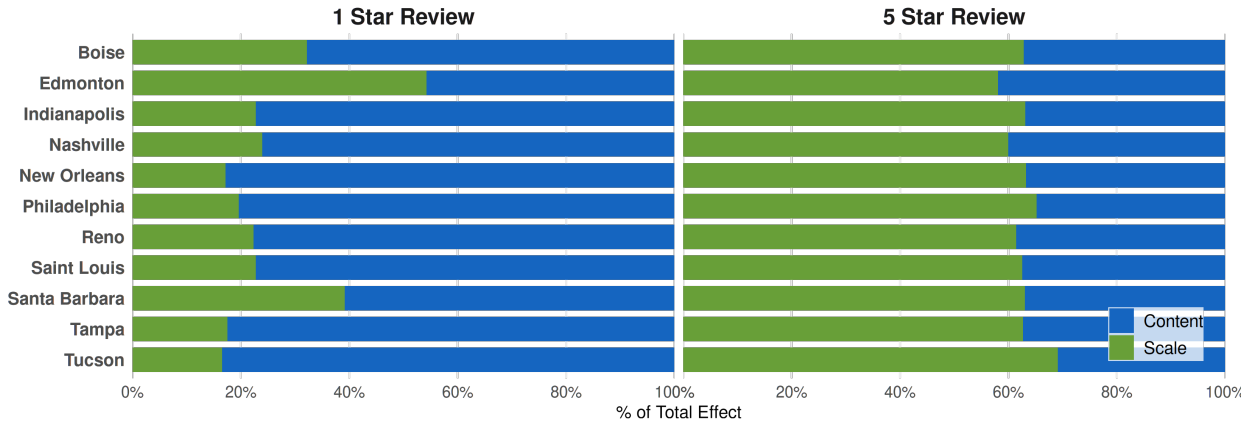


Figure 14: Decomposition of Scale and Content Effects for Major Markets from 2012 to 2021

to 2021. A more detailed breakdown is provided in Table 19 of Web Appendix B. The figure reveals a consistent pattern: for 5-star reviews, the scale effect is the main driver, contributing 58% to 69% of the increase in 5-star shares. For 1-star reviews, the content effect dominates, contributing 46% to 84% of the increase. When separating the two types of scale effects, we find that the scale effect from year trends contributes substantially to both 1-star and 5-star increases (about 42% and 54% on average, respectively). The scale effect from reviewer/business heterogeneity contributes moderately to 5-star increases (about 9% on average), but actually reduces the share of 1-star ratings in most markets (averaging a negative contribution of about -16%).

6 Impact on Informativeness and Managerial Implications

The existence of the scale effect suggests that ratings are influenced not only by consumer experiences, but also by changing rating standards. In this section, we examine how the scale effect impacts the usefulness of reviews and how it can distort the distribution of average business ratings. Drawing on our findings, we offer recommendations for platforms to enhance the informativeness,

Table 6: Impact of Scale Effect on Usefulness of Reviews

Dependent Variable: Useful	**Estimate**	**Std. Error**	**P-Value**
Scale Effect	−0.035	0.004	0.000
Content-based Predicted Rating	−0.132	0.006	0.000
Controls		✓	
N		4,260,492	
Year×Month FE		120	
Business FE		51,178	

Note: Control variables include user gender, cohort, tenure, one-time-user status, weekend or holiday effect, and review sequence.

consistency, and accuracy of ratings.

6.1 Impact of Scale Effect on Usefulness of Reviews

To assess how the scale effect impacts the usefulness of reviews, we regress an indicator for whether review i received a useful vote on both the scale and content components of the review. The regression model is specified as:

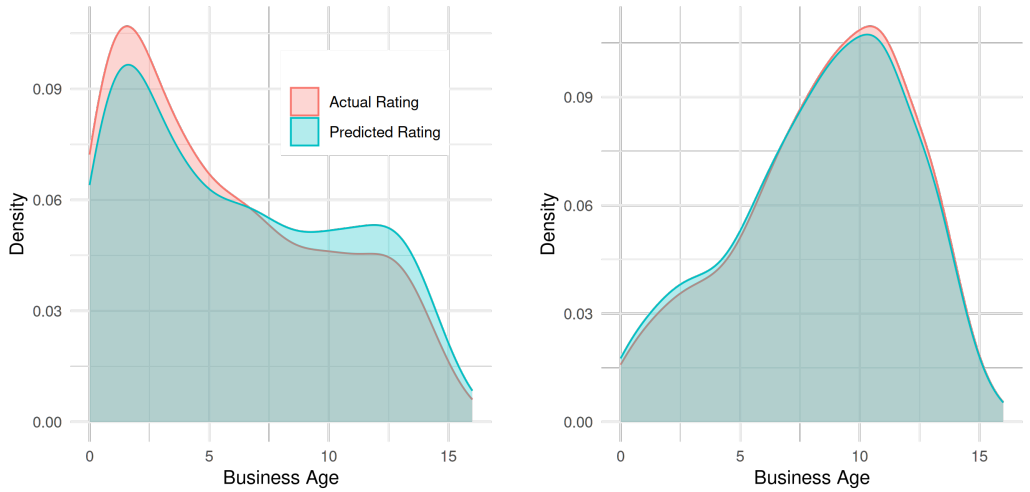
$$Useful_i = \phi_s \Delta Rating_i + \phi_c \widehat{Rating}_i + \boldsymbol{X}\boldsymbol{\phi} + \varepsilon_i \tag{7}$$

where $Useful_i$ is a binary variable indicating whether review i was marked as useful. $\widehat{Rating}_i$ is the predicted rating for review i based on its text, using the 2012 baseline $(\bar{\boldsymbol{\alpha}}_{\boldsymbol{b}}, \bar{\sigma}_b)$, representing the content effect. $\Delta Rating_i$ is the difference between the displayed rating and the content-based predicted rating, capturing the scale effect. $\boldsymbol{X}$ is a vector of control variables, including all variables from equation (2), year-month fixed effects, and business fixed effects. These controls account for time trends, reviewer and business heterogeneity, and the duration since the review was posted (considering that newer reviews had less opportunity to be marked as useful). ε_i is the error term.

Table 6 presents the results. The scale effect is negative and statistically significant, indicating that as the gap between the displayed rating and the content-based predicted rating increases by 1 star, the probability of a review being marked as useful drops by 3.5%. The content effect also has a negative and significant coefficient, suggesting that reviews with higher predicted ratings based on content are less likely to be considered useful. These results imply that discrepancies between displayed and content-based ratings, i.e., the scale effect, can reduce the perceived value of reviews.

6.2 Impact of Scale Effect on Rating Distribution by Business Age

Our results demonstrate that changes in the rating scale over time have a substantial impact on the observed numerical ratings on Yelp. Consequently, comparing ratings across years or between



(a) Top 20% Business Average Ratings (b) Bottom 20% Business Average Ratings

Figure 15: Distribution of Businesses Over Age Among Top and Bottom 20% Ratings in 2021, Philadelphia

businesses that entered the market at different times can be misleading. Specifically, because the scale effect associated with year trends has significantly increased the share of 5-star ratings in recent years, older businesses are systematically rated lower than newer ones, even when their service quality is similar. This puts older businesses at a competitive disadvantage.

To illustrate this, we conduct a counterfactual analysis by calculating average business ratings using only the content-based predicted ratings. Focusing on all businesses active in 2021 in Philadelphia, Figure 15 displays the age distribution of businesses among the top and bottom 20% of average ratings, using both displayed and content-based predicted ratings. The results show that businesses with the top 20% displayed average ratings tend to be newer, while those with the top 20% content-based predicted ratings are more likely to be older. This suggests that, among high-quality businesses, older establishments are more likely to receive lower displayed ratings than newer ones, despite offering comparable service. For the bottom 20% of average ratings, because the scale effect has minimal impact on the increase in the share of 1-star ratings over time, the age distribution of businesses in the lowest rating group is nearly identical whether based on displayed ratings or content-based predicted ratings.

Distortions in the distribution of business ratings are not only unfair to businesses but can also mislead consumers. Since average ratings are often the first information consumers see when comparing businesses online, such distortions may obscure the true quality of businesses and hinder

consumers' ability to identify the best options. Our findings therefore underscore the need for platforms like Yelp to reconsider how ratings are presented to users.

6.3 Recommendations: Scale Adjustment and AI-generated Scores

The importance of maintaining consistent and reliable rating standards has been increasingly recognized by online review platforms. It's worth noting that major platforms such as Yelp, Google, and Amazon have already implemented various measures to improve rating accuracy and credibility. Yelp filters reviews for reliability, Google now requires written comments and author identification for ratings, and Amazon uses algorithms that weigh review recency and helpfulness. These steps help ensure that displayed ratings better reflect genuine customer experiences and reduce bias or manipulation (Yelp, 2024; Search Engine Journal, 2021; Land, 2025; Bishop, 2015; Amazon, 2024).

The scale effect we identify represents an additional significant challenge to this goal, as it introduces systematic biases that can mislead both consumers and businesses. Our findings suggest that temporal shifts in rating standards continue to challenge the informativeness and fairness of online review systems, highlighting the need for more systematic solutions. We propose leveraging recent advances in LLMs to enhance the consistency and reliability of review systems, as these models can process nuanced review content and help mitigate distortions caused by shifting rating standards over time.

Specifically, we propose two main strategies. First, platforms could implement time-consistent rating systems by correcting for the scale effect we have identified. Since our framework tracks how rating standards have shifted over time, platforms could recalibrate ratings to a common reference point, such as the baseline scale used in our study. This would involve estimating the base year scale parameters, predict the content-based ratings, and then adjusting displayed ratings to reflect what they would have been under a fixed rating standard. Such normalization would remove temporal biases from ratings, allowing for fairer comparisons between businesses regardless of when they entered the market or when reviews were posted. Alternatively, platforms could implement a model-light yet effective approach by down-weighting ratings from one-time and inexperienced reviewers, whose rating scales are significantly more extreme. These approaches would help mitigate the systematic disadvantage of older businesses due to rating inflation over time.

Second, platforms could implement content-based AI-generated scores using LLMs. By analyzing the actual review texts rather than relying solely on numerical ratings, these models can extract consistent customer satisfaction measures regardless of when and by whom the review was written.

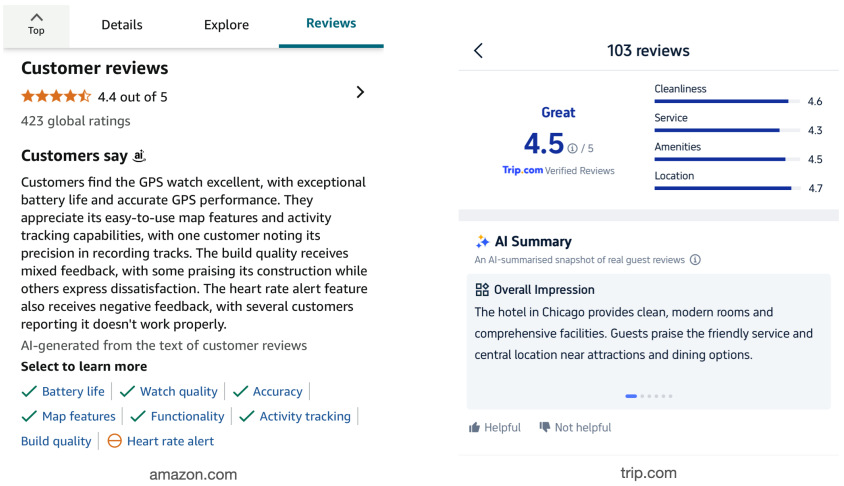


Figure 16: Example of AI-generated Review Summaries (Amazon and Trip.com)

Major platforms like Amazon (starting in August 2023) and Trip.com have recently started to use AI-generated customer review summaries on their websites (see Figure 16 for examples).

We conducted an illustrative experiment using ChatGPT (specifically, the GPT-4.1-mini model) to generate content-based scores for businesses in Philadelphia. For each business, we provided the model with all of its reviews history containing the original review texts without the numerical ratings and asked it to summarize the overall customer experience into a numerical score on a 5-point scale for each review message and then provide an aggregated score for the business. In addition, we ask the model to provide a textual summary of the review history, identifying recurring themes and customer satisfaction levels. See Figure 21 of Web Appendix B for the detailed prompt we used. This practice is similar to the AI summaries adopted on platforms like Amazon and Trip.com.

Our analysis revealed that these AI-generated content-based scores demonstrated a stronger correlation with the average of the content-based predicted ratings than with the average of the actual displayed ratings. Specifically, when regressing the AI-generated scores against the average content-based predicted ratings, we obtained an R^2 of 0.85, compared to an R^2 of 0.82 when regressing against the average displayed ratings. This difference of 0.03 in R^2 values, while modest, suggests that the AI model's interpretation of review content aligns more closely with our content-based predicted ratings, which control for the scale effect. This finding provides additional evidence that LLMs can effectively extract consistent quality signals from review text, potentially mitigating the distortions introduced by shifting rating standards over time. These results suggest

that platforms could leverage similar AI-based approaches to generate more consistent and reliable business quality indicators that are less susceptible to temporal biases in rating standards.

7 Conclusion

This paper examines the growing phenomenon of review polarization on online consumer review platforms, focusing specifically on Yelp restaurant reviews between 2012 and 2021. Utilizing advanced language processing techniques and a robust econometric framework, we decompose observed rating polarization into distinct scale and content effects. Our analysis reveals that the dramatic increase in polarized ratings is driven differentially by these two factors: the surge in 5-star reviews largely reflects changes in rating standards (scale effect), whereas the rise in 1-star reviews predominantly captures genuine deteriorations in consumer experiences (content effect).

We document significant shifts in rating standards over time, identifying year trends as a major contributor. Additionally, reviewer and business heterogeneity substantially amplify this polarization, particularly through the increasing prevalence of one-time reviewers who disproportionately contribute extreme ratings. Our results underscore critical implications for review informativeness and business competitiveness. We find that reviews with significant scale effects, where numerical ratings diverge from review content, are perceived as less useful, reducing the overall credibility of the platform. Furthermore, rating scale inflation systematically favors newer businesses, unfairly disadvantaging established businesses by distorting average ratings independent of true quality.

The findings presented here yield several managerial and policy implications. Review platforms should consider mitigating the adverse impacts of scale drift by reweighting ratings to emphasize experienced reviewers or by adopting AI-generated, content-based numerical scores to represent business quality more reliably. Our preliminary experiments demonstrate that content-based scores more accurately reflect consumer experiences, offering a feasible approach to restoring consistency and credibility of online ratings.

Future research could extend this analysis to other industries and platforms to further validate our findings. Moreover, investigating interventions aimed at reducing the prevalence of extreme ratings, such as reviewer incentives, improved review solicitation strategies, or enhanced reviewer identification, could yield valuable insights. Addressing review polarization is crucial not only for preserving the integrity and informativeness of online ratings but also for ensuring fair and competitive market outcomes.

References

Amazon (2024). Customer reviews. Available at: https://www.amazon.com/gp/help/customer/display.html?nodeId=G8UYX7LALQC8V9KA. Accessed: 2025-03-14.

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*. Available at: https://epi.washington.edu/news/large-language-models-outshine-traditional-natural-language-processing-methods-for-identifying-rare-circumstances/. Accessed: 2024-01-01.

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*.

Bishop, T. (2015). Amazon changes its influential formula for calculating product ratings. *GeekWire*. Available at: https://www.geekwire.com/2015/amazon-changes-its-influential-formula-for-calculating-product-ratings/. Accessed: 2025-03-14.

Bondi, T., Rossi, M., and Stevens, R. L. (2024). The good, the bad and the picky: Consumer heterogeneity and the reversal of product ratings. *Management Science*.

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.

Brucks, M. and Toubia, O. (2025). Prompt architecture induces methodological artifacts in large language models. *PLoS One*, 20(4):e0319159.

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.

Donati, D. (2025). The end of tourist traps: The impact of review platforms on quality upgrading. *Marketing Science*.

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

Filippas, A., Horton, J. J., and Golden, J. M. (2022). Reputation inflation. *Marketing Science*, 41(4):733–745.

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.

Fradkin, A., Grewal, E., and Holtz, D. (2021). Reciprocity and unveiling in two-sided reputation systems: Evidence from an experiment on Airbnb. *Marketing Science*, 40(6):1013–1029.

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*.

He, S., Hollenbeck, B., Overgoor, G., Proserpio, D., and Tosyali, A. (2022). Detecting fake-review buyers using network structure: Direct evidence from amazon. *Proceedings of the National Academy of Sciences*, 119(47):e2211932119.

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.

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*.

Land, P. (2025). Google adds requirement for review comments and author names in rating systems. Available at: https://ppc.land/google-adds-requirement-for-review-comments-and-author-names-in-rating-systems/. Accessed: 2025-03-14.

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*.

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.

Pera, R., Viglia, G., Grazzini, L., and Dalli, D. (2019). When empathy prevents negative reviewing behavior. *Annals of Tourism Research*, 75:265–278.

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.

Roberts, D. (2013). Yelp's fake review problem. *Fortune*. Available at: https://fortune.com/2013/09/26/yelps-fake-review-problem/. Accessed: 2025-03-14.

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.

Search Engine Journal (2021). Google star ratings guide: How they work and how to get them. *Search Engine Journal*. Available at: https://www.searchenginejournal.com/google-star-ratings-guide/416951/. Accessed: 2025-03-14.

Simchon, A., Brady, W. J., and Van Bavel, J. J. (2022). Troll and divide: the language of online polarization. *PNAS nexus*, 1(1):pgac019.

Singhal, M., Pacheco, J., Moosavi Khorzooghi, S. M. S., Debi, T., Asudeh, A., Das, G., and Nilizadeh, S. (2023). Auditing yelp's business ranking and review recommendation through the lens of fairness. *arXiv preprint arXiv:2308.02129*. Accepted at ICWSM 2025.

Sweeney, E. (2018). Dunkin' drops 'donuts' from its name in new branding. *Marketing Dive*. Available at: https://www.marketingdive.com/news/dunkin-drops-donuts-from-its-name-in-new-branding/533243/. Accessed: 2025-03-14.

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.

Taylor, K. (2016). Starbucks is launching a new drink that could change how you order coffee forever. *Business Insider*. Available at: https://www.businessinsider.com/starbucks-launches-nitro-cold-brew-2016-5. Accessed: 2025-03-14.

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*.

Yelp (2024). Recommendation software. Available at: https://trust.yelp.com/recommendation-software/. Accessed: 2025-03-14.

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.

Appendix

A Language Use Over Time

To investigate the stability of language use norms in Yelp reviews across our sample period (2012-2022), we employ two complementary analytical approaches. First, we examine the distribution of review language embeddings for restaurants with consistent product and service quality. Our rationale is that if systematic language norm changes were occurring, we would expect to observe shifts in language features (as captured by embeddings) even for businesses maintaining stable quality standards over time. Second, we conduct a temporal analysis of the vocabulary used in review texts across different years. If language norms were evolving systematically, we would anticipate corresponding changes in word usage patterns. The two subsections below present the detailed methodologies and results of these two approaches. Our comprehensive analysis reveals that language use in Yelp reviews remained remarkably consistent throughout the 2012-2022 period.

A.1 Language Use in Coffee Shop Reviews

To examine restaurants with consistent product and service quality, we focus on chain restaurant categories that have accumulated a substantial number of reviews. For this analysis, we select the coffee chain category, specifically Starbucks and Dunkin' Donuts (which together account for over $8,300$ reviews). Although both chains implemented menu upgrades during our sample period, these changes were primarily incremental refinements rather than fundamental category transformations (Taylor, 2016; Sweeney, 2018).

We analyze the temporal distributions of review language embeddings by presenting yearly distributions for two sets of principal components: the top 10 components capturing the greatest variation in the LLM embeddings, and the top 10 components having the strongest influence on rating predictions. As discussed in Section 4, these embedding dis-

tributions may be affected by selection mechanisms as well as reviewer and business heterogeneity. To account for these confounding factors, we regress each principal component on all variables in equation (2) except for the year indicators, and incorporate business fixed effects to control for outlet-specific characteristics.

The distributions of the resulting residuals are displayed in Figures 17 and 18. (See Web Appendix B for Figure 18.) Notably, these distributions remain remarkably stable across years for all principal components illustrated, where the value ranges remain similar across all years and the distributions are very similar in shape. This suggests that the language use norms remained consistent for the coffee chain category over the sample period.

A.2 Analysis of Word Usage in Reviews

A second dimension of language stability is revealed through examining the vocabulary consumers employ when describing their experiences in Yelp reviews. To systematically assess temporal changes in word usage, we analyzed the evolution of review vocabulary across years. To reduce noise from fluctuating business names and product offernings (which naturally vary as different establishments are reviewed each year), we focused specifically on adjectives, which more directly capture sentiment and evaluation. For each pair of years in our dataset, we identified adjectives that appear exclusively in one year but not the other, then calculated the percentage of reviews in each year containing these unique words. This pairwise comparison methodology provides a more nuanced perspective on language shifts, effectively capturing natural variability in word choice across time periods. Table 7 presents these results comprehensively. Notably, the percentage of reviews containing year-specific adjectives is remarkably small, ranging between just 0.19% and 2.36%. This minimal lexical differentiation across years strongly reinforces our conclusion that the vocabulary used in Yelp reviews has maintained remarkable consistency throughout the entire study period.

B Alternative Prediction Models

We evaluated several alternative predictive models to understand the relationship between review text and ratings, including the multinomial logit (MNL) and a deep neural network (DNN). The ordered logit model (OLM) using 404 principal components demonstrated comparable predictive performance to these alternatives. Table 8 summarizes the

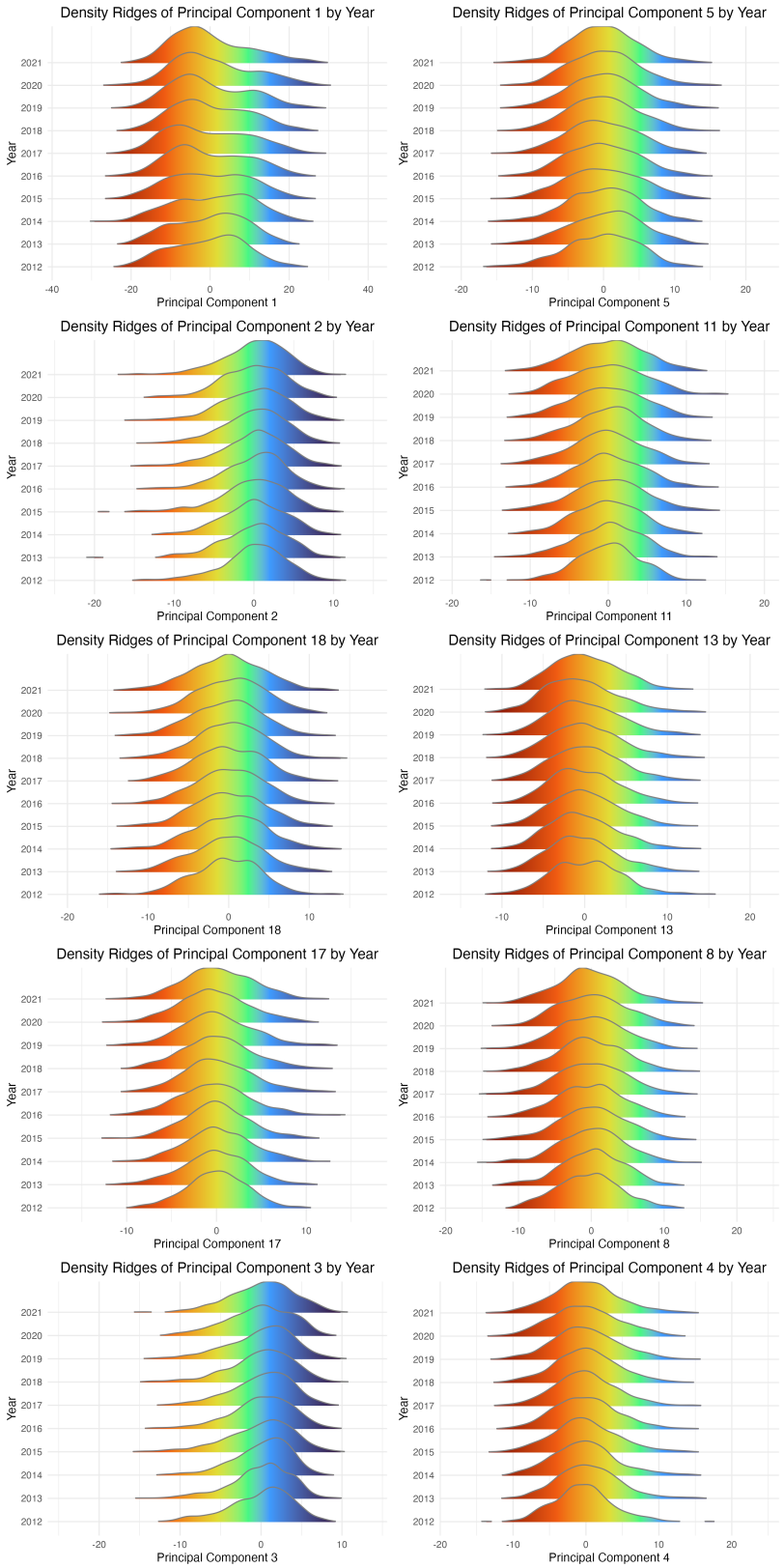


Figure 17: Distribution of Top 10 Most Impactful Principal Components in Rating Prediction for Coffee Shop Reviews, 2012-2022

Table 7: Percentage of Reviews with Different Adjective Words Between Year Pairs

Year 1	Year 2	% Year 1 Reviews with Unique Words	% Year 2 Reviews with Unique Words	Year 1	Year 2	% Year 1 Reviews with Unique Words	% Year 2 Reviews with Unique Words
2012	2013	0.55%	0.56%	2015	2016	0.26%	0.25%
2012	2014	0.55%	0.48%	2015	2017	0.27%	0.23%
2012	2015	0.47%	0.42%	2015	2018	0.26%	0.22%
2012	2016	0.48%	0.41%	2015	2019	0.30%	0.21%
2012	2017	0.51%	0.39%	2015	2020	0.61%	2.25%
2012	2018	0.48%	0.38%	2015	2021	0.51%	1.63%
2012	2019	0.53%	0.37%	2016	2017	0.25%	0.21%
2012	2020	1.08%	2.36%	2016	2018	0.26%	0.21%
2012	2021	0.90%	1.76%	2016	2019	0.27%	0.19%
2013	2014	0.45%	0.40%	2016	2020	0.53%	2.24%
2013	2015	0.38%	0.34%	2016	2021	0.45%	1.61%
2013	2016	0.39%	0.31%	2017	2018	0.22%	0.21%
2013	2017	0.41%	0.29%	2017	2019	0.24%	0.20%
2013	2018	0.40%	0.28%	2017	2020	0.48%	2.22%
2013	2019	0.45%	0.27%	2017	2021	0.41%	1.61%
2013	2020	0.95%	0.28%	2018	2019	0.23%	0.19%
2013	2021	0.79%	0.40%	2018	2020	0.44%	0.21%
2014	2015	0.31%	0.31%	2018	2021	0.38%	0.32%
2014	2016	0.31%	0.30%	2019	2020	0.42%	0.20%
2014	2017	0.32%	0.26%	2019	2021	0.34%	0.19%
2014	2018	0.31%	0.26%	2020	2021	0.28%	0.36%
2014	2019	0.35%	0.25%				
2014	2020	0.73%	0.25%				
2014	2021	0.62%	0.25%				

Note: The table presents the percentage of reviews containing adjective words not found in the other year for each year pair between 2012 and 2021. "Unique Words" refers to adjective words that appear in one year but not the other.

performance metrics across all models.

The results reveal that both the MNL and DNN models using the full set of 3,072 embeddings perform worse on all metrics compared to their counterparts using 404 principal components. This pattern likely reflects the challenges of high dimensionality, which increases the variance of model estimates. We observe a clear variance-bias trade-off: while full embedding models theoretically incorporate more information, they underperform in practice because the additional dimensions introduce noise rather than valuable signal for rating prediction. For the OLM models specifically, the version with full embeddings shows slightly lower accuracy but marginally higher recall and F1 scores, suggesting both approaches yield comparable predictive performance. These findings confirm that the 404 principal components effectively capture the essential information needed for accurate rating prediction.

When comparing across model types using principal components, the OLM achieves similar accuracy to MNL and DNN, though with recall and F1 scores approximately 0.025 to 0.035 lower. However, the OLM offers substantially greater interpretability than either alternative model. In particular, it is visually straightfoward to check the changes in the

rating thresholds over time. We therefore select the OLM for our analysis, accepting a modest reduction in predictive performance in exchange for significantly enhanced interpretability of results.

Table 8: Performance Comparison of Different Predictive Models, Philadelphia Market

Model	Accuracy	Precision	Recall	F1
DNN with Full Embeddings	0.698	0.644	0.647	0.647
DNN with Principal Components	0.720	0.675	0.680	0.680
MNL with Full Embeddings	0.688	0.611	0.611	0.611
MNL with Principal Components	0.717	0.663	0.670	0.670
OLM with Full Embeddings	0.706	0.660	0.663	0.663
OLM with Principal Components	0.707	0.639	0.643	0.643

Note: DNN refers to deep neural network, MNL refers to multinomial logit, and OLM refers to ordered logit model. For the DNN model, we use an architecture with 2 hidden layers with 128 and 32 units each. The full embeddings are the 3,072-dimensional embeddings of the text. Principal Components refers to the use of the 404 principal components identified through the PCA. Performance metrics are based on out-of-sample predictions using 5-fold cross-validation.

C Justification for Constant β Across Years

To test whether the relationship between review text and ratings changes over time, we estimated a model with full interactions between year indicators (a dummary variable for each year) and principal components included in the model. This specification, which includes 4,040 interaction terms, not only fails to improve prediction accuracy but actually performs worse than the more parsimonious model without interactions. The computational cost is substantial as well—estimating this interaction model took more than 8 hours on a high-performance computer with 128GB of memory, compared to less than 10 minutes for the model without interactions. Table 9 presents the comparative performance metrics for both models. The results clearly demonstrate that introducing the year-content interactions leads to decreased performance across all evaluation metrics, suggesting that the relationship between content and ratings remains largely stable over our study period.

Table 9: Performance Comparison of Models With Full Interactions and Without Any Interactions, Philadelphia Market

Model	Accuracy	Precision	Recall	F1
Model with full interactions (varying βs)	0.702	0.632	0.636	0.636
Model without year interactions (constant β)	0.707	0.639	0.643	0.643

Note: Performance metrics are based on out-of-sample predictions using 5-fold cross-validation.

D Regression Results for Thresholds

Table 10 presents parameter estimates for α in regression of the thresholds for the Philadelphia market, with thresholds between adjacent rating levels labeled in $n|n+1$ format.

Year Effects The first panel of Table 10 shows the 1|2 threshold rising over time, while higher thresholds, especially 4|5, decline—confirming the increasing polarization in ratings.

Cohort Effects Most cohort fixed effects are insignificant for lower thresholds but show significant negative effects for higher thresholds among recent cohorts. For 3|4, cohorts from 2011 onward have increasingly negative coefficients, while for 4|5, cohorts from 2012 onward show significant negative effects. This pattern confirms newer cohorts are more likely to give higher ratings, contributing to rating polarization.

Cohort-specific Trends The interaction terms between cohort and tenure reveal that for 2010-2016 cohorts, all thresholds decline with reviewer tenure, indicating reviewers become more positive with experience. For 2017-2020 cohorts, most coefficients for lower thresholds are significant and negative, suggesting experienced reviewers are less likely to give negative reviews. Overall, platform experience tends to reduce the likelihood of extremely negative ratings.

Effects of Reviewer and Business Characteristics Several clear patterns emerge: (1) Male reviewers have higher thresholds for giving low ratings but lower thresholds for high ratings compared to female reviewers, indicating greater polarization; (2) One-time reviewers show even stronger polarization, with much larger coefficients; (3) Later reviews for a business show decreased polarization, with lower thresholds for 2-4 star ratings and higher thresholds for 5 stars; (4) Older businesses receive more negative reviews across all thresholds; and (5) Weekend/holiday reviews show increased polarization.

Decomposing Contribution to Variation in α_i The final panel quantifies each factor's contribution to threshold variation, with year effects accounting for 48.3%, one-time reviewer 19.1%, cohort and tenure 11.5%, review sequence 8.7%, business age 6.9%, reviewer gender 5.3%, and weekend/holiday 0.32%. Table 11 presents these decompositions for all metro markets.

Table 10: Contributing Factors to Threshold Changes, Philadelphia

Variable	1|2	2|3	3|4	4|5
Intercept	-7.295 (0.165)	-4.257 (0.118)	-0.483 (0.068)	2.876 (0.076)
year2013	0.041 (0.036)	-0.041 (0.033)	0.042 (0.026)	-0.004 (0.020)
year2014	0.075 (0.045)	-0.116 (0.041)	-0.080 (0.033)	-0.184 (0.025)
year2015	0.092 (0.053)	-0.209 (0.048)	-0.159 (0.040)	-0.324 (0.030)
year2016	0.157 (0.061)	-0.187 (0.055)	-0.256 (0.045)	-0.504 (0.035)
year2017	0.183 (0.068)	-0.286 (0.061)	-0.322 (0.050)	-0.608 (0.039)
year2018	0.241 (0.073)	-0.354 (0.066)	-0.441 (0.054)	-0.782 (0.043)
year2019	0.251 (0.078)	-0.376 (0.070)	-0.509 (0.058)	-0.886 (0.046)
year2020	0.355 (0.083)	-0.390 (0.075)	-0.688 (0.063)	-1.111 (0.051)
year2021	0.332 (0.086)	-0.468 (0.078)	-0.769 (0.066)	-1.132 (0.053)
cohort2011	-0.081 (0.088)	-0.017 (0.079)	-0.198 (0.066)	0.032 (0.047)
cohort2012	0.064 (0.093)	0.047 (0.085)	-0.360 (0.072)	-0.200 (0.050)
cohort2013	-0.003 (0.108)	-0.119 (0.099)	-0.580 (0.085)	-0.394 (0.058)
cohort2014	0.134 (0.121)	-0.143 (0.110)	-0.693 (0.096)	-0.461 (0.065)
cohort2015	0.144 (0.130)	-0.136 (0.120)	-0.643 (0.103)	-0.523 (0.071)
cohort2016	0.127 (0.139)	-0.193 (0.128)	-0.784 (0.112)	-0.539 (0.076)
cohort2017	0.055 (0.147)	-0.256 (0.136)	-0.795 (0.119)	-0.631 (0.081)
cohort2018	0.098 (0.153)	-0.138 (0.141)	-0.744 (0.124)	-0.602 (0.084)
cohort2019	0.186 (0.158)	-0.151 (0.147)	-0.795 (0.131)	-0.709 (0.089)
cohort2020	0.121 (0.170)	-0.025 (0.164)	-0.669 (0.150)	-0.637 (0.103)
cohort2021	-0.049 (0.173)	-0.179 (0.167)	-0.626 (0.157)	-0.814 (0.110)
cohort2010×tenure	-0.168 (0.067)	-0.159 (0.060)	-0.313 (0.052)	-0.188 (0.036)
cohort2011×tenure	-0.081 (0.054)	-0.162 (0.050)	-0.247 (0.042)	-0.260 (0.030)
cohort2012×tenure	-0.168 (0.042)	-0.224 (0.039)	-0.199 (0.034)	-0.168 (0.023)
cohort2013×tenure	-0.104 (0.037)	-0.152 (0.035)	-0.101 (0.031)	-0.079 (0.021)
cohort2014×tenure	-0.180 (0.033)	-0.128 (0.033)	-0.033 (0.030)	-0.069 (0.020)
cohort2015×tenure	-0.151 (0.032)	-0.126 (0.033)	-0.086 (0.031)	-0.046 (0.020)
cohort2016×tenure	-0.151 (0.034)	-0.106 (0.036)	-0.012 (0.036)	-0.085 (0.023)
cohort2017×tenure	-0.086 (0.041)	-0.087 (0.046)	-0.028 (0.046)	-0.032 (0.029)
cohort2018×tenure	-0.118 (0.050)	-0.175 (0.057)	-0.040 (0.060)	0.033 (0.037)
cohort2019×tenure	-0.158 (0.069)	0.011 (0.080)	0.010 (0.085)	-0.100 (0.056)
cohort2020×tenure	-0.269 (0.141)	-0.432 (0.170)	-0.061 (0.185)	-0.141 (0.125)
cohort2021×tenure	0.396 (0.320)	0.585 (0.396)	0.193 (0.440)	1.185 (0.296)
male	0.178 (0.012)	0.198 (0.012)	-0.064 (0.010)	-0.131 (0.008)
gender unknown	0.195 (0.018)	0.161 (0.018)	-0.041 (0.017)	-0.098 (0.012)
one time reviewer	0.523 (0.017)	0.453 (0.018)	-0.260 (0.017)	-0.559 (0.015)
review sequence	-0.156 (0.006)	-0.122 (0.005)	-0.032 (0.004)	0.079 (0.003)
business age	0.041 (0.002)	0.034 (0.002)	0.025 (0.002)	0.004 (0.001)
weekend or holiday	0.054 (0.010)	0.041 (0.011)	-0.032 (0.010)	-0.049 (0.007)
Decomposition of α_i				
Year	48.27%			
Cohort	3.91%			
Tenure	7.59%			
Gender	5.25%			
One-time reviewer	19.10%			
Weekend/Holiday	0.32%			
Review sequence	8.67%			
Business age	6.89%			

*Note:*Each cell shows Estimate (Std. Error). Gender female is the baseline category. All values rounded to three decimals.

Table 11: Decomposition of $\boldsymbol{\alpha}$ across Metropolitan Areas

Metro	year	cohort	tenure	gender	one time reviewer	weekend/holiday	review sequence	business age
Boise	41.37%	2.63%	14.26%	6.46%	13.98%	0.27%	14.92%	6.11%
Edmonton	34.13%	4.95%	19.22%	8.41%	19.47%	0.86%	8.16%	4.80%
Indianapolis	46.29%	3.15%	7.63%	7.17%	15.26%	0.45%	13.66%	6.39%
Nashville	43.80%	0.90%	10.33%	3.77%	18.61%	0.34%	15.80%	6.45%
New Orleans	45.63%	1.58%	9.61%	2.40%	24.96%	0.26%	11.67%	3.88%
Philadelphia	48.27%	3.91%	7.59%	5.25%	19.10%	0.32%	8.67%	6.89%
Reno	41.91%	6.57%	13.95%	4.16%	16.89%	0.52%	6.34%	9.67%
Saint Louis	45.09%	1.46%	12.69%	2.96%	13.79%	0.17%	17.07%	6.78%
Santa Barbara	39.99%	2.41%	16.93%	4.88%	12.25%	0.26%	12.03%	11.23%
Tampa	41.95%	4.40%	13.69%	4.52%	16.77%	0.30%	13.75%	4.62%
Tucson	47.55%	5.20%	9.65%	3.97%	13.12%	0.10%	13.12%	7.30%

Notes: The decomposition illustrates the contributions of different factors to the overall variation in α_i. It considers the correlation between these factors. The decomposition is based on ANOVA.

E Regression Results for Content Score

We analyze how review content varies based on the factors in equation (2). We use the same regression framework but with the content score r_i^* as the dependent variable and include business fixed effects. This approach reveals how content varies with temporal trends and reviewer characteristics, with results presented in Table 12.

Year coefficients show a significant positive trend, indicating review content has become more positive over time when controlling for other factors. Conversely, cohort coefficients exhibit a significant negative trend, with more recent cohorts writing more negative content than earlier ones. Cohort-specific tenure effects show interesting patterns: cohorts from 2010-2018 demonstrate negative and significant trends, writing increasingly negative content with platform experience, while 2019-2021 cohorts display positive trends, becoming more positive with experience. Gender effects reveal male reviewers produce more negative content than females, with reviewers of unidentified gender being the most negative. One-time reviewers write more negative content than repeat reviewers, and weekend/holiday reviews contain more negative content, consistent with findings by Bayerl et al. (2023). For a given business, review sequence has a negative and significant coefficient, showing that later reviews tend to contain more negative content, aligning with previous research by Godes and Silva (2012).

Table 12: Content Score Regression, Philadelphia

Year and Cohort Effects				Demographics and Interactions			
Variable	Estimate	Std. Error	P-Value	Variable	Estimate	Std. Error	P-Value
Intercept	-0.183	0.035	0.000	cohort2010	-0.436	0.068	0.000
year2013	0.098	0.029	0.001	male	-0.187	0.009	0.000
year2014	0.271	0.036	0.000	gender unknown	-0.633	0.016	0.000
year2015	0.329	0.042	0.000	one time reviewer	-0.402	0.018	0.000
year2016	0.302	0.049	0.000	review sequence	-0.301	0.014	0.000
year2017	0.347	0.054	0.000	weekend or holiday	-0.253	0.009	0.000
year2018	0.333	0.058	0.000				
year2019	0.353	0.062	0.000				
year2020	0.472	0.067	0.000				
year2021	0.290	0.071	0.000	cohort2010×tenure	-0.437	0.047	0.000
cohort2011	-0.438	0.078	0.000	cohort2011×tenure	-0.451	0.038	0.000
cohort2012	-0.522	0.083	0.000	cohort2012×tenure	-0.441	0.030	0.000
cohort2013	-0.707	0.096	0.000	cohort2013×tenure	-0.355	0.027	0.000
cohort2014	-0.648	0.108	0.000	cohort2014×tenure	-0.282	0.024	0.000
cohort2015	-0.637	0.117	0.000	cohort2015×tenure	-0.252	0.023	0.000
cohort2016	-0.731	0.124	0.000	cohort2016×tenure	-0.262	0.025	0.000
cohort2017	-0.676	0.132	0.000	cohort2017×tenure	-0.298	0.030	0.000
cohort2018	-0.658	0.111	0.000	cohort2018×tenure	-0.117	0.048	0.015
cohort2019	-0.870	0.116	0.000	cohort2019×tenure	0.092	0.069	0.183
cohort2020	-1.075	0.131	0.000	cohort2020×tenure	0.197	0.146	0.178
cohort2021	-1.545	0.137	0.000	cohort2021×tenure	1.782	0.345	0.000

Note: The table shows how review content changes with year, cohort, and reviewer characteristics. Business fixed effects are included. All values rounded to three decimals.

Web Appendices

These materials have been supplied by the authors to aid in the understanding of their paper. The AMA is sharing these materials at the request of the authors.

A Web Appendix A: Specification and Estimates of σ

Since the scale parameter σ_{it} must be positive, we specify it as an exponential function of the same factors included in equation (2):

$$\sigma_{it} = \exp\Bigg(\sum_t \theta_t YEAR_{t(i)} + \sum_c \theta_c COHORT_{ci} + \sum_c \theta_r COHORT_{ci} \times TENURE_{ci} + \theta_g GENDER_i + \theta_1 ONCE_i + \theta_o ORDER_i + \theta_a BIZ_AGE_i + \theta_w WKND_i\Bigg), \tag{8}$$

where all variables maintain the same definitions as in equation (2). Table 13 presents the coefficient estimates for the scale parameter (σ_{it}) in the Philadelphia market, while Table 14 summarizes the distribution of these estimated values across reviews.

Most parameter estimates are relatively small in magnitude, suggesting limited influence on the scale of rating variation. However, we observe a notable increasing trend across user cohorts, indicating that reviews from more recent cohorts exhibit greater noise in the relationship between text content and ratings. This suggests that newer Yelp users' ratings are less predictable from their review text compared to users from earlier cohorts. In addition, the estimates for the tenure of reviewers are mostly negative, indicating that reviewers are more likely to provide more robust ratings as they gain more experience on the platform. This pattern remains consistent across all metropolitan areas examined in our study, though those detailed results are omitted for brevity.

Table 13: Parameter Estimates for σ, Philadelphia

Year and Cohort Effects				Demographics and Interactions			
Variable	Estimate	Std. Error	P-Value	Variable	Estimate	Std. Error	P-Value
Intercept	-0.102	0.018	0.000	cohort2010	0.001	0.018	0.956
year2013	-0.022	0.008	0.006	male	0.027	0.003	0.000
year2014	-0.015	0.010	0.134	gender unknown	0.053	0.004	0.000
year2015	-0.024	0.012	0.046	one time reviewer	0.008	0.004	0.046
year2016	-0.031	0.013	0.017	review sequence	0.001	0.000	0.000
year2017	-0.050	0.015	0.001	business age	0.010	0.000	0.000
year2018	-0.049	0.016	0.002	weekend or holiday	-0.001	0.003	0.740
year2019	-0.039	0.017	0.022				
year2020	-0.034	0.019	0.073				
year2021	-0.049	0.019	0.010	cohort2010×tenure	0.043	0.015	0.004
cohort2011	0.044	0.020	0.028	cohort2011×tenure	0.022	0.012	0.067
cohort2012	0.079	0.021	0.000	cohort2012×tenure	0.000	0.010	1.000
cohort2013	0.099	0.025	0.000	cohort2013×tenure	-0.016	0.009	0.076
cohort2014	0.091	0.027	0.001	cohort2014×tenure	-0.004	0.008	0.617
cohort2015	0.108	0.030	0.000	cohort2015×tenure	-0.015	0.008	0.061
cohort2016	0.142	0.032	0.000	cohort2016×tenure	-0.044	0.009	0.000
cohort2017	0.153	0.034	0.000	cohort2017×tenure	-0.061	0.011	0.000
cohort2018	0.135	0.035	0.000	cohort2018×tenure	-0.053	0.014	0.000
cohort2019	0.150	0.036	0.000	cohort2019×tenure	-0.045	0.020	0.024
cohort2020	0.148	0.040	0.000	cohort2020×tenure	-0.056	0.042	0.182
cohort2021	0.153	0.041	0.000	cohort2021×tenure	-0.017	0.097	0.861

Note: Gender female is the baseline category. All values rounded to three decimals.

Table 14: Summary Statistics of σ_{it}

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.922	1.022	1.054	1.055	1.085	1.297

B Web Appendix B: Additional Tables and Figures

We present additional tables and figures that are not included in the main document. Below is a list of the tables and figures that are referred to in the main text.

- Figure 18: Distribution of Top 10 Principal Components of Content Embeddings for Coffee Shop Reviews, 2012-2022. This figure shows the distribution of the top 10 principal components of content embeddings for coffee shop reviews across years.
- Figure 19: Yelp Review Counts Over Time. This justifies the use of year 2012 to 2021 in our analysis.
- Figure 20: Scree Plot of Principal Component Analysis. This figure shows over 50% of the variance in the content embeddings is captured by the first 80 principal components, while 404 principal components capture over 80% of the variance.
- Figure 21: Prompt for AI-Generated Content-Based Score. The prompt used to generate the content-based score in our managerial implications.
- Table 15: Primary Category Review Counts. The chosen restaurant category contributes 67.5% of total reviews in the Philadelphia market.
- Table 16: Recall and F1 Score for Content-Based Score, showing additional model performance metrics.
- Table 17: Hit Rate for Philadelphia Market, 2012-2021 (Oversampled).
- Table 18: LIWC Categories. The LIWC categories used in the paper.
- Table 19: Decomposition of Scale and Content Effects for All Metro Markets.
- Table 20: Relation between Principal Components and LIWC Categories. The table demonstrates the most important LIWC categories explaining the variations in each of the top 10 principal components.

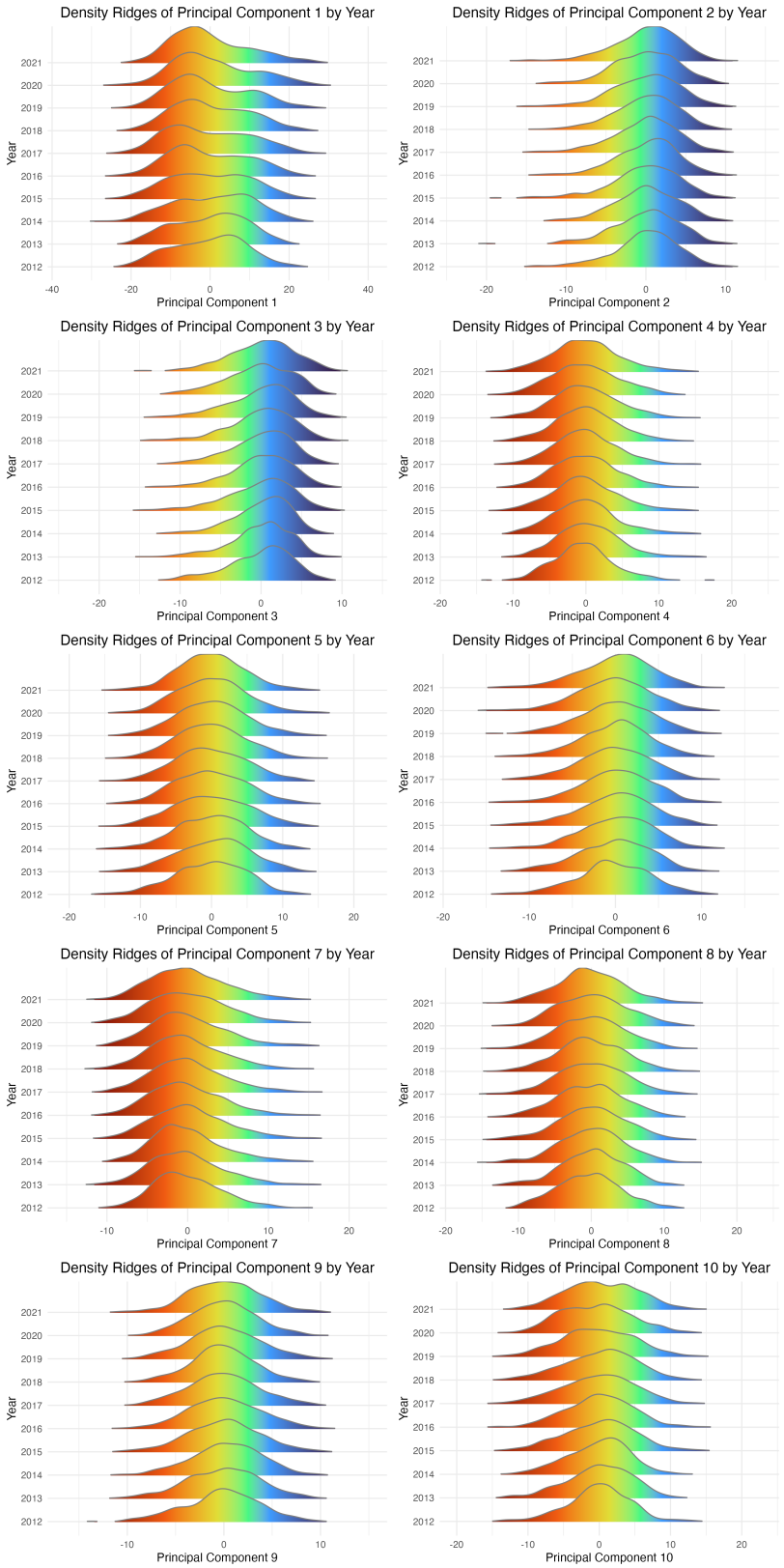


Figure 18: Distribution of Top 10 Principal Components of Content Embeddings for Coffee Shop Reviews, 2012-2022

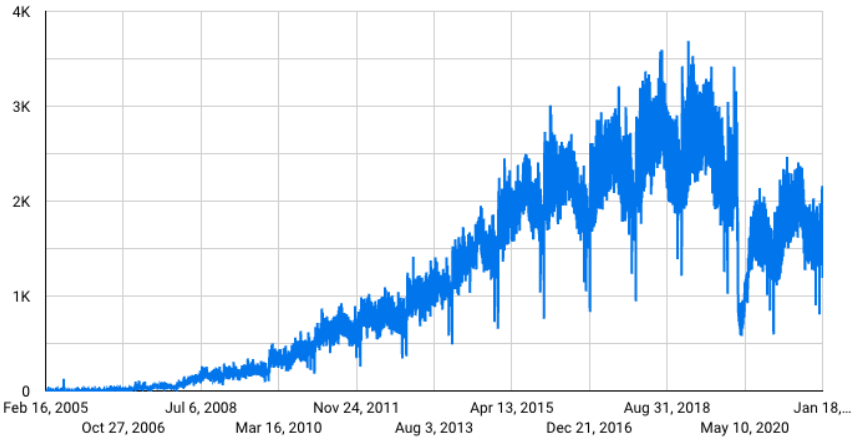


Figure 19: Yelp Review Counts Over Time

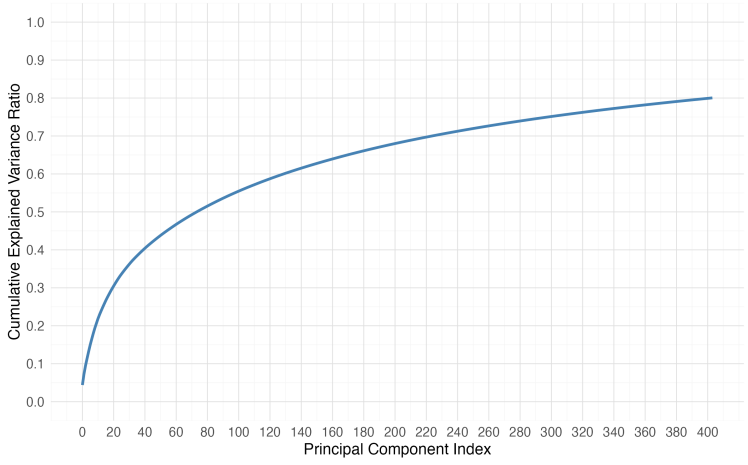


Figure 20: Scree Plot of Principal Component Analysis

Table 15: Summary of Review Counts by Yelp Primary Category

Primary Category	Total Reviews	Percentage of Reviews
Restaurants	4,722,190	67.58%
Food	401,658	5.75%
Beauty Spas	352,998	5.05%
Shopping	300,487	4.3%
Automotive	224,893	3.22%
Hotels Travel	183,597	2.63%
Arts Entertainment	156,320	2.24%
Home Services	156,056	2.23%
Active Life	131,098	1.88%
Health Medical	99,009	1.42%
Nightlife	72,274	1.03%
Pets	72,148	1.03%
Local Services	54,683	0.78%
Event Planning Services	15,494	0.22%
Financial Services	14,318	0.2%
Public Services Government	8,966	0.13%
Education	8,343	0.12%
Professional Services	5,001	0.07%
Local Flavor	4,098	0.06%
Religious Organizations	1,801	0.03%
Mass Media	1,480	0.02%
Other	689	0.01%

Note: The table presents the number and percentage of reviews across different primary categories of businesses on Yelp.

Table 16: LLM Model Performance Metrics Across Metropolitan Areas

Metro	Weighted Recall	Weighted F1 Score	Recall	F1 Score
Boise	0.730	0.720	0.629	0.635
Edmonton	0.685	0.682	0.653	0.656
Indianapolis	0.713	0.706	0.635	0.640
Nashville	0.729	0.721	0.637	0.642
New Orleans	0.722	0.714	0.630	0.636
Philadelphia	0.707	0.700	0.639	0.643
Reno	0.731	0.721	0.629	0.636
Saint Louis	0.720	0.713	0.638	0.643
Santa Barbara	0.743	0.734	0.628	0.635
Tampa	0.733	0.724	0.630	0.636
Tucson	0.714	0.706	0.629	0.634
Average	0.717	0.709	0.634	0.640

Note: Weighted Recall and weighted F1 scores are weighted by the number of observations in each star rating category. The scores are calculated based on out-of-sample predictions, where the LLM model is trained on 80% random samples and evaluated on the remaining 20%. Recall and F1 scores are unweighted.

Table 17: Confusion Matrix for Model Predictions of Philadelphia Reviews (Frequency-weighted Sampling)

	Predicted Star Rating				
Actual Stars	**1**	**2**	**3**	**4**	**5**
1	0.759	0.220	0.019	0.001	0.000
2	0.236	0.584	0.175	0.005	0.000
3	0.020	0.208	0.596	0.168	0.009
4	0.001	0.006	0.133	0.609	0.252
5	0.000	0.001	0.010	0.230	0.759

Note: Each row shows how reviews with a given actual star rating are distributed among the model's predicted ratings. Diagonal entries represent correct predictions for each rating. The model is trained on a balanced sample of reviews across all rating categories by oversampling the minority rating categories to ensure that each rating category has an equal number of observations.

Table 18: Linguistic Inquiry and Word Count (LIWC) Categories

Category	Description	Example Words
WC	Total word count	
WPS	Average words per sentence	
Analytic	Analytical thinking, metric of logical, formal thinking	because, therefore, evaluate, recommend
Clout	Language of leadership, status	definitely, always, lead, control
Authentic	Perceived honesty, genuineness	I, smell, feel so, actually
Tone	Degree of positive (negative) tone	good, well, bad, wrong
BigWords	Percent words 7 letters or longer	appetizer, unfriendly, expensive, ambience
Linguistic	Linguistic dimensions, including pronouns, negations, quantities etc.	I, you, my, me, not, no, never, all, one
Drives	Motivational and goal-directed language	we, help, work, better, own, allow, power
Cognition	Words reflecting mental activity and cognitive processing	all, but, not, if, because, make, really, of course
Affect	Words expressing emotion or emotional tone	good, new, happy, hate, tired, mad
Social	Words related to social relationships and interactions	we, care, thank, please, honor, tell, kill
Culture	Words related to social institutions, customs, education, cultural values	united states, govern, french, chinese, car, email
Lifestyle	Words related to people's activities, responsibilities, well-being and behavioral context	game, fun, play, work, home, pay, price, church
Physical	Words related to body, physical states, health, substances	medic, health, sick, gym, beer, food, drink, death
Perception	Words indicating sensory awareness	look, see, eye, feel, sound
Conversation	Words related to spoken discourse markers and informal speech	um, oh, lol, ok, sooo, youknow

Note: These LIWC categories are used to analyze the linguistic content of Yelp reviews. LIWC (Linguistic Inquiry and Word Count) is a text analysis program that categorizes words into psychologically meaningful categories.

Table 19: Scale and Content Effects by Metro

	1 Star Review				5 Star Review			
	Proportion	Scale Effect		Content	Proportion	Scale Effect		Content
	Change	Year Trend	Heterogeneity		Change	Year Trend	Heterogeneity	
Boise	56.7%	65.2%	-33.0%	67.8%	63.5%	41.3%	21.5%	37.2%
Edmonton	62.7%	49.7%	4.6%	45.8%	94.5%	44.8%	13.3%	41.9%
Indianapolis	65.7%	53.5%	-30.7%	77.2%	75.2%	66.6%	-3.5%	36.9%
Nashville	93.1%	41.1%	-17.1%	76.0%	78.9%	51.2%	8.8%	40.0%
New Orleans	60.6%	36.8%	-19.6%	82.8%	68.2%	55.8%	7.4%	36.8%
Philadelphia	82.4%	18.1%	1.6%	80.4%	69.2%	48.6%	16.6%	34.8%
Reno	66.7%	25.0%	-2.6%	77.6%	63.9%	50.4%	11.0%	38.6%
Saint Louis	100.3%	37.5%	-14.7%	77.2%	74.8%	55.2%	7.4%	37.4%
Santa Barbara	41.5%	68.2%	-29.0%	60.8%	65.1%	56.2%	6.8%	37.0%
Tampa	78.4%	42.3%	-24.7%	82.5%	68.9%	54.2%	8.5%	37.3%
Tucson	82.4%	21.6%	-5.1%	83.5%	60.5%	67.2%	1.9%	30.9%

Note: The table shows the decomposition of changes in 1-star and 5-star review proportions into scale effects (driven by year and reviewer heterogeneity) and content effects across different metropolitan areas. For 1-star reviews, content effects dominate in most markets (ranging from 60.8% to 83.5%), while for 5-star reviews, scale effects are more prominent (ranging from 58.1% to 69.1%). The heterogeneity in scale effects across markets suggests that local factors may influence rating behavior.

Table 20: Relationship Between Top Principal Components and LIWC Scores

	Principal Component									
	1	5	2	11	18	13	17	8	3	4
Avg Impact Weight[a]	4.3479	0.3499	0.2930	0.2762	0.2645	0.2825	0.2428	0.2455	0.2155	0.1992
ANOVA[b]										
ln(WC)	**0.0761**	**0.0501**	**0.0324**	**0.1922**	0.0091	**0.0426**	0.0011	0.0007	0.0021	0.0077
ln(WPS)	0.0116	0.0000	0.0102	0.0013	0.0000	0.0019	0.0010	0.0018	0.0018	0.0012
Analytic	0.0123	0.0148	0.0066	0.0097	0.0086	0.0014	0.0032	0.0030	0.0083	0.0000
Clout	0.0079	**0.0733**	**0.0546**	0.0010	**0.0410**	0.0042	**0.0081**	0.0005	0.0064	**0.0504**
Authentic	0.0077	0.0057	0.0098	**0.0131**	0.0095	0.0039	0.0000	0.0015	**0.0308**	0.0190
Tone	**0.3603**	0.0058	0.0246	0.0032	0.0111	0.0148	0.0002	**0.0248**	0.0000	0.0004
BigWords	0.0003	0.0195	0.0188	0.0002	0.0012	0.0006	0.0032	0.0034	0.0002	0.0000
Linguistic	0.0058	0.0048	0.0004	0.0008	0.0061	0.0000	0.0012	0.0055	0.0010	0.0001
Drives	0.0024	0.0258	0.0142	0.0000	0.0000	0.0000	0.0000	0.0082	0.0064	0.0075
Cognition	0.0042	0.0207	0.0004	0.0000	0.0020	0.0011	0.0029	0.0005	0.0004	0.0007
Affect	0.0010	0.0016	0.0053	0.0031	0.0083	0.0016	0.0021	0.0000	0.0092	0.0077
Social	0.0079	0.0238	0.0020	0.0011	0.0004	0.0012	**0.0051**	0.0000	0.0194	0.0240
Culture	0.0008	0.0000	0.0130	0.0000	0.0000	**0.0278**	0.0000	0.0007	0.0179	0.0001
Lifestyle	0.0087	0.0000	0.0029	0.0003	0.0109	0.0001	0.0038	0.0001	0.0152	0.0122
Physical	0.0002	0.0000	0.0054	0.0009	**0.0146**	0.0008	0.0038	**0.0102**	**0.0287**	**0.0325**
Perception	0.0001	0.0004	0.0001	0.0003	0.0004	0.0239	0.0001	0.0042	0.0064	0.0001
Conversation	0.0000	0.0001	0.0002	0.0000	0.0027	0.0000	0.0011	0.0016	0.0000	0.0004
R-Squared[c]	0.507	0.246	0.201	0.227	0.126	0.126	0.037	0.067	0.154	0.164

[a] Avg Impact Weight is the average of the impact weights across all metropolitan markets.
[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 bolded values represent the highest proportion for each principal component.
[c] R-Squared measures the proportion of the variance of each principal component explained by all the LIWC scores in linear regressions.

```
You are an AI assistant tasked with analyzing and summarizing a set of customer reviews for a restaurant, similar to those
found on Yelp.

**Instructions:**

1. **Estimate Review Ratings:**
   For each review, estimate an expected rating score from 1 to 5, based solely on the content and language of the review
(ignore any explicit or external star ratings). Please consider the nature of online review distribution where numerical
ratings could be polarized, and pay attention to reviews that use strongly positive or negative language, and do not
"average out" extreme opinions.

2. **Report Score Distribution:**
   Count and report the number of reviews you rated as positive (5), neutral/mixed (3–4), and negative (1–2).

3. **Identify Common Themes:**
   List the most frequently mentioned aspects of the restaurant experience (e.g., food quality, service, ambiance, price,
specific menu items, cleanliness, etc.). For each theme, briefly explain its relevance and provide an assessment of the
overall sentiment expressed in the reviews regarding that theme.

4. **Assign a Numeric Sentiment Score:**
   Based on your analysis of all review texts, assign an overall numeric sentiment score to the restaurant from 1.0 (very
negative) to 5.0 (very positive), using one decimal point. This score should reflect your internal scoring and the overall
tone of the reviews.

5. **Generate a Summary Statement:**
   Write a concise, factual summary (around 5 sentences) that captures the overall impression of the restaurant based on
*all* the reviews. The summary should be helpful to potential customers to quickly understand the overall experience from
all these reviews, reflect the main themes and sentiment distribution, and be consistent with your calculated scores.
Avoid generic statements; instead, highlight specific strengths, weaknesses, or polarizing aspects as revealed by the
reviews. Do not explicitly mention the rating scores in the summary.

   **When reporting the summary and overall sentiment, ensure that the diversity and polarization of opinions are
accurately represented.**

   Please also present a summary of the common themes in one line and use checkmark (good), warning sign (mixed), and
cross mark (bad) to indicate the sentiment of the theme. Separate by space.

**Output Format:**

Your response must be in valid JSON format as follows:

```json
{
  "review_score_distribution": {
    "positive_count": X,
    "neutral_mixed_count": Y,
    "negative_count": Z,
    "average_rating": X.X,
    "standard_deviation": X.X
  },
  "common_themes": [
    {
      "theme": "Theme 1",
      "explanation": "Brief explanation of theme",
      "sentiment": "positive/mixed/negative",
      "symbol": "✓/⚠/✗"
    },
    {
      "theme": "Theme 2",
      "explanation": "Brief explanation of theme",
      "sentiment": "positive/mixed/negative",
      "symbol": "✓/⚠/✗"
    }
  ],
  "overall_sentiment": {
    "score": X.X,
    "standard_deviation": X.X
  },
  "summary_statement": "Concise factual summary helpful for potential customers...",
  "theme_summary": "theme 1: ✓ theme 2: ⚠ theme 3: ✗ ..."
}
```

IMPORTANT:
- Do NOT use any explicit or external star ratings—score each review based solely on its content.
- Your calculated average and summary should reflect your internal scoring and the actual content of the reviews, not
outside information.
- Ensure your summary is consistent with the sentiment distribution and highlights any polarization or consensus present
in the reviews.
- Your response MUST be a valid JSON object as shown in the format above.
- Use ✓ for positive themes, ⚠ for mixed themes, and ✗ for negative themes.
```

Figure 21: Prompt for AI-Generated Content-Based Score