university of groningen / faculty of economics and business

From Words to Action: Modeling the Cognitive and Behavior Paths from Review Features to Consumer Engagement

MSc Thesis for Marketing Analytics and Data Science

Faculty of Economics and Business

University of Groningen

Name: Yichong Tao

Student Number: S5065623

Submission Date: 16.06.2025

1st Supervisor: Dr. Qiong Tang

2nd Supervisor: Dr. Alec Minnema

Abstract

This study examined how review content features and reviewer characteristics influenced other customers subsequent engagement behavior, focusing on the role of review usefulness and valence. The study employed a two-stage empirical strategy. In Study 1, structural equation modeling was applied to a subset of 5,000 Yelp reviews to assess how argument quality and source credibility affect review useful vote. Study 2 used negative binomial regression on the full 170,615 dataset to validate behavioral pathways. Findings indicate that valence has a strong positive effect on future engagement, whereas the effect of review usefulness requires control variables to become positive. In addition, there was a significant negative interaction between valence and usefulness, suggesting that useful negative reviews are more likely to trigger further user engagement. These results demonstrated the joint influence of cognitive and affective factors on consumer behavior and provide practical implications for restaurant review management.

Keywords: Review Usefulness, Valence, Consumer Engagemen, eWOM, Online Reviews

Tables

Abbreviations

The following abbreviations are used throughout the thesis:

Abbreviation	Full Term
RU	Review Usefulness
EB	Engagement Behavior
AQ	Argument Quality
SC	Source Credibility
NBR	Negative Binomial Regression
SEM	Structural Equation Modeling
CFA	Confirmatory Factor Analysis
IAM	Information Adoption Model
TAM	Technology Acceptance Model
ELM	Elaboration Likelihood Model

1 Introduction

As the digital economy flourishes, consumers increasingly rely on online reviews to obtain first-hand information about goods or services. Online reviews have not only become an information intermediary between companies and consumers, but also have largely shaped the cognitive structure and purchase path of consumers. As a typical form of electronic word-of-mouth (eWOM), online reviews have been shown to have a significant impact on consumer decision-making due to their extensiveness, spontaneity, and information transparency (Chevalier & Mayzlin, 2006; Dellarocas et al., 2007). According to Dimensional research (2013), online reviews have been shown to evoke impressions of advertisements before 60-70% of consumers make a purchase, and 90% of these consumers admitted that their purchasing behavior was influenced by these impressions. Yelp, as the world's leading business review platform, aggregates a large amount of real review data, which is an important data field for studying consumer information adoption behaviors and behavioral responses.

Many studies have shown that the quality of the review content (e.g., clarity, comprehensiveness, relevance) and the credibility of the reviewer (e.g., professionalism, history of behavior) significantly influence the subjective assessment of review usefulness by consumers (Cheung et al., 2008; Bhattacherjee & Sanford, 2006; Mudambi & Schuff, 2010). This cognitive process is highly consistent with the mechanism proposed by the Information Adoption Model (IAM), which suggests that when individuals are exposed to information, they make judgments based on argument quality and source credibility, and then decide whether to to adopt the information (Sussman & Siegal, 2003). The Elaboration Likelihood Model (ELM) also suggests that the quality of information plays a key role in high-involvement contexts, that is, where the

perception of the impact of the decision on the outcome is high, while in low-involvement contexts, consumers tend to rely on peripheral cues for heuristic processing (Petty & Cacioppo, 1986).

For the existing studies, I think there are two main research gaps that exist. The first one is the lack of effective path analysis from cognition to behavior. Existing research has primarily examined how different elements in reviews (e.g., quality, credibility) (Cheung et al., 2008; Bhattacherjee & Sanford, 2006; Kim et al., 2018; Srivastava & Kalro, 2019) affect perceived usefulness—a cognitive evaluation stage. However, much less attention has been paid to how this cognitive evaluation translates into actual consumer behaviors, such as visiting the business, writing follow-up reviews, or engaging further with the platform. Meanwhile, most of the existing studies rely on self-reported behavioral data through surveys and experimental methods, which primarily capture consumers' attitudes or stated intentions rather than actual behaviors (Cheung et al., 2008; Bhattacherjee & Sanford, 2006; Kim et al., 2018; Liu et al., 2007; Park et al., 2007; Park & Lee, 2008). These methods often fail to trace the real behavioral outcomes that follow from exposure to online reviews, because self-reported data are often subject to biases such as social desirability (Fisher, 1993), recall errors (Coughlin, 1990), and the intention-behavior gap (Sheeran, 2002). In this paper, I extend the IAM model, in the traditional IAM model, information adoption intention or behavior is only controlled by information usefulness. In this paper, I link the perceived usefulness with Valence, and introduce the influence of interaction terms on subsequent behavior, in order to reveal more realistically the cognitive processing and behavioral response mechanism of consumers in the eWOM environment.

In addition to this, I believe that another gap in this area of research is the inconsistency of face validation due to the neglect of the inclusion of valence, that is, people usually intuitively perceive negative useful information as having a negative effect on subsequent behavior, whereas

the IAM model, which emphasizes perceived usefulness as a key cognitive driver of behavior, does not incorporate the interaction between valence and usefulness. Most of the previous studies have used valence as a direct predictor of consumer behavior, while only partially considering the interaction between valence and usefulness. For example, some studies have shown that positive valence reviews increase purchase intentions (Chevalier & Mayzlin, 2006; Jia & Liu, 2018; Coursaris et al., 2018). Another study admits an interaction between valence and perceptions, but employs a complex structure, splitting usefulness into dual perceptions of usefulness and risk, which may diminish the clarity of the explanation for the moderating effect (Xiao & Li, 2019). A study has even proposed the opposite mechanism, considering utility as a moderator rather than a predictor (Park & Lee, 2008). In contrast, this study adopts a reverse moderating structure, considering perceived usefulness as a moderator of the effect of valence on EB, thus providing a more precise explanation of consumers' responses in emotional eWOM situations.

Based on the above research background and gap identification, this paper intends to construct an extended version of the IAM model to explore how review features in eWOM jointly affect EB through Review Usefulness (RU) and Valence. The three main objectives of the study are: first, to examine whether the argument quality (AQ) and source credibility (SC) of a review significantly affects RU. Second, examine whether Valence directly affects EB. Third, to explore whether Valence and RU significantly interact to influence EB.

This study is based on the Yelp open-source dataset, which includes business-level information, user-generated reviews, and reviewer attributes. This study focuses on the restaurant industry in Yelp and analyzes online review data for restaurant-based businesses. As one of the most active and richly reviewed industries on Yelp, restaurants are highly representative of consumer engagement and behavior (Zhang & Luo, 2023). Restaurant reviews tend to contain

more sensory language and emotional expressions, which can stimulate more effective consumer attention and interactive responses, and thus have significant research value in eWOM behavioral studies (Li et al., 2019; Ariyasriwatana & Quiroga, 2016).

A two-stage empirical design is implemented. Structural Equation Modeling (SEM) was chosen as the empirical analysis method for this study, because of its ability to simultaneously perform factor analysis and path analysis on the model within a single framework. In this study, a CFA will be used to assess whether the selected observational variables effectively reflect the underlying constructs to ensure the reliability and validity of the latent variables such as AQ and SC. For the path analysis component, SEM can not only estimate the directional relationship between variables, including the relationship between latent and observed variables and even between different observed variables, but also examine the moderating effect. A comprehensive examination can be made regarding the path of information adoption from information properties to the cognitive evaluation of information recipients, as well as the cognitive behavioral path of the cognitive moderating effect on behavioral outcomes.

Since the outcome variable in this study (EB) is a count variable (number of new reviews in next quarter), applying SEM directly to such a dependent variable may lead to unstable or biased estimation results. Therefore, I introduced a second-stage analysis to address this limitation. Specifically, Study 2 employed Negative Binomial Regression (NBR). This model is particularly appropriate for count data that exhibit hyperdispersion (where the variance exceeds the mean), which is a common feature of online behavioral data. The model allows me to robustly test the interaction of RU and sentimental value on actual EB, while appropriately addressing the distributional characteristics of the dependent variable.

By integrating SEM and NBR, this two-stage design provides a robust and complementary strategy for testing complex theoretical models-especially in the eWOM environment, where the interaction of information quality and sentiment plays a critical role in driving consumer behavior.

This study has two main contributions. First, in terms of the measurement of latent variables, this study extracts multidimensional linguistic features from user reviews based on Natural Language Processing (NLP) technology to measure AQ and SC, and systematically verifies the validity of the dimensions through SEM. This measure breaks through the limitation of traditional eWOM research that relies on questionnaires to obtain attitude and intention data. The results revealed that some important dimensions traditionally considered to measure AQ, such as clarity, relevance and comprehensiveness of information, lacked validity, reflecting the limitations of traditional studies in selecting measurement dimensions, and providing empirical experience for future modeling of consumer research based on textual data. Second, in terms of research design and theory expansion, this study integrates SEM and NBR to construct a two-stage empirical study. This paper not only confirms the positive effects of AQ and SC on RU, but also finds that negative reviews with high RU are more likely to motivate subsequent engegement behaviors such as writing new review. By constructing the interaction term of RU × valence, this study extends the explanatory power of the IAM on the behavioral level, and reveals the synergistic mechanism of cognition and emotion in the formation of consumer behavior.

2 Research Background and Hypotheses

In the section, I will introduce the IAM, the core theory of this study, and explore the definitions of key concepts such as AQ, SC, RU, Valence, the EB, and the definitions of key concepts and their implications for information processing.

2.1 Theoretical Background

2.1.1 Information Adoption Model

The IAM was first proposed by Sussman and Siegal in 2003 to explain how individuals evaluate and adopt information in computer-mediated communication environments. The theoretical foundation of the IAM combines two classic models: the Technology Acceptance Model (TAM) and the ELM. The theoretical foundation of IAM combines two classic models of TAM and ELM.

TAM, which was proposed by Davis (1986), is based on Ajzen & Fishbein's Theory of Reasoned Action (TRA), which is used to explain individual acceptance of information technology and emphasizes the influence of perceived usefulness and perceived ease of use on intention and behavior, while ELM is a model used in social psychology to explain the acceptance of information technology, which suggests that individuals may process information in persuasive situations through either the central route (focusing on information quality) or the peripheral route (focusing on illuminating cues), thus affecting their attitudes toward information, and is often used to study perceived usefulness of information in eWOM research (Petty & Cacioppo, 1986).

IAM is developed based on these two models, absorbing the idea of "dual-path processing" in ELM, replacing Central Route with argument quality and Peripheral Route with source credibility, and incorporating the emphasis of TAM on perceived usefulness, proposing that

individuals judge whether information is worth adopting based on perceived usefulness as a mediator. The main structure of the model is the influence of argument quality and Source Credibility, which together affect Perceived Usefulness and then Adoption Intention.

While TAM emphasizes information adoption and behavior, and ELM emphasizes attitude change, IAM is a more comprehensive model that incorporates the cognitive and behavioral mechanisms in the information dissemination environment, and is used to explain how users make adoption decisions in the face of a large amount of complex information. It has been believed that the IAM is applicable to a variety of domains such as social media, electronic word-of-mouth (eWOM), and online reviews (Cheung et al., 2008).

In previous studies, Park et al. (2007) examined the acceptance of a document management system by employees in an Eastern European governmental organization based on IAM, and found that argument quality and source credibility significantly influenced users' perceived usefulness, which in turn predicted their intentions to use the system. In consumer behavior research, Bhattacherjee and Sanford (2006) used the IAM to explore the influence of online reviews on purchase intentions, further validating the explanatory power of the model in digital contexts.

2.1.2 Argument Quality

Argument quality is often defined as the combination of information value, logical structure, clarity, and topical relevance of the review itself, and is regarded as a core latent variable representing central path processing in both the IAM and ELM theories (Petty & Cacioppo, 1986). It has been widely found that high argument quality leads to stronger persuasive effects and increased perceptions of review usefulness, which in turn promote information adoption (Bhattacherjee & Sanford, 2006; Cheung et al., 2008). However, due to the unstructured and

affective nature of user-generated text, argument quality cannot be directly observed but is instead modeled as a latent construct measured through multiple observable indicators extracted from review content.

Different scholars have proposed various dimensions to capture this variable. For example, Srivastava and Kalro (2019) proposed four dimensions of descriptiveness, product focus, review structure, and emotional intensity to capture the implicit connotations of argument quality. Kim et al. (2018), on the other hand, emphasized the role of clarity, concreteness and subjectivity on the perception of usefulness. Liu et al. (2007) even suggested that information comprehensiveness and emotional objectiveness can be used to predict the persuasive efficacy of reviews in terms of syntactic features. In addition to these, word count has been widely adopted in prior ELM-based studies as a proxy for argument richness and informational depth (Kim et al., 2018; Liu & Park, 2015; Cheng & Ho, 2015), longer reviews are considered to include more detailed and contributes more to argument quality.

2.1.3 Source Credibility

Source credibility is the subjective trust and expertise perception of the reviewer by consumers, which is a latent variable like argument quality, and is a typical peripheral path variable in ELM models (Chaiken, 1980). In the eWOM context, consumers usually judge whether a reviewer is trustworthy based on indirect cues such as the reviewer's identity information, historical behavior, and interactive performance. Traditionally, this variable is often composed of two dimensions, expertise and trustworthiness (Petty et al., 1981; Bhattacherjee & Sanford, 2006). And some studies have attempted to incorporate more social cues to capture its more complex psychological underpinnings. For example, Srivastava and Kalro (2019) proposed that review

consistency, sincerity of expression, and experience of engagement are the three major observable variables of source credibility. López-López and Parra (2016) emphasize that consumers judge the adoptability of a reviewer based on whether he or she belongs to the mainstream. Mudambi and Schuff (2010), on the other hand, point out that labels assigned by platforms (e.g., "Top reviewer" or "Elite") can be seen as community proxies signaling credibility.

2.1.4 Valence and Its Positive Effect on Engagement Behavior

Valence refers to the emotional orientation of a review, indicating whether the message is positive, negative, or neutral. It is one of the most prominent variables in eWOM literature, influencing consumer attitudes and behavioral responses. In some IAM or ELM-based studies, valence is considered as part of the observables of argument quality, alongside clarity, relevance and structure. However, this study conceptually separates emotional positivity and negativity from the quality of a review, considering that a review can be usefulness good regardless of whether it expresses a positive or negative point of view. Thus, valence is defined as an emotional attribute independent of review quality.

From a psychological perspective, valence is associated with the well-documented "negativity bias," which suggests that consumers respond more strongly to negative than to positive information (Baumeister et al., 2001). In marketing research, negative reviews have been shown to reduce consumer trust, lower product evaluations, and decrease purchase intentions. Nonetheless, recent studies indicate that the effects of valence are context-dependent, and may be moderated by review quality, consumer involvement, and individual expectations. The interpretation of emotional tone is often shaped by subjective and situational factors, making valence a highly dynamic construct in eWOM environments.

Valence has also been extensively studied as a predictor of downstream consumer behaviors, particularly engagement behaviors such as visiting, reviewing, or interacting with a business. Prior research has confirmed that review valence significantly influences consumer behavioral responses. For example, Luca (2011) found that each one-star increase in a restaurant's average Yelp rating was associated with a 5–9% increase in revenue. Similarly, Anderson and Magruder (2012) reported that higher Yelp scores led to a greater likelihood of restaurant bookings. These findings reflect the broader notion that positively valenced reviews enhance behavioral engagement by signaling satisfaction and reducing perceived risk. Given that this relationship has been well established, the current study incorporates the positive effect of valence on engagement behavior as a baseline condition, rather than as a hypothesis to be re-tested.

2.1.5 Review Usefulness

Review usefulness refers to the extent to which consumers perceive the information in a review as helpful for their decision-making. It is a key mediating variable in IAM and TAM, representing users' assessment of the functional value of information. A review perceived as useful typically provides valuable insights, reduces uncertainty, or enhances the consumer's understanding of the product. Based on the dual-path theory of the ELM, review usefulness is influenced both by argument quality along the central route and source credibility along the peripheral route.

Prior research shows that review usefulness is not a static property, and its effect on decision-making may vary under different contextual boundaries. For example, usefulness may operate through different mechanisms under positive and negative valence. Xiao and Li (2019) proposed that usefulness has a dual role: enhancing confidence in decision-making or amplifying

risk perception. In online review platforms like Yelp, this construct is often operationalized through the number of useful votes a review receives, for example in the study of Mudambi and Schuff (2010), reflecting how well the information is accepted by the community. These studies have demonstrated that higher review quality or reviewer influence tends to yield higher perceived usefulness. In this study, I not only examine the main effect of review usefulness but also focus on its interaction with valence to address the emerging trend of dual-path models that integrate cognition and emotion in eWOM research.

2.1.6 Engagement Behavior

Engagement behavior, as the outcome variable in this study, captures the consumer's behavioral response after being exposed to review information. Unlike traditional studies that use behavioral intention or purchase intention, I adopted actual observed behaviors from the platform. Specifically, I used the number of newly generated reviews per quarter as a proxy for consumer engagement with businesses. This shift from intention to observable behavior strengthens the model's external validity and enhances causal inference.

This behavioral measure was inspired by Chevalier and Mayzlin (2006), who found a significant relationship between review activity and sales, showing that reviews can drive measurable consumer responses. Therefore, engagement behavior serveed as a quantifiable indicator of review adoption and user activation, reflecting whether a consumer is motivated to interact, review, or revisit a business after reading existing reviews.

2.1.7 Conceptual model Development

This study is also based on IAM, first investigates how argument quality and source credibility affect RU, and then moderately extends the model on this basis. To adapt the data structure of the review data from Yelp, I used the number of new reviews added per quarter (EB) as a proxy for user behavioral interaction, thus capturing the impact in actual platform interactions.

In addition, Valence has been shown to significantly predict consumer behavior such as purchase intention or sales performance in eWOM contexts (Chevalier & Mayzlin, 2006; Jia & Liu, 2018; Coursaris et al., 2018), while the direct influence of Valence on consumer behavior has already been widely established in prior research, this study still incorporates this relationship into the model structure as a theoretical baseline, without formulating a new hypothesis for it.

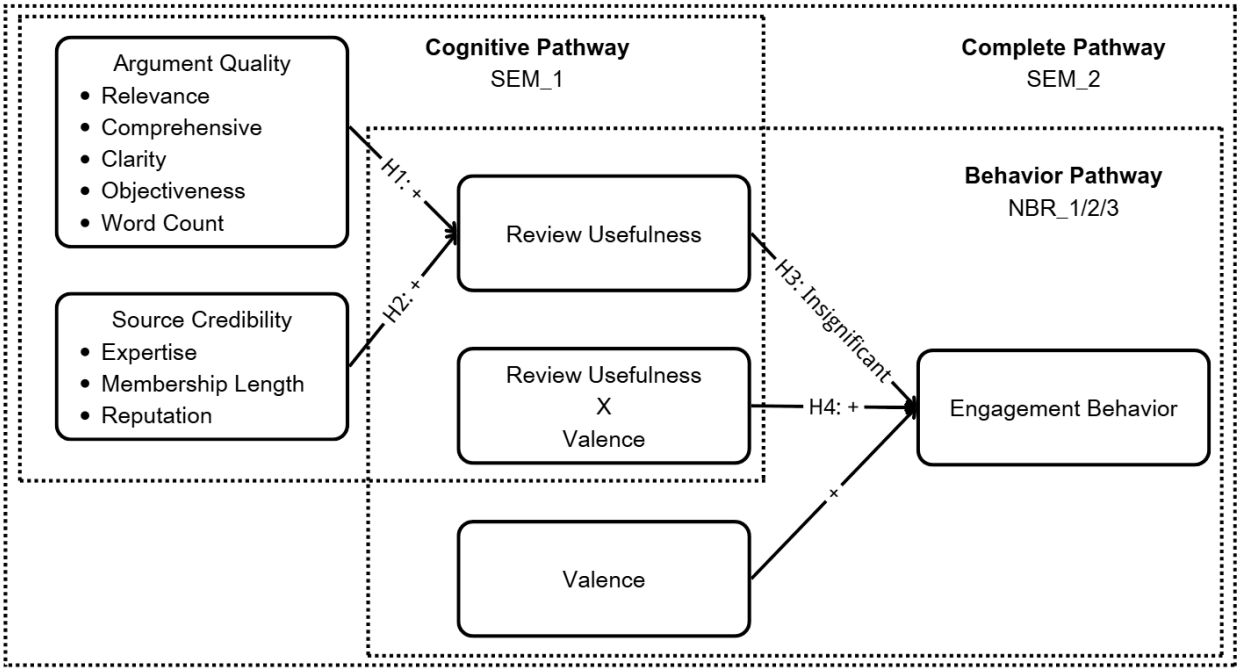
However, the question of how perceived usefulness interacts with valence to jointly influence consumer behavior has not been fully examined. Some studies have provided initial insights into this interaction, but there are limitations. For example, Coursaris et al. (2018) explored the influence of valence on purchase intentions under conditions of high perceived usefulness, but did not cover the case of low usefulness. Jia and Liu (2018) found that usefulness strengthens the impact of positive reviews, no effect on the negative reviews, but they also only used useful information for their study. Xiao and Li (2019) attempted to explain the valence impacts by splitting usefulness into usefulness and perceived risk, but this decomposition may weaken the interpretability of the interaction. Park and Lee (2008) proposed an inverse model that treats valence as a moderator of perceived usefulness, and tested it only in the context of positive reviews, without directly modeling the interaction.

Given these gaps, the present study focuses on the moderating role of perceived usefulness, how it moderates the established relationship between valence and EB. This structure better

reflects how consumers integrate emotional and cognitive cues when responding to online reviews, and offers a refined extension of the IAM model in emotionally complex eWOM environments.

Therefore, instead of continuing the traditional path of direct prediction of behavior by perceived usefulness, this study introduces the interaction between valence and usefulness as the core research object, and further hypothesizes that valence itself influences behavior, while usefulness plays a moderating role in it. This setting not only responds to the recent research trend of cognition behavior interaction mechanism, but also more closely matches the logic of consumers' decisions in the real context of online reviews, thus providing theoretical basis and empirical support for the further expansion of the IAM model in the user eWOM environment.

Fig 1. Conceptual Model



2.2 Hypotheses

2.2.1 Argument Quality influence on Review Usefulness

Based on the integration of the above literature, this paper constructs an indicator of observational variables of argumentation quality from five dimensions: relevance, which assesses whether a review is focused on the product or service theme; comprehensiveness, which reflects the extent to which multiple aspects of the experience are covered; clarity, which indicates the coherence and readability of the language; objectiveness, which evaluates the extent to which the content avoids emotionally charged or highly subjective expressions that could undermine rational judgment; and word count, while a longer content of review will be considerd to include more details. These dimensions are not only relevant to consumer expression, but are also highly quantifiable in natural language processing, making them suitable operationalizations for measuring AQ in this context.

H1: Argument quality of a review has a positive effect on review usefulness.

2.2.2 Source Credibility influence on Review Usefulness

In this study, I consider the theoretical structure and empirical observability, and select three observational variables as the main dimensions of SC: expertise refers to whether the reviewer demonstrates in-depth knowledge and experience of the product, and the label given by the platform is used as an indicator of whether the reviewer's professionalism and engagement are recognized by the platform; reputation reflects whether the reviewer's previous reviews are considered valuable by other platform users, and the number of useful votes received by the reviewer is used as an indicator; membership length is used as an indicator of whether the

reviewer's previous reviews have been recognized by other platform users. The platform recognizes the reviewer's professionalism and participation. Membership length, as an alternative of behavioral history metric, represents the length of time reviewers have been active on the platform and their historical trajectory of engagement, and is a reflection of sustained engagement. Together, these three factors represent the SC accumulated by reviewers in the process of long-term interaction, which is closer to the psychological evaluation path of consumers in the actual decision-making process than the explicit labels.

H2: Source credibility of a reviewer has a positive effect on review usefulness.

2.2.3 Interaction effect of Valance and Review Usefulness on Engagement Behavior

While perceived usefulness is often positioned as a direct driver of information adoption in the IAM framework, this study challenges this assumption by suggesting that usefulness itself may not directly lead to EB. For example, a negative review may be perceived as useful but still trigger avoidance rather than engagement. In contrast, positive reviews will only motivate EB if they are perceived as credible and informative. In this sense, the role of usefulness is more of a filter than a behavioral predictor, it determines whether valance of review will be taken seriously and acted upon. Based on this logic, I argue that perceived usefulness does not work in isolation but moderates the effect of emotion on consumer engagement.

H3: Review usefulness does not directly affect engagement behavior.

H4: Review usefulness moderates the effect of valence on engagement behavior.

3 Methodology

This chapter introduces the data sources, variable measurements, and analytical methods employed in this study. First, the data selection strategy is explained, followed by a detailed description of the data origin and preprocessing procedures. Then, the operational definitions and measurement logic for each variable are outlined, and finally, the statistical techniques used to test the hypotheses are presented.

This study uses the open-source review dataset provided by Yelp as the primary empirical basis. Yelp is one of the most widely used, structurally complete, and highly interactive review platforms globally. In 2021, Yelp received approximately 800 million monthly visits and ranked 9th among the most visited websites globally (RankRanger, 2021). Furthermore, Yelp reviews contain diverse data types, including ratings, textual content, timestamps, usefulness votes, and reviewer attributes, enabling analysis of review timeliness and its impact on usefulness. In the US market, Yelp also has a high level of market penetration, and its reviews exert real influence on consumer decisions, making it a suitable environment to study information adoption behavior.

This research further focuses on the restaurant sector within Yelp, for three main reasons. First, consumer decision-making in the food service industry relies heavily on online reviews, since dining experiences are categorized as experience goods, meaning their quality cannot be verified prior to consumption and thus depends more on peer evaluations. Second, restaurant reviews make up the largest portion of Yelp content. From 2017 to 2021, restaurants accounted for 40.7 percent of all business listings in the US Yelp dataset, providing a robust and diverse data foundation for modeling. Third, prior studies have shown that sensory cues embedded in restaurant reviews—such as descriptions of taste, smell, appearance, and texture—are key triggers of consumer EBs. These descriptions enhance the vividness and appeal of reviews, increase reader

resonance and attention, and ultimately affect consumer interaction and rating behavior. Therefore, the restaurant industry offers both reliability and interpretability for behavioral modeling.

3.1 Data source

The datasets used in this study are from three parts of Yelp open source, the reviews dataset, the business dataset, and the user dataset, with a geographic scope of the United States and a timeframe that covers data from January 1, 2017, to December 31, 2021. The original review dataset contains 3,806,440 reviews across all business categories. I filtered the dataset to retain only the restaurant data, and the processed data contained a total of 44,218 restaurants in the U.S. and 2,538,886 reviews left by 953,124 Yelp users. Next, I removed reviews that lacked complete user metadata, such as elite status, length of membership, or behavioral history. Additionally, I removed reviews from three U.S. states that had only one restaurant review to avoid extreme scarcity and geographic imbalance.

For the business data, I had access to the address, location, average stars, and total number of reviews, while for the review dataset I have the time of review, text data, and useful votes, and for the user data set I have the number of useful votes by individual level, the time of elite, length of registration, and the total number of reviews left.

Finally, because the dependent variable measures the number of new reviews generated in the three months following each focal review, I excluded all reviews from the last three months of the dataset (October through December 2021) because future review activity outside of the dataset timeframe could not be observed. After all screening steps, the final analyzed sample consisted of 170,615 restaurant reviews.

3.2 Measures

This study constructs and measures five core variables: AQ, SC, RU, valence, and EB. Each variable is defined and extracted using both theoretical references and natural language processing techniques to align with the structure of Yelp data. In addition, to control the influence of other variables on the dependent variable and to improve the predictive accuracy of the model, this study introduces three variables as control variables, including the total number of reviews received by each restaurant, the state where the restaurant is located, and the year and quarter of the reviews. These three variables control the variation of the dependent variable in three dimensions, including the popularity of the restaurant itself, the degree of development of the city, and the variation in the time dimension, which strengthens the explanatory power of the model.

The empirical analysis is divided into two stages. In Study 1, AQ and SC are conceptualized as latent independent variables, each operationalized through several observable indicators, and together predict consumers perceived RU. This stage mainly uses SEM to examine how AQ and SC jointly influence RU in online review settings.

In Study 2, valence is treated as an established predictor of EB, and the analysis focuses on whether RU moderates this relationship. In other words, NBR model tests whether the impact of valence on consumer engagement varies depending on the level of RU. Moderation analysis is applied to assess the interaction effects.

3.2.1 Argument Quality

To measure AQ, this study follows the four-dimensional framework proposed by Srivastava and Kalro (2019), which includes descriptiveness, product focus, structural integrity, and emotional intensity, and incorporates elements from Kim et al. (2018), who emphasize clarity,

concreteness, and subjectivity, as well as from Liu et al. (2007), who assess coverage and polarity using syntactic features. Based on these references, AQ is operationalized using four indicators: relevance, comprehensiveness, clarity, and objectiveness.

Relevance was measured based on the distribution of themes extracted from the Latent Dirichlet Allocation (LDA). The main keywords associated with each topic are illustrated in Fig 2. A weighted summation method was used, and this study assigned weights to each topic based on its semantic closeness to core aspects of the restaurant experience. Topics directly describing food quality, price, or service were assigned the highest weight (1.0) and had the highest relevance. Topics related to the menu, the ordering process, or to drinks were assigned a medium weight (0.8), reflecting partial relevance. Lower weights (0.6) were given to topics about subjective emotional expressions and locations, as they contributed limited information content to the reviews. The specific weights assigned to each topic are presented in

table 1. This approach allows for a more nuanced assessment, e.g., a review that includes only words such as “very good” will only be of general interest to other consumers, whereas a more comprehensive review such as “steak is very good” will receive a higher score, besides avoiding the binary categorization that would have directly deleted certain topics from contributing to the relevance score.

Fig 2. Top Words of Each Topics

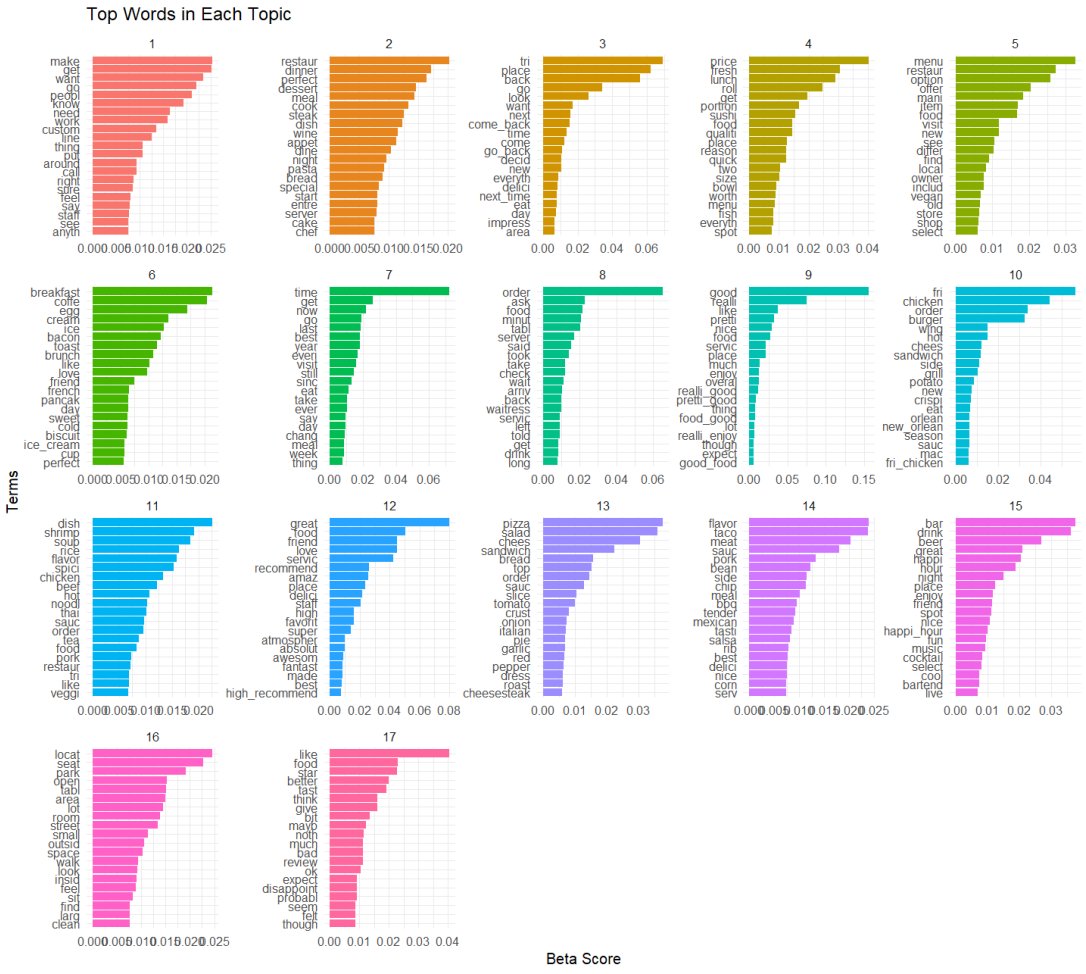


table 1. Weight of Topics for Relevance Calculation

Topic Label	Main Keywords	Weight	Rationale
Tp1_Recommendation_and_Service	recommend, service, make, friend	1	Directly related to core customer experience features such as recommendation and service.
Tp2_Dinner_and_Steak	dinner, meat, steak, restaurant	1	Describes main dishes; clearly aligned with food quality.

Topic Label	Main Keywords	Weight	Rationale
Tp3_Return_and_Impression	place, back, come, impression	0.8	Reflects revisit intention and overall impression, indirectly relevant.
Tp4_Price_and_Quality	price, portion, quality, fresh	1	Clearly refers to price and quality assessment.
Tp5_Menu_and_Restaurant	menu, item, restaurant, option	0.8	Discusses menu structure, partially static in nature.
Tp6_Breakfast_and_Bacon	breakfast, bacon, sausage, egg	1	Clearly refers to breakfast items and food categories.
Tp7_Timing_and_Arrival	time, early, day, now	0.6	Focuses on arrival timing, peripheral to dining experience.
Tp8_Order_and_Waffle	order, get, waffle, wait	0.8	Related to ordering process, partially includes operational context.
Tp9_Positive_Emotion	love, good, amazing	0.6	Generalized praise lacking detailed or concrete information.
Tp10_Burgers_and_Fries	burger, fries, sandwich, chicken	1	Specific food categories are clearly mentioned.
Tp11_Soups_and_Asian_Food	noodle, soup, Asian, dish	1	Specific dish types, relevant to content.
Tp12_Great_Atmosphere_and_Staff	great, atmosphere, friendly, staff	1	Specific reviews on ambiance and staff, core to experience.
Tp13_Pizza_and_Cheesesteak	pizza, salad, cheesesteak	1	Identifies specific food items; clearly structured.

Topic Label	Main Keywords	Weight	Rationale
Tp14_Mexican_Flavors	flavor, taco, Mexican	1	Regional cuisine references, strongly tied to food experience.
Tp15_Bar_and_Drinks	bar, drink, dinner	0.8	Refers to drink experience, relevant but not central to core meals.
Tp16_Local_and_place	local, clear, food	0.6	Mentions locality in a vague manner, lacks specificity.
Tp17_Subjective_Feelings	feel, like, want	0.6	Highly subjective impressions with limited substantive content.

Comprehensiveness is the number of different topics discussed in a review. The topic was present if the probability of each topic exceeded a threshold of 8% in each review. The total number of these topics is then summed to reflect the breadth of information in the review. This binary aggregation method captures the diversity of the content of a review, independent of the weighting of each topic.

Clarity was measured using a dictionary-based spelling error detection method. The hunspell package in R was used to calculate spelling errors as a percentage of the total number of words, and the inverse of the error rate was used as the clarity score. This metric reflects how linguistically clean and readable a review is, if clearer reviews contain fewer spelling errors and are easier to cognitively process.

Objectiveness is measured using sentiment scores derived from the pre-trained model nlptown/bert-base-multilingual-uncased-sentiment, which is a deep-learning model that allows a better understanding of the context, scoring the sentiment of reviews in English using Bert has a

67% probability of being the same as the manual scoring, and a 95% probability of having an error of 1 or less (Hugging Face, n.d.). The model assigns a sentiment score to each review, ranging from 1 (very negative) to 5 (very positive). In this study, a score of 3 (neutral) is considered as maximum objectiveness and extreme values (1 or 5) are considered as minimum objectiveness. I used the following formula to convert the sentiment score to a continuous objectiveness score: 1 - (abs(score - 3) / 2). To ensure compliance with CFA, which requires continuous observed variables, I introduced a small amount of random noise using the jitter() function. This allows for smoother estimation during CFA while retaining the explanatory logic of the objectiveness measure.

Counting words is a common metric for observing the quality of an argument. In this study the original (untokenized) text of the reviews was used to count the number of word units separated by whitespace without removing stop words or punctuation. This method preserves the full narrative structure of the review and is consistent with the way readers perceive length in natural reading contexts.

3.2.2 Source Credibility

SC, as a key peripheral variable, reflects consumers' perceptions of a reviewer's trustworthiness and expertise. Drawing on the dual-dimensional structure from Petty et al. (1981) and Bhattacherjee and Sanford (2006), and incorporating the engagement-based approach from Srivastava and Kalro, this study uses three observable indicators: expertise, reputation, and membership length.

Expertise is measured using the Yelp Elite badge. Yelp awards this badge to users who consistently contribute high-quality content and engage in the community. The total number of

elite years is used to represent expertise level, with a higher number indicating greater recognized expertise.

Reputation is measured by the total number of useful votes a reviewer has received historically, indicating how much their contributions are valued by others and representing their social capital on the platform.

Membership length is calculated as the time span (in months) between the reviewer's registration and the date of the review. This metric reflects the reviewer's historical activity and sustained presence, which contributes to perceived credibility.

3.2.3 Review Usefulness

RU serves as the key mediating variable in the information adoption process. It refers to the extent to which consumers perceive review content as valuable and helpful for decision-making. On the Yelp platform, this variable is operationalized using the number of useful votes each review receives, which reflects both its observability and user-based assessment.

This study not only examines the main effect of RU but also focuses on its interaction with valence to explore how cognitive and emotional components jointly influence consumer behavior. RU, as a measure of perceived information value, acts as a central bridge between review features and user EB. In addition to functioning as a mediator, it is also treated as a moderator to test whether the perceived usefulness of a review strengthens or weakens the behavioral effects of emotional valence.

3.2.4 Valence

Valence refers to the overall emotional orientation expressed in a review, categorized as positive, neutral, or negative. In this study, valence is constructed as an independent cognitive input variable to capture consumers' emotional information processing.

For measurement, I employ the same pre-trained sentiment analysis model used for sentiment extremity: nlptown/bert-base-multilingual-uncased-sentiment. This model, based on the BERT architecture, supports multilingual emotion classification and has been validated across various eWOM research contexts. Each review is mapped to an integer score from 1 to 5, with higher values indicating more positive sentiment. I then classify the scores into three valence categories: reviews rated 1 or 2 are coded as negative, those rated 3 as neutral, and those rated 4 or 5 as positive. This classification yields a categorical variable representing the overall emotional direction of each review.

3.2.5 Engagement Behavior

EB is used as the outcome variable in this study, capturing the actual behavioral response of consumers after being exposed to review content. Compared to traditional measures like intention to adopt or intention to purchase, this study adopts the number of new reviews posted per quarter as a proxy for consumer EB, indicating whether the consumer is motivated to further interact, express, or respond.

This behavioral indicator reflects the real-world impact of reviews on consumer behavior and enhances both the causal inference and external validity of the proposed model. Inspired by the measurement strategies of Zhang et al. and Chevalier and Mayzlin, I define EB at the business-quarter level, using each restaurant as the unit of analysis. At each time point t+1, the number of

new reviews is recorded as the dependent variable, while aggregated features from the previous period t (e.g., average RU, valence distribution) are used as predictors.

3.2.6 Control Variables

To account for potential disturbing factors in the modeling process, this study included three control variables: year_quarter, state, and review count (total for each restaurant). Introducing these variables into the model helps to control the effects of seasonality, degree of development in the region, and store popularity on consumer EB. Thus, the explanatory power of the effects of the core variables on the dependent variable are isolated more accurately.

First, year_quarter dummies are introduced to control temporal heterogeneity, including seasonality and macroeconomic changes. Previous research has shown that the visibility and perceived usefulness of reviews may vary depending on the time of posting, especially during high-traffic periods (Zhang et al., 2014). By incorporating quarterly controls, this study ensures that temporal fluctuations in platform activity or consumer behavior do not bias the results.

Second, state-level dummy variables were included to control for regional differences in the degree of regional development, review culture, and Yelp market penetration. While previous studies have not always directly included geographic differences, Zhang et al. (2015) emphasized the importance of accounting for contextual differences across platforms and regions in eWOM studies. Thus, controlling for states allows the model to capture potential cultural and structural heterogeneity across the United States.

Third, the total number of reviews for each business was controlled to adjust for the overall exposure or popularity of each restaurant. Businesses that accumulate a higher number of reviews may benefit from a visibility advantage, making users more likely to engage regardless of review

content. Pan and Zhang (2011) show that the number of reviews significantly affects the perceived usefulness of a review, reinforcing the need to control for this effect when analyzing review-level determinants.

3.3 Model

To empirically examine how review features affect consumers EB, this paper adopted a two-stage empirical study combining SEM and NBR. In Study 1, SEM was used to estimate cognitive and behavioral pathways, and in Study 2, NBR was used to re-examine the behavioral pathways in order to address the distributional characteristics of the dependent variable.

To ensure the accuracy of the measurement and structural estimation, Study 1 followed a three-step SEM process to progressively build and evaluate the theoretical model. The first step was to validate the latent variable structure using CFA, the second step was to model the cognitive paths of the RUs using SEM_1, and the third step was to estimate the EBs using SEM_2 with the introduction of control variables. This process is supported by Anderson and Gerbing (1988) who advocate separating the measurement and structural modeling steps because adding structural paths changes the factor loadings and also affects the path parameter estimates generating bias. A similar suggestion was given by Kline (2023) that separating measurements from structural assessment could improve the interpretability of the model. Therefore in this paper, I tested the robustness of the parameter estimates by incrementally adding to the model in three steps.

However, SEM has some known limitations that affect its applicability to EB analysis. First, SEM assumes that continuous variables are normally distributed (Kline, 2023), which conflicts with the highly skewed and highly dispersed nature of the original RU and EB variables. Therefore, in Study 1, a dummy encoding was applied to the RU and a logarithmic transformation

was applied to the EB to approximate a normal distribution. These transformations reduce the interpretability and prediction accuracy of the structural paths. Second, SEM does not inherently support count-dependent outcomes, which are better handled by generalized linear models such as Poisson regression (Hilbe, 2011). Third, the pre-modeling diagnostics showed strong hyperdispersion of the EB variables (mean = 18.26, variance = 815.37, dispersion $\approx$ 45), NBR is appropriate for variables with dispersion rates higher than 1, which makes NBR a more suitable method to be used in the study of EB compared to Poisson regression (Long, 1997; Hilbe, 2011).

Therefore, Study 2 employed NBR to estimate the effects of review valence and RU on EB. Specifically, three models were constructed, progressively adding more variables, main effects (NB_1), interaction effects (NB_2), and the effects of contextual control variables (e.g., location, time of day, and the restaurant itself) (NB_3). This stepwise modeling strategy is common in applied regression analyses (Menard, 2002) and is particularly useful for exploring interaction effects and assessing the robustness of models constructed on realistic data (Aiken & West, 1991; Long, 1997).

3.3.1 Study 1: Structural Equation Modeling

To ensure the reliability of the model estimates, I randomly selected a subset of 5,000 reviews from the full dataset. The data were randomly divided into five equal folds (n = 1,000 per fold) to repeat the CFA and SEM estimation and to ensure that the model fit and path estimation remained stable across subsamples.

CFA was first conducted to evaluate the measurement model. CFA includes two latent constructs: AQ and SC. AQ was measured by five observable indicators: clarity, word count, relevance, objectiveness, and comprehensiveness. SC was measured by user average usefulness,

elite years, and membership length. Prior to estimation, all observed indicators of the latent variables were standardized, which can improve comparability and numerical stability in model estimation, a common practice in CFA and SEM (Kline, 2023). Based on the results of the CFA test, I can determine which observed variables were suitable for measuring latent variables.

In SEM_1, structural path part was combined, the influence of AQ and SC on RU were estimated. To meet the assumption of normality in SEM, the highly skewed and zero-inflated RU was dummy-coded (0 = not useful, 1 = useful) before modelling. Robust maximum likelihood estimation (MLR) was used to further accommodate non-normality of variables, as it provides robust standard errors and chi-square tests even under violation of distributional assumptions (Brown, 2015; Muthén & Muthén, 2017). The SEM_1 model was expressed as:

$$RU = \beta_0 + \beta_1 \cdot AQ + \beta_2 \cdot SC + \varepsilon_1 \quad (1)$$

Note. RU = Review Usefulness (dummy-coded)

AQ = Argument Quality

SC = Source Credibility

ε_1 = error term.

SEM_2 extends this by adding behavior pathways, including valence, interaction between valence and RU, EB and control vairable. Valence was mean-centered, and the interaction term was constructed using the centered valence and dummy-coded RU. To conform to the normality assumption required by SEM, EB was log-transformed. In addition, since EB I calculated was the number of new reviews within three months after the focus review, EB represents the state in the next quarter (t+1), while all other variables refer to the values at t. Review count was added to the

SEM_2 model as a standardized control variable to adjust for baseline restaurant popularity. The SEM_2 model was expressed as:

$$EB_{t+1} = \beta_0 + \beta_1 \cdot Valence_t + \beta_2 \cdot RU_t + \beta_3 \cdot (Valence_t \times RU_t) + \beta_4 \cdot Review\ Count_t + \varepsilon_2 \quad (2)$$

Note. EB= Engagement Behavior (log-transformed new reviews)

RU = Review Usefulness (dummy-coded)

Valence × RU = Interaction term between centered valence and dummy-coded review usefulness

Review Count = Standardized number of past reviews for the business

ε_2 = error term

t = Quarter of the focal review posting

t+1 = the subsequent quarter used to capture behavioral response.

These models were estimated using the lavaan package in R, across five folds, ensuring the robustness and reliability of the SEM results. This model allowed me to test whether the commonly used observable variables of AQ and SC in previous studies, as measured using NLP techniques, were still valid in the context of eWOM. This model also allows for the examination of the cognitive pathways of these two latent variables affecting RU, followed by an initial examination of the behavioral pathways.

table 2. Definition of variables, and coding manual for variable derivation for Study 1

Variable	Definition	Type	Derivation	Reference
Argument Quality	Latent variable capturing the content quality of a review	Latent Variable	Constructed from five observed variables (clarity, word_count, relevant, bert_objective, comprehensive)	Petty & Cacioppo (1986); Bhattacherjee & Sanford (2006); Cheung et al. (2008); Srivastava & Kalro (2019); Kim et al. (2018); Liu et al. (2007); Liu & Park (2015); Cheng & Ho (2015)
Source Credibility	Latent variable capturing perceived reviewer trustworthiness	Latent Variable	Constructed from three observed variables (user_avg_useful, elite_years, membership_length)	Chaiken (1980); Petty et al. (1981); Bhattacherjee & Sanford (2006); Srivastava & Kalro (2019); López-López & Parra (2016); Mudambi & Schuff (2010)
Clarity	Spelling accuracy as a proxy for clarity	Standardized Numeric (-21.29~0.76)	Standardized version of calculated as 1 - spelling error rate using R package `hunspell`	Kim et al. (2018)
Word count	Stanardized length of the review text	Numeric (-1.11~7.41)	Standardized version of number of words in raw review text before tokenization	Kim et al. (2018); Liu & Park (2015); Cheng & Ho (2015)
Relevance	Stanardized topic relevance of the review content	Numeric (-4.51~3.30)	Standardized version of weighted sum based on LDA topic distributions	Srivastava & Kalro (2019)
Objectiveness (bert_objective)	Stanardized degree of objectivity in review sentiment	Numeric (-1.14~2.06)	Standardized version of transformed from BERT sentiment score using linear distance to neutral (score 3)	Liu et al. (2007)
Comprehensive	Stanardized number of distinct topics mentioned	Integer (-3.22~3.40)	Standardized version of count of LDA topics above threshold 0.08	Liu et al. (2007)

Variable	Definition	Type	Derivation	Reference
Reputation (user_avg_useful)	Stanardized average helpfulness votes per user	Numeric (-0.26~27.64)	Standardized version of average useful vote based on history	Mudambi & Schuff (2010)
Expertise (elite_years)	Stanardized number of years as Yelp Elite	Numeric (-1.00~2.87)	Standardized version of elite years counted	Mudambi & Schuff (2010)
Membership Length	Stanardized reviewer account age (in days)	Numeric (-2.94~3.46)	Standardized version of days between review date and yelping_since	Srivastava & Kalro (2019)
Review Count	Standardized of total number of past reviews of the business	Numeric (-0.58~9.88)	Standardized version of total review received by restaurant	Pan & Zhang (2011)
Review Usefullness (useful_dummy)	Whether the review received any usefulness vote	Binary (0 or 1)	If useful vote > 0 then 1, else 0	Mudambi & Schuff (2010); Xiao & Li (2019)
Valence	Centered sentiment value of the review	Numeric (-1.53~0.46)	Mean-centered sentiment from bert_score	Baumeister et al. (2001); Luca (2011); Anderson & Magruder (2012)
Review Usefullness × Valence	Interaction between usefulness and valence	Numeric (-1.53~0.46)	useful_dummy × valence_c	/
Engagement Behavior (new_reviews_3m_ln)	Log of new reviews within 3 months	Numeric (0.00~5.93)	Log-transformed version of new_reviews_3m	self-constructed variable (no prior reference)

3.3.2 Study 2: Negative Binomial Regression

In Experiment 2, I first divided the complete dataset into a training set and a test set according to the ratio of 80%/20%, which allowed me to test the overfitting and prediction ability of the model at the end. Study 2 used the training data set (80% of full Yelp dataset) to construct

NBR models that were applied to the hyperdispersion present in the counting variables, consistent with the nature of the number of new reviews. The structure of the model was consistent with the behavioral pathways established in SEM_2, replacing the SEM estimation with NBR to better handle the nature of the dependent variable.

To ensure systematic testing and to separate the effects of interaction and environmental variables, I used a stepwise modeling strategy (Menard, 2002; Aiken & West, 1991). Three nested models were constructed to progressively assess main effects, interaction effects, and control variables. This allowed for a clearer explanation of the contribution of the interaction term to the behavioral outcome and for robustness checks with increasing model complexity.

The dependent variable is the number of new reviews generated per business in quarter t+1, and independent variables included RU, valence, and their interaction, which were generated at t. Following the recommendations of Aiken and West (1991), valence and RU were mean centered prior to calculating the interaction terms to reduce multicollinearity between the main effects and the interaction terms in the regression model. Their interaction was then calculated as a product of the centered variables. In addition, review count was standardized to facilitate interpretation and to ensure comparability of predictors. Three models were constructed:

NBR_1: Main Effects Model

$$EB_{t+1} = \beta_0 + \beta_1 \cdot Valence_t + \beta_2 \cdot RU_t + \varepsilon_3 \tag{3}$$

NBR_2: Interaction Model

$$EB_{t+1} = \beta_0 + \beta_1 \cdot Valence_t + \beta_2 \cdot RU_t + \beta_3 \cdot (Valence_t \times RU_t) + \varepsilon_4 \tag{4}$$

NBR_3: Full Model with Controls

$$EB_{t+1} = \beta_0 + \beta_1 \cdot Valence_t + \beta_2 \cdot RU_t + \beta_3 \cdot (Valence_t \times RU_t) + \beta_4 \cdot Review\ Count_t + \mathbf{\sum \delta\_state_t + \sum \delta\_year\_quarter_t} + \varepsilon_5 \quad (5)$$

EB = Engagement Behavior (count of new reviews in a quarter)

RU = Review Usefulness (mean-centered)

Valence × RU = Interaction term between centered valence and review usefulness

Review Count= Standardized number of historical reviews for the restaurant

∑δ_state = State dummy variables to control regional effects

∑δ_year_quarter = Year-quarter dummy variables to control seasonal/temporal effects

ε_3, ε_4, ε_5 = Error terms

t = Quarter when the focal review is posted

t+1 = Next quarter used to measure behavioral engagement.

Full model also included total number of restaurant reviews, state dummy variables, and year-quarter dummy variables to control for the popularity of the restaurant itself, and for regional and temporal heterogeneity in review behavior. This modeling strategy ensured that the behavioral findings were robust to generalized environment. Besides, this model allowed me to test whether RU acts as an amplifier or an inhibitor of emotional influences on behavioral engagement. Significant interaction terms indicate that the psychological mechanisms driving behavior differ across emotional contexts and levels of perceived RU, providing empirical support for the extended IAM framework proposed in this study.

table 3. Definition of variables, and coding manual for variable derivation for Study 2

Variable	Definition	Type	Derivation	Derivation
Review Usefullness	Centered useful vote received of review	Numeric (-2.57~137.43)	Mean-centered useful vote received of review	Mudambi & Schuff (2010); Xiao & Li (2019)
Valence	Centered sentiment value of the review	Numeric (-1.53~0.46)	Mean-centered sentiment from bert_score	Baumeister et al. (2001); Luca (2011); Anderson & Magruder (2012)
Review Usefullness × Valence	Interaction between usefulness and valence	Numeric (-205.44~63.27)	useful_c × valence_c	/
Engagement Behavior (new_reviews_3m)	New reviews within 3 months	Numeric (0.00~452.00)	New_reviews_3m	self-constructed variable (no prior reference)
Review Count	Standardized of total number of past reviews of the business	Numeric (-0.58~9.88)	Standardized version of total review received by restaurant	Pan & Zhang (2011)
State	Dummy variables for U.S. states	Dummies	Derived from state variable; dummy coded; first category dropped for reference group	Zhang et al. (2015)
Year-Quarter	Dummy variables for time periods (quarter-level)	Dummies	Derived from review date; dummy coded; first quarter dropped as reference	Zhang et al. (2015)

4 Results

4.1 Descriptive Analysis

4.1.1 Descriptive statistical analysis of SEM samples (n = 5000)

The statistical analyses in this section were based on a subset of 5,000 samples taken from the full dataset to construct the SEM analyses. The SEM subsample covering 4122 unique restaurant, 3239 unique users, and spanning across 14 U.S. states.

Fig 3. Boxplot of Core Variables of Subset of the Data from Study 1

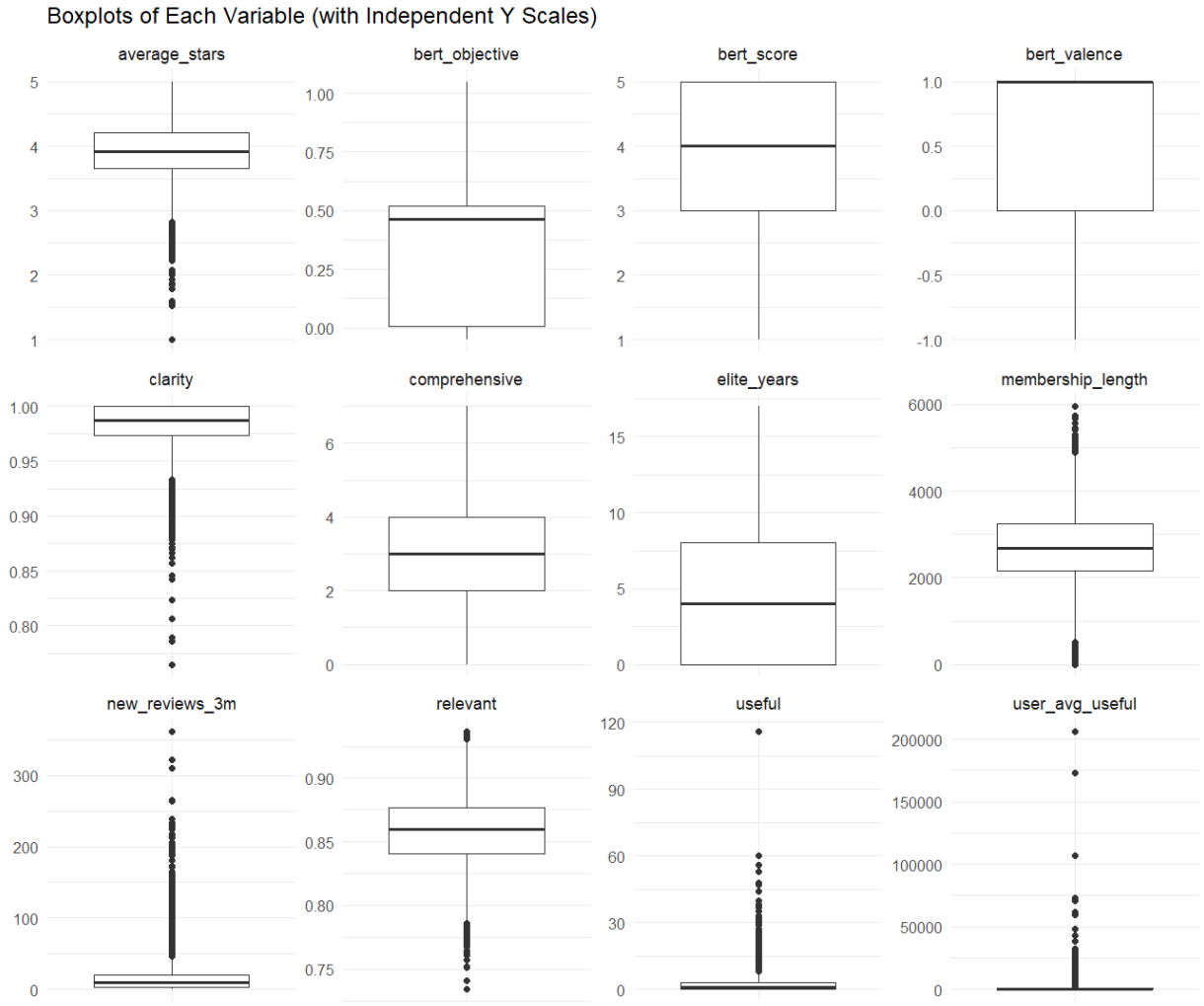


table 4. Descriptive Statistics of Core Variables of Subset of the Data from Study 1

Variable	Min	Max	Mean	SD
useful (RU)	0.000	116.000	2.430	4.810
new_reviews_3m (EB)	0.000	362.000	17.580	27.960
relevant	0.734	0.937	0.858	0.028
comprehensive	0.000	7.000	3.090	1.070
clarity	0.765	1.000	0.982	0.021
bert_score	1.000	5.000	3.860	1.220
bert_objective	-0.050	1.050	0.339	0.347
bert_valence	-1.000	1.000	0.536	0.764
membership_length	0.000	5950.000	2670.450	906.460
user_avg_useful (Reputation)	0.000	206296.000	1893.220	6869.560
average_stars	1.000	5.000	3.890	0.460
elite_years (Expertise)	0.000	17.000	4.380	4.340

As can be seen from table 4 & Fig 3, for RU, the mean was 2.43 with a standard deviation of 4.81, suggesting that most of the comments were only voted as useful once or twice, but there were a few comments that received a very large number of useful votes, and the distribution of the data was not very even. This feature also appeared in the number of new reviews added in three months, with a mean of 17.58 and a standard deviation of 27.96, suggesting that most of the reviews had less impact on the subsequent reviews of other users, but individual reviews may have resulted in a higher number of interactions.

Among the several variables that measure the quality of comment content, clarity had a mean value close to 1 and a small standard deviation, suggesting that the majority of comments were judged by the model to be very clear in their language. However, this metric was so concentrated that it would be hard to distinguish differences between reviews in the model. In contrast, the mean values of comprehensive and relevant were 3.09 and 0.86, respectively, and the standard deviations were relatively larger, suggesting that the comments varied more in content coverage and topic relevance, and that these metrics could better differentiate the quality of different reviews.

In terms of comment sentiment, the bert_score averages 3.86, indicating that the overall sentiment of most comments was on the positive side . The bert_valence was based on the bert_score, reflecting the same situation. The mean value of bert_objective is 0.34, indicating that both subjective expressions and objective descriptions were distributed in the reviews with large variations.

Strong variation was observed in the user-level data. The average length of time a user had been active on Yelp (membership length) is 2,670 days, which is equal to about 7.3 years, suggesting that the sample contains both new and experienced users. Elite years (Expertise) had a mean value of 4.38, representing some users with high activity or authority. User_avg_useful (Reputation) had a mean of 1893, but the standard deviation was very large, indicating that most users receive very few useful votes, but there were a very small number of very influential users.

4.1.2 Descriptive statistical analysis of Full Dataset

The descriptive analysis in this section is based on the total dataset, which is also the dataset used for Study 2. The full dataset comprises a total of 170,615 reviews, originating from 30,374 distinct businesses and 19,929 individual users, distributed across 14 different U.S. states.

Fig 4. Distribution of Review Usefulness

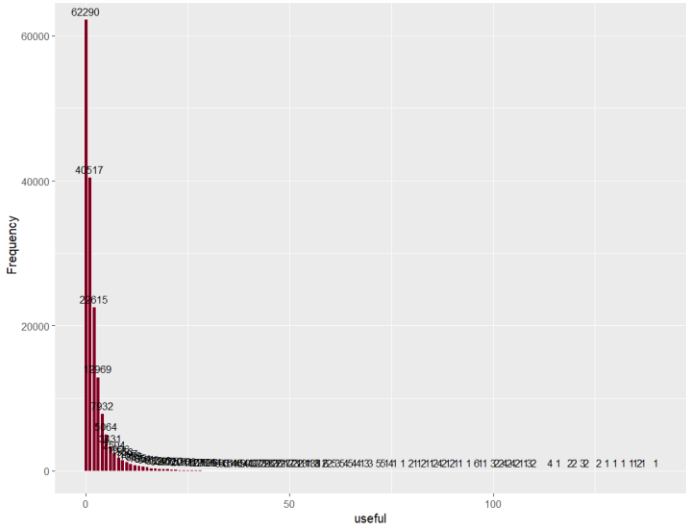
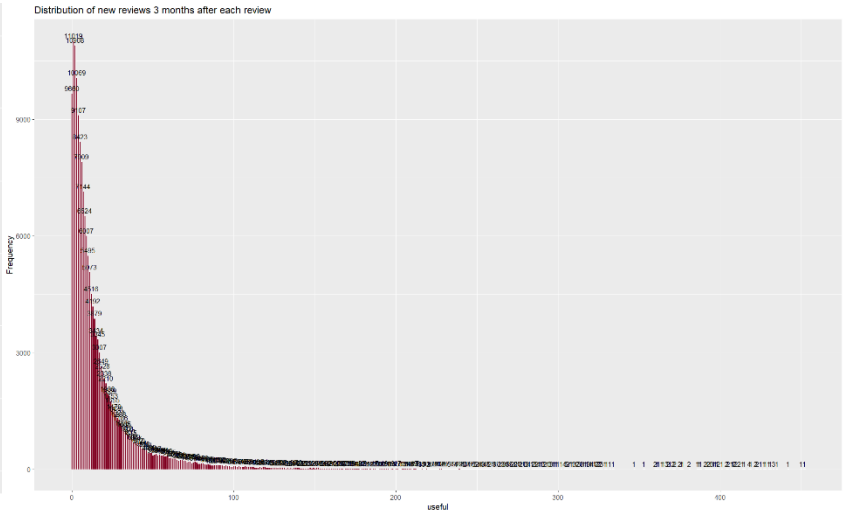


Fig 5. Distribution of Engagement Behavior



From Fig 4, the distribution of RU can be seen, the distribution of usefulness scores of reviews was extremely unbalanced. Most of the reviews had usefulness scores between 0 and 2, with the largest number of reviews rated 0, amounting to 62,290 reviews, accounting for about 36% of the total. Very few reviews received dozens or even hundreds of useful votes, pulling up the overall mean (2.57) and standard deviation (5.46). This reflects the fact that only a small number of reviews are recognized and liked by a large number of users in the platform.

Fig 5 shows the distribution of the number of new reviews (EB) brought by each review over a 3-month period, also showing a severe right skewed pattern. Although the mean was 18.32, most of the reviews did not trigger many follow-up comments, and a very small number of reviews generated a high driving effect, up to 362, suggesting that a few reviews may have created a strong topic lead on the platform.

Fig 6 illustrates the valence distribution of the reviews. The reviews were dominated by positive sentiment, accounting for about 70% of the comments, with a relatively small number of neutral and negative comments.

Fig 7 showes the distribution of the total number of reviews per state. PA had the highest number of reviews, followed by FL, LA, IN, and TN, indicating that platform users were more active in these regions.

Fig 8 showes the average number of follow-up comments (EB) generated in each state. LA leaded the way with 36.4 reviews in terms of follow-up additions, indicating that its reviews have a stronger lead effect, followed by TN and CA. In contrast, states such as AB and IL had weaker average interactions.

Fig 9 presentes the average usefulness scores of reviews by state. States NV had the highest average rating (4.4), followed by CA and NJ, while states TN, IL, and AB had a lower average usefulness of about 1.5. This may be related to the rating habits of users, the culture of platform engagement, and the quality of reviews in different regions.

Fig 6. Distribution of Valence

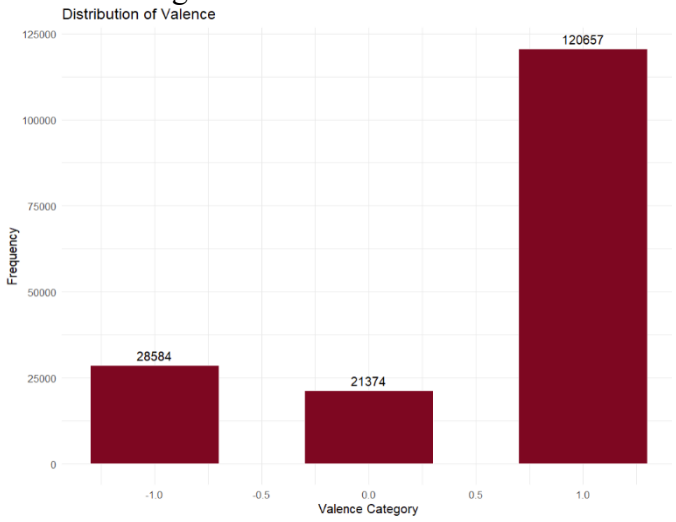


Fig 7. Number of Total Reviews per State

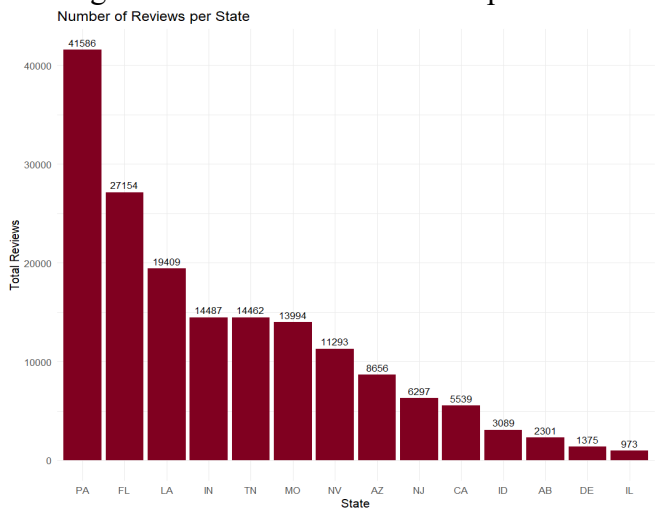


Fig 8. Average Engagement Behavior per State

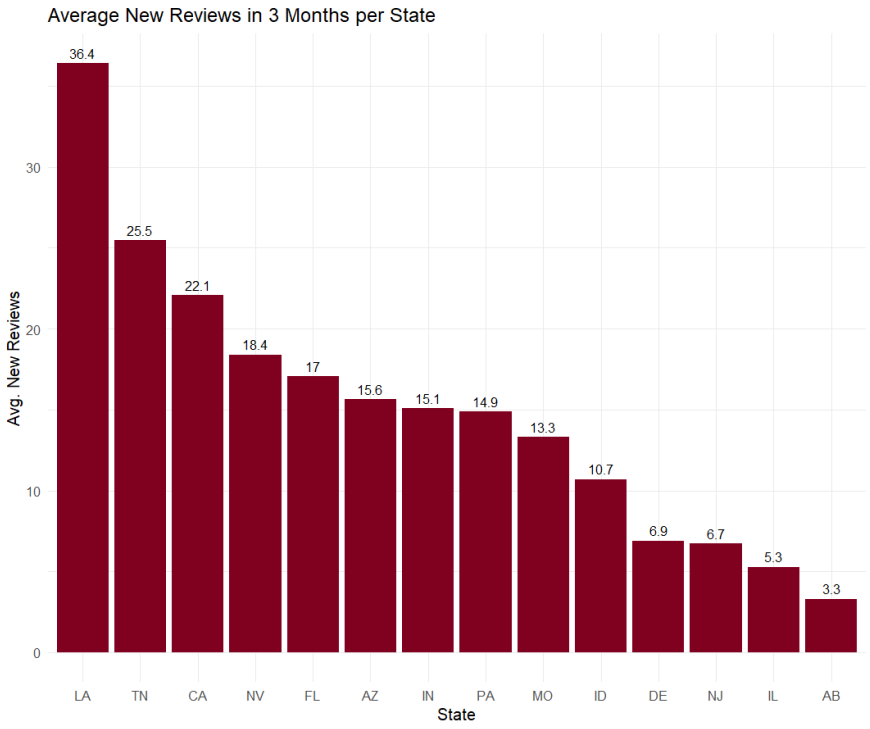
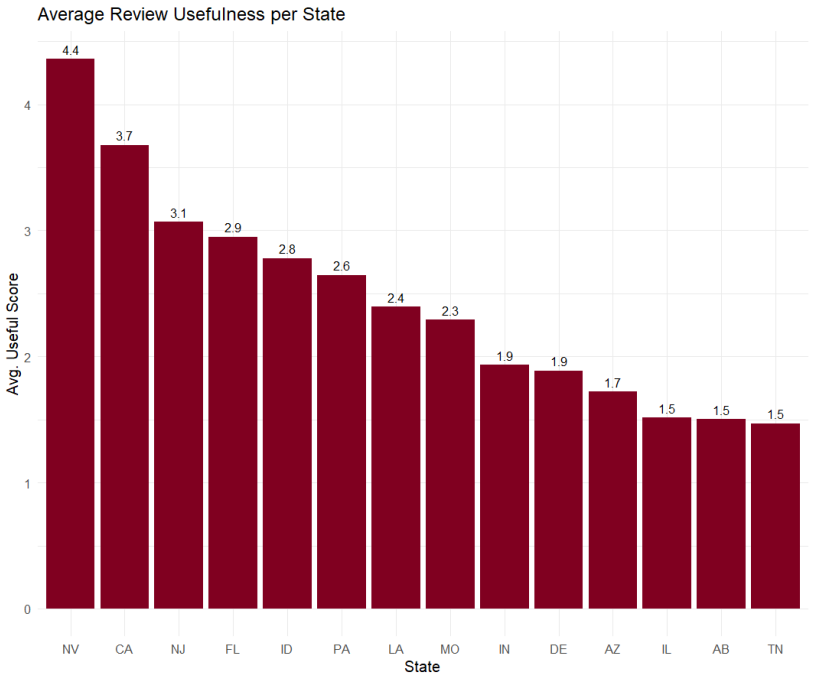


Fig 9. Average Review Usefullness per State



4.2 Study 1: Validation of Cognitive Pathways and Initial Validation of Behavioral Pathways

To examine the influence of latent variables in the cognitive pathways on the RU, as well as to preliminarily explore the pathways of their effects on user behavior, Study 1 conducted CFA and two sets of model estimation analyses based on SEM.

Four fit metrics were used to evaluate CFA and SEM, which were Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), Root Mean Square Error of Approximation (RMSEA), and Standardized Root Mean Square Residual (SRMR). These four indices are widely accepted as the standard overall fit statistics for the SEM. Hu and Bentler (1999) suggested that the CFI and TLI (also known as the unstandardized fit indices) should be combined with the RMSEA and SRMR, arguing that together they provide a more reliable assessment of overall model fit. Specifically, they proposed CFI and TLI ≥ 0.95 and SRMR and RMSEA ≤ 0.08 as acceptable fit thresholds based on extensive simulation studies. Kline (2023) also endorsed the reporting of this set of indices and suggested that RMSEA, CFI, and SRMR should be used as the minimum fit statistics required for SEM reporting. Although Kline argues that the TLI is highly correlated with the CFI, it is still frequently used and adopted in practice as a complement to the CFI.

As shown in table 5, the fit metrics for the five data folds performed similarly, indicating good robustness of the structure of the model measurements. However, an inspection of the factor loading and path estimates for each fold reveals subtle differences in the significance and direction of certain relationships. Since Fold 3 performed similarly to Folds 2, 4, and 5 in these respects, and the scores of its four model fit metrics were higher, I consider Fold 3 to be the most representative in this study, and Fold 3 will be chosen for subsequent model interpretation to report

the main results. Tables and visual comparison of the factor loading and structural paths across all five folds is provided in Appendix II & Appendix III.

table 5. 5-fold fit metrics of CFA & SEM

	CFA				SEM_1				SEM_2			
Fold	**CFI**	**TLI**	**RMSEA**	**SRMR**	**CFI**	**TLI**	**RMSEA**	**SRMR**	**CFI**	**TLI**	**RMSEA**	**SRMR**
1	0.991	0.987	0.011	0.023	0.998	0.997	0.006	0.022	0.870	0.837	0.051	0.055
2	0.941	0.913	0.034	0.032	0.963	0.946	0.028	0.029	0.862	0.828	0.054	0.055
3	0.964	0.928	0.034	0.031	0.964	0.948	0.034	0.030	0.846	0.808	0.049	0.052
4	0.889	0.837	0.044	0.037	0.923	0.889	0.040	0.036	0.834	0.794	0.054	0.055
5	0.956	0.936	0.028	0.028	0.956	0.937	0.031	0.029	0.859	0.824	0.051	0.052

$p < .001$ *** $p < .01$ ** $p < .05$ * $p < .10$. Not significant otherwise.

4.2.1 Results of CFA

For the test of CFA model, as can be seen in table 6, for the factor analysis of AQ, the factor loadings of Word Count (std.all = 0.647, P < 0.000), Objectiveness (std.all = 0.365, p < 0.001), were highly significant and positive, these results suggested that longer text length and objectivity of sentiment expression were two important factors when consumers were judging the RU. Secondly, the factor loading Clarity (std.all = 0.141, p = 0.002) showed a significant positive relationship, suggesting that higher linguistic clarity were important indicators when consumers evaluated the usefulness of a review. In contrast, the loading of Relevant (std.all = -0.054, p = 0.209), and Comprehensive (std.all = 0.012, p = 0.778) were statistically insignificant, This suggests that the two metrics I obtained from using the NLP tool, that is, the relevance of the topic to the restaurant and the coverage of multiple topics in the review, are not consistently recognized by consumers as a measure of the quality of reviews. In subsequent examinations, both variables were retained in the structural model, but their low explanatory power suggests that their contribution to the underlying construct of argument quality is limited.

4.2.2 Results of SEM_1

SEM_1 was used to test the effects of two latent variables, AQ and SC, on RU, representing the cognitive pathway in the conceptual framework. As can be seen in table 6 and Fig 10, the factor analysis part of SEM_1 was similar to most of the results of CFA and therefore will not be repeated. And the path analysis part contains the cognitive pathways of the conceptual model.

In the structural path analysis, both latent variables demonstrated significant and positive effects on RU. SC had a stronger effect on RU (est = 0.156, $p < 0.000$), and the results examined Hypothesis 2 of this paper was supported. This suggests that when the source of information was perceived to be more reputable, i.e., when the reviewer has higher professional labeling, better past performance, and longer years of membership, the published reviews were more likely to be perceived as useful. Meanwhile, AQ also showed a significant positive effect on RU ($\beta = 0.083$, $p = 0.003$), again indicating that Hypothesis 1 of this study was true. This suggests that people are more likely to find the content of reviews with clear language, objectivity, and longer word counts to be more helpful. However, the effect size of AQ was slightly lower than SC. This implies that the quality of the review itself, compared to the credibility of the source of the reviewer, has a relatively weaker effect, which may be due to a less appropriate choice of observational variables, resulting in a lower path coefficient.

4.2.3 Results of SEM_2

SEM_2 expanded on SEM_1 by incorporating the behavioral pathways proposed in the conceptual model. Valence and the interaction term between RU and Valence, as well as the control variable Review Count was introduced.

From table 6 and Fig 10, a significant positive effect on of Valence on EB can be seen (est = 0.168, p = 0.008), suggesting that emotionally positive reviews were more likely to trigger EB such as writing new reviews by other customers. The main effect of RU also showed a significant positive effect (β = 0.128, p = 0.030), which contradicted Hypothesis 3. Which meant that reviews perceived as useful had a direct influence on follow-up behavior. In addition, the interaction term Valence × RU showed a significant negative effect (β = -0.168, p = 0.030), which was contrary to Hypothesis 4, and might indicate that high usefulness of negative valence reviews may instead increase consumer motivation to follow up, but the evidence was not sufficiently strong now to support a definitive conclusion.

In conclusion, hypotheses 1,2 of this study proved to be true and hypothesis 3, 4 were false in Study 1. Study 1 provided supportive evidence about cognitive pathways through the SEM model, and preliminary tests for behavioral pathways. It also laid part of the theoretical foundation for the further introduction of NBR analyses in Study 2 to test the main effects and moderating effects of the engagement variables.

Fig 10. Estimated Structural Equation Model (SEM) Path Diagram

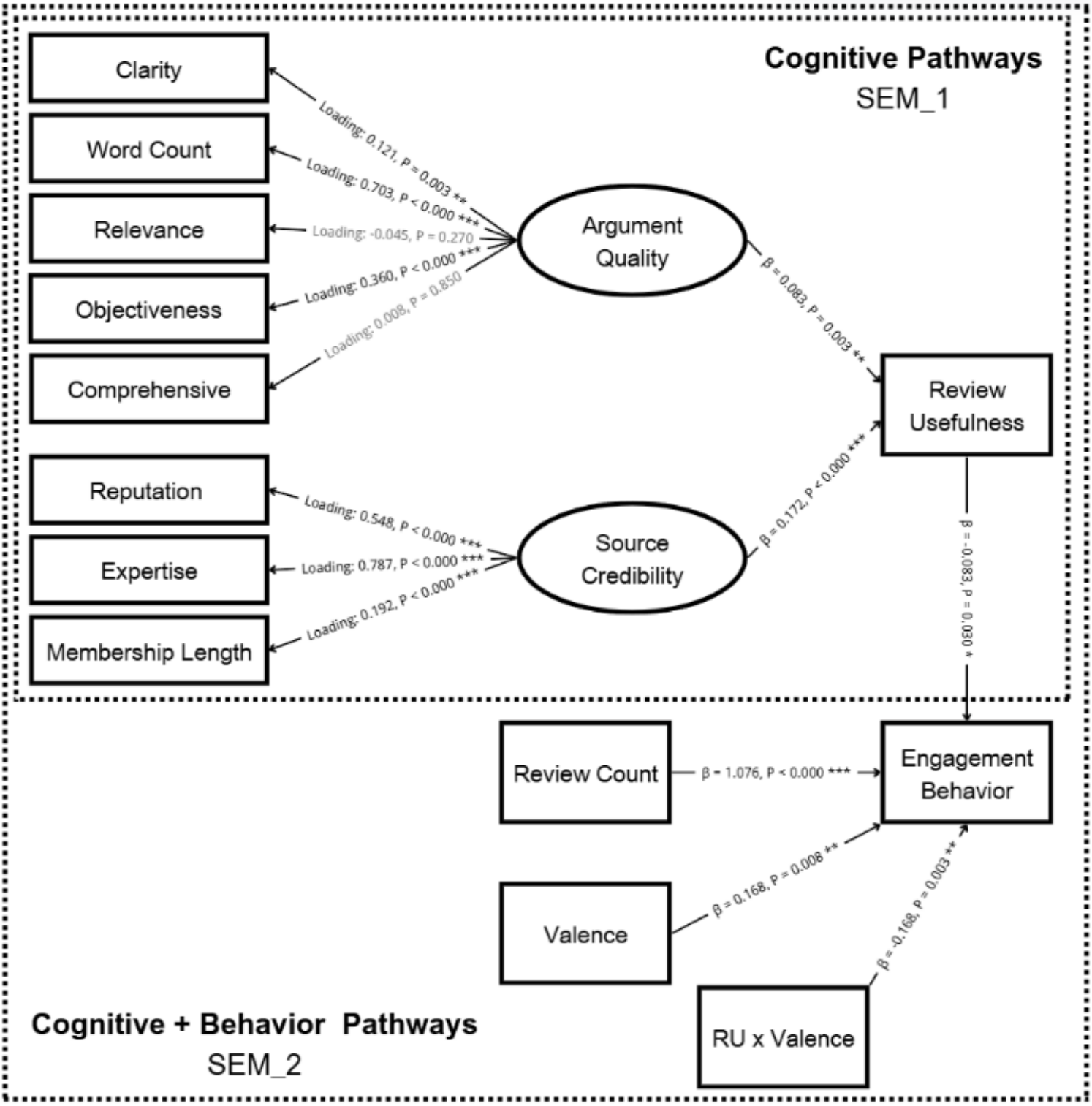


table 6. Summary of 3 models in Study 1 (CFA / SEM_1 / SEM_2)

Variable/Path	CFA Loading (std.all)	Sig. in CFA	SEM_1 Loading (std.all) / β (est)	Sig. in SEM_1	SEM_2 Loading (std.all) / β (est)	Sig. in SEM_2
AQ → Clarity	0.141	0.002 **	0.132	0.003 **	0.121	0.003 **
AQ → Word Count	0.647	0.000 ***	0.672	0.000 ***	0.703	0.000 ***
AQ → Relevant	-0.054	0.209	-0.047	0.270	-0.045	0.270
AQ → Objectiveness	0.365	0.000 ***	0.352	0.000 ***	0.360	0.000 ***
AQ → Comprehensive	0.012	0.778	0.008	0.850	0.008	0.850
SC → Reputation	0.342	0.000 ***	0.334	0.000 ***	0.548	0.000 ***
SC → Expertise	0.758	0.000 ***	0.779	0.000 ***	0.787	0.000 ***
SC → Membership Length	0.198	0.000 ***	0.190	0.000 ***	0.192	0.000 ***
AQ → RU	—	—	0.083	0.003 **	0.083	0.003 **
SC → RU	—	—	0.172	0.000 ***	0.172	0.000 ***
RU → EB	—	—	—	—	0.128	0.030 *
Valence → EB	—	—	—	—	0.168	0.008 **
Valence × RU → EB	—	—	—	—	-0.168	0.030 **
Review Count → EB	—	—	—	—	1.076	0.000 ***

p < .001 *** p < .01 ** p < .05 * p < .10 . Not significant otherwise.

4.3 Study 2: Validation of Behavioral Pathways

To further validate the behavioral pathways mechanism of Valence and RU, and to deal with the feature of overdispersion of the dependent variable, Study 2 adopted the NBR model for modeling and analysis of the training sample (80% of full dataset). Before modeling, this study first pre-tested the distribution of the dependent variable EB (New_Reviews_3m), and found that its mean was 18.26 while the variance reached 815.37, which was much more dispersed than the consistency assumption required by the Poisson distribution, and the computed dispersion ratio was nearly 45, which is significantly larger than 1, which is a cut-off value suggested by Hilbe (2011) and Long (1997) to adopt Negative Binomial Model (table 7). At the same time, the percentage of zero values was only 5.67%, as suggested by Ridout, Demétrio, and Hinde (1998) that zero-inflated model is needed necessarily when the precentage of zero value is over 20-30%, therefore the zero-inflated model structure was not used.

table 7. Summary of the dependent variable Behavior Engagement

Mean	Variance	dispersion_ratio	Proportion of 0
18.256	815.37	45	5.67%

A total of three sets of models were developed for Study 2, the main effects model (NBR_1), the interaction term model (NBR_2), and the extended model with control variables (NBR_3), the results of them were shown in table 9. Variables contained in the NBR_1 model included the centered RU (Useful_C), Valence (Valence_C). An additional interaction term between RU and Valence (Int_Use_Val) was introduced in the NBR_2 model. The standardized Review Count, and

dummy variables for State and Year-Quarter were further introduced as control variables in NBR_3.

To assess the overall explanatory power of NBR models, I used two widely used pseudo-R^2 measures: the McFadden R^2 and the Nagelkerke R^2. These metrics were typically used in models involving count dependent variables, which do not fit OLS-based R^2 (Long, 1997; Menard, 2002). McFadden R^2 is based on log-likelihood ratios, values between 0.02 and 0.05 are considered weak, values around 0.1 indicate modest fit, and values above 0.2 suggest a good model fit (McFadden, 1974; Louviere et al. ,2000) The Nagelkerke R^2 is a standardized version that rescales of the McFadden R^2 to a range of 0 to 1, and values above 0.2 are often interpreted as acceptable (Nagelkerke, 1991).

table 8. Pseudo-R^2 Fit Statistics for Negative Binomial Regression Models

Model	McFadden R^2	Nagelkerke R^2
NBR_1 (Main Effects)	0.0006	0.0049
NBR_2 (Interaction)	0.0006	0.0051
NBR_3 (Full with Controls)	0.0782	0.4572

As shown in , the values of both indices for NBR_1 (main effect) and NBR_2 (interaction) were very low (McFadden $R^2 = 0.0006$; Nagelkerke $R^2 \approx 0.005$), which suggests that the explanatory power was minimal when using only the review features as predictors. However, with the inclusion of control variables, the explanatory power of NBR_3 increased substantially, with McFadden R^2 rising to 0.0782 and Nagelkerke R^2 reaching 0.4572. McFadden R^2 values above 0.05 are generally considered acceptable for behavioral models (McFadden, 1974), and the corresponding Nagelkerke R^2 was greater than 0.02, hence the model was moderately good fit, i.e., nearly half (45%) of the pseudo-variance in the participation behavior can be explained when

restaurant heat, time, and geographic heterogeneity were controlled for. These results demonstrate the validity of the stepwise modeling approach and emphasize the importance of contextual variables in predicting consumer EB beyond the content characteristics of reviews.

table 9. Summary of Negative Binomial Regression

Variable	NBR_1 Est	NBR_1 P	NBR_2 Est	NBR_2 P	NBR_3 Est	NBR_3 P
(Intercept)	2.909	0.000***	2.910	0.000***	1.546	0.000***
useful_c	-0.002	0.000***	-0.002	0.000***	0.001	0.000***
valence_c	0.112	0.000***	0.112	0.000***	0.079	0.000***
int_use_val	—	—	-0.004	0.000***	-0.004	0.000***
review_count	—	—	—	—	0.703	0.000***
State dummies	—	—	—	—	Included	—
Year-quarter dummies	—	—	—	—	Included	—

table 9 shows the summary of NBR_1, Valence has a significant positive main effect on EB (β= 0.112, P < 0.000 ***), while RU has a significant negative main effect (β= -0.002, P < 0.000 ***). Both were lower than the 0.001 significance level. From this result, it can be seen that the more positive the valence of a review was, the more likely it was to trigger subsequent review participation by other customers, which was the same as the result of previous studies. While the effect of RU on the positivity of subsequent reviews was significant, but the effect was smaller than that of Valence.

Subsequently, after adding the interaction term (Int_Use_Val) to the NBR_2 model, I found that the coefficients and significance of the effects of RU and Valence on EB did not change in table 9. The effect of the interaction term on EB was significantly negative (β = -0.004, P < 0.000***), suggesting a useful negative review, instead, stimulates the willingness of other users to engage subsequently. This result was consistent with the direction of the effect in SEM_2 of Study 1, but it complemented the weakly significant estimate in SEM.

To enhance the explanatory power of the model, NBR_3 introduced control variables, and the results were shown in table 9. The direction of the core variables was affected to different degrees, indicating that the behavioral mechanism has a more cleansing explanatory power after controlling for temporal, regional and restaurant popularity heterogeneity. First, the effect of Valence on EB was still significantly positive (β= 0.112, P < 0.000***), but it was reduced from 0.112 to 0.079, which means that partly of the variation in EB that was explained by Valence was also explained by the control variables, and thus the effect was reduced. It was worth noting that the effect of RU on EB is positive and significant (β= 0.001, P < 0.000***) after the introduction of the control variable, which means that the more useful reviews were more likely to cause the subsequent customers to engage, and this was more consistent with face validation compared with the results of NBR_2, and therefore, the H3 of the study was tested to be false. Finally, it was the interaction term of RU and Valence on EB remains unchanged and was significantly negative (β= -0.004, P < 0.000***), thus Hypothesis 4 of this study was similarly rejected. From Fig 11, the trend of RU in moderating EB at different Valence levels can be seen. Under the four levels of RU of 0, 40, 80 and 120 respectively, the higher the Valence was, the corresponding effect on EB gradually decreases or even reverses its effect, which further confirms the negative effect mechanism of the interaction term.

Fig 11. Visualization of the interaction of RU and Valence on EB

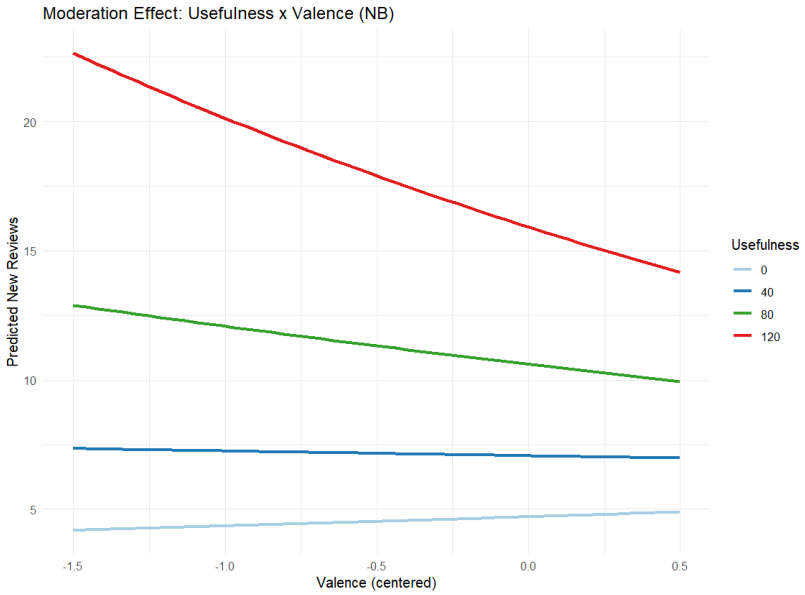
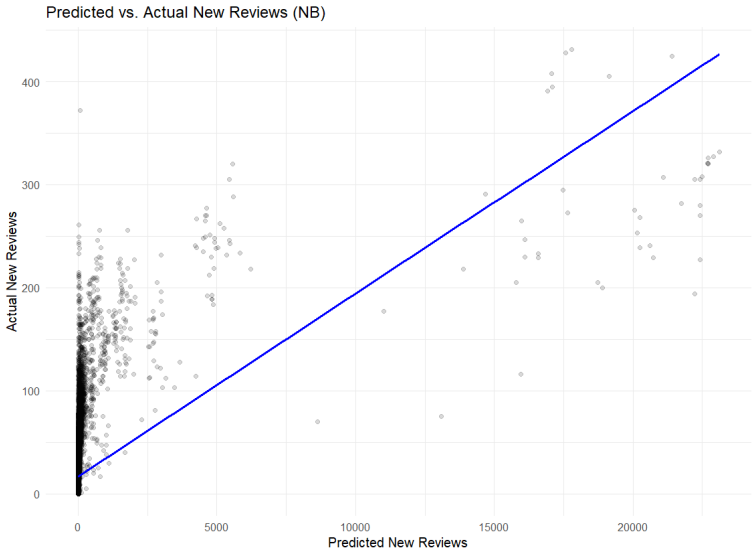


Fig 12. QQ-plot of NBR_3 predictions on the test set



To assess the validity of the model, this study applied the model to the test dataset (20% of full dataset), which was used to examine the degree of match between the predictions and the observations. In this study, the residuals of the QQ plots of EB (new_reviews_3m) were plotted. In Fig 12, if all points strictly fit the diagonal line, it indicates that the model predicted distribution

was in perfect alignment with the actual observations. However, in this study, it was observed that most of the points deviate from the line to some extent, especially at the extreme value intervals. This discrepancy can be attributed to the severe hyperdispersion inherent in EB (Mean = 18.26, Variance = 815.37), which makes it more difficult to obtain a desirable fit where the points fit the line perfectly in this type of high variance count data. This observation was further supported by the residual plot (Fig 14), where the model tends to underestimate high levels of EB and overestimate low levels, highlighting the challenge of extreme values in count prediction.

For further verifying the generalization ability and actual prediction effect of the model, this paper evaluates the performance of the NBR_3 model on the training set and the test set. As shown in the table 10, on the training set, the mean absolute error (MAE) of the model was 56.82, the root mean square error (RMSE) was 848.46, and the Hit Rate was 74.28%; on the test set, the MAE was 71.91, the RMSE was 984.75, and the Hit Rate was 74.19%. Although the error of the model on the test set increased slightly, the overall difference was not significant. Besides, the Hit-Rate of the model on both datasets were similar and high, indicating that the model did not have obvious overfitting problems and had strong predictive ability.

table 10. Generalization and Predictive Ability Tests for NBR Models

	MAE	RMSE	Hit-Rate
Training Dataset	56.82	848.46	0.7428
Testing Dataset	71.91	984.75	0.7419

Finally, I also performed diagnostics on the model, and the results indicated the structural fit of the model. First, Fig 13 shows low correlation between each pair of independent variables, a subsequent VIF test showed that the VIF values for the four explanatory variables (Useful, Valence, Interaction Term, Review_Count) had VIF values less than 1.02 in table 11, indicating that the

correlation between the model's explanatory variables was acceptable and the model did not contain multicollinearity.

Fig 13. Correlation Matrix of Core Variables

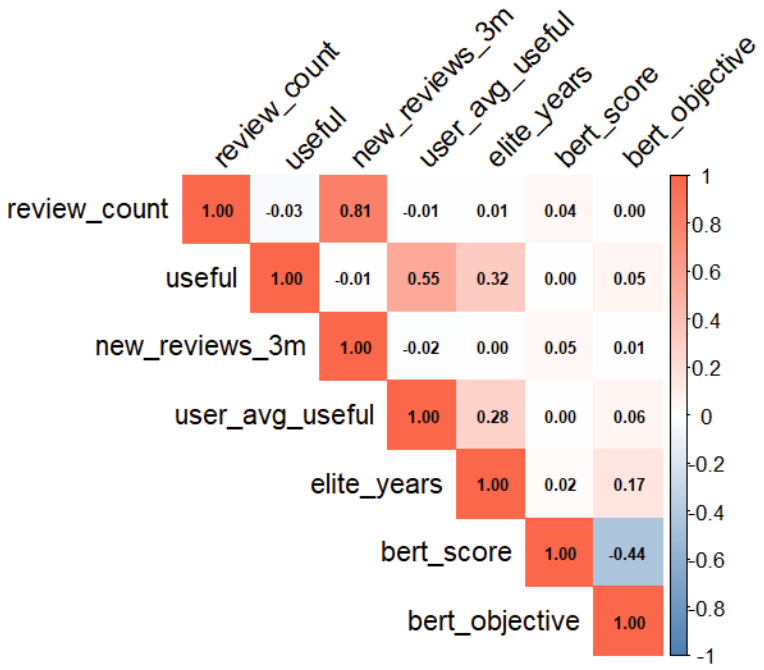


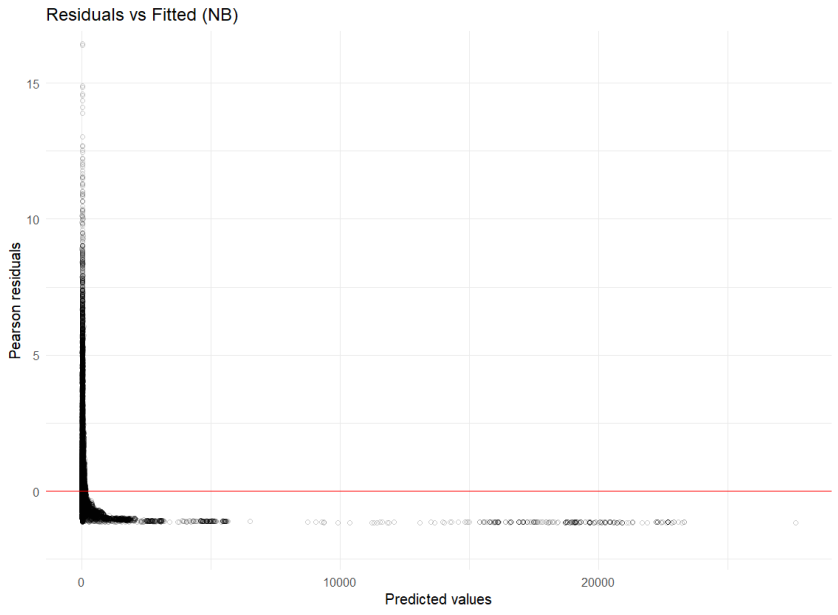
table 11. VIF Check of Core Variables

Useful	valence	interaction term	review_count
1.012209	1.002745	1.011421	1.003022

The residual plots further reveal the structural characteristics of the model errors. Fig 14 shows that in the lower ranges of EB (e.g., less than 20), the Pearson residuals have a lot of positive values, implying that the model was likely to overestimate the number of actual new reviews. The residuals in the higher ranges of EB show many negative values, suggesting that the model underestimates the number of new reviews. This may indicate that although NBR outperformed the Poisson model in dealing with excessive dispersion, its predictions were still conservative on

samples with extremely high review volume, and future research on EB may need to choose a nonlinear model for a better explanation. However, the overall trend direction was still consistent with theoretical expectations, indicating that the model still has some explanatory power in fitting the overall structure.

Fig 14. Plot of predicted residuals of NBR_3



In conclusion, Study 2 validates the main effects of RU and Valence based on the behavioral pathways perspective, and also provides empirical support for the extended IAM theory through significant interaction terms. These results were partly the same as the results of the SEM in Study 1, but there were also differences. Finally, in Study 2, the two hypotheses H3 and H4 proposed in this paper were rejected, but the model of NBR was more suitable for large-sample counting-type variables, and therefore the results were more applicable and have more explanatory value.

5 Conclusion and limitations

5.1 Conclusion

The purpose of this study was to explore how review features in eWOM affect consumer behavioral responses from cognition and emotion through the integrated pathway of cognitive and behavioral pathways. By integrating both SEM and NBR methods, this paper constructed a two-stage empirical analysis framework to quantitatively study real review data on the Yelp platform.

table 12. Validity of measurement indicators for observed variables

Latent Variables	Observed Variables	Validity
AQ	Clarity	valid
AQ	Word Count	valid
AQ	Relevant	Invalid
AQ	Objectiveness	valid
AQ	Comprehensive	Invalid
SC	Reputation	valid
SC	Expertise	valid
SC	Membership Length	valid

In the validity assessment of the measurement model, the study conducted factor loadings and significance tests on each of the observed variables corresponding to the two latent variables AQ and SC, to determine their measurability in structural equation modeling. As shown in table 12, the three observed variables under SC of Reputation, Expertise and Membership Length, were all tested as valid indicators, indicating that SC as a latent variable was effectively measured in in this study. This indicates that users refer to objective information such as the reviewer's experience level, professional label, and seniority on the platform when judging the credibility of the review source.

In contrast, among the five observed variables of AQ, only Clarity, Word Count and Objectiveness showed valid measurement loadings, and these three metrics share the common characteristics of being relatively clear, operational, and directly and objectively modeled by NLP methods. Specifically, Word Count, as a basic linguistic feature, had a natural correlation with information density and degree of argumentation development, which was easy to quantify and interpret. Objectiveness, on the other hand, was based on the sentiment scores generated by the BERT model, which measured the objectivity of a statement in terms of its expressive polarity. As this measure was based on the deep learning of a large pre-trained language model, it showed stable validity. Validity of Clarity also reached a statistically significant level, although clarity was subjective, but features such as lexical spelling error rate can indirectly characterize this notion to a certain extent, the hunspell package in R capture of text readability can reflect judgment of the quality of information of customers.

On the contrary, Relevant and Comprehensive failed the validity test and were determined to be invalid indicators. The result was notable because it may reflect three dimensions of problems. On the one hand, from the perspective of the concept itself, these two features are abstract and subjective, which are not as easy to judge directly as the number of words or clarity, especially in the case without manual annotation and relying entirely on machine extraction, it was difficult for the system to accurately grasp whether a review was truly Relevant or Comprehensive. On the other hand, in terms of the implementation of the model, this study indirectly measured these two dimensions by means of LDA topic modeling, followed by the fact that the relevance in this study was weighted by the authors based on the percentage of LDA topics as well as the high-frequency words of the topics themselves, which was relatively subjective and lacked theoretical support, which was innovative, but may also have the problem of not being stable or having weak degree

of differentiation. On the last hand, previous studies measured AQ usually by means of questionnaires, allowing respondents to subjectively judge the relevance and comprehensiveness of the information. This study adopted an NLP approach to extract features directly from the review text itself. While this method had the advantage of objectivity and automation, it may also deviate from human perception. Therefore, the two methods may not be completely equivalent in the selection of indicators for measuring AQ, which in turn affects the statistical validity performance.

Taken together, this measurement validity analysis suggests that observables generated through NLP should be selected and validated more judiciously in future studies, especially for variables that are difficult to quantify directly, it is necessary to consider the introduction of manual coding or mixed methods for cross-validation to enhance the measurement stability and interpretability of complex models.

table 13. Summary of studies findings

Variable/Path	**Hypothesis**	**Study 1**	**Study 2**
AQ → RU	H1: Argument quality of a review has a positive effect on review usefulness	Accepted	Not tested
SC → RU	H2: Source credibility of a reviewer has a positive effect on review usefulness.	Accepted	Not tested
RU → EB	H3: Review usefulness does not directly affect engagement behavior.	Rejected	Rejected
Valence → EB	Proven by previous study	—	—
Valence × RU → EB	H4: Review usefulness moderates the effect of valence on engagement behavior.	Rejected *	Rejected *

Rejected* = Significant, but the effect was contrary to the assumptions.

In the test of the path structure hypothesis, the main four findings of this study can be obtained from table 12 & Fig 15. First, both H1 and H2 are verified in Study 1, indicating that both the AQ and SC significantly enhance the perception of the usefulness of the review by other users and the behavior of voting the review as useful. In particular, the effect of SC on RU is slightly stronger than that of AQ (path coefficients of 0.172 and 0.083, respectively), and this difference may derive from the fact that platform users, when faced with the context of information overload, prefer relying on the labels of the people rather than the content itself to make their initial judgments. This phenomenon can be explained by the Heuristic-Systematic Model proposed by Chaiken, S. (1980) in psychology. That is, the ability to think is a resource that can be consumed, and in daily decision making, people often choose to conserve their thinking energy and react instinctively by using heuristic cues as a basis for judgment. For example, a reviewe platform badge can lead to an immediate assumption that a review must be of high quality, rather than taking time to examine the content of the review itself.

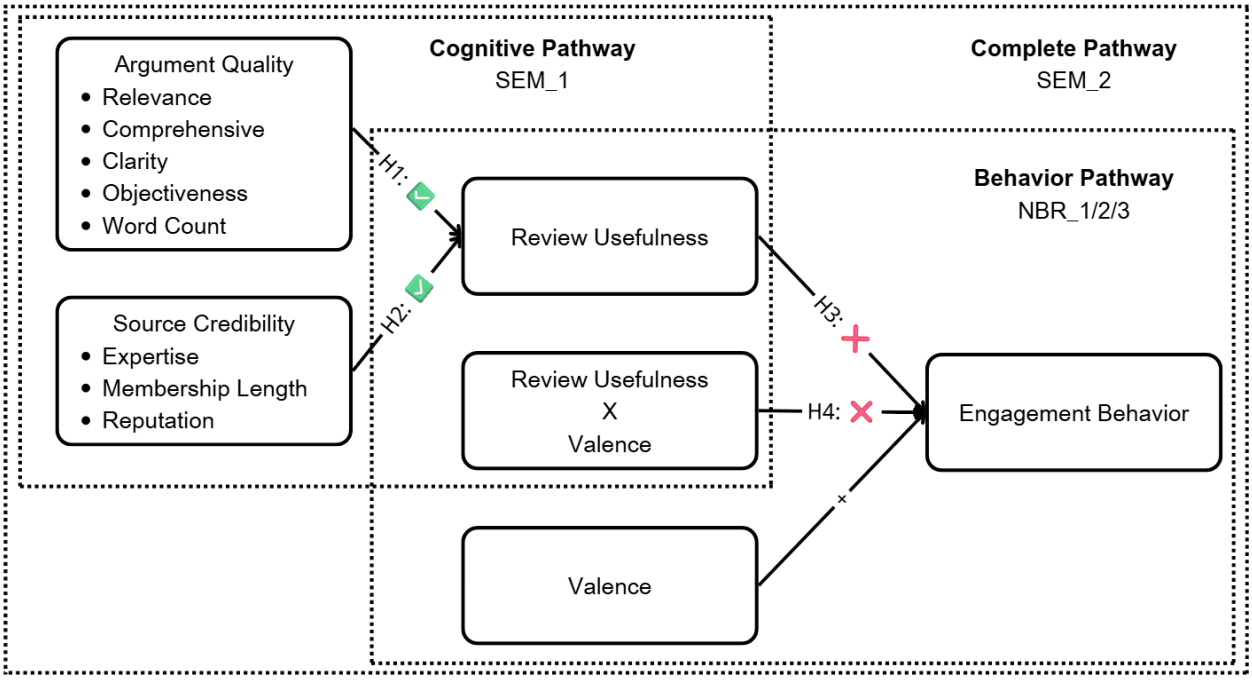
Following this, H3 was rejected in both studies, RU had a significantly positive effect on EB, contradicting the original hypothesis of no significant effect. This result suggests that in a reality platform environment, without considering the valence of reviews, useful reviews may have a strong social incentive effect, encouraging others to participate in the conversation as well. The more useful reviews a restaurant has, the more trust and curiosity other potential customers have about the restaurant, thus promoting an increase in the number of subsequent review participants.

Most notably, H4, while reaching statistical significance in both studies, went in the opposite direction of what was expected and was thus labeled rejected. RU did not reinforce the role of positive valence in promoting EB. Instead, reviews with lower valence (i.e., negative reviews) and higher usefulness were more likely to motivate other users' EB. The first one is the

negative review defense mechanism, where other consumers (e.g., fans, regulars, or people with different attitudes) may respond to a useful negative review for the purpose of defending the restaurant's reputation or expressing a different point of view, thus increasing the frequency of engagement. The other is the positive review default mechanism, which means that when other consumers find a very useful and positive review, they feel that there is nothing more to add, thus decreasing the number of subsequent new reviews.

Overall, this study provides solid support for H1 and H2, emphasizing the critical role of AQ and SC in consumers' judgments of review usefulness. The inverse results of H3 and H4 suggest that when studying EB, we cannot simply assume that the relationship between usefulness and valence is linearly enhanced, but need to further explore the complex social interaction mechanisms and psychological response paths.

Fig 15. Graphical summary of the validation of the study



This paper verified and expanded the applicability of the IAM framework in eWOM from the level of theoretical research, pointed out that the intertwining of cognitive and behavioral paths constitutes the mechanistic basis of user's responses to online reviews, and provides a new theoretical basis and modeling ideas for online review management and platform review recommendation.

At the practical level, the findings provided specific insights into the review management and content recommendation mechanisms of online platforms. First, the study found that SC (e.g., professional labeling, historical reputation) and AQ (e.g., clarity vs. objectivity) jointly affect the degree to which a review is perceived as useful. Therefore, platforms should consider both review content characteristics and reviewer characteristics when sorting and recommending reviews to enhance information credibility and user trust.

More importantly, the study found a significant negative interaction between RU and Valence on EB. When a review is perceived as very useful, subsequent users are more likely to agree and refrain from reviewing if it is a positive sentiment, leading to a disruption in review accumulation. Therefore, if platforms want to promote the sustainability of comment increase, they can dynamically adjust the display strategy of comments. Firstly, for positive and useful comments, after obtaining a certain number of useful votes, the display priority should be adjusted downwards to avoid suppressing other users' willingness to express their opinions. For negative reviews, restaurants can increase the visibility of negative reviews, while ensuring that the content is appropriate, to activate customers' willingness to discuss.

In summary, this study not only theoretically deepens the understanding of user information processing mechanisms, but also provides platform providers with actionable strategy suggestions to optimize user engagement on online review platforms.

5.2 Limitations and Future Research

Although this study had made important findings in validating the influence of cognitive and behavior pathways on consumer behavior in eWOM through a two-stage modeling strategy, there are still several limitations that provide room for future research.

First, on the results of the interaction effect between valence and RU, Study 2 found that the interaction term between RU and Valence was negatively significant, with useful negative reviews being more likely to trigger subsequent EBs than useful positive reviews.The underlying mechanism for this phenomenon may be that, when a positive review that is widely recognized as being useful is present, subsequent consumers are more likely to have a tendency to agree that there is no need for additional reviews, thereby reducing the generation of new reviews. In contrast, when a useful negative review is posted, it may trigger maintenance behavior among restaurant fans who voluntarily post positive responses, thus pulling in a new round of active reviews. Since the existing analysis cannot reveal the emotional tendency of new reviews, future research can further analyze the valence of new reviews based on the number of new reviews, and explore whether negative RU will trigger positive responses or defensive responses, to reveal the adversarial communication mechanism in eWOM.

Second, this paper used the number of new reviews in the quarterly dimension as a proxy for EB, which is representative, but failed to incorporate the actual sales data to capture the transformation of consumer behavior more directly, limiting the explanatory power of the results on the actual economic impact of consumer behavior. A future study that incorporates platform APIs or merchant cooperation to obtain sales data will further enhance the model validity and industry application value.

In addition, the current model only controlled business popularity, temporal and regional heterogeneity, but did not yet account for the possibility that the characterization of reviews prior to their posting may interfere with the number of new reviews. That is, while this study implicitly assumes that focal reviews are the driving factor for the number of new reviews, there may be a trend effect, the current focal reviews are only a part of the trend, not its starting point. Future research could introduce pre-review characteristics as control variables, such as the number of reviews and sentiment distribution in the week or month prior to the focal review, in order to more accurately capture the net effect of the focal review on behavioral variables and improve the rigor of causal identification.

Lastly, in terms of variable construction and measurement, there were still some deviations and theoretical vacancies. On the one hand, Relevant and Comprehensive in AQ failed to show good explanatory power in NLP process, indicating that semantic latent variables are still limited by the accuracy of current text processing technology. Furthermore, the weighting construction of Relevant in this study was customized by the researcher and lacked systematic theoretical support and external validation, which might affect its stability as an observed variable in SEM. Future research can combine manually labeled data, more sophisticated semantic identification methods, and a clear theoretical foundation to improve the consistency and accuracy of latent variable measurement.

6 References

Aiken, L. S., & West, S. G. (1991). Multiple regression: Testing and interpreting interactions. Sage Publications.

Anderson, M., & 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. https://doi.org/10.1111/j.1468-0297.2012.02512.x

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, 103(3), 411–423. https://doi.org/10.1037/0033-2909.103.3.411

Ariyasriwatana, W., & Quiroga, L. M. (2016). A thousand ways to say delicious: Categorizing expressions of deliciousness from restaurant reviews on the social network site Yelp. *Appetite*, 104, 18–32. https://doi.org/10.1016/j.appet.2016.01.002

Arun, R., Suresh, V., Madhavan, C. E. V., & Murthy, M. N. (2010). On finding the natural number of topics with latent dirichlet allocation: Some observations. In Proceedings of the Machine Learning and Applications (ICMLA), 2010 Ninth International Conference (pp. 638–643). IEEE.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, 5(4), 323–370. https://doi.org/10.1037/1089-2680.5.4.323

Bhattacherjee, A., & Sanford, C. (2006). Influence processes for information technology acceptance: An elaboration likelihood model. *MIS Quarterly*, 30(4), 805–825. https://doi.org/10.2307/25148755

Brown, T. A. (2015). Confirmatory factor analysis for applied research (2nd ed.). Guilford Press.

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing*, 72(7–9), 1775–1781.

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology*, 39(5), 752–766. https://doi.org/10.1037/0022-3514.39.5.752

Cheng, Y., & Ho, H. (2015). Social influence's impact on reader perceptions of online reviews. *Journal of Business Research*, 68(4), 883–887. https://doi.org/10.1016/j.jbusres.2014.11.046

Cheung, C. M. K., Lee, M. K. O., & Rabjohn, N. (2008). The impact of electronic word-of-mouth. *Internet Research*, 18(3), 229–247. https://doi.org/10.1108/10662240810883290

Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, 43(3), 345–354. https://doi.org/10.1509/jmkr.43.3.345

Coughlin, S. S. (1990). Recall bias in epidemiologic studies. Journal of Clinical Epidemiology, 43(1), 87–91. https://doi.org/10.1016/0895-4356(90)90060-3

Coursaris, C., Van Osch, W., & Albini, A. (2018). Antecedents and consequents of information usefulness in user-generated online reviews: A multi-group moderation analysis of review valence. *AIS Transactions on Human-Computer Interaction*, 10(1), 1–25. https://doi.org/10.17705/1thci.00102

Dellarocas, C., Zhang, X. M., & 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. https://doi.org/10.1002/dir.20087

Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique*, 17(1), 61–84.

Dimensional Research. (2013). Customer service and business results: a survey of customer service from mid-size companies. *Dimensional Research report*, Sunnyvale, CA.

Fisher, R. J. (1993). Social desirability bias and the validity of indirect questioning. *Journal of Consumer Research*, 20(2), 303–315. https://doi.org/10.1086/209351

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1), 5228–5235.

Hilbe, J. M. (2011). Negative binomial regression (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511973420

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6(1), 1–55. https://doi.org/10.1080/10705519909540118

Hugging Face. (n.d.). *nlptown/bert-base-multilingual-uncased-sentiment* [Computer software]. https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment

Jia, Y., & Liu, I. L. B. (2018). Do consumers always follow 'useful' reviews? The interaction effect of review valence and review usefulness on consumers' purchase decisions. *Journal of the Association for Information Science and Technology*, 69(11), 1304–1317. https://doi.org/10.1002/asi.24050

Kim, S. J., Maslowska, E., & Malthouse, E. C. (2018). Understanding the effects of different review features on purchase probability. *International Journal of Advertising*, 37(1), 29–53. https://doi.org/10.1080/02650487.2017.1340928

Kline, R. B. (2023). Principles and practice of structural equation modeling (Fifth edition). The Guilford Press.

Li, H., Wang, C. (R.), Meng, F., & Zhang, Z. (2019). Making restaurant reviews useful and/or enjoyable? The impacts of temporal, explanatory, and sensory cues. *International Journal of Hospitality Management*, 83, 257–265. https://doi.org/10.1016/j.ijhm.2018.11.002

Liu, J., Cao, Y., Lin, C. Y., Huang, Y., & Zhou, M. (2007). Low-quality product review detection in opinion summarization. In J. Eisner (Ed.), Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL) (pp. 334–342). Association for Computational Linguistics. https://aclanthology.org/D07-1035/

Liu, Z., & Park, S. (2015). What makes a useful online review? Implication for travel product websites. *Tourism Management*, 47, 140–151. https://doi.org/10.1016/j.tourman.2014.09.020

Long, J. S. (1997). Regression models for categorical and limited dependent variables. Sage Publications.

Louviere, J. J., Hensher, D. A., & Swait, J. D. (2000). Stated choice methods: Analysis and applications. Cambridge University Press.

Luca, M. (2011). Reviews, reputation, and revenue: The case of Yelp.com. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1928601

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., ... & Adam, S. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, 12(2–3), 93–118.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in Econometrics* (pp. 105–142). Academic Press.

Menard, S. (2002). Applied logistic regression analysis (2nd ed.). Sage Publications.

Mudambi, S. M., & Schuff, D. (2010). What makes a helpful online review? A study of customer reviews on Amazon.com. *MIS Quarterly*, 34(1), 185–200. https://doi.org/10.2307/20721420

Muthén, L. K., & Muthén, B. O. (2017). Mplus User's Guide (8th ed.). Muthén & Muthén.

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, 78(3), 691–692.

Pan, Y., & Zhang, J. Q. (2011). Born unequal: A study of the helpfulness of user-generated product reviews. *Journal of Retailing*, 87(4), 598–612. https://doi.org/10.1016/j.jretai.2011.05.002

Park, D. H., & Lee, J. (2008). eWOM overload and its effect on consumer behavioral intention depending on consumer involvement. *Electronic Commerce Research and Applications*, 7(4), 386–398. https://doi.org/10.1016/j.elerap.2007.11.004

Park, D. H., Lee, J., & Han, I. (2007). The effect of on-line consumer reviews on consumer purchasing intention: The moderating role of involvement. *International Journal of Electronic Commerce*, 11(4), 125–148.

Petty, R. E., & Cacioppo, J. T. (1984). The effects of involvement on responses to argument quantity and quality: Central and peripheral routes to persuasion. *Journal of Personality and Social Psychology*, 46(1), 69–81. https://doi.org/10.1037/0022-3514.46.1.69

Petty, R. E., Cacioppo, J. T., & Goldman, R. (1981). Personal involvement as a determinant of argument-based persuasion. *Journal of Personality and Social Psychology*, 41(5), 847–855. https://doi.org/10.1037/0022-3514.41.5.847

RankRanger. (2021). Top websites by traffic. Retrieved December 15, 2021, from https://www.rankranger.com/top-websites

Ridout, M., Demétrio, C. G. B., & Hinde, J. (1998). Models for count data with many zeros. *International Biometric Conference*, 19, 179–192.

Sheeran, P. (2002). Intention—behavior relations: A conceptual and empirical review. *European Review of Social Psychology*, 12(1), 1–36. https://doi.org/10.1080/14792772143000003

Srivastava, V., & Kalro, A. D. (2019). Enhancing the helpfulness of online consumer reviews: The role of latent (content) factors. *Journal of Interactive Marketing*, 48, 33–50. https://doi.org/10.1016/j.intmar.2018.12.003

Sussman, S. W., & Siegal, W. S. (2003). Informational influence in organizations: An integrated approach to knowledge adoption. *Information Systems Research*, 14(1), 47–65. https://doi.org/10.1287/isre.14.1.47.14767

Xiao, L., & Li, Y. (2019). Examining the effect of positive online reviews on consumers' decision making: The valence framework. *Journal of Global Information Management*, 27(3), 159–181. https://doi.org/10.4018/JGIM.2019070109

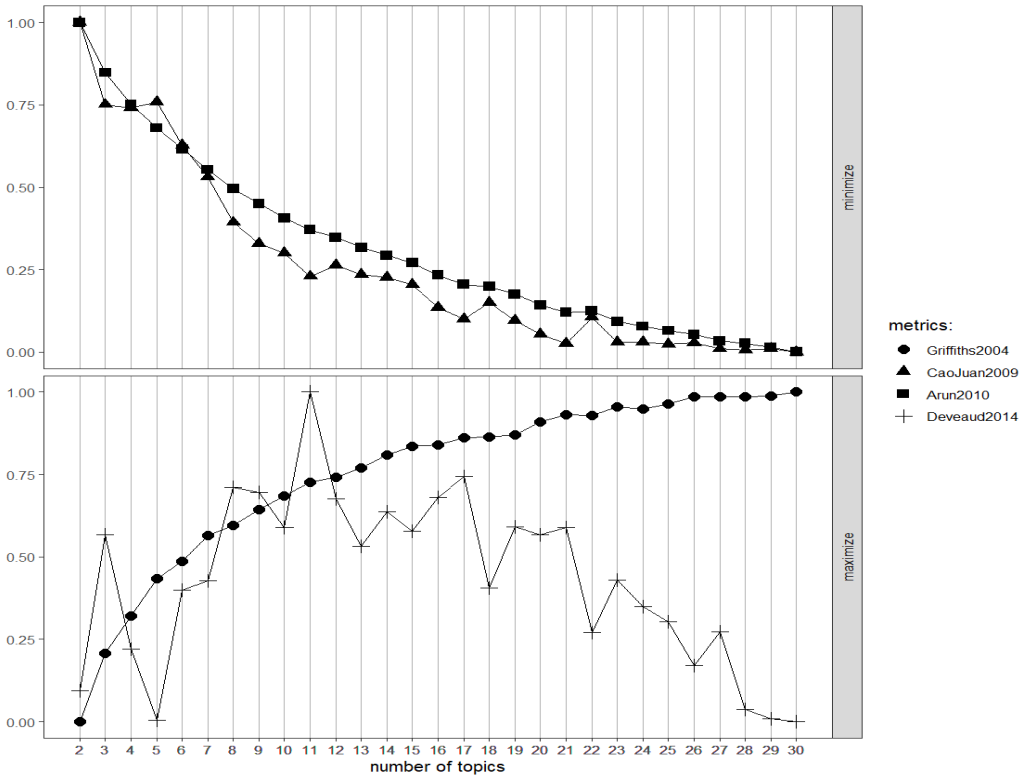
Zhang, K. Z. K., Zhao, S. J., Cheung, C. M. K., & Lee, M. K. O. (2015). Examining the influence of online reviews on consumers' decision-making: A heuristic–systematic model. *Decision Support Systems*, 67, 78–89. https://doi.org/10.1016/j.dss.2014.08.005

Zhang, L., Peng, T.-Q., Zhang, Y.-P., Wang, X.-H., & Zhu, J. J. H. (2014). Content or context: Which matters more in information processing on microblogging sites? *Computers in Human Behavior*, 31, 242–249. https://doi.org/10.1016/j.chb.2013.10.031

Zhang, M., & Luo, L. (2023). Can consumer-posted photos serve as a leading indicator of restaurant survival? Evidence from Yelp. *Management Science*, 69(1), 25–50. https://doi.org/10.1287/mnsc.2022.4359

Appendix I: LDA, Topic Number Diagnostics

Appendix I lists diagnostics for selecting the optimal number of topics in the topic modeling process, which is the most commonly used method for determining the number of topics in LDA methods. I adopted four widely used evaluation metrics to assess the number of topics.This method of combining different metrics to decide the number of topics is also suggested by Maier et al. (2018).These four metrics include Griffiths2004, based on log-likelihood, where higher values indicate better model fit (Griffiths & Steyvers, 2004). CaoJuan2009, calculates the average cosine similarity between topics, with lower values indicating better topic uniqueness (Cao et al, 2009). Arun2010, used KL divergence between document-topic and topic-word distributions, with lower values being preferred (Arun et al, 2010). Deveaud2014, relied on Jensen-Shannon divergence to assess the distance between topics, with higher values indicating a more pronounced separation (Deveaud et al, 2014).



Appendix II: Graphs Generated in Study 1 of CFA, SEM_1 & SEM_2

Appendix II contains the three sets of factor loadings and path struccture graphs generated in Study 1 from the CFA, SEM_1 and SEM_2 models.

Fig 16. CFA-generated graph (Fold 1)

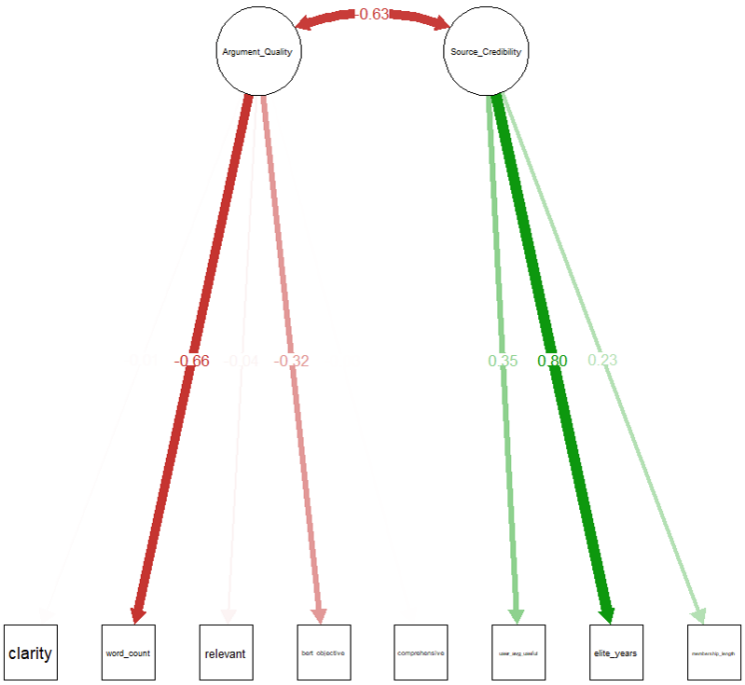


Fig 17. CFA-generated graph (Fold 2)

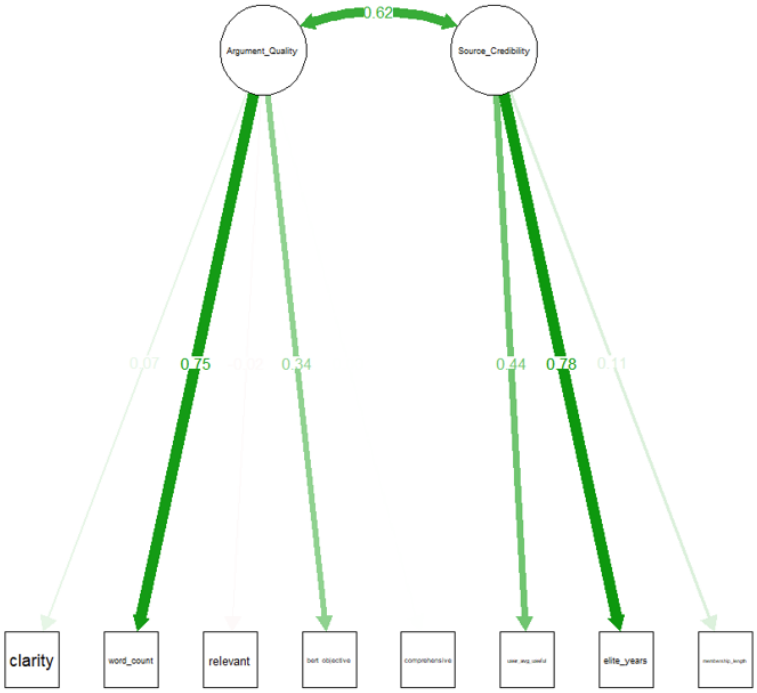


Fig 18. CFA-generated graph (Fold 3)

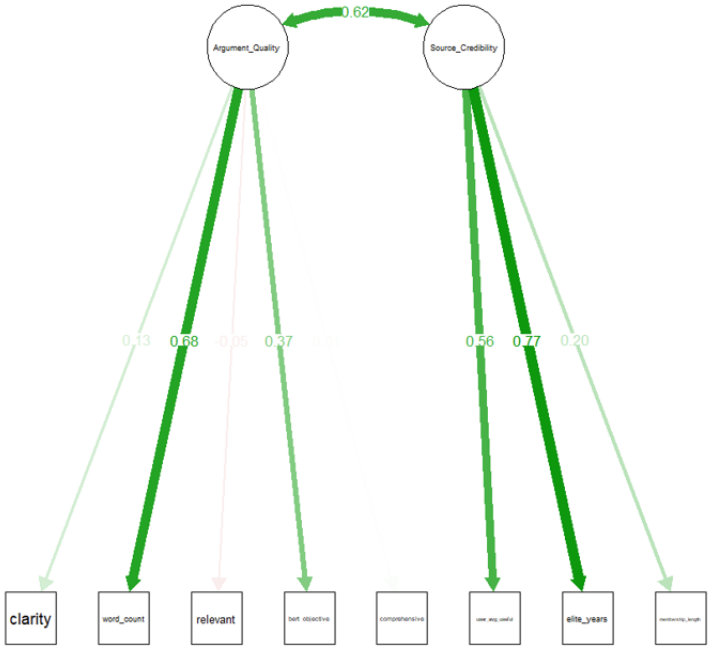


Fig 19. CFA-generated graph (Fold 4)

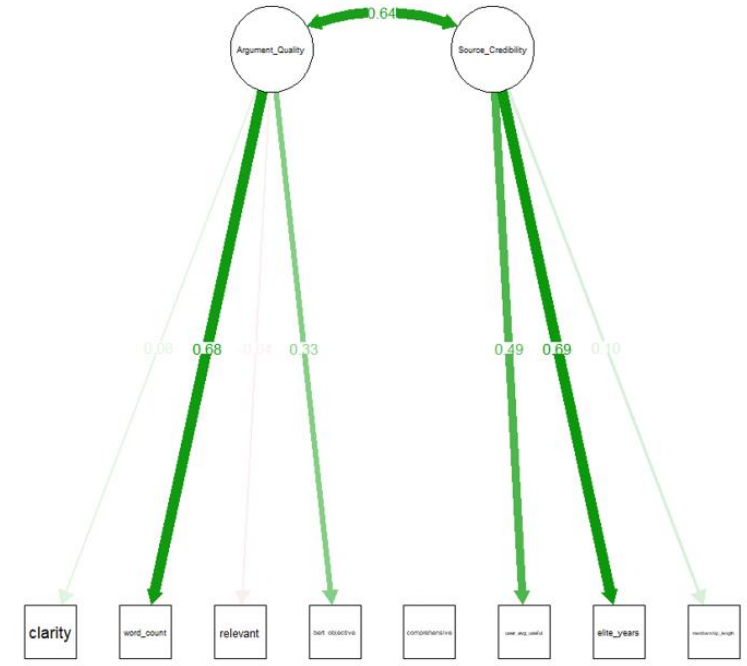


Fig 20. CFA-generated graph (Fold 5)

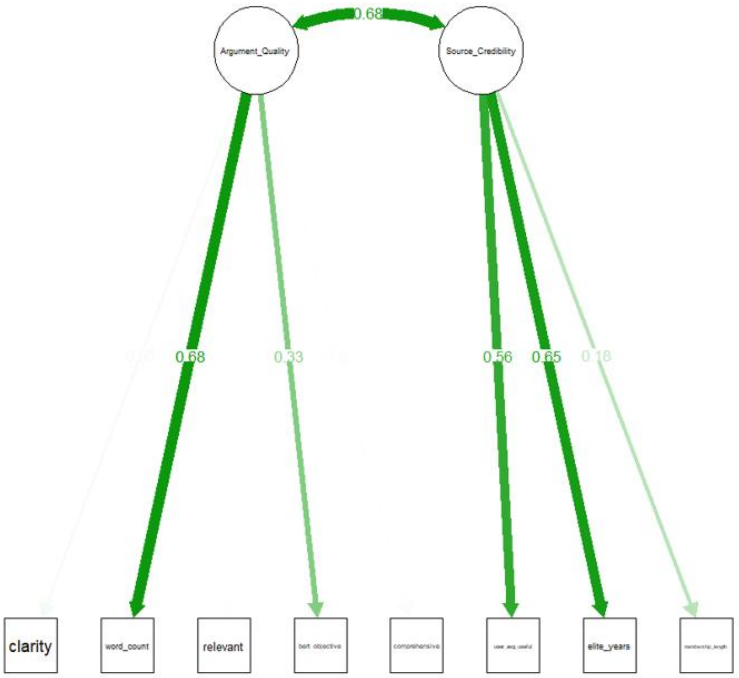


Fig 21. SEM_1-generated graph (Fold 1)

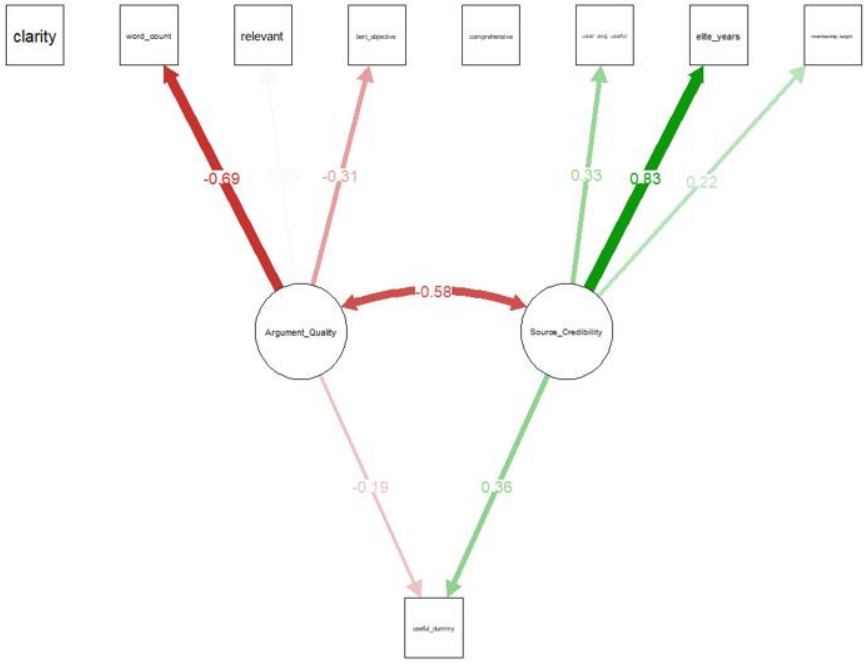


Fig 22. SEM_1-generated graph (Fold 2)

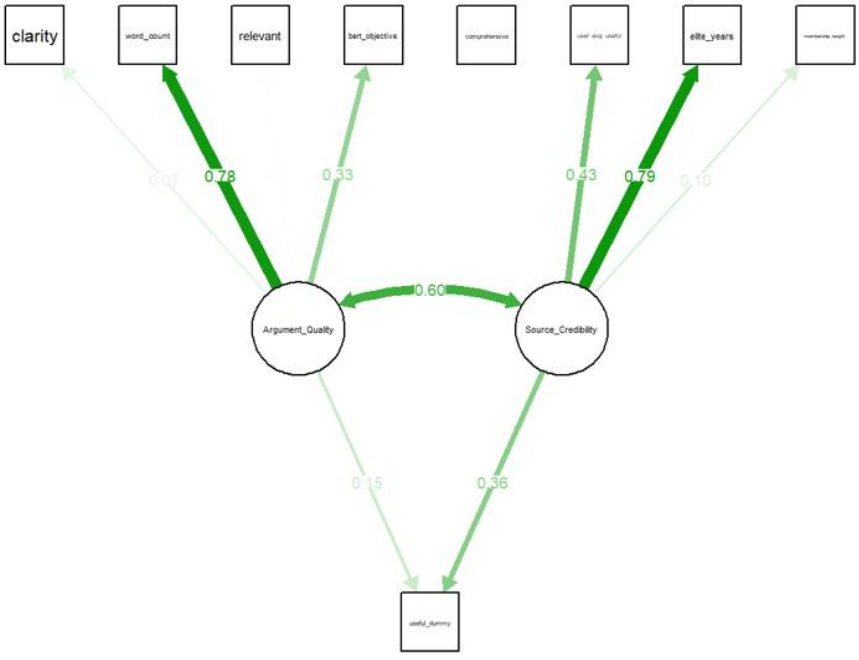


Fig 23. SEM_1-generated graph (Fold 3)

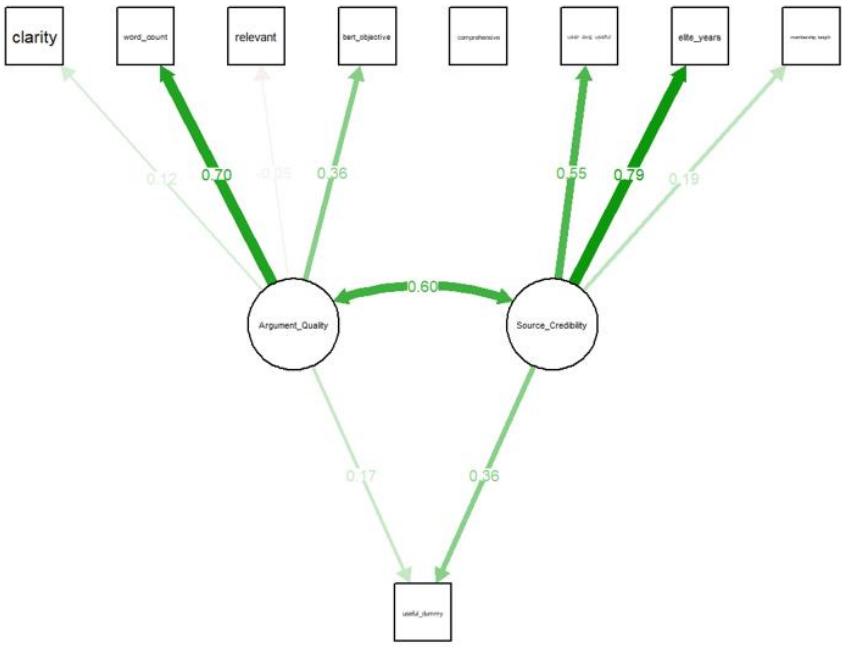


Fig 24. SEM_1-generated graph (Fold 4)

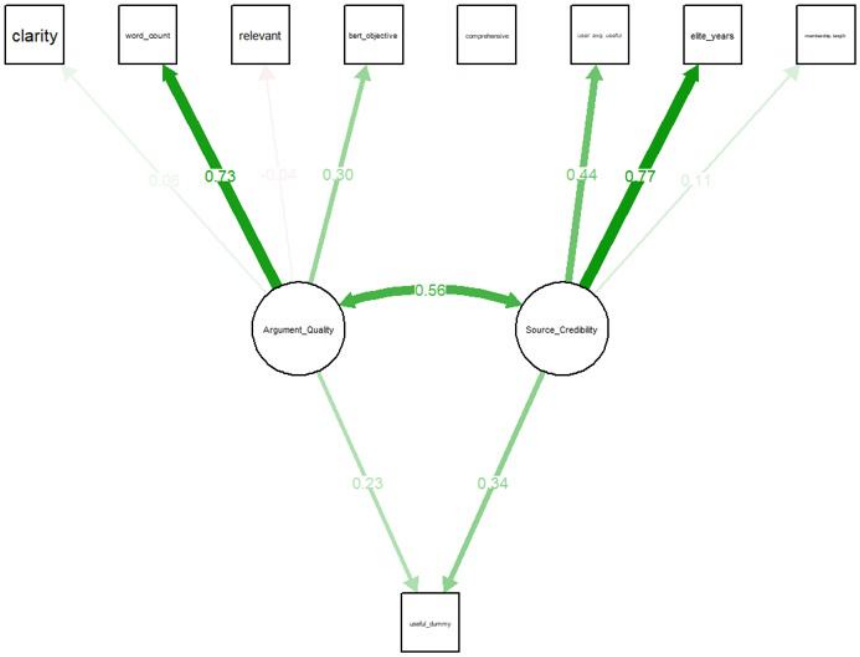


Fig 25. SEM_1-generated graph (Fold 5)

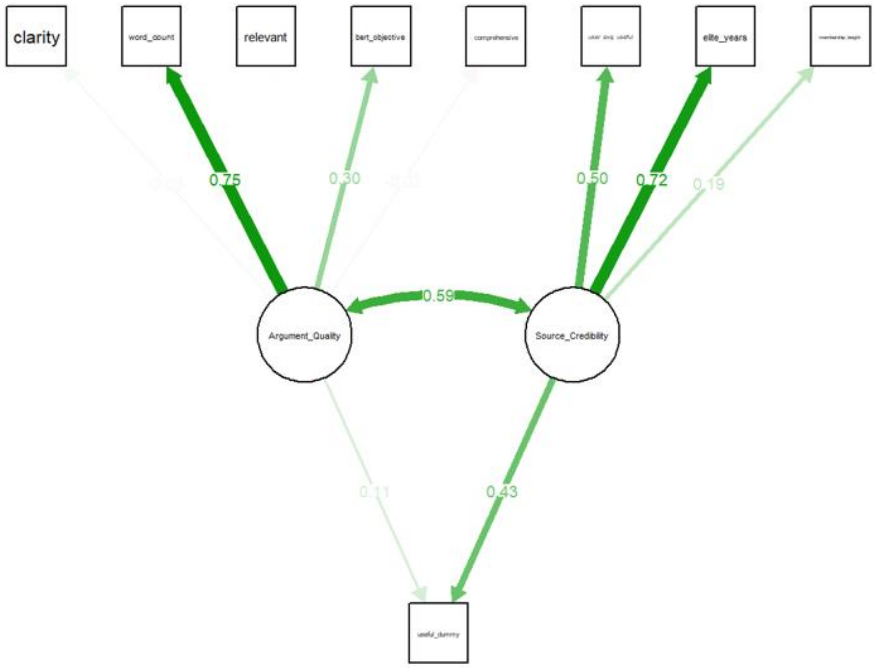


Fig 26. SEM_2-generated graph (Fold 1)

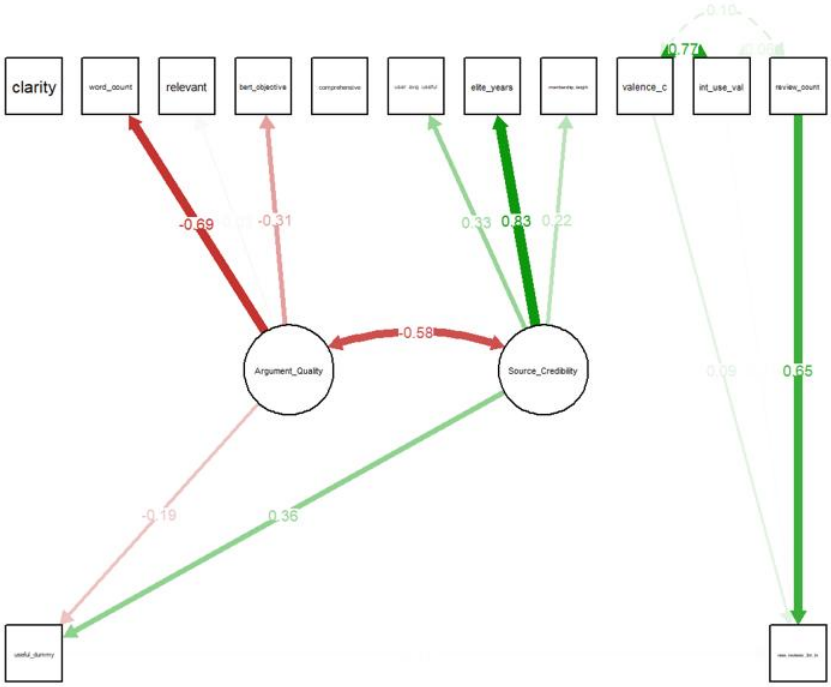


Fig 27. SEM_2-generated graph (Fold 2)

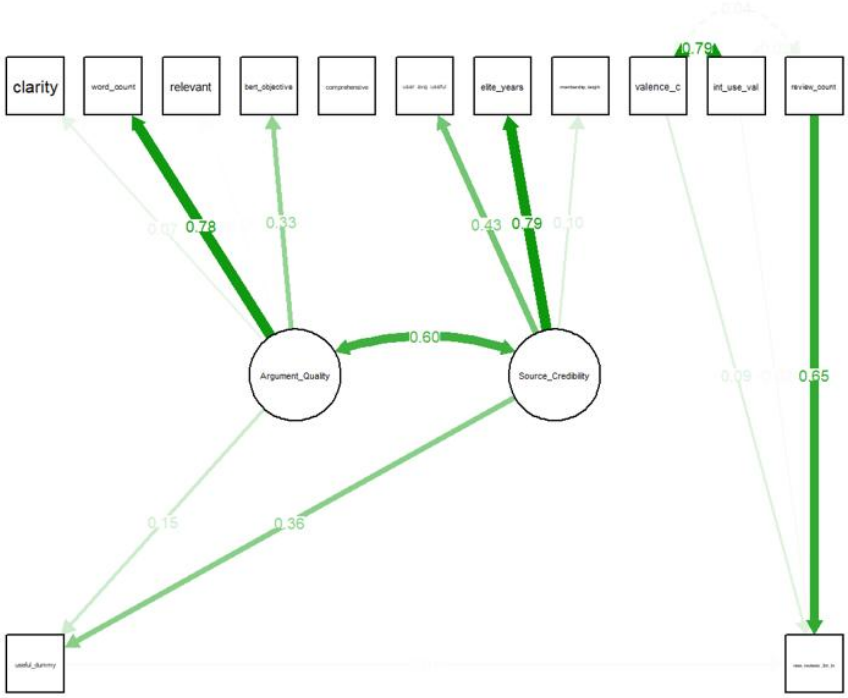


Fig 28. SEM_2-generated graph (Fold 3)

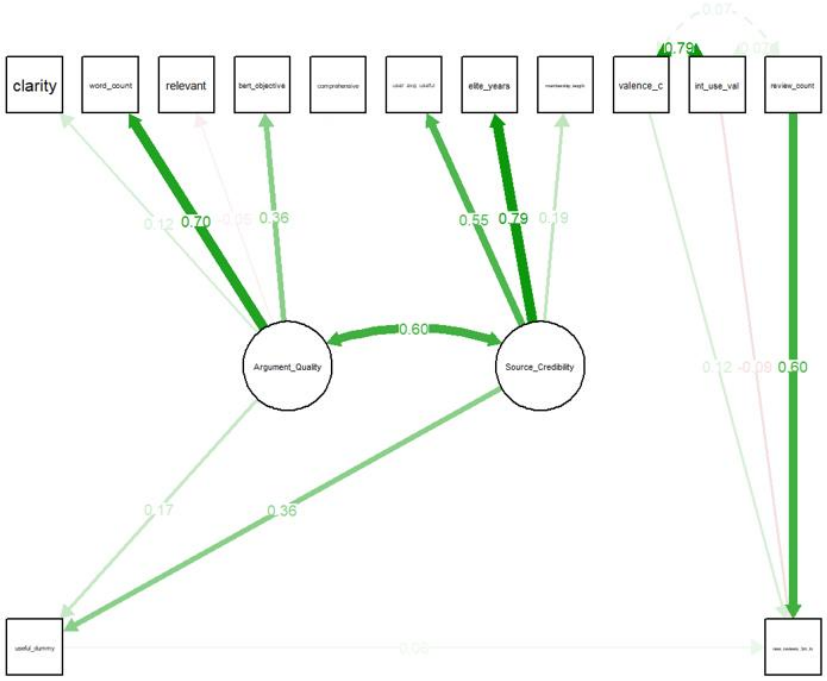


Fig 29. SEM_2-generated graph (Fold 4)

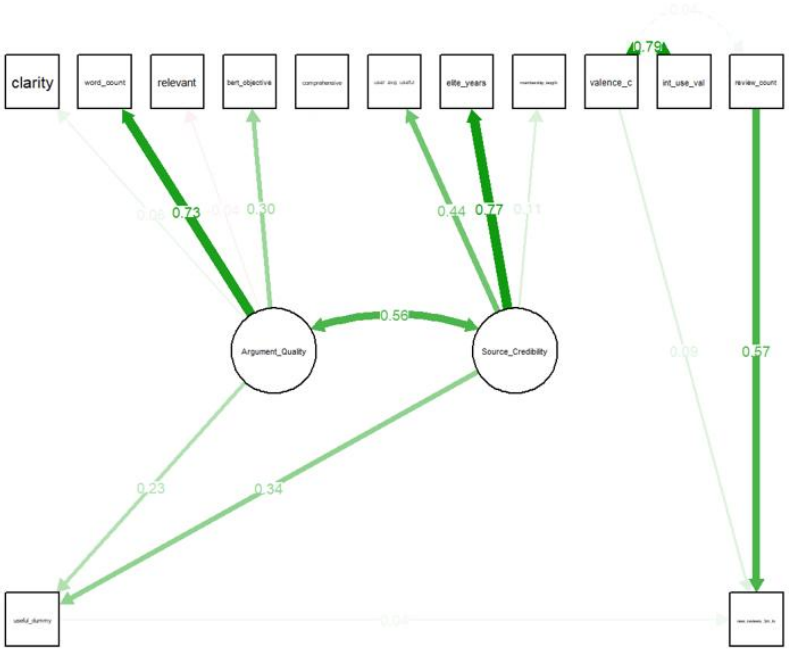
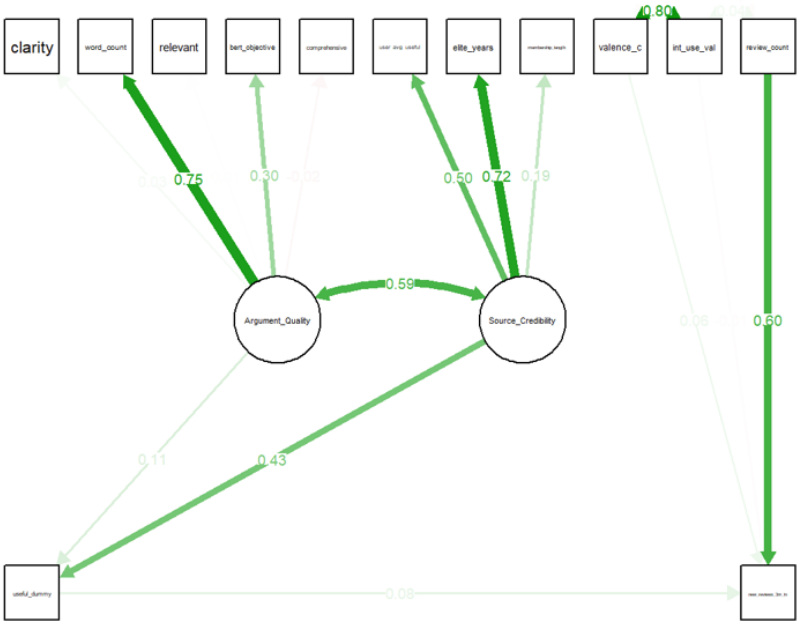


Fig 30. SEM_2-generated graph (Fold 5)



Appendix III: Tables of Results for CFA, SEM_1 and SEM_2

Appendix III contains three tables showing the complete results for CFA, SEM_1 and SEM_2 from Study 1, containing information for all five folds. The rightmost column of the icon shows the models, each representing a subset of the data for one of the five folds. All models were estimated using the R package lavaan. The following are explanations of some of the metrics in these three tables.

Lhs:	left-hand side, the left-hand side of the equation.
op:	operator indicating the type of relationship between the variables:
=~:	factor loadings (used to measure the model)
~:	regression paths (structural paths)
~~:	covariance/residual variance
rhs:	right-hand side, the right-hand side of the equation.
est:	Unstandardized estimates, i.e., path coefficients at the original scale.
se:	Standard Error
z:	z = est / se, used to test for significance
pvalue:	p-value, used to determine if the path/relationship is significant
ci.lower:	lower bound of confidence interval (usually 95% CI)
ci.upper:	upper bound of the confidence interval

std.lv : standardized path coefficient (based on latent variable variance of 1)

std.all: fully standardized path coefficient (most commonly used explanatory indicator)

std.nox: estimate based on exogenous variable standardization (not commonly used)

table 14. CFA Complete Results

lhs	op	rhs	est	se	z	P value	ci. lower	ci. upper	std. lv	std. all	model
Argument_Quality	=~	clarity	-0.008	0.042	-0.200	0.842	-0.091	0.074	-0.008	-0.008	Model 1
Argument_Quality	=~	word_count	-0.666	0.072	-9.301	0.000	-0.806	-0.525	-0.666	-0.662	Model 1
Argument_Quality	=~	relevant	-0.034	0.041	-0.840	0.401	-0.115	0.046	-0.034	-0.036	Model 1
Argument_Quality	=~	bert_objective	-0.317	0.044	-7.240	0.000	-0.402	-0.231	-0.317	-0.321	Model 1
Argument_Quality	=~	comprehensive	-0.005	0.042	-0.107	0.915	-0.088	0.079	-0.005	-0.005	Model 1
Source_Credibility	=~	user_avg_useful	0.541	0.066	8.252	0.000	0.413	0.670	0.541	0.350	Model 1
Source_Credibility	=~	elite_years	0.799	0.068	11.759	0.000	0.666	0.932	0.799	0.795	Model 1
Source_Credibility	=~	membership_length	0.233	0.040	5.831	0.000	0.154	0.311	0.233	0.232	Model 1
clarity	~~	clarity	0.994	0.044	22.359	0.000	0.907	1.081	0.994	1.000	Model 1
word_count	~~	word_count	0.567	0.091	6.214	0.000	0.388	0.746	0.567	0.561	Model 1
relevant	~~	relevant	0.927	0.042	22.338	0.000	0.846	1.008	0.927	0.999	Model 1
bert_objective	~~	bert_objective	0.874	0.044	19.924	0.000	0.788	0.960	0.874	0.897	Model 1
comprehensive	~~	comprehensive	0.998	0.045	22.360	0.000	0.911	1.086	0.998	1.000	Model 1
user_avg_useful	~~	user_avg_useful	2.104	0.106	19.827	0.000	1.896	2.312	2.104	0.878	Model 1
elite_years	~~	elite_years	0.371	0.101	3.659	0.000	0.172	0.570	0.371	0.368	Model 1
membership_length	~~	membership_length	0.951	0.044	21.584	0.000	0.864	1.037	0.951	0.946	Model 1
Argument_Quality	~~	Argument_Quality	1.000	0.000			1.000	1.000	1.000	1.000	Model 1
Source_Credibility	~~	Source_Credibility	1.000	0.000			1.000	1.000	1.000	1.000	Model 1
Argument_Quality	~~	Source_Credibility	-0.629	0.079	-7.981	0.000	-0.784	-0.475	-0.629	-0.629	Model 1

lhs	op	rhs	est	se	z	P value	ci. lower	ci. upper	std. lv	std. all	model
Argument_Quality	=~	clarity	0.071	0.039	1.819	0.069	-0.006	0.148	0.071	0.072	Model 2
Argument_Quality	=~	word_count	0.748	0.070	10.624	0.000	0.610	0.886	0.748	0.754	Model 2
Argument_Quality	=~	relevant	-0.020	0.040	-0.493	0.622	-0.098	0.059	-0.020	-0.020	Model 2
Argument_Quality	=~	bert_objective	0.346	0.044	7.873	0.000	0.260	0.433	0.346	0.336	Model 2
Argument_Quality	=~	comprehensive	0.002	0.040	0.062	0.950	-0.075	0.080	0.002	0.002	Model 2
Source_Credibility	=~	user_avg_useful	0.409	0.039	10.382	0.000	0.332	0.486	0.409	0.438	Model 2
Source_Credibility	=~	elite_years	0.779	0.058	13.340	0.000	0.664	0.893	0.779	0.779	Model 2
Source_Credibility	=~	membership_length	0.107	0.039	2.729	0.006	0.030	0.183	0.107	0.105	Model 2
clarity	~~	clarity	0.975	0.044	22.294	0.000	0.889	1.061	0.975	0.995	Model 2
word_count	~~	word_count	0.426	0.099	4.285	0.000	0.231	0.620	0.426	0.432	Model 2
relevant	~~	relevant	1.022	0.046	22.356	0.000	0.933	1.112	1.022	1.000	Model 2
bert_objective	~~	bert_objective	0.943	0.047	19.980	0.000	0.850	1.035	0.943	0.887	Model 2
comprehensive	~~	comprehensive	1.004	0.045	22.361	0.000	0.916	1.092	1.004	1.000	Model 2
user_avg_useful	~~	user_avg_useful	0.705	0.039	18.183	0.000	0.629	0.782	0.705	0.808	Model 2
elite_years	~~	elite_years	0.393	0.083	4.734	0.000	0.230	0.556	0.393	0.393	Model 2
membership_length	~~	membership_length	1.017	0.046	22.237	0.000	0.927	1.107	1.017	0.989	Model 2
Argument_Quality	~~	Argument_Quality	1.000	0.000			1.000	1.000	1.000	1.000	Model 2
Source_Credibility	~~	Source_Credibility	1.000	0.000			1.000	1.000	1.000	1.000	Model 2
Argument_Quality	~~	Source_Credibility	0.623	0.069	9.084	0.000	0.489	0.758	0.623	0.623	Model 2
Argument_Quality	=~	clarity	0.141	0.046	3.091	0.002	0.052	0.231	0.141	0.129	Model 3
Argument_Quality	=~	word_count	0.647	0.058	11.114	0.000	0.533	0.762	0.647	0.677	Model 3
Argument_Quality	=~	relevant	-0.054	0.043	-1.258	0.209	-0.139	0.030	-0.054	-0.052	Model 3
Argument_Quality	=~	bert_objective	0.365	0.042	8.587	0.000	0.282	0.448	0.365	0.373	Model 3
Argument_Quality	=~	comprehensive	0.012	0.041	0.282	0.778	-0.069	0.093	0.012	0.012	Model 3
Source_Credibility	=~	user_avg_useful	0.342	0.025	13.448	0.000	0.292	0.392	0.342	0.562	Model 3
Source_Credibility	=~	elite_years	0.758	0.048	15.798	0.000	0.664	0.852	0.758	0.765	Model 3
Source_Credibility	=~	membership_length	0.198	0.038	5.263	0.000	0.124	0.272	0.198	0.201	Model 3
clarity	~~	clarity	1.182	0.054	22.080	0.000	1.077	1.286	1.182	0.983	Model 3
word_count	~~	word_count	0.495	0.071	7.005	0.000	0.357	0.634	0.495	0.542	Model 3
relevant	~~	relevant	1.079	0.048	22.316	0.000	0.984	1.174	1.079	0.997	Model 3
bert_objective	~~	bert_objective	0.823	0.043	19.174	0.000	0.739	0.908	0.823	0.861	Model 3

lhs	op	rhs	est	se	z	P value	ci. lower	ci. upper	std. lv	std. all	model
comprehensive	~~	comprehensive	0.986	0.044	22.358	0.000	0.899	1.072	0.986	1.000	Model 3
user_avg_useful	~~	user_avg_useful	0.254	0.017	14.995	0.000	0.220	0.287	0.254	0.684	Model 3
elite_years	~~	elite_years	0.407	0.063	6.411	0.000	0.283	0.531	0.407	0.415	Model 3
membership_length	~~	membership_length	0.931	0.042	21.906	0.000	0.848	1.014	0.931	0.960	Model 3
Argument_Quality	~~	Argument_Quality	1.000	0.000			1.000	1.000	1.000	1.000	Model 3
Source_Credibility	~~	Source_Credibility	1.000	0.000			1.000	1.000	1.000	1.000	Model 3
Argument_Quality	~~	Source_Credibility	0.622	0.060	10.412	0.000	0.505	0.739	0.622	0.622	Model 3
Argument_Quality	=~	clarity	0.075	0.042	1.782	0.075	-0.007	0.157	0.075	0.076	Model 4
Argument_Quality	=~	word_count	0.681	0.074	9.261	0.000	0.537	0.825	0.681	0.677	Model 4
Argument_Quality	=~	relevant	-0.040	0.041	-0.970	0.332	-0.122	0.041	-0.040	-0.041	Model 4
Argument_Quality	=~	bert_objective	0.327	0.045	7.202	0.000	0.238	0.417	0.327	0.326	Model 4
Argument_Quality	=~	comprehensive	0.004	0.043	0.089	0.929	-0.080	0.088	0.004	0.004	Model 4
Source_Credibility	=~	user_avg_useful	0.436	0.041	10.764	0.000	0.357	0.515	0.436	0.488	Model 4
Source_Credibility	=~	elite_years	0.691	0.056	12.426	0.000	0.582	0.801	0.691	0.689	Model 4
Source_Credibility	=~	membership_length	0.102	0.042	2.434	0.015	0.020	0.183	0.102	0.100	Model 4
clarity	~~	clarity	0.979	0.044	22.258	0.000	0.893	1.066	0.979	0.994	Model 4
word_count	~~	word_count	0.547	0.096	5.711	0.000	0.359	0.735	0.547	0.541	Model 4
relevant	~~	relevant	0.960	0.043	22.331	0.000	0.875	1.044	0.960	0.998	Model 4
bert_objective	~~	bert_objective	0.903	0.046	19.669	0.000	0.813	0.993	0.903	0.894	Model 4
comprehensive	~~	comprehensive	1.032	0.046	22.360	0.000	0.942	1.123	1.032	1.000	Model 4
user_avg_useful	~~	user_avg_useful	0.609	0.038	15.965	0.000	0.534	0.683	0.609	0.762	Model 4
elite_years	~~	elite_years	0.528	0.071	7.461	0.000	0.389	0.667	0.528	0.525	Model 4
membership_length	~~	membership_length	1.030	0.046	22.208	0.000	0.939	1.121	1.030	0.990	Model 4
Argument_Quality	~~	Argument_Quality	1.000	0.000			1.000	1.000	1.000	1.000	Model 4
Source_Credibility	~~	Source_Credibility	1.000	0.000			1.000	1.000	1.000	1.000	Model 4
Argument_Quality	~~	Source_Credibility	0.642	0.075	8.511	0.000	0.494	0.790	0.642	0.642	Model 4
Argument_Quality	=~	clarity	0.026	0.038	0.664	0.507	-0.050	0.101	0.026	0.028	Model 5
Argument_Quality	=~	word_count	0.703	0.074	9.496	0.000	0.558	0.849	0.703	0.679	Model 5
Argument_Quality	=~	relevant	0.006	0.042	0.148	0.882	-0.076	0.089	0.006	0.006	Model 5
Argument_Quality	=~	bert_objective	0.327	0.044	7.407	0.000	0.241	0.414	0.327	0.329	Model 5
Argument_Quality	=~	comprehensive	-0.010	0.041	-0.252	0.801	-0.092	0.071	-0.010	-0.011	Model 5

lhs	op	rhs	est	se	z	P value	ci. lower	ci. upper	std. lv	std. all	model
Source_Credibility	=~	user_avg_useful	0.417	0.033	12.692	0.000	0.353	0.481	0.417	0.560	Model 5
Source_Credibility	=~	elite_years	0.645	0.047	13.677	0.000	0.553	0.737	0.645	0.647	Model 5
Source_Credibility	=~	membership_length	0.177	0.040	4.468	0.000	0.099	0.255	0.177	0.182	Model 5
clarity	~~	clarity	0.837	0.037	22.347	0.000	0.763	0.910	0.837	0.999	Model 5
word_count	~~	word_count	0.580	0.099	5.831	0.000	0.385	0.775	0.580	0.540	Model 5
relevant	~~	relevant	1.003	0.045	22.360	0.000	0.915	1.091	1.003	1.000	Model 5
bert_objective	~~	bert_objective	0.885	0.045	19.786	0.000	0.797	0.972	0.885	0.892	Model 5
comprehensive	~~	comprehensive	0.975	0.044	22.359	0.000	0.890	1.061	0.975	1.000	Model 5
user_avg_useful	~~	user_avg_useful	0.382	0.027	14.258	0.000	0.329	0.434	0.382	0.687	Model 5
elite_years	~~	elite_years	0.578	0.055	10.448	0.000	0.470	0.687	0.578	0.582	Model 5
membership_length	~~	membership_length	0.917	0.042	21.846	0.000	0.835	1.000	0.917	0.967	Model 5
Argument_Quality	~~	Argument_Quality	1.000	0.000			1.000	1.000	1.000	1.000	Model 5
Source_Credibility	~~	Source_Credibility	1.000	0.000			1.000	1.000	1.000	1.000	Model 5
Argument_Quality	~~	Source_Credibility	0.676	0.074	9.181	0.000	0.532	0.820	0.676	0.676	Model 5

table 15. SEM_1 Complete Results

lhs	op	rhs	est	se	z	P value	ci. lower	ci. upper	std. lv	std. all	model
Argument_Quality	=~	clarity	-0.001	0.041	-0.035	0.972	-0.082	0.079	-0.001	-0.001	Model 1
Argument_Quality	=~	word_count	-0.697	0.072	-9.678	0	-0.838	-0.556	-0.697	-0.693	Model 1
Argument_Quality	=~	relevant	-0.025	0.036	-0.68	0.496	-0.095	0.046	-0.025	-0.025	Model 1
Argument_Quality	=~	bert_objective	-0.304	0.041	-7.473	0	-0.384	-0.224	-0.304	-0.308	Model 1
Argument_Quality	=~	comprehensive	0.002	0.041	0.054	0.957	-0.078	0.083	0.002	0.002	Model 1
Source_Credibility	=~	user_avg_useful	0.516	0.09	5.744	0	0.34	0.693	0.516	0.334	Model 1
Source_Credibility	=~	elite_years	0.837	0.047	17.788	0	0.745	0.93	0.837	0.833	Model 1
Source_Credibility	=~	membership_length	0.224	0.036	6.293	0	0.154	0.293	0.224	0.223	Model 1
useful_dummy	~	Argument_Quality	-0.095	0.031	-3.045	0.002	-0.156	-0.034	-0.095	-0.194	Model 1
useful_dummy	~	Source_Credibility	0.175	0.028	6.34	0	0.121	0.229	0.175	0.357	Model 1
clarity	~~	clarity	0.994	0.116	8.588	0	0.767	1.221	0.994	1	Model 1
word_count	~~	word_count	0.524	0.111	4.735	0	0.307	0.741	0.524	0.519	Model 1
relevant	~~	relevant	0.928	0.047	19.92	0	0.836	1.019	0.928	0.999	Model 1
bert_objective	~~	bert_objective	0.882	0.04	22.273	0	0.804	0.96	0.882	0.905	Model 1
comprehensive	~~	comprehensive	0.998	0.043	23.478	0	0.915	1.082	0.998	1	Model 1
user_avg_useful	~~	user_avg_useful	2.131	1.032	2.065	0.039	0.108	4.153	2.131	0.889	Model 1
elite_years	~~	elite_years	0.308	0.072	4.255	0	0.166	0.45	0.308	0.305	Model 1
membership_length	~~	membership_length	0.955	0.047	20.26	0	0.863	1.047	0.955	0.95	Model 1
useful_dummy	~~	useful_dummy	0.181	0.007	27.544	0	0.168	0.193	0.181	0.755	Model 1
Argument_Quality	~~	Argument_Quality	1	0			1	1	1	1	Model 1
Source_Credibility	~~	Source_Credibility	1	0			1	1	1	1	Model 1
Argument_Quality	~~	Source_Credibility	-0.582	0.075	-7.795	0	-0.728	-0.436	-0.582	-0.582	Model 1
Argument_Quality	=~	clarity	0.066	0.037	1.776	0.076	-0.007	0.14	0.066	0.067	Model 2
Argument_Quality	=~	word_count	0.771	0.07	11.072	0	0.634	0.907	0.771	0.777	Model 2
Argument_Quality	=~	relevant	-0.011	0.038	-0.282	0.778	-0.085	0.064	-0.011	-0.011	Model 2
Argument_Quality	=~	bert_objective	0.337	0.041	8.22	0	0.257	0.417	0.337	0.327	Model 2
Argument_Quality	=~	comprehensive	0	0.038	-0.006	0.995	-0.075	0.075	0	0	Model 2
Source_Credibility	=~	user_avg_useful	0.404	0.049	8.206	0	0.308	0.501	0.404	0.433	Model 2
Source_Credibility	=~	elite_years	0.788	0.05	15.603	0	0.689	0.887	0.788	0.788	Model 2

lhs	op	rhs	est	se	z	P value	ci. lower	ci. upper	std. lv	std. all	model
Source_Credibility	=~	membership_length	0.106	0.04	2.62	0.009	0.027	0.185	0.106	0.104	Model 2
useful_dummy	~	Argument_Quality	0.073	0.03	2.438	0.015	0.014	0.132	0.073	0.15	Model 2
useful_dummy	~	Source_Credibility	0.176	0.028	6.268	0	0.121	0.231	0.176	0.362	Model 2
clarity	~~	clarity	0.976	0.09	10.854	0	0.799	1.152	0.976	0.996	Model 2
word_count	~~	word_count	0.391	0.104	3.77	0	0.188	0.594	0.391	0.397	Model 2
relevant	~~	relevant	1.023	0.047	21.744	0	0.93	1.115	1.023	1	Model 2
bert_objective	~~	bert_objective	0.949	0.043	22.163	0	0.865	1.033	0.949	0.893	Model 2
comprehensive	~~	comprehensive	1.004	0.043	23.522	0	0.92	1.088	1.004	1	Model 2
user_avg_useful	~~	user_avg_useful	0.709	0.192	3.703	0	0.334	1.085	0.709	0.813	Model 2
elite_years	~~	elite_years	0.379	0.073	5.186	0	0.236	0.522	0.379	0.379	Model 2
membership_length	~~	membership_length	1.017	0.049	20.738	0	0.921	1.113	1.017	0.989	Model 2
useful_dummy	~~	useful_dummy	0.184	0.006	30.312	0	0.172	0.196	0.184	0.78	Model 2
Argument_Quality	~~	Argument_Quality	1	0			1	1	1	1	Model 2
Source_Credibility	~~	Source_Credibility	1	0			1	1	1	1	Model 2
Argument_Quality	~~	Source_Credibility	0.602	0.068	8.837	0	0.469	0.736	0.602	0.602	Model 2
Argument_Quality	=~	clarity	0.132	0.045	2.952	0.003	0.045	0.22	0.132	0.121	Model 3
Argument_Quality	=~	word_count	0.672	0.064	10.42	0	0.546	0.798	0.672	0.703	Model 3
Argument_Quality	=~	relevant	-0.047	0.042	-1.104	0.27	-0.13	0.036	-0.047	-0.045	Model 3
Argument_Quality	=~	bert_objective	0.352	0.042	8.35	0	0.27	0.435	0.352	0.36	Model 3
Argument_Quality	=~	comprehensive	0.008	0.04	0.189	0.85	-0.07	0.085	0.008	0.008	Model 3
Source_Credibility	=~	user_avg_useful	0.334	0.031	10.755	0	0.273	0.395	0.334	0.548	Model 3
Source_Credibility	=~	elite_years	0.779	0.038	20.726	0	0.706	0.853	0.779	0.787	Model 3
Source_Credibility	=~	membership_length	0.19	0.033	5.682	0	0.124	0.255	0.19	0.192	Model 3
useful_dummy	~	Argument_Quality	0.083	0.028	2.985	0.003	0.028	0.137	0.083	0.171	Model 3
useful_dummy	~	Source_Credibility	0.172	0.024	7.201	0	0.126	0.219	0.172	0.357	Model 3
clarity	~~	clarity	1.184	0.121	9.761	0	0.946	1.422	1.184	0.985	Model 3
word_count	~~	word_count	0.463	0.087	5.349	0	0.293	0.632	0.463	0.506	Model 3
relevant	~~	relevant	1.08	0.054	20.064	0	0.974	1.185	1.08	0.998	Model 3
bert_objective	~~	bert_objective	0.832	0.04	20.702	0	0.754	0.911	0.832	0.87	Model 3
comprehensive	~~	comprehensive	0.986	0.041	23.893	0	0.905	1.067	0.986	1	Model 3
user_avg_useful	~~	user_avg_useful	0.259	0.04	6.427	0	0.18	0.338	0.259	0.699	Model 3

lhs	op	rhs	est	se	z	P value	ci. lower	ci. upper	std. lv	std. all	model
elite_years	~~	elite_years	0.374	0.052	7.137	0	0.272	0.477	0.374	0.381	Model 3
membership_length	~~	membership_length	0.934	0.046	20.273	0	0.844	1.025	0.934	0.963	Model 3
useful_dummy	~~	useful_dummy	0.18	0.006	30.816	0	0.169	0.192	0.18	0.77	Model 3
Argument_Quality	~~	Argument_Quality	1	0			1	1	1	1	Model 3
Source_Credibility	~~	Source_Credibility	1	0			1	1	1	1	Model 3
Argument_Quality	~~	Source_Credibility	0.599	0.065	9.231	0	0.471	0.726	0.599	0.599	Model 3
Argument_Quality	=~	clarity	0.061	0.044	1.385	0.166	-0.025	0.147	0.061	0.061	Model 4
Argument_Quality	=~	word_count	0.73	0.072	10.072	0	0.588	0.872	0.73	0.726	Model 4
Argument_Quality	=~	relevant	-0.039	0.039	-1.005	0.315	-0.116	0.037	-0.039	-0.04	Model 4
Argument_Quality	=~	bert_objective	0.304	0.044	6.856	0	0.217	0.391	0.304	0.303	Model 4
Argument_Quality	=~	comprehensive	0.006	0.042	0.144	0.885	-0.077	0.089	0.006	0.006	Model 4
Source_Credibility	=~	user_avg_useful	0.392	0.059	6.681	0	0.277	0.507	0.392	0.438	Model 4
Source_Credibility	=~	elite_years	0.77	0.054	14.157	0	0.663	0.877	0.77	0.768	Model 4
Source_Credibility	=~	membership_length	0.112	0.044	2.519	0.012	0.025	0.198	0.112	0.109	Model 4
useful_dummy	~	Argument_Quality	0.111	0.031	3.553	0	0.05	0.173	0.111	0.233	Model 4
useful_dummy	~	Source_Credibility	0.161	0.026	6.081	0	0.109	0.212	0.161	0.336	Model 4
clarity	~~	clarity	0.981	0.123	7.985	0	0.74	1.222	0.981	0.996	Model 4
word_count	~~	word_count	0.478	0.113	4.236	0	0.257	0.699	0.478	0.473	Model 4
relevant	~~	relevant	0.96	0.045	21.521	0	0.872	1.047	0.96	0.998	Model 4
bert_objective	~~	bert_objective	0.918	0.044	20.892	0	0.831	1.004	0.918	0.908	Model 4
comprehensive	~~	comprehensive	1.032	0.044	23.427	0	0.946	1.119	1.032	1	Model 4
user_avg_useful	~~	user_avg_useful	0.645	0.165	3.903	0	0.321	0.969	0.645	0.808	Model 4
elite_years	~~	elite_years	0.413	0.077	5.333	0	0.261	0.565	0.413	0.411	Model 4
membership_length	~~	membership_length	1.028	0.048	21.273	0	0.933	1.123	1.028	0.988	Model 4
useful_dummy	~~	useful_dummy	0.17	0.006	26.676	0	0.158	0.183	0.17	0.745	Model 4
Argument_Quality	~~	Argument_Quality	1	0			1	1	1	1	Model 4
Source_Credibility	~~	Source_Credibility	1	0			1	1	1	1	Model 4
Argument_Quality	~~	Source_Credibility	0.562	0.083	6.782	0	0.399	0.724	0.562	0.562	Model 4
Argument_Quality	=~	clarity	0.026	0.037	0.698	0.485	-0.046	0.098	0.026	0.028	Model 5
Argument_Quality	=~	word_count	0.774	0.096	8.061	0	0.586	0.962	0.774	0.747	Model 5
Argument_Quality	=~	relevant	0.007	0.038	0.183	0.855	-0.068	0.081	0.007	0.007	Model 5

lhs	op	rhs	est	se	z	P value	ci. lower	ci. upper	std. lv	std. all	model
Argument_Quality	=~	bert_objective	0.297	0.043	6.972	0	0.214	0.381	0.297	0.298	Model 5
Argument_Quality	=~	comprehensive	-0.024	0.042	-0.572	0.567	-0.105	0.058	-0.024	-0.024	Model 5
Source_Credibility	=~	user_avg_useful	0.375	0.041	9.136	0	0.295	0.456	0.375	0.503	Model 5
Source_Credibility	=~	elite_years	0.718	0.04	18.021	0	0.64	0.796	0.718	0.72	Model 5
Source_Credibility	=~	membership_length	0.181	0.042	4.31	0	0.099	0.263	0.181	0.186	Model 5
useful_dummy	~	Argument_Quality	0.055	0.028	1.972	0.049	0	0.109	0.055	0.114	Model 5
useful_dummy	~	Source_Credibility	0.209	0.025	8.496	0	0.16	0.257	0.209	0.434	Model 5
clarity	~~	clarity	0.837	0.092	9.074	0	0.656	1.017	0.837	0.999	Model 5
word_count	~~	word_count	0.476	0.151	3.14	0.002	0.179	0.773	0.476	0.443	Model 5
relevant	~~	relevant	1.003	0.052	19.444	0	0.902	1.104	1.003	1	Model 5
bert_objective	~~	bert_objective	0.903	0.041	22.233	0	0.824	0.983	0.903	0.911	Model 5
comprehensive	~~	comprehensive	0.975	0.044	22.132	0	0.889	1.061	0.975	0.999	Model 5
user_avg_useful	~~	user_avg_useful	0.415	0.115	3.593	0	0.188	0.641	0.415	0.747	Model 5
elite_years	~~	elite_years	0.479	0.052	9.208	0	0.377	0.581	0.479	0.482	Model 5
membership_length	~~	membership_length	0.916	0.047	19.464	0	0.824	1.008	0.916	0.966	Model 5
useful_dummy	~~	useful_dummy	0.17	0.006	27.028	0	0.158	0.183	0.17	0.74	Model 5
Argument_Quality	~~	Argument_Quality	1	0			1	1	1	1	Model 5
Source_Credibility	~~	Source_Credibility	1	0			1	1	1	1	Model 5
Argument_Quality	~~	Source_Credibility	0.587	0.09	6.499	0	0.41	0.763	0.587	0.587	Model 5

table 16. SEM_2 Complete Results

lhs	op	rhs	est	se	z	P value	ci. lower	ci. upper	std. lv	std. all	model
Argument_Quality	=~	clarity	-0.001	0.041	-0.035	0.972	-0.082	0.079	-0.001	-0.001	Model 1
Argument_Quality	=~	word_count	-0.697	0.072	-9.678	0	-0.838	-0.556	-0.697	-0.693	Model 1
Argument_Quality	=~	relevant	-0.025	0.036	-0.68	0.496	-0.095	0.046	-0.025	-0.025	Model 1
Argument_Quality	=~	bert_objective	-0.304	0.041	-7.473	0	-0.384	-0.224	-0.304	-0.308	Model 1
Argument_Quality	=~	comprehensive	0.002	0.041	0.054	0.957	-0.078	0.083	0.002	0.002	Model 1
Source_Credibility	=~	user_avg_useful	0.516	0.09	5.744	0	0.34	0.693	0.516	0.334	Model 1
Source_Credibility	=~	elite_years	0.837	0.047	17.787	0	0.745	0.93	0.837	0.833	Model 1
Source_Credibility	=~	membership_length	0.224	0.036	6.293	0	0.154	0.293	0.224	0.223	Model 1
useful_dummy	~	Argument_Quality	-0.095	0.031	-3.045	0.002	-0.156	-0.034	-0.095	-0.194	Model 1
useful_dummy	~	Source_Credibility	0.175	0.028	6.34	0	0.121	0.229	0.175	0.357	Model 1
new_reviews_3m_ln	~	useful_dummy	0.023	0.057	0.398	0.691	-0.088	0.133	0.023	0.01	Model 1
new_reviews_3m_ln	~	valence_c	0.142	0.059	2.395	0.017	0.026	0.258	0.142	0.088	Model 1
new_reviews_3m_ln	~	int_use_val	-0.018	0.08	-0.222	0.824	-0.174	0.139	-0.018	-0.008	Model 1
new_reviews_3m_ln	~	review_count	0.865	0.047	18.409	0	0.773	0.957	0.865	0.65	Model 1
clarity	~~	clarity	0.994	0.116	8.588	0	0.767	1.221	0.994	1	Model 1
word_count	~~	word_count	0.524	0.111	4.735	0	0.307	0.741	0.524	0.519	Model 1
relevant	~~	relevant	0.928	0.047	19.92	0	0.836	1.019	0.928	0.999	Model 1
bert_objective	~~	bert_objective	0.882	0.04	22.273	0	0.804	0.96	0.882	0.905	Model 1
comprehensive	~~	comprehensive	0.998	0.043	23.478	0	0.915	1.082	0.998	1	Model 1
user_avg_useful	~~	user_avg_useful	2.131	1.032	2.065	0.039	0.108	4.153	2.131	0.889	Model 1
elite_years	~~	elite_years	0.308	0.072	4.255	0	0.166	0.45	0.308	0.305	Model 1
membership_length	~~	membership_length	0.955	0.047	20.26	0	0.863	1.047	0.955	0.95	Model 1
useful_dummy	~~	useful_dummy	0.181	0.007	27.544	0	0.168	0.193	0.181	0.755	Model 1
new_reviews_3m_ln	~~	new_reviews_3m_ln	0.753	0.032	23.226	0	0.69	0.817	0.753	0.561	Model 1
Argument_Quality	~~	Argument_Quality	1	0			1	1	1	1	Model 1
Source_Credibility	~~	Source_Credibility	1	0			1	1	1	1	Model 1
Argument_Quality	~~	Source_Credibility	-0.582	0.075	-7.795	0	-0.728	-0.436	-0.582	-0.582	Model 1
valence_c	~~	valence_c	0.51	0			0.51	0.51	0.51	1	Model 1
valence_c	~~	int_use_val	0.306	0			0.306	0.306	0.306	0.773	Model 1

lhs	op	rhs	est	se	z	P value	ci. lower	ci. upper	std. lv	std. all	model
valence_c	~~	review_count	0.061	0			0.061	0.061	0.061	0.097	Model 1
int_use_val	~~	int_use_val	0.308	0			0.308	0.308	0.308	1	Model 1
int_use_val	~~	review_count	0.03	0			0.03	0.03	0.03	0.063	Model 1
review_count	~~	review_count	0.757	0			0.757	0.757	0.757	1	Model 1
Argument_Quality	=~	clarity	0.066	0.037	1.776	0.076	-0.007	0.14	0.066	0.067	Model 2
Argument_Quality	=~	word_count	0.771	0.07	11.072	0	0.634	0.907	0.771	0.777	Model 2
Argument_Quality	=~	relevant	-0.011	0.038	-0.282	0.778	-0.085	0.064	-0.011	-0.011	Model 2
Argument_Quality	=~	bert_objective	0.337	0.041	8.22	0	0.257	0.417	0.337	0.327	Model 2
Argument_Quality	=~	comprehensive	0	0.038	-0.006	0.995	-0.075	0.075	0	0	Model 2
Source_Credibility	=~	user_avg_useful	0.404	0.049	8.206	0	0.308	0.501	0.404	0.433	Model 2
Source_Credibility	=~	elite_years	0.788	0.05	15.603	0	0.689	0.887	0.788	0.788	Model 2
Source_Credibility	=~	membership_length	0.106	0.04	2.62	0.009	0.027	0.185	0.106	0.104	Model 2
useful_dummy	~	Argument_Quality	0.073	0.03	2.438	0.015	0.014	0.132	0.073	0.15	Model 2
useful_dummy	~	Source_Credibility	0.176	0.028	6.268	0	0.121	0.231	0.176	0.362	Model 2
new_reviews_3m_ln	~	useful_dummy	0.035	0.058	0.599	0.549	-0.079	0.148	0.035	0.015	Model 2
new_reviews_3m_ln	~	valence_c	0.142	0.063	2.255	0.024	0.019	0.265	0.142	0.094	Model 2
new_reviews_3m_ln	~	int_use_val	-0.041	0.077	-0.538	0.591	-0.192	0.109	-0.041	-0.022	Model 2
new_reviews_3m_ln	~	review_count	0.764	0.052	14.56	0	0.661	0.867	0.764	0.65	Model 2
clarity	~~	clarity	0.976	0.09	10.854	0	0.799	1.152	0.976	0.996	Model 2
word_count	~~	word_count	0.391	0.104	3.77	0	0.188	0.594	0.391	0.397	Model 2
relevant	~~	relevant	1.023	0.047	21.744	0	0.93	1.115	1.023	1	Model 2
bert_objective	~~	bert_objective	0.949	0.043	22.163	0	0.865	1.033	0.949	0.893	Model 2
comprehensive	~~	comprehensive	1.004	0.043	23.522	0	0.92	1.088	1.004	1	Model 2
user_avg_useful	~~	user_avg_useful	0.709	0.192	3.703	0	0.334	1.085	0.709	0.813	Model 2
elite_years	~~	elite_years	0.379	0.073	5.186	0	0.236	0.522	0.379	0.379	Model 2
membership_length	~~	membership_length	1.017	0.049	20.738	0	0.921	1.113	1.017	0.989	Model 2
useful_dummy	~~	useful_dummy	0.184	0.006	30.312	0	0.172	0.196	0.184	0.78	Model 2
new_reviews_3m_ln	~~	new_reviews_3m_ln	0.76	0.037	20.438	0	0.687	0.833	0.76	0.566	Model 2
Argument_Quality	~~	Argument_Quality	1	0			1	1	1	1	Model 2
Source_Credibility	~~	Source_Credibility	1	0			1	1	1	1	Model 2
Argument_Quality	~~	Source_Credibility	0.602	0.068	8.837	0	0.469	0.736	0.602	0.602	Model 2

lhs	op	rhs	est	se	z	P value	ci. lower	ci. upper	std. lv	std. all	model
valence_c	~~	valence_c	0.594	0			0.594	0.594	0.594	1	Model 2
valence_c	~~	int_use_val	0.366	0			0.366	0.366	0.366	0.785	Model 2
valence_c	~~	review_count	0.033	0			0.033	0.033	0.033	0.043	Model 2
int_use_val	~~	int_use_val	0.367	0			0.367	0.367	0.367	1	Model 2
int_use_val	~~	review_count	0.014	0			0.014	0.014	0.014	0.024	Model 2
review_count	~~	review_count	0.974	0			0.974	0.974	0.974	1	Model 2
Argument_Quality	=~	clarity	0.132	0.045	2.952	0.003	0.045	0.22	0.132	0.121	Model 3
Argument_Quality	=~	word_count	0.672	0.064	10.42	0	0.546	0.798	0.672	0.703	Model 3
Argument_Quality	=~	relevant	-0.047	0.042	-1.104	0.27	-0.13	0.036	-0.047	-0.045	Model 3
Argument_Quality	=~	bert_objective	0.352	0.042	8.35	0	0.27	0.435	0.352	0.36	Model 3
Argument_Quality	=~	comprehensive	0.008	0.04	0.189	0.85	-0.07	0.085	0.008	0.008	Model 3
Source_Credibility	=~	user_avg_useful	0.334	0.031	10.755	0	0.273	0.395	0.334	0.548	Model 3
Source_Credibility	=~	elite_years	0.779	0.038	20.726	0	0.706	0.853	0.779	0.787	Model 3
Source_Credibility	=~	membership_length	0.19	0.033	5.682	0	0.124	0.255	0.19	0.192	Model 3
useful_dummy	~	Argument_Quality	0.083	0.028	2.985	0.003	0.028	0.137	0.083	0.171	Model 3
useful_dummy	~	Source_Credibility	0.172	0.024	7.201	0	0.126	0.219	0.172	0.357	Model 3
new_reviews_3m_ln	~	useful_dummy	0.128	0.059	2.171	0.03	0.012	0.243	0.128	0.056	Model 3
new_reviews_3m_ln	~	valence_c	0.168	0.063	2.667	0.008	0.044	0.291	0.168	0.118	Model 3
new_reviews_3m_ln	~	int_use_val	-0.168	0.077	-2.17	0.03	-0.319	-0.016	-0.168	-0.093	Model 3
new_reviews_3m_ln	~	review_count	1.076	0.085	12.685	0	0.91	1.242	1.076	0.598	Model 3
clarity	~~	clarity	1.184	0.121	9.761	0	0.946	1.422	1.184	0.985	Model 3
word_count	~~	word_count	0.463	0.087	5.349	0	0.293	0.632	0.463	0.506	Model 3
relevant	~~	relevant	1.08	0.054	20.064	0	0.974	1.185	1.08	0.998	Model 3
bert_objective	~~	bert_objective	0.832	0.04	20.702	0	0.754	0.911	0.832	0.87	Model 3
comprehensive	~~	comprehensive	0.986	0.041	23.893	0	0.905	1.067	0.986	1	Model 3
user_avg_useful	~~	user_avg_useful	0.259	0.04	6.427	0	0.18	0.338	0.259	0.699	Model 3
elite_years	~~	elite_years	0.374	0.052	7.137	0	0.272	0.477	0.374	0.381	Model 3
membership_length	~~	membership_length	0.934	0.046	20.272	0	0.844	1.025	0.934	0.963	Model 3
useful_dummy	~~	useful_dummy	0.18	0.006	30.816	0	0.169	0.192	0.18	0.77	Model 3
new_reviews_3m_ln	~~	new_reviews_3m_ln	0.781	0.033	23.461	0	0.716	0.846	0.781	0.632	Model 3
Argument_Quality	~~	Argument_Quality	1	0			1	1	1	1	Model 3

lhs	op	rhs	est	se	z	P value	ci. lower	ci. upper	std. lv	std. all	model
Source_Credibility	~~	Source_Credibility	1	0			1	1	1	1	Model 3
Argument_Quality	~~	Source_Credibility	0.599	0.065	9.231	0	0.471	0.726	0.599	0.599	Model 3
valence_c	~~	valence_c	0.616	0			0.616	0.616	0.616	1	Model 3
valence_c	~~	int_use_val	0.381	0			0.381	0.381	0.381	0.787	Model 3
valence_c	~~	review_count	0.035	0			0.035	0.035	0.035	0.072	Model 3
int_use_val	~~	int_use_val	0.381	0			0.381	0.381	0.381	1	Model 3
int_use_val	~~	review_count	0.027	0			0.027	0.027	0.027	0.07	Model 3
review_count	~~	review_count	0.381	0			0.381	0.381	0.381	1	Model 3
Argument_Quality	=~	clarity	0.061	0.044	1.385	0.166	-0.025	0.147	0.061	0.061	Model 4
Argument_Quality	=~	word_count	0.73	0.072	10.072	0	0.588	0.872	0.73	0.726	Model 4
Argument_Quality	=~	relevant	-0.039	0.039	-1.005	0.315	-0.116	0.037	-0.039	-0.04	Model 4
Argument_Quality	=~	bert_objective	0.304	0.044	6.856	0	0.217	0.391	0.304	0.303	Model 4
Argument_Quality	=~	comprehensive	0.006	0.042	0.144	0.885	-0.077	0.089	0.006	0.006	Model 4
Source_Credibility	=~	user_avg_useful	0.392	0.059	6.681	0	0.277	0.507	0.392	0.438	Model 4
Source_Credibility	=~	elite_years	0.77	0.054	14.157	0	0.663	0.877	0.77	0.768	Model 4
Source_Credibility	=~	membership_length	0.112	0.044	2.519	0.012	0.025	0.198	0.112	0.109	Model 4
useful_dummy	~	Argument_Quality	0.111	0.031	3.553	0	0.05	0.173	0.111	0.233	Model 4
useful_dummy	~	Source_Credibility	0.161	0.026	6.081	0	0.109	0.212	0.161	0.336	Model 4
new_reviews_3m_ln	~	useful_dummy	0.097	0.063	1.543	0.123	-0.026	0.219	0.097	0.039	Model 4
new_reviews_3m_ln	~	valence_c	0.139	0.064	2.173	0.03	0.014	0.265	0.139	0.091	Model 4
new_reviews_3m_ln	~	int_use_val	0	0.082	0.003	0.997	-0.16	0.16	0	0	Model 4
new_reviews_3m_ln	~	review_count	0.505	0.034	14.84	0	0.438	0.571	0.505	0.571	Model 4
clarity	~~	clarity	0.981	0.123	7.985	0	0.74	1.222	0.981	0.996	Model 4
word_count	~~	word_count	0.478	0.113	4.236	0	0.257	0.699	0.478	0.473	Model 4
relevant	~~	relevant	0.96	0.045	21.521	0	0.872	1.047	0.96	0.998	Model 4
bert_objective	~~	bert_objective	0.918	0.044	20.892	0	0.831	1.004	0.918	0.908	Model 4
comprehensive	~~	comprehensive	1.032	0.044	23.427	0	0.946	1.119	1.032	1	Model 4
user_avg_useful	~~	user_avg_useful	0.645	0.165	3.903	0	0.321	0.969	0.645	0.808	Model 4
elite_years	~~	elite_years	0.413	0.077	5.333	0	0.261	0.565	0.413	0.411	Model 4
membership_length	~~	membership_length	1.028	0.048	21.273	0	0.933	1.123	1.028	0.988	Model 4
useful_dummy	~~	useful_dummy	0.17	0.006	26.676	0	0.158	0.183	0.17	0.745	Model 4

lhs	op	rhs	est	se	z	P value	ci. lower	ci. upper	std. lv	std. all	model
new_reviews_3m_ln	~~	new_reviews_3m_ln	0.933	0.038	24.524	0	0.858	1.007	0.933	0.66	Model 4
Argument_Quality	~~	Argument_Quality	1	0			1	1	1	1	Model 4
Source_Credibility	~~	Source_Credibility	1	0			1	1	1	1	Model 4
Argument_Quality	~~	Source_Credibility	0.562	0.083	6.782	0	0.399	0.724	0.562	0.562	Model 4
valence_c	~~	valence_c	0.61	0			0.61	0.61	0.61	1	Model 4
valence_c	~~	int_use_val	0.377	0			0.377	0.377	0.377	0.786	Model 4
valence_c	~~	review_count	0.042	0			0.042	0.042	0.042	0.04	Model 4
int_use_val	~~	int_use_val	0.377	0			0.377	0.377	0.377	1	Model 4
int_use_val	~~	review_count	0.005	0			0.005	0.005	0.005	0.006	Model 4
review_count	~~	review_count	1.808	0			1.808	1.808	1.808	1	Model 4
Argument_Quality	=~	clarity	0.026	0.037	0.698	0.485	-0.046	0.098	0.026	0.028	Model 5
Argument_Quality	=~	word_count	0.774	0.096	8.061	0	0.586	0.962	0.774	0.747	Model 5
Argument_Quality	=~	relevant	0.007	0.038	0.183	0.855	-0.068	0.081	0.007	0.007	Model 5
Argument_Quality	=~	bert_objective	0.297	0.043	6.972	0	0.214	0.381	0.297	0.298	Model 5
Argument_Quality	=~	comprehensive	-0.024	0.042	-0.572	0.567	-0.105	0.058	-0.024	-0.024	Model 5
Source_Credibility	=~	user_avg_useful	0.375	0.041	9.136	0	0.295	0.456	0.375	0.503	Model 5
Source_Credibility	=~	elite_years	0.718	0.04	18.021	0	0.64	0.796	0.718	0.72	Model 5
Source_Credibility	=~	membership_length	0.181	0.042	4.31	0	0.099	0.263	0.181	0.186	Model 5
useful_dummy	~	Argument_Quality	0.055	0.028	1.972	0.049	0	0.109	0.055	0.114	Model 5
useful_dummy	~	Source_Credibility	0.209	0.025	8.496	0	0.16	0.257	0.209	0.434	Model 5
new_reviews_3m_ln	~	useful_dummy	0.179	0.061	2.959	0.003	0.061	0.298	0.179	0.076	Model 5
new_reviews_3m_ln	~	valence_c	0.084	0.064	1.303	0.193	-0.042	0.21	0.084	0.056	Model 5
new_reviews_3m_ln	~	int_use_val	-0.016	0.078	-0.21	0.834	-0.17	0.137	-0.016	-0.009	Model 5
new_reviews_3m_ln	~	review_count	0.664	0.04	16.533	0	0.585	0.742	0.664	0.601	Model 5
clarity	~~	clarity	0.837	0.092	9.074	0	0.656	1.017	0.837	0.999	Model 5
word_count	~~	word_count	0.476	0.151	3.14	0.002	0.179	0.773	0.476	0.443	Model 5
relevant	~~	relevant	1.003	0.052	19.444	0	0.902	1.104	1.003	1	Model 5
bert_objective	~~	bert_objective	0.903	0.041	22.233	0	0.824	0.983	0.903	0.911	Model 5
comprehensive	~~	comprehensive	0.975	0.044	22.132	0	0.889	1.061	0.975	0.999	Model 5
user_avg_useful	~~	user_avg_useful	0.415	0.115	3.593	0	0.188	0.641	0.415	0.747	Model 5
elite_years	~~	elite_years	0.479	0.052	9.208	0	0.377	0.581	0.479	0.482	Model 5

lhs	op	rhs	est	se	z	P value	ci. lower	ci. upper	std. lv	std. all	model
membership_length	~~	membership_length	0.916	0.047	19.464	0	0.824	1.008	0.916	0.966	Model 5
useful_dummy	~~	useful_dummy	0.17	0.006	27.028	0	0.158	0.183	0.17	0.74	Model 5
new_reviews_3m_ln	~~	new_reviews_3m_ln	0.812	0.035	22.924	0	0.742	0.881	0.812	0.632	Model 5
Argument_Quality	~~	Argument_Quality	1	0			1	1	1	1	Model 5
Source_Credibility	~~	Source_Credibility	1	0			1	1	1	1	Model 5
Argument_Quality	~~	Source_Credibility	0.587	0.09	6.499	0	0.41	0.763	0.587	0.587	Model 5
valence_c	~~	valence_c	0.578	0			0.578	0.578	0.578	1	Model 5
valence_c	~~	int_use_val	0.373	0			0.373	0.373	0.373	0.803	Model 5
valence_c	~~	review_count	-0.009	0			-0.009	-0.009	-0.009	-0.012	Model 5
int_use_val	~~	int_use_val	0.373	0			0.373	0.373	0.373	1	Model 5
int_use_val	~~	review_count	0.026	0			0.026	0.026	0.026	0.041	Model 5
review_count	~~	review_count	1.053	0			1.053	1.053	1.053	1	Model 5

Appendix IV: Declaration of Artificial Intelligence Use

In the process of writing this thesis, I moderately used ChatGPT-4o, a generative artificial intelligence (AI), as an auxiliary tool to enhance research efficiency and clarity of expression. The use of AI was in three dimensions, firstly, in the early stages of conceptualizing my research direction, I brainstormed with the AI to stimulate my initial understanding of the field and the feasibility of my experimental design ideas. Using prompts such as: "Can NLP research on reviews incorporate machine learning?" and "Is it feasible for me to use the number of new comments in three months as a behavioral test?"

Second, following the data analysis phase, I used AI moderately to assist in modifying R code and solve the error reporting problems.

Finally, during the writing process, AI was used to improve the quality of the writing, in particular to correct errors in the logic of the expression, to improve the clarity and logic of the statements. Prompts used such as "Is there any ambiguity in the meaning I am expressing in this sentence and how can I make enhancements?"

It should be emphasized that all core components of this thesis, including literature review, hypothesis formulation, methodology design, and data interpretation were done independently by me, and AI was not used to generate large sections of text or to replace my independent thinking and research work. I take full academic responsibility for the content and results of this thesis.

Appendix V: R code for Data Analysis

```
# Data loading

Review <- read.csv("yelp_academic_dataset_review_short.csv", header = TRUE)

Business <- read.csv("yelp_academic_dataset_business_short.csv", header = TRUE)

User <- read.csv("yelp_academic_dataset_user.csv", header = TRUE)

head(Review)

head(Business)

head(User)


#################

# 0️⃣ Data Cleaning -----

#################

library(tidyr)

library(dplyr)

library(ggplot2)

library(stringr)

library(lubridate)

library(data.table)

library(quanteda)

library(quanteda.textstats)

library(quanteda.textplots)

library(stopwords)

library(wordcloud2)
```

```
library(topicmodels)
library(ldatuning)
library(tidytext)
library(hunspell)
library(corrplot)
library(reticulate)
library(lavaan)
library(semPlot)
library(fastDummies)
library(pscl)
library(MASS)
library(car)

# Initial exploration
summary(Business)
colSums(is.na(Business))

# Check time range
min(Review$date)
max(Review$date)

## 0.1 NAs & empty string removal ----
# Looks like missing values, but they might be empty strings
```

```
sum(Business == "", na.rm = TRUE)

# Replace empty strings with NA
Business <- Business %>%
  mutate(across(everything(), ~na_if(trimws(as.character(.)), "")))
summary(Business)
colSums(is.na(Business))  # Some businesses have no categories – remove them; other NAs can be ignored (filtered later)

Review <- Review %>%
  mutate(across(everything(), ~na_if(trimws(as.character(.)), "")))
colSums(is.na(Review))  # No missing strings or NAs in Review

Review <- Review %>% arrange(desc(date))
summary(Review)

## 0.2 Data type transformation ----
Review_clean <- Review %>% mutate(
  across(c(stars,useful,funny,cool),~ as.numeric(.x))) %>%
  mutate(date = as.Date(date))

Business_clean <- Business %>% mutate(
  across(c(postal_code,latitude,longitude,stars,review_count),~ as.numeric(.x))) %>%
```

```
  rename(business_stars = stars)


User_clean <- User %>%
select(user_id,yelping_since,useful,elite,friends,fans,average_stars) %>%
  mutate(yelping_since = as.Date(yelping_since),
         elite_years = ifelse(elite == "", 0, str_count(elite, ",") + 1),
         friend_count = ifelse(friends == "", 0, str_count(friends, ",") + 1)) %>%
  rename(user_avg_useful = useful)


colSums(is.na(Review_clean))
colSums(is.na(Business_clean))
colSums(is.na(User_clean))


## 0.3 Create DV: New reviews per quarter ----
Review_clean <- Review_clean %>%
  mutate(
    year = format(as.Date(date), "%Y"),
    month = as.numeric(format(as.Date(date), "%m")),
    quarter = case_when(
      month %in% 1:3 ~ "Q1",
      month %in% 4:6 ~ "Q2",
      month %in% 7:9 ~ "Q3",
      month %in% 10:12 ~ "Q4"
```

```
    ),
    year_quarter = paste0(year, "-", quarter)
  )

# Create a column of date 3 months after each review
Review_clean <- Review_clean %>%
  mutate(date_plus_3m = date %m+% months(3))

# Create a new dataframe for calculating new review count
review_growth <- Review_clean %>%
  select(review_id, business_id, date, date_plus_3m) %>%
  arrange(business_id,date)

# Use data.table to calculate number of new reviews in the following 3 months
setDT(review_growth)
review_growth[, new_reviews_3m := 0]

# Process each business_id separately
review_growth[, new_reviews_3m := {
  n_total <- .N
  res <- integer(n_total)
  for (i in 1:n_total) {
    t1 <- date[i]
```

```
    t2 <- date_plus_3m[i]
    # Count number of reviews within 3 months after the current one (excluding itself)
    res[i] <- sum(date > t1 & date <= t2)
  }
  res
}, by = business_id]

# Merge back to the original dataframe
Review_full <- Review_clean %>%
  left_join(select(review_growth, review_id, new_reviews_3m),
            by = "review_id")

# Filter: only keep reviews before or on 2021-09-30
Review_full <- Review_full[Review_full$date <= as.Date("2021-09-30"), ]
# Ensure date_plus_3m does not exceed dataset's max date, e.g. 2021-12-31
Review_full <- Review_full[Review_full$date_plus_3m <= as.Date("2021-12-31"), ]

# Calculate membership length = review date - yelping_since
Review_full$membership_length <- as.numeric(as.Date(Review_full$date)-
as.Date(Review_full$yelping_since))

# Calculate word count of each review
Review_full$word_count <- sapply(strsplit(Review_full$text, "\\s+"), length)
```

```
Review_full$word_count <- as.numeric(scale(Review_full$word_count, center = TRUE, scale = TRUE))

## 0.4 Data filtering & merging ----

# Check restaurant percentage among all businesses
Business_clean$contains_restaurant <- grepl("restaurant", Business_clean$categories, ignore.case = TRUE)
sum(Business_clean$contains_restaurant)/nrow(Business_clean)*100

# Filter: keep only restaurants and create a new dataframe
restaurant_business <- Business_clean %>% filter(contains_restaurant)

# Merge business with review
restaurant_review <- restaurant_business %>% left_join(Review_full , by = "business_id")
colSums(is.na(restaurant_review))

# Check the share of restaurant reviews
nrow(restaurant_review)/nrow(Review_clean)  # 67% of all reviews are for restaurants

# Keep only reviews with available user info by using inner join
restaurant_review <- restaurant_review %>% inner_join(User_clean , by = "user_id")
colSums(is.na(restaurant_review))
```

```
## 0.5 Data Exploration ----


# Visualize the distribution of new reviews within 3 months
ggplot(restaurant_review, aes(new_reviews_3m))+
  geom_histogram(fill = "#800020", color = "white")+
  geom_text(stat = "bin", aes(label = ..count..), vjust = -0.5, size = 3.2)+
  labs(title = "Distribution of New Reviews Within 3 Months Following Each Review",
       y = "Frequency",
       x = "Number of New Reviews")


colSums(is.na(restaurant_review))


# Check how many reviews are missing text
sum(is.na(restaurant_review$text))


# Check how many users left restaurant reviews
sum(n_distinct(restaurant_review$user_id))


# Visualize how many reviews each user wrote
restaurant_review %>% group_by(user_id) %>% slice(review_count) %>%
  ggplot(aes(review_count))+
  geom_histogram(fill = "#800020", color = "white")+
```

```
  geom_text(stat = "bin", aes(label = ..count..), vjust = -0.5, size = 3)+
  labs(title = "Distribution of Reviews by Each User",
       y = "Frequency",
       x = "Number of Reviews")

# Check review count by state
restaurant_review %>% group_by(state) %>%
  summarize(review_count = n())

# Remove rare states (HI, MT, NC), each with only one sample
rare_states <- c("HI", "MT", "NC")

# Remove those observations
restaurant_review <- restaurant_review[!(restaurant_review$state %in% rare_states), ]

# Save the filtered data
write.csv(restaurant_review,file = "restaurant_review.csv", row.names = FALSE)
restaurant_review <- read.csv("restaurant_review.csv",header = TRUE)

############################################################
# 2️⃣ NLP Step 1 -- Create DFM --------------------
############################################################
```

```
# Create corpus
review_corpus <- corpus(restaurant_review$text)
docnames(review_corpus) <- restaurant_review$review_id

# Tokenization
review_tokens <- tokens(review_corpus,
                        what = "word",
                        remove_punct = TRUE,
                        remove_numbers = TRUE,
                        remove_symbols = TRUE,
                        remove_url = TRUE,
                        remove_separators = TRUE)

# Remove stopwords to avoid unhelpful n-grams like "of the restaurant"
review_tokens <- tokens_remove(review_tokens, stopwords("en"))

# Create n-grams (unigrams, bigrams, trigrams)
review_tokens <- tokens_ngrams(review_tokens, n = 1:3)

# Remove custom stopwords (common but non-informative words)
custom_stopwords <- c(
  "just", "one", "us", "also", "got", "came", "can", "well", "even",
  "always", "first", "come", "wait", "never", "little", "went","definitely")
```

```
review_tokens <- tokens_remove(review_tokens, custom_stopwords)

# Create stemmed version of tokens
review_tokens_stem <- tokens_wordstem(review_tokens)

# Create document-feature matrices
review_dfm <- dfm(review_tokens)
review_dfm_stem <- dfm(review_tokens_stem)

# Remove documents with zero tokens in the DFM
kept_review_ids <- docnames(review_dfm)[rowSums(review_dfm) > 0]
kept_stem_review_ids <- docnames(review_dfm)[rowSums(review_dfm) > 0]
setequal(kept_review_ids, kept_stem_review_ids)  # Verify both DFMs filtered same documents

# Save filtered DFM
write.csv(kept_review_ids,file = "kept_review_ids.csv", row.names = FALSE)
kept_review_ids <- read.csv("kept_review_ids.csv",header = TRUE)

review_dfm <- review_dfm[kept_review_ids, ]
review_dfm_stem <- review_dfm_stem[kept_review_ids, ]
```

```
# Save DFM objects for reuse
saveRDS(review_dfm, file = "review_dfm.rds")
saveRDS(review_dfm_stem, file = "review_dfm_stem.rds")
review_dfm <- readRDS("review_dfm.rds")
review_dfm_stem <- readRDS("review_dfm_stem.rds")

# Filter original data based on DFM-kept reviews
restaurant_review_clean <- restaurant_review %>%
  filter(review_id %in% kept_review_ids)

# Save cleaned version of full dataset
write.csv(restaurant_review_clean,file = "restaurant_review_clean.csv", row.names = FALSE)
restaurant_review_clean <- read.csv("restaurant_review_clean.csv",header = TRUE)

# Top frequent features
top_words <- textstat_frequency(review_dfm, n = 100)
top_words_stem <- textstat_frequency(review_dfm_stem, n = 100)

# Plot top 20 frequent words (unstemmed)
ggplot(head(top_words,20), aes(x = reorder(feature, frequency), y = frequency)) +
  geom_col(fill = "#800020") +
  coord_flip() +
```

```
  labs(title = "Top 20 Most Frequent Features",
       x = "Feature",
       y = "Frequency") +
  theme_minimal()

# Plot top 20 frequent words (stemmed)
ggplot(head(top_words_stem,20), aes(x = reorder(feature, frequency), y = frequency)) +
  geom_col(fill = "#800020") +
  coord_flip() +
  labs(title = "Top 20 Most Frequent Features (stem)",
       x = "Feature",
       y = "Frequency") +
  theme_minimal()
############################################################
# 3️⃣ NLP Step 2 -- LDA Topic Modeling Using Stemmed Tokens ----------
############################################################

# Sample 5000 reviews
set.seed(1234)
sampled_ids <- sample(kept_review_ids, size = 5000)
review_sample <- restaurant_review_clean %>% filter(review_id %in% sampled_ids)
review_dfm_stem_sample <- review_dfm_stem[sampled_ids, ]
```

```
# Trim dfm by removing rare terms (term frequency < 5)
review_dfm_stem_sample_trim <- dfm_trim(review_dfm_stem_sample, min_termfreq =
5)

# Convert dfm to dtm for topicmodels
dtm_stem_sample <- convert(review_dfm_stem_sample_trim, to = "topicmodels")

# Save DTM for future use
saveRDS(dtm_stem_sample, file = "dtm_stem_sample.rds")
dtm_stem_sample <- readRDS("dtm_stem_sample.rds")

# Determine optimal number of topics
# (Note: ldatuning requires R 4.3+ and may be unavailable after upgrade)
result <- FindTopicsNumber(
  dtm_stem_sample,
  topics = seq(2, 30, by = 1),
  metrics = c("Griffiths2004", "CaoJuan2009", "Arun2010", "Deveaud2014"),
  method = "Gibbs",
  control = list(seed = 77),
  mc.cores = 1,
  verbose = TRUE
)
```

```
FindTopicsNumber_plot(result)


# Run LDA with 17 topics
lda_model <- LDA(dtm_stem_sample, k = 17, method = "Gibbs",
                 control = list(seed = 1234))


saveRDS(lda_model, file = "lda_model.rds")
lda_model <- readRDS("lda_model.rds")


# Analyze topic distribution (gamma matrix)
gamma_values <- tidy(lda_model, matrix = "gamma")


# Get most dominant topic per document
grouped_gammas <- gamma_values %>%
  group_by(document) %>%
  arrange(desc(gamma)) %>%
  slice(1) %>%
  group_by(topic)


# Count documents by dominant topic
grouped_gammas %>%
  tally(topic, sort=TRUE)
```

```
# Extract top 20 terms per topic (beta matrix)
top_terms <- tidy(lda_model, matrix = "beta") %>%
  group_by(topic) %>%
  slice_max(order_by = beta, n = 20) %>%
  ungroup()

print(top_terms)

# Visualize top terms by topic
top_terms %>%
  ggplot(aes(x = reorder_within(term, beta, topic),
             y = beta,
             fill = as.factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  scale_x_reordered() +
  labs(title = "Top Words in Each Topic",
       x = "Terms", y = "Beta Score") +
  theme_minimal()

# Assign custom topic labels
topic_labels <- c(
```

```
    "Tp1_Recommendation_and_Service",          # make, friend, recommend, always
    "Tp2_Dinner_and_Steak",                    # restaur, dinner, meat, steak
    "Tp3_Return_and_Impression",               # tri, place, back, come
    "Tp4_Price_and_Quality",                   # price, fresh, portion, quality
    "Tp5_Menu_and_Restaurant",                 # menu, restaur, option, item
    "Tp6_Breakfast_and_Bacon",                 # breakfast, bacon, egg, sausage
    "Tp7_Timing_and_Arrival",                  # time, now, early, day
    "Tp8_Order_and_Waffle",                    # order, get, waffle, wait
    "Tp9_Positive_Emotion",                    # good, love, amazing
    "Tp10_Burgers_and_Fries",                  # chicken, fri, burger, sandwich
    "Tp11_Soups_and_Asian_Food",               # dish, soup, noodle, asian
    "Tp12_Great_Atmosphere_and_Staff",         # great, friend, recommend, atmosphere
    "Tp13_Pizza_and_Cheesesteak",              # pizza, salad, cheesesteak
    "Tp14_Mexican_Flavors",                    # flavor, taco, mexican
    "Tp15_Bar_and_Drinks",                     # bar, dinner, drink
    "Tp16_Local_and_place",                    # local, food, clear
    "Tp17_Subjective_Feelings"                 # like, feel, though
)

# Notes:
# The following topics are considered less relevant when appearing alone,
# because they express general sentiment or subjective impressions without concrete reference to service/food quality:
```

```
# Tp9: Generalized praise (e.g., love, best, amazing) – lacks specific detail
# Tp17: Highly subjective words (like, want, feel) – not specific to service/food
# Tp15: General intent or behavior (like, go, eat) – vague rather than descriptive
# These topics may still provide useful signals when combined with other themes, but are considered weakly relevant on their own.

# Extract topic proportions (theta matrix)
theta_all <- as.data.frame(posterior(lda_model)$topics)
colnames(theta_all) <- topic_labels

# Merge with original sampled review data
sample_topic <- cbind(review_sample, theta_all)

write.csv(sample_topic, file = "sample_topic.csv", row.names = FALSE)
sample_topic <- read.csv("sample_topic.csv", header = TRUE)

# Compute a relevance score using weighted topic contribution
# Assign topic importance weights (manual weighting based on interpretability and relevance)
topic_weights <- c(
  1.0,  # Tp1: Recommendation and service
  1.0,  # Tp2: Main dinner dishes
  0.8,  # Tp3: Return intent / overall impression
```

```
  1.0,  # Tp4: Food price and quality
  0.8,  # Tp5: Menu content
  1.0,  # Tp6: Breakfast mains
  0.6,  # Tp7: Time / arrival-related topics
  0.8,  # Tp8: Order process
  0.6,  # Tp9: General positive emotions (not specific)
  1.0,  # Tp10: Burgers and fries
  1.0,  # Tp11: Asian food / soups
  1.0,  # Tp12: Atmosphere and staff service
  1.0,  # Tp13: Pizza / cheesesteaks
  1.0,  # Tp14: Mexican cuisine
  0.8,  # Tp15: Bar / drinks
  0.6,  # Tp16: Locality and location
  0.6   # Tp17: Subjective, vague personal impressions
)

# Calculate relevance score as weighted sum
topic_cols <- grep("^Tp", names(sample_topic), value = TRUE)
sample_topic$relevant <- as.numeric(as.matrix(sample_topic[, topic_cols]) %*% topic_weights)

# Visualize distribution of relevance
ggplot(sample_topic, aes(relevant))+
```

```
  geom_histogram(fill= "#800020",color="white")+
  labs(title = "Distribution of Relevance of Each Review",
       y = "Frequency",
       x = "Relevance Score")

# Example topic distribution visualizations
ggplot(sample_topic, aes(Tp1_Recommendation_and_Service))+
  geom_histogram(fill= "#800020",color="white")

ggplot(sample_topic, aes(Tp2_Dinner_and_Steak))+
  geom_histogram(fill= "#800020",color="white")

# Convert topic probabilities to binary flags (>= 0.08)
for (tp in topic_labels) {
  sample_topic[[tp]] <- if_else(sample_topic[[tp]] >= 0.08, 1, 0)
}

# Define "comprehensive" as the number of activated topics per review
sample_topic <- sample_topic %>%
  mutate(comprehensive = rowSums(across(all_of(topic_labels))))

# Visualize comprehensive topic distribution
ggplot(sample_topic, aes(comprehensive))+
```

```
  geom_bar(fill= "#800020",color="white")+
  geom_text(stat = "count", aes(label = after_stat(count)),
            vjust = -0.5, size = 3.2)+
  labs(title = "Distribution of Comprehensive Score per Review",
       y = "Frequency",
       x = "Number of Topics")


# Drop unnamed column
sample_topic <- sample_topic %>% select(-Unnamed..0)


sample_step1_done <- sample_topic
colSums(is.na(sample_step1_done))


# Clarity score via spell-checking
spelling_errors <- hunspell(sample_step1_done$text)
error_count <- sapply(spelling_errors, length)
word_count <- sapply(strsplit(sample_step1_done$text, "\\s+"), length)
spelling_error_rate <- error_count / word_count
sample_step1_done$clarity <- 1 - spelling_error_rate


############################################################
# 4️⃣ NLP step 3 -- Sentiment Analysis using BERT
############################################################
```

```
# Load reticulate package for Python integration
library(reticulate)

# Install necessary Python packages
# reticulate::install_miniconda()
py_install(c("transformers", "torch"))

# Load pre-trained multilingual BERT sentiment model
transformers <- import("transformers")
torch <- import("torch")

tokenizer <- transformers$AutoTokenizer$from_pretrained("nlptown/bert-base-multilingual-uncased-sentiment")
model <- transformers$AutoModelForSequenceClassification$from_pretrained("nlptown/bert-base-multilingual-uncased-sentiment")

# Define Python function to predict sentiment using BERT
reticulate::py_run_string("
from transformers import AutoTokenizer, AutoModelForSequenceClassification
import torch
import torch.nn.functional as F
```

```
tokenizer = AutoTokenizer.from_pretrained('nlptown/bert-base-multilingual-uncased-sentiment')
model = AutoModelForSequenceClassification.from_pretrained('nlptown/bert-base-multilingual-uncased-sentiment')

def predict_single_sentiment(text):
    inputs = tokenizer(text, return_tensors='pt', truncation=True)
    with torch.no_grad():
        outputs = model(**inputs)
        probs = F.softmax(outputs.logits, dim=1)
        probs = probs.numpy().flatten()
        label = probs.argmax() + 1  # labels are 1-indexed
    return label, probs.tolist()
")

# Wrapper function in R to predict sentiment scores for a list of texts
predict_sentiment <- function(texts) {
  results <- lapply(texts, function(txt) {
    out <- reticulate::py$predict_single_sentiment(txt)
    list(label = out[[1]], probs = unlist(out[[2]]))
  })
  return(results)
```

```
}

# Apply sentiment prediction to all reviews
texts <- sample_step1_done$text
sentiment_results <- predict_sentiment(texts)

# Save results for reuse
saveRDS(sentiment_results, file = "sentiment_results.rds")
sentiment_results <- readRDS("sentiment_results.rds")

# Add sentiment results to dataframe
sample_step1_done <- sample_step1_done %>%
  mutate(
    bert_score = sapply(sentiment_results, function(x) x$label),
    prob_1star = sapply(sentiment_results, function(x) x$probs[1]),
    prob_2star = sapply(sentiment_results, function(x) x$probs[2]),
    prob_3star = sapply(sentiment_results, function(x) x$probs[3]),
    prob_4star = sapply(sentiment_results, function(x) x$probs[4]),
    prob_5star = sapply(sentiment_results, function(x) x$probs[5])
  )

# Construct "bert_objective" as a proximity score to neutral sentiment
# Score ranges from 1 (neutral = 3 stars) to 0 (very negative or very positive)
```

```
sample_step1_done$bert_objective <- 1 - (abs(sample_step1_done$bert_score - 3) / 2)

# Add jitter to bert_objective to increase variance for linear modeling
set.seed(1234)
sample_step1_done$bert_objective <- jitter(sample_step1_done$bert_objective, amount =
0.05)

# Drop raw probability columns to reduce redundancy
sample_step1_done <- sample_step1_done %>%
  select(-prob_1star, -prob_2star, -prob_3star, -prob_4star, -prob_5star)

# Save intermediate result
sample_step2_done <- sample_step1_done

# Visualize distribution of BERT sentiment scores
ggplot(sample_step2_done,aes(bert_score))+
  geom_bar(fill= "#800020", color="white")+
  geom_text(stat = "count", aes(label = after_stat(count)),
            vjust = -0.5, size = 3.2)+
  labs(title = "Distribution of bert_score",
       y = "Frequency",
       x = "bert_score")
```

```
# Construct bert_valence as a categorical sentiment valence indicator
# Positive (1), Neutral (0), Negative (-1)
sample_step2_done <- sample_step2_done %>%
  mutate(bert_valence = case_when(
    bert_score > 3 ~  1,
    bert_score < 3 ~ -1,
    bert_score == 3 ~  0,
    TRUE ~ NA
  ))

# Check key variable summaries
summary(sample_step2_done[, c("comprehensive", "clarity", "bert_score",
"bert_objective", "bert_valence")])

############################################################
# 5️⃣ Descriptive Analysis ----------
############################################################

# Remove unnecessary columns before analysis
sample_step2_done <- sample_step2_done %>%
  select(-address, -postal_code, -latitude, -longitude, -is_open,
         -attributes, -hours)
```

```
# Check for missing values
colSums(is.na(sample_step2_done))


# Summary statistics for key numerical variables
summary(sample_step2_done[, c(
  "useful", "new_reviews_3m",
  "relevant", "comprehensive", "clarity", "bert_score",
  "bert_objective", "bert_valence", "membership_length"
)])


# Calculating SD
descriptive_stats <- sample_step2_done %>%
  select(all_of(c("useful", "new_reviews_3m",
                  "relevant", "comprehensive", "clarity", "bert_score",
                  "bert_objective", "bert_valence", "membership_length",
                  "user_avg_useful", "average_stars", "elite_years"))) %>%
  summarise(across(everything(), list(mean = ~mean(., na.rm = TRUE),
                                      sd = ~sd(., na.rm = TRUE))))


# Summary statistics for distinct users (one row per user)
summary(sample_step2_done %>%
          select(user_id, user_avg_useful, fans, average_stars, elite_years, friend_count) %>%
          distinct(user_id, .keep_all = TRUE))
```

```
# ---------- Visualization ----------
# Reshape data into long format for multi-variable boxplots
df_long <- sample_step2_done %>%
  select(all_of(c("useful", "new_reviews_3m",
                  "relevant", "comprehensive", "clarity", "bert_score",
                  "bert_objective", "bert_valence", "membership_length",
                  "user_avg_useful", "average_stars", "elite_years"))) %>%
  pivot_longer(cols = everything(),
               names_to = "variable",
               values_to = "value")

# Create independent boxplots for each variable
ggplot(df_long, aes(x = "", y = value)) +
  geom_boxplot() +
  facet_wrap(~ variable, scales = "free_y") +  # Independent Y-axis per plot
  theme_minimal() +
  labs(title = "Boxplots of Each Variable (with Independent Y Scales)", x = "", y = "") +
  theme(strip.text = element_text(size = 10))

# ---------- Correlation Matrix ----------
# Select numerical columns for correlation
selected_cols <- sample_step2_done %>%
```

```
  select(as.numeric(all_of(c("useful", "new_reviews_3m",
                             "relevant", "comprehensive", "clarity", "bert_objective", "word_count",
                             "bert_valence", "membership_length", "elite_years", "user_avg_useful"))))

# Convert logical variables to numeric if any (just in case)
selected_cols <- selected_cols %>%
  mutate(across(where(is.logical), as.numeric))

# Compute pairwise correlations (ignoring NA)
cor_matrix <- cor(selected_cols, use = "pairwise.complete.obs")

# Visualize correlation matrix
corrplot(cor_matrix, method = "color", type = "upper",
         tl.col = "black", tl.srt = 45, addCoef.col = "black",
         number.cex = 0.6, col = colorRampPalette(c("steelblue", "white", "tomato"))(200))

descriptive_stats
# Save final data for modeling
sample_step3_done <- sample_step2_done

############################################################
# 6️⃣ Step 6: CFA & SEM Modeling ----------
############################################################
```

```
#  Data Preparation: Centering + Interaction Terms
# Create useful_dummy to mitigate zero-inflation and skew of "useful" variable
sample_step3_done <- sample_step3_done %>% mutate(
  useful_dummy = if_else(useful == 0, 0, 1)
)

# Create centered valence and interaction term
sample_step3_done <- sample_step3_done %>%
  mutate(
    valence_c = scale(bert_valence, center = TRUE, scale = FALSE),
    int_use_val = useful_dummy * valence_c
  )

# Standardize observed variables for SEM estimation
sample_step3_done <- sample_step3_done %>%
  mutate(
    clarity = as.numeric(scale(clarity, center = TRUE, scale = TRUE)),
    comprehensive = as.numeric(scale(comprehensive, center = TRUE, scale = TRUE)),
    user_avg_useful = as.numeric(scale(user_avg_useful, center = TRUE, scale = TRUE)),
    elite_years = as.numeric(scale(elite_years, center = TRUE, scale = TRUE)),
    membership_length = as.numeric(scale(membership_length, center = TRUE, scale = TRUE)),
    bert_objective = as.numeric(scale(bert_objective, center = TRUE, scale = TRUE)),
    stars = as.numeric(scale(stars, center = TRUE, scale = TRUE)),
```

```
  review_count = as.numeric(scale(review_count, center = TRUE, scale = TRUE)),
  relevant = as.numeric(scale(relevant, center = TRUE, scale = TRUE))
 )

# Log-transform DV to reduce skew
sample_step3_done$new_reviews_3m_ln <- log1p(sample_step3_done$new_reviews_3m)

# Convert text variables to factor type
sample_step3_done$name <- as.factor(sample_step3_done$name)
sample_step3_done$state <- as.factor(sample_step3_done$state)
sample_step3_done$year_quarter <- as.factor(sample_step3_done$year_quarter)

summary(sample_step3_done)

# ================================
# Step 1: Split Dataset
# ================================
set.seed(123)
n <- nrow(sample_step3_done)
split_indices <- split(1:n, cut(1:n, breaks = 5, labels = FALSE))
data_list <- lapply(split_indices, function(idx) sample_step3_done[idx, ])

# ================================
```

```
# Step 2: Loop Over CFA Estimation
# ================================

# Create empty containers
cfa_results <- list()
cfa_fitmeasures <- data.frame()
cfa_results_plot <- list()

for (i in 1:5) {
  data_i <- data_list[[i]]

  cfa_model <- '
    Argument_Quality =~ clarity + word_count + relevant + bert_objective + comprehensive
    Source_Credibility =~ user_avg_useful + elite_years + membership_length
  '

  fit_cfa <- cfa(cfa_model, data = data_i, std.lv = TRUE)
  cfa_results[[i]] <- parameterEstimates(fit_cfa, standardized = TRUE)
  cfa_results_plot[[i]] <- fit_cfa

  fitm <- fitMeasures(fit_cfa, c("cfi", "tli", "rmsea", "srmr"))
  cfa_fitmeasures <- rbind(cfa_fitmeasures,
                           data.frame(set = i,
```

```
                        cfi = fitm["cfi"],
                        tli = fitm["tli"],
                        rmsea = fitm["rmsea"],
                        srmr = fitm["srmr"]))
}

# Output parameter estimates and fit indices
cfa_results[[1]]
cfa_results[[2]]
cfa_results[[3]]
cfa_results[[4]]
cfa_results[[5]]

cfa_fitmeasures

# ================================
# Step 3: CFA Visualization
# ================================

for (i in seq_along(cfa_results_plot)) {
  cat("\n📊 Displaying SEM path diagram for subset", i, "...\n")

  semPaths(
```

```
  cfa_results_plot[[i]],
  what = "std",              # Standardized path coefficients
  layout = "tree",           # Tree structure
  style = "lisrel",          # LISREL-style layout
  nCharNodes = 0,            # Show full variable names
  residuals = FALSE,         # Hide residual arrows
  title = FALSE,             # Hide diagram title
  sig = 0.05,                # Significance threshold for path display
  edge.label.sig = TRUE      # Add significance stars
 )

 # Pause to review each diagram
 readline(prompt = "Press Enter to continue...")
}

# ================================
# Step 4: SEM_1 Analysis Loop
# ================================

# Create empty data structures
sem_results <- list()
sem_results_plot <- list()
sem_fitmeasures <- data.frame()
```

```
for (i in 1:5) {
  data_i <- data_list[[i]]

  sem_model <- '
    Argument_Quality =~ clarity  +  word_count + relevant + bert_objective + comprehensive
    Source_Credibility =~ user_avg_useful + elite_years + membership_length
    useful_dummy ~ Argument_Quality + Source_Credibility
  '
  # Due to the non-normality of useful, a Robust Maximum Likelihood method (MLR) was selected
  fit_sem <- sem(sem_model, data = data_i, std.lv = TRUE, estimator = "MLR")

  sem_results[[i]] <- parameterEstimates(fit_sem, standardized = TRUE)
  # Save the complete model object (not just summary or parameters)
  sem_results_plot[[i]] <- fit_sem

  # Store fit indices
  fitm <- fitMeasures(fit_sem, c("cfi", "tli", "rmsea", "srmr"))
  sem_fitmeasures <- rbind(sem_fitmeasures,
                           data.frame(set = i,
                                      cfi = fitm["cfi"],
                                      tli = fitm["tli"],
                                      rmsea = fitm["rmsea"],
```

```
                          srmr = fitm["srmr"]))
}

# Fit indices
sem_fitmeasures

# Parameter estimates
sem_results[[1]]
sem_results[[2]]
sem_results[[3]]
sem_results[[4]]
sem_results[[5]]

# Visualization
for (i in seq_along(sem_results_plot)) {
  cat("\n📊 Displaying SEM Path Diagram for Subset", i, "...\n")

  semPaths(
    sem_results_plot[[i]],
    what = "std",             # Use standardized path coefficients
    layout = "tree",          # Tree layout
    style = "lisrel",         # LISREL style
    nCharNodes = 0,           # Do not truncate variable names
```

```
  residuals = FALSE,         # Do not show residual arrows
  title = FALSE,             # Do not display title
  sig = 0.05,                # Significance threshold for asterisks
  edge.label.sig = TRUE      # Display significance asterisks
 )

 # Optional: Pause to view each diagram
 readline(prompt = "Press Enter to view the next diagram...")
}

###################################
# Step 5: Full SEM Model
###################################
# Initialize storage lists
sem_full_results <- list()
sem_full_results_plot <- list()
sem_full_fitmeasures <- data.frame()

for (i in 1:5) {
 data_i <- data_list[[i]]

 # SEM Model: includes structural path from useful to new_reviews_3m
 sem_model_full <- '
```

```
  # Measurement model
  Argument_Quality =~ clarity + word_count + relevant + bert_objective + comprehensive
  Source_Credibility =~ user_avg_useful + elite_years + membership_length

  # Structural model
  useful_dummy ~ Argument_Quality + Source_Credibility
  new_reviews_3m_ln ~ useful_dummy + valence_c + int_use_val + review_count
'

# Robust estimator (MLR) used due to count-type outcome
fit_sem_full <- sem(sem_model_full, data = data_i, std.lv = TRUE, estimator = "MLR")

# Save results
sem_full_results[[i]] <- parameterEstimates(fit_sem_full, standardized = TRUE)
sem_full_results_plot[[i]] <- fit_sem_full

# Extract fit indices
fitm <- fitMeasures(fit_sem_full, c("cfi", "tli", "rmsea", "srmr"))
sem_full_fitmeasures <- rbind(sem_full_fitmeasures,
                              data.frame(set = i,
                                         cfi = fitm["cfi"],
                                         tli = fitm["tli"],
                                         rmsea = fitm["rmsea"],
```

```
                                srmr = fitm["srmr"]))
}

# Fit indices
sem_full_fitmeasures

# Parameter estimates
sem_full_results[[1]]
sem_full_results[[2]]
sem_full_results[[3]]
sem_full_results[[4]]
sem_full_results[[5]]

# Visualization of full SEM path diagrams
for (i in seq_along(sem_full_results_plot)) {
  cat("\n📊 Displaying Full SEM Path Diagram for Subset", i, "...\n")

  semPaths(
    sem_full_results_plot[[i]],
    what = "std",             # Use standardized path coefficients
    layout = "tree",          # Tree layout
    style = "lisrel",         # LISREL style
    nCharNodes = 0,           # Do not truncate variable names
```

```
   residuals = FALSE,         # Do not show residual arrows
   title = FALSE,             # Do not display title
   sig = 0.05,                # Significance threshold for asterisks
   edge.label.sig = TRUE      # Display significance asterisks
  )

  readline(prompt = "Press Enter to view the next diagram...")
}

summary(sample_step3_done)
sd(sample_step3_done$word_count)
mean(sample_step3_done$word_count)

############################################################
# Step 7: Negative Binomial Regression (Full Dataset) ----------
############################################################
# 7.1 Analyzing the full dataset with bert -------
## Pre-trained Deep Learning (BERT) NLP Model ------------
# Install reticulate package
# install.packages("reticulate")
library(reticulate)

# reticulate::install_miniconda()
```

```
py_install(c("transformers", "torch"))

#  Load Python Model
transformers <- import("transformers")
torch <- import("torch")

tokenizer <- transformers$AutoTokenizer$from_pretrained("nlptown/bert-base-multilingual-uncased-sentiment")
model <- transformers$AutoModelForSequenceClassification$from_pretrained("nlptown/bert-base-multilingual-uncased-sentiment")

reticulate::py_run_string("
from transformers import AutoTokenizer, AutoModelForSequenceClassification
import torch
import torch.nn.functional as F

tokenizer = AutoTokenizer.from_pretrained('nlptown/bert-base-multilingual-uncased-sentiment')
model = AutoModelForSequenceClassification.from_pretrained('nlptown/bert-base-multilingual-uncased-sentiment')

def predict_single_sentiment(text):
    inputs = tokenizer(text, return_tensors='pt', truncation=True)
    with torch.no_grad():
```

```
        outputs = model(**inputs)
        probs = F.softmax(outputs.logits, dim=1)
        probs = probs.numpy().flatten()
        label = probs.argmax() + 1  # labels are 1-indexed
    return label, probs.tolist()
")

# Set up sentiment prediction function
predict_sentiment <- function(texts) {
  results <- lapply(texts, function(txt) {
    out <- reticulate::py$predict_single_sentiment(txt)
    list(label = out[[1]], probs = unlist(out[[2]]))
  })
  return(results)
}

texts <- restaurant_review_clean$text

# Use BERT for sentiment analysis
sentiment_results_full <- predict_sentiment(texts)
restaurant_review_full_with_bert <- restaurant_review %>%
  mutate(
    bert_score = sapply(sentiment_results_full, function(x) x$label),
```

```
    prob_1star = sapply(sentiment_results_full, function(x) x$probs[1]),
    prob_2star = sapply(sentiment_results_full, function(x) x$probs[2]),
    prob_3star = sapply(sentiment_results_full, function(x) x$probs[3]),
    prob_4star = sapply(sentiment_results_full, function(x) x$probs[4]),
    prob_5star = sapply(sentiment_results_full, function(x) x$probs[5])
  )
# 7.2 Data preperation ----------
# Construct objectiveness based on bert_score
restaurant_review_full_with_bert$bert_objective <- 1 -
(abs(restaurant_review_full_with_bert$bert_score - 3) / 2)
set.seed(1234)
restaurant_review_full_with_bert$bert_objective <-
jitter(restaurant_review_full_with_bert$bert_objective, amount = 0.05)

restaurant_review_full_with_bert <- restaurant_review_full_with_bert %>%
  select(-prob_1star, -prob_2star, -prob_3star, -prob_4star, -prob_5star)

# Visualization of bert_score distribution
ggplot(restaurant_review_full_with_bert, aes(bert_score)) +
  geom_bar(fill= "#800020", color="white") +
  geom_text(stat = "count", aes(label = after_stat(count)),
            vjust = -0.5, size = 3.2) +
  labs(title = "Distribution of bert_score",
```

```
      y = "Frequency",
      x = "bert_score")


# Classify valence as positive (1), neutral (0), or negative (-1)
restaurant_review_full_with_bert <- restaurant_review_full_with_bert %>%
  mutate(bert_valence = case_when(bert_score > 3 ~  1,
                                  bert_score < 3 ~ -1,
                                  bert_score == 3 ~  0,
                                  TRUE ~ NA))


# 7.3 Descriptive Analysis ----------
# Select specified columns
selected_cols <- sample_step2_done %>%
  select(as.numeric(all_of(c("useful", "new_reviews_3m",
                             "relevant", "comprehensive", "clarity", "bert_objective", "word_count",
                             "bert_valence", "membership_length","elite_years", "user_avg_useful"))))


# Convert logical variables to numeric
selected_cols <- selected_cols %>%
  mutate(across(where(is.logical), as.numeric))


# Compute correlation matrix (automatically excludes NA)
cor_matrix <- cor(selected_cols, use = "pairwise.complete.obs")
```

```
# Draw correlation plot
corrplot(cor_matrix, method = "color", type = "upper",
         tl.col = "black", tl.srt = 45, addCoef.col = "black",
         number.cex = 0.6, col = colorRampPalette(c("steelblue", "white", "tomato"))(200))


# Check distribution of 'useful'
ggplot(restaurant_review_full_with_bert,aes(useful))+
  geom_bar(fill= "#800020",color="white")+
  geom_text(stat = "count", aes(label = after_stat(count)),
            vjust = -0.5, size = 3.2)+
  labs(title = "Distribution of useful",
       y = "Frequency",
       x = "useful")


# Check distribution of Engagement Behavior (EB)
ggplot(restaurant_review_full_with_bert,aes(new_reviews_3m))+
  geom_bar(fill= "#800020",color="white")+
  geom_text(stat = "count", aes(label = after_stat(count)),
            vjust = -0.5, size = 3.2)+
  labs(title = "Distribution of new reviews 3 months after each review",
       y = "Frequency",
       x = "useful")
```

```
summary(restaurant_review_full_with_bert$new_reviews_3m)

# Visualize the distribution of word count
ggplot(restaurant_review_full_with_bert, aes(word_count)) +
  geom_histogram(fill= "#800020", color="white") +
  labs(title = "Distribution of Word Count",
       y = "Frequency",
       x = "Useful")

summary(restaurant_review_full_with_bert$new_reviews_3m)

# Visualize the distribution of valence
ggplot(restaurant_review_full_with_bert, aes(x = bert_valence)) +
  geom_bar(fill = "#800020", color = "white", width = 0.6) +
  geom_text(stat = "count", aes(label = after_stat(count)),
            vjust = -0.5, size = 4) +
  labs(title = "Distribution of Valence",
       x = "Valence Category",
       y = "Frequency") +
  theme_minimal()

# Count the number of reviews per state and sort in descending order
state_review_counts <- restaurant_review_full_with_bert %>%
```

```
  group_by(state) %>%
  summarise(review_count = n()) %>%
  arrange(desc(review_count))

# Plot the number of reviews per state
ggplot(state_review_counts, aes(x = reorder(state, -review_count), y = review_count)) +
  geom_bar(stat = "identity", fill = "#800020") +
  geom_text(aes(label = review_count),
            vjust = -0.5, size = 3.2) +
  labs(title = "Number of Reviews per State",
       x = "State",
       y = "Total Reviews") +
  theme_minimal()

# Plot average business stars per state
ggplot(state_means, aes(x = reorder(state, -avg_business_stars), y = avg_business_stars)) +
  geom_bar(stat = "identity", fill = "#800020") +
  geom_text(aes(label = round(avg_business_stars, 2)), vjust = -0.5, size = 3.2) +
  labs(title = "Average Business Stars per State",
       x = "State",
       y = "Average Stars") +
  theme_minimal()
```

```
# Plot average new reviews (3 months) per state
ggplot(state_means, aes(x = reorder(state, -avg_new_reviews_3m), y = avg_new_reviews_3m)) +
  geom_bar(stat = "identity", fill = "#800020") +
  geom_text(aes(label = round(avg_new_reviews_3m, 1)), vjust = -0.5, size = 3.2) +
  labs(title = "Average New Reviews in 3 Months per State",
       x = "State",
       y = "Avg. New Reviews") +
  theme_minimal()

# Plot average review usefulness per state
ggplot(state_means, aes(x = reorder(state, -avg_useful), y = avg_useful)) +
  geom_bar(stat = "identity", fill = "#800020") +
  geom_text(aes(label = round(avg_useful, 1)), vjust = -0.5, size = 3.2) +
  labs(title = "Average Review Usefulness per State",
       x = "State",
       y = "Avg. Useful Score") +
  theme_minimal()

# Center useful and valence, and create interaction term
restaurant_review_full_with_bert <- restaurant_review_full_with_bert %>%
  mutate(
    useful_c = scale(useful, center = TRUE, scale = FALSE),
    valence_c = scale(bert_valence, center = TRUE, scale = FALSE),
```

```
    int_use_val = useful_c * valence_c
  )

# Standardize observed variables
restaurant_review_full_with_bert <- restaurant_review_full_with_bert %>%
  mutate(
    review_count = as.numeric(scale(review_count, center = TRUE, scale = TRUE)),
  )

# Convert text variables to factor
restaurant_review_full_with_bert$state <- as.factor(restaurant_review_full_with_bert$state)
restaurant_review_full_with_bert$year_quarter <-
as.factor(restaurant_review_full_with_bert$year_quarter)
str(restaurant_review_full_with_bert)
summary(restaurant_review_full_with_bert)

# One-hot encode state and year_quarter (17 + 20 dummy variables)
#install.packages("fastDummies")
restaurant_review_full_with_bert <- fastDummies::dummy_cols(
  restaurant_review_full_with_bert,
  select_columns = c("state", "year_quarter"),
  remove_selected_columns = TRUE,  # Remove original columns
  remove_first_dummy = TRUE        # Avoid multicollinearity (leave reference group)
```

```
)

# Replace "-" with "_" in all column names (only column names, not content)
names(restaurant_review_full_with_bert) <- gsub("-", "_", names(restaurant_review_full_with_bert))
colSums(restaurant_review_full_with_bert[,37:70])

# 7.4 Pre-test for NBR ----------
# Mean vs variance test (check for overdispersion)
set.seed(1234)  # Reproducible splitting
n_total <- nrow(restaurant_review_full_with_bert)
train_indices <- sample(1:n_total, size = 0.8 * n_total)

train_df <- restaurant_review_full_with_bert[train_indices, ]
test_df  <- restaurant_review_full_with_bert[-train_indices, ]

mean(train_df$new_reviews_3m)
var(train_df$new_reviews_3m)
dispersion_ratio <- var(train_df$new_reviews_3m) / mean(train_df$new_reviews_3m)
# If Variance >> Mean (e.g., ratio > 1.5~2), it indicates significant overdispersion.
Dispersion_Ratio = 45 → passed

# Proportion of zeros (check for zero inflation)
```

```
zero_rate <- mean(train_df$new_reviews_3m == 0)
cat("Proportion of 0s:", round(zero_rate * 100, 2), "%\n")
# Only 5.67% are zeros → no need for zero-inflated model

# Compare Poisson vs Negative Binomial models
pois_model <- glm(new_reviews_3m ~ useful_c + valence_c + int_use_val,
                  family = poisson, data = train_df)
nb_model <- glm.nb(new_reviews_3m ~ useful_c + valence_c + int_use_val,
                   data = train_df)
AIC(pois_model, nb_model)

# Compare NB vs Zero-Inflated NB
zinb <- zeroinfl(
  new_reviews_3m ~ useful_c + valence_c + int_use_val | 1,
  data = train_df,
  dist = "negbin"
)
vuong(zinb, nb_model)

# Raw Vuong test is not significant → can't distinguish between models;
# AIC/BIC adjusted test is highly significant:
# z value is negative → NB is better;
# p < 0.001 → very significant difference;
```

```
# Negative Binomial is more appropriate; ZINB may overfit


# 7.5 Build models ----------
# NBR_1: Main effects only
nb1 <- glm.nb(new_reviews_3m ~ useful_c + valence_c, data = train_df)
summary(nb1)


# NBR_2: Add interaction term
nb2 <- glm.nb(new_reviews_3m ~ useful_c + valence_c + int_use_val, data = train_df)
summary(nb2)


# NBR_3: Add control variables and dummies
nb3 <- glm.nb(new_reviews_3m ~ useful_c + valence_c + int_use_val + review_count +
              state_AZ + state_CA + state_DE + state_FL + state_ID + state_IL + state_IN +
              state_LA + state_MO +  state_NJ + state_NV + state_PA + state_TN +
              year_quarter_2017_Q2 + year_quarter_2017_Q3 + year_quarter_2017_Q4 +
              year_quarter_2018_Q1 + year_quarter_2018_Q2 + year_quarter_2018_Q3 +
year_quarter_2018_Q4 +
              year_quarter_2019_Q1 + year_quarter_2019_Q2 + year_quarter_2019_Q3 +
year_quarter_2019_Q4 +
              year_quarter_2020_Q1 + year_quarter_2020_Q2 + year_quarter_2020_Q3 +
year_quarter_2020_Q4 +
              year_quarter_2021_Q1 + year_quarter_2021_Q2 + year_quarter_2021_Q3,
```

```
            data = train_df)
summary(nb3)


# 7.6 Model fit check ----------------
# Calculate McFadden R² + Nagelkerke R²
# Model 1: Main effects
cat(" Model 1 (Main Effects):\n")
r2_nb1 <- pR2(nb1)
cat("  McFadden's R²: ", round(r2_nb1["McFadden"], 4), "\n")
cat("  Nagelkerke R²:", round(r2_nb1["r2CU"], 4), "\n\n")  # r2CU is Nagelkerke's R²

# Model 2: With interaction
cat(" Model 2 (Interaction):\n")
r2_nb2 <- pR2(nb2)
cat("  McFadden's R²: ", round(r2_nb2["McFadden"], 4), "\n")
cat(" Nagelkerke R²:", round(r2_nb2["r2CU"], 4), "\n\n")

# Model 3: Full model with controls and dummies
cat(" Model 3 (Full Model with Controls):\n")
r2_nb3 <- pR2(nb3)
cat("  McFadden's R²: ", round(r2_nb3["McFadden"], 4), "\n")
cat("  Nagelkerke R²:", round(r2_nb3["r2CU"], 4), "\n\n")
```

```
# 7.6 Model generalizability testing ----------------

# Predict on test/train set ------------------------------

test_df$predicted_reviews <- predict(nb3, newdata = test_df, type = "response")

train_df$predicted_reviews <- predict(nb3, newdata = train_df, type = "response")


# Visualize prediction accuracy

ggplot(test_df, aes(x = predicted_reviews, y = new_reviews_3m)) +

  geom_point(alpha = 0.15) +

  geom_smooth(method = "lm", color = "blue", se = FALSE) +

  labs(title = "Predicted vs. Actual New Reviews (NB)",

       x = "Predicted New Reviews",

       y = "Actual New Reviews") +

  theme_minimal()


# MAE, RMSE, Hit Rate --------------------------

# MAE, RMSE, Hit Rate for training set

mae_train <- mae(train_df$new_reviews_3m, train_df$predicted_reviews)

rmse_train <- rmse(train_df$new_reviews_3m, train_df$predicted_reviews)

hit_train <- mean(ifelse(train_df$new_reviews_3m > median(train_df$new_reviews_3m), 1, 0)

==

                  ifelse(train_df$predicted_reviews > median(train_df$predicted_reviews), 1, 0))


# MAE, RMSE, Hit Rate for test set
```

```
mae_test <- mae(test_df$new_reviews_3m, test_df$predicted_reviews)
rmse_test <- rmse(test_df$new_reviews_3m, test_df$predicted_reviews)
hit_test <- mean(ifelse(test_df$new_reviews_3m > median(test_df$new_reviews_3m), 1, 0) ==
                   ifelse(test_df$predicted_reviews > median(test_df$predicted_reviews), 1, 0))

# Output
cat(" Train MAE:", round(mae_train, 2), " | Test MAE:", round(mae_test, 2), "\n")
cat(" Train RMSE:", round(rmse_train, 2), " | Test RMSE:", round(rmse_test, 2), "\n")
cat(" Train Hit Rate:", round(hit_train, 4), " | Test Hit Rate:", round(hit_test, 4), "\n")

# 7.7 Interaction plot  Visualization ----------------------------------
interaction_grid <- expand.grid(
  valence_c = seq(-1.5, 0.5, length.out = 50),
  useful_c = c(0, 40, 80, 120)
)
interaction_grid$int_use_val <- interaction_grid$useful_c * interaction_grid$valence_c
interaction_grid$review_count <- mean(train_df$review_count, na.rm = TRUE)
for (col in grep("^state_|^year_quarter_", names(train_df), value = TRUE)) {
  interaction_grid[[col]] <- 0
}
interaction_grid$predicted_reviews <- predict(nb3, newdata = interaction_grid, type = "response")
interaction_grid$useful_level <- factor(interaction_grid$useful_c)
```

```
ggplot(interaction_grid, aes(x = valence_c, y = predicted_reviews, color = useful_level)) +
  geom_line(size = 1.2) +
  labs(
    title = "Moderation Effect: Usefulness x Valence (NB)",
    x = "Valence (centered)",
    y = "Predicted New Reviews",
    color = "Usefulness"
  ) +
  theme_minimal() +
  scale_color_manual(values = c("0" = "#a6cee3", "40" = "#1f78b4", "80" = "#33a02c", "120" =
"#e31a1c"))

# 7.7 Post-hoc model check----------------------------------
# Residual plot (NB) ----------------------------------
resid_df <- data.frame(
  fitted = predict(nb3, type = "response"),
  resid = residuals(nb3, type = "pearson")
)

ggplot(resid_df, aes(x = fitted, y = resid)) +
  geom_point(alpha = 0.2, shape = 1) +
  geom_hline(yintercept = 0, color = "red") +
  coord_cartesian(ylim = c(-2, 16)) +
```

```
  labs(
    title = "Residuals vs Fitted (NB)",
    x = "Predicted values",
    y = "Pearson residuals"
  ) +
  theme_minimal()


# Multicollinearity check --------------------------------
vif(glm.nb(new_reviews_3m ~ useful_c + valence_c + int_use_val + review_count, data = train_df))
```