uc3m | Universidad Carlos III de Madrid

Master Degree in Computational Social Science
2024-2025 Academic Year

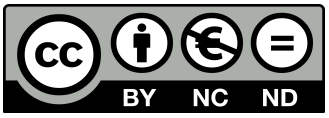
Final's Master Thesis

"Why Do We Find Some Reviews More Helpful? The Role of Verified Status and Emotional Tone"

Ekin Kizildas

Sebastian Daza

Madrid and 19.06.2025



ABSTRACT

This study investigates what makes some online product reviews more helpful than others. Based on a dataset of 160,000 Amazon reviews written in English, it focuses on two main factors: whether the reviewer is a verified purchaser and the emotional tone of the review. Sentiment was measured using three established lexicons which are AFINN, BING, and NRC and combined with other characteristics like review length, rating, price, and product category.

The results show that verified purchase status is the most consistent factor associated with helpfulness. Emotional expression also plays a role, especially when the review is written by someone verified. Reviews that are longer and more detailed tend to receive more helpful votes, while extremely positive or emotionally flat reviews are seen as less useful. These effects vary by product category, suggesting that context matters.

Together, these findings highlight the importance of trust, emotional tone, and content richness in shaping how helpful a review is perceived. They offer practical insights for platforms and sellers aiming to promote more useful user-generated content.

Keywords: *Verified Purchase, Emotional Tone, Sentiment Analysis, Consumer Behavior, Helpfulness, Online Reviews*

ACKNOWLEDGEMENT

I would like to thank everyone who contributed to the completion of this project, especially my mentor Sebastian Daza, valuable guidance, and constructive feedback throughout this thesis. Also, I would like to thank my colleagues for their collaboration and support.

My mother (Dilek) and father (Mahsuni), thank you for your endless love, patience, and for always supporting me. To my brother (Umur), thank you for being my silent strength and for always standing by my side. This achievement is as much yours as it is mine.

Aurora, Gina, Gur, and Isabel, thank you for the friendship, encouragement, and subtle reminders that I was never alone on this journey.

TABLE OF CONTENTS

TABLE OF ILLUSTRATIONS

1. INTRODUCTION

In the past decade, online shopping has changed the way people decide what to buy. Since it's not possible to see or try products in person, many consumers now rely heavily on online reviews to help them make informed choices (Mudambi & Schuff,

2010). Instead of just trusting ads or brand messages, they turn to what other users have shared about their experiences. These reviews fill the gap by offering real feedback in a digital space and help build a sense of trust when buying something online.

Although the academic literature has explored various structural factors such as length, star rating, readability, and opinion extremity that influence review helpfulness, comparatively less attention has been paid to qualitative and psychological dimensions. Specifically, two elements remain underexplored: the emotional tone conveyed in the review text and the reviewer's verified purchase status. Emotional expressions in reviews can help readers connect with the message or perceive it as more sincere. Similarly, having a verified purchase label signals real product usage, which may enhance trust in the review (Yin et al., 2014; Luca, 2016). Despite their apparent relevance, the specific and joint impacts of these factors on helpfulness ratings remain insufficiently understood.

Psychologically, having a verified purchase label fits with what's known as source credibility theory. According to this theory, people could tend to trust and be persuaded more easily by those they perceive as knowledgeable and reliable (Hovland, Janis, & Kelley, 1953). Emotional language in a review can also make it more engaging by triggering feelings of empathy or urgency in the reader (Yin, Bond, & Zhang, 2014). However, how that emotional tone is received may depend on the reviewer's credibility. When a review comes from a verified buyer, strong emotions might be seen as genuine and convincing. In contrast, similar expressions from an unverified source could come across as exaggerated or even manipulative.

To address this gap, this study seeks to answer the following research question: *How do verified purchase status and emotional tone influence the perceived helpfulness of online product reviews?*

In addition to this main question, the research explores whether this relationship is moderated by contextual variables, including product price, review age, and product rating. These moderators may influence how much weight readers assign to verification or emotion. For example, high-priced products may make readers more sensitive to both factors, while older reviews may lose credibility regardless of content. To examine these

questions, this research analyzes approximately 160,000 English-language reviews from the Amazon Reviews 2023 dataset, covering eight product categories.

Sentiment scores are computed using three lexicons that are NRC, Bing, and AFINN while verified status is coded as a binary indicator. The number of helpful votes serves as the dependent variable. The Negative Binomial regression model is used due to the count nature of the data, with ZIF (Zero Inflation Model) and cross-validation applied for robustness. This thesis aims to contribute to the literature by integrating emotional and credibility dimensions into review analysis, offering insights for platforms aiming to promote helpful user content and for vendors seeking to encourage more impactful customer feedback.

1.1 Literature Review

The growing reliance on online reviews in e-commerce has driven a surge in scholarly interest in what makes certain reviews more "helpful" than others. Early studies focused on structural attributes such as review length, star rating, and extremity. For instance, Mudambi and Schuff (2010) demonstrated that moderately positive reviews of longer length are perceived as more helpful, especially for experience goods. Subsequent research expanded on these findings, showing that readability (Ghose & Ipeirotis, 2011), spelling accuracy (Liu et al., 2008), and even formatting style can influence how readers perceive the utility of a review. These studies established that surface-level cues shape first impressions and cognitive processing of textual information.

As research on online reviews has developed, it's become clear that surface-level features like length or star rating don't fully explain why some reviews are seen as more helpful. Researchers have started to pay more attention to what's being said in the review and who it's coming from. People don't just not only look for facts but they also respond to emotional tone and how authentic a review feels. This shift shows a move away from simply describing online content, toward understanding how people interpret and connect with it on a psychological level. Emotional expression in reviews plays a multifaceted role. On the one hand, it engages readers cognitively and effectively, making the review feel more vivid and memorable (Zhu & Zhang, 2010). On the other, it may serve as a signal of honesty or involvement, particularly when the emotion is

aligned with the reviewer's narrative. Studies have shown that emotionally rich reviews are more likely to be rated as helpful when they include narrative elements such as regret, excitement, or frustration (Yin et al., 2014). These emotional cues help the reader simulate the reviewer's experience of a process consistent with the theory of transportation in narrative persuasion (Green & Brock, 2000). However, emotional tone can be a double-edged sword. When emotions in a review come across as too intense or aren't supported with clear explanations, they can actually make the review seem less trustworthy or useful. For instance, a vague comment like "I absolutely hated it!!!" might be seen as just venting. In contrast, something like "It stopped working after two uses, which really upset me" offers context and feels more sincere. This shows that emotional language doesn't always have the same effect and it depends on how and where it's used, and whether it contributes meaningfully to the review's credibility.

In this evaluative process, verified purchase status emerges as a potentially moderating factor. Building on the Source Credibility Framework (Hovland et al., 1953), verified status can be seen as a proxy for both expertise (the reviewer has used the product) and trustworthiness (they are less likely to be a fake reviewer or competitor). Several studies have shown that verified reviews are more trusted (Chevalier & Mayzlin, 2006), and that platforms like Amazon increasingly use this signal in their algorithms. The Source Credibility Theory emphasizes the importance of perceived trustworthiness and expertise in the evaluation of information. A verified purchase label often functions as a signal of both product familiarity and reviewer authenticity, aligning with classical notions of credibility, which emphasize expertise and trust as central components of persuasive communication (Hovland et al., 1953; Metzger et al., 2010) These elements can enhance the credibility of the message and this situation could make it more likely to be trusted and valued by readers. This aligns with persuasion studies showing that credible sources tend to improve information acceptance and perceived usefulness, which ultimately influence how helpful a review is considered to be (Hovland et al., 1953; Metzger et al., 2010).

Emotional language in reviews can influence how people respond by creating a sense of empathy, urgency, or personal connection. According to the Elaboration Likelihood Model (Petty & Cacioppo, 1986), emotional content plays a stronger role when readers are not fully focused, such as during casual browsing on a phone. Even

when the argument itself isn't particularly strong, emotional expression can increase engagement and make a review seem more trustworthy (Yin et al., 2014). Thus, emotional tone not only enriches the narrative quality of reviews but may also substitute for factual depth, especially when the source is perceived as credible.

Recent research suggests that the persuasive power of emotional language in reviews is context-dependent. Emotional expressions are more likely to be accepted as sincere when the reviewer is verified, whereas similar expressions from unverified users might be perceived as manipulative or attention-seeking (Otterbacher, 2009; Willemsen et al., 2011). This interaction underlines the importance of analyzing both structural and psychological features together rather than in isolation. As such, the interplay between credibility cues and emotional tone may significantly influence how helpful a review is perceived to be, a phenomenon that remains empirically underexplored in current literature.

Yet, most empirical studies treat verified status as a control variable or a simple binary marker. Very few analyze its interaction with the emotional content of the review. Theoretically, verified users should be granted more leeway to express intense emotion without losing credibility, while unverified reviewers might trigger suspicion when writing emotionally charged reviews. This suggests that the impact of emotional tone on perceived helpfulness may depend on the presence of verification, a proposition that has received little direct investigation.

Integrating these two strands that are emotional tone and verification requires a more holistic framework for understanding review helpfulnessOne way to understand this is through the idea of diagnosticity. According to this concept, a review is seen as helpful not just because it shows emotion or comes from a verified user, but because the combination of who wrote it and how it's written helps the reader draw meaningful conclusions(Filieri, 2015). For example, when a verified reviewer expresses regret, it may come across as a genuine and trustworthy warning. The same tone from someone unverified, however, might feel exaggerated or less credible.. This interaction remains underexplored in the literature, despite being central to how consumers assess trust and usefulness.

Moreover, recent work has emphasized the importance of contextual moderators that include product price, review age, and product popularity in shaping how review characteristics are processed (Filieri, 2015; Forman et al., 2008). High priced products may increase consumers' reliance on both emotional cues and source credibility, while reviews of widely known products might be judged differently than those for niche items. These findings suggest that any model of review helpfulness must be context-aware, not static.

Taken together, existing research provides valuable insights into the structural and psychological factors that influence review helpfulness. However, there is a clear theoretical and empirical gap in understanding how emotional tone and verified status interact, and how this relationship is shaped by the surrounding review environment. By focusing on this intersection, the current study seeks to contribute to both theory and practice: offering a refined model of perceived helpfulness and providing practical implications for e-commerce platforms and review algorithms.

All in all, as the theoretical foundation of this study has been established, it is now appropriate to formally articulate the hypotheses that guide the empirical analysis. Drawing on prior research in consumer behavior, source credibility, and sentiment processing, the following hypotheses were developed to examine how verified status, emotional tone, and contextual factors influence review helpfulness.

1.2 Hypotheses

H1: *Reviews with verified purchase status are more likely to be perceived as helpful compared to non-verified reviews.*

H2: *Reviews with a more positive emotional tone are more likely to receive higher helpfulness votes.*

H3: *The positive effect of emotional tone on helpfulness could be stronger when the review is from a verified purchaser.*

H4: *Product-related characteristics, such as price and review age, could moderate the relationships between emotional tone, verified status, and perceived helpfulness.*

2. METHODOLOGY

This section outlines the methodological approach used to investigate what makes certain Amazon product reviews appear more helpful than others. The focus was on understanding how verified purchase status, emotional tone, textual features, product category, and control variables such as price and rating affect the number of helpful

votes a review receives. The study followed a quantitative research design and relied on a large dataset containing 160,000 English language reviews collected from the Amazon Reviews 2023 archive.

To ensure the dataset was complete and reliable, only reviews with full metadata which included review text, star rating, review date, helpful vote count, verified purchase indicator, and product price were included. Reviews were sampled from eight broad product categories that are Appliances, Beauty, Beauty & Personal Care, Books, Electronics, Fashion, Pet Supplies, and Software. This ensured a diverse representation of product types and consumer expectations. To maintain balance across categories and reduce category-specific bias, an equal number of observations was selected from each product category. Furthermore, the sample was stratified by verified purchase status to reduce sampling bias and enable valid group comparisons. During the data selection phase, an equal number of reviews from verified and unverified purchasers was retained.

Data preprocessing involved cleaning and preparing the text data. This included removing duplicate entries and reviews with missing values, followed by basic text processing such as tokenization, lemmatization, and removal of stopwords, special characters, and extremely short reviews. After cleaning, sentiment scores were calculated using three established lexicons. Sentiment analysis was conducted using three distinct lexicons. AFINN assigns integer values to sentiment laden words to quantify polarity. BING uses binary classification to label terms as either positive or negative. NRC goes further by mapping terms to specific emotional states such as joy, trust, and anger. Sentiment scores were aggregated at the review level and normalized by review length to reduce length-related bias.

The main outcome variable, the number of helpful votes, is a non-negative count variable characterized by overdispersion and a high number of zeros. To address these features, multiple count models were applied. First, a Negative Binomial (NB) regression was used to account for overdispersion. Then, interaction terms especially between verified status and sentiment were included to test moderation effects. Finally, a Zero-Inflated Negative Binomial (ZINB) model was used to account for the structural presence of zero-vote reviews that may result from limited visibility rather than

perceived unhelpfulness. Key predictors in the models included verified purchase status, sentiment and emotion scores, review length, review age, star rating, product price, and product category.

2.1 Ethical Considerations and Data Limitations

All data used in this study were publicly available and anonymized, containing no personally identifiable information. Nonetheless, platform-provided metadata may contain biases in how helpfulness votes are collected and displayed. It is possible that early-posted reviews accumulate more votes due to visibility advantages (the "early bird" effect), or that verified reviews are algorithmically prioritized, creating feedback loops that affect helpfulness scores. Additionally, emotional expression might differ based on demographic or cultural factors not captured in the dataset.

While verified status is assumed to reflect product ownership, it may not guarantee actual use or full experience. Similarly, emotion lexicons capture sentiment with general linguistic rules, which may not fully reflect the nuanced tone of consumer narratives. Despite these limitations, the study employs rigorous controls, balanced sampling, and robust modeling techniques to enhance reliability and validity.

3. RESULTS

This section presents the empirical results from an extensive analysis of approximately 160,000 Amazon product reviews. The objective was to explore how verified purchase status, emotional tone, textual characteristics, product types, and additional control variables such as price and rating influence the number of helpful

votes a review receives. Before moving on to the regression results, some descriptive patterns in the data are worth noting. Nearly 68% of reviews received no helpful votes, pointing to a strong imbalance in user engagement. Reviews also differed greatly in both length and emotional tone. Some were brief and neutral, while others were more detailed and rich in emotional expression. These contrasts suggest that some reviews are far more likely to be seen and rated as helpful than others. Thus, a combination of Negative Binomial (NB), interaction models, and Zero-Inflated Negative Binomial (ZINB) regressions were used to address the overdispersion and prevalence of zero helpful votes in the dataset. Also, all reported coefficients in the main results section are standardized to facilitate comparison across variables.

Among all variables, verified purchase status emerged as the most consistent and significant factor associations helpfulness. In the baseline NB model (see Table B1 Model 1), this effect was quantified with a standardized coefficient of $\beta = 0.555$ ($p < 0.001$), indicating a strong positive association with helpful votes. This robust finding confirms that source credibility, as indicated by verified purchase status, enhances the perceived utility of reviews. Notably, the advantage of verified status was observed across all product categories and remained significant in more complex model specifications.

Review length ($\beta = 0.76$) and review age ($\beta = 0.39$) were also positively associated with helpfulness. In contrast, rating had a negative association ($\beta = -0.18$), indicating that higher-rated products did not necessarily receive more helpful reviews.

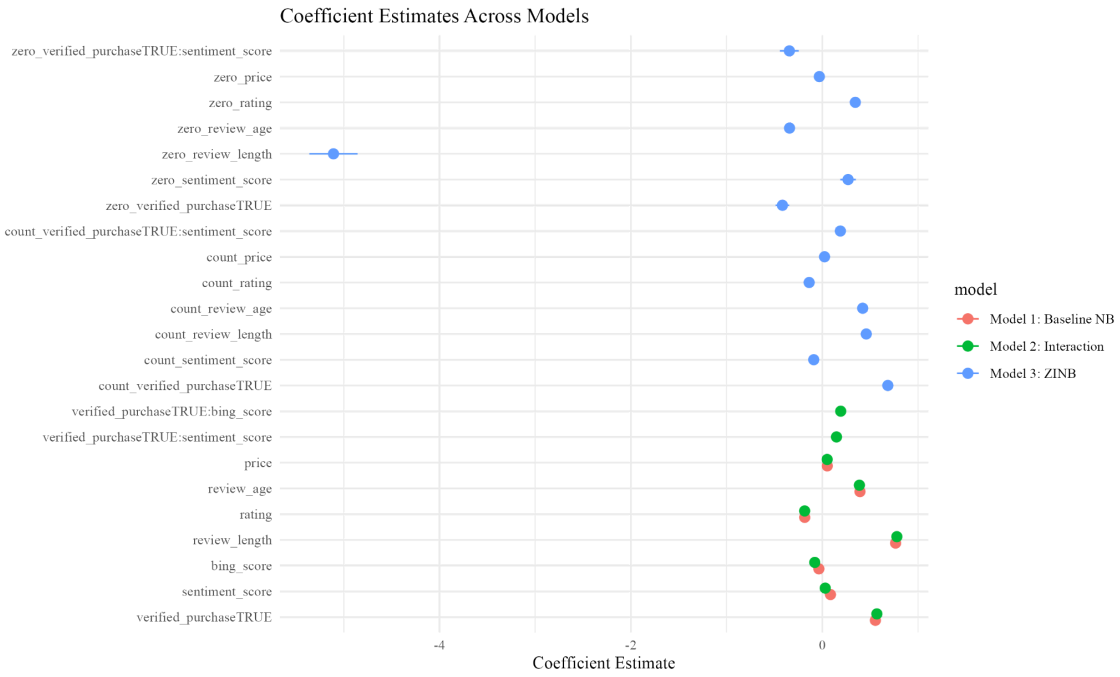


Figure 1: Coefficient Estimates Across Key Models (Baseline, Interaction, ZINB)"

This plot compares standardized coefficient estimates across three core models: Model 1 (Baseline NB), Model 2 (Interaction), and Model 3 (Zero-Inflated NB). It highlights the robustness and consistency of the predictors such as verified status, sentiment, and review features across different model specifications.

The data indicated that when emotional expression was accompanied by a verified purchase tag, it led to higher helpfulness scores. This combination enhanced the review's persuasive strength, as evidenced by a statistically significant interaction effect between verified status and sentiment score ($\beta = 0.147$, $p < 0.001$; Table B1 (Model 2)). Reviews that combined high sentiment scores with verified status received more helpful votes than their counterparts. This supports the hypothesis that emotional expressions are more persuasive when delivered by credible reviewers, reinforcing the moderating role of verification in the sentiment-helpfulness relationship.

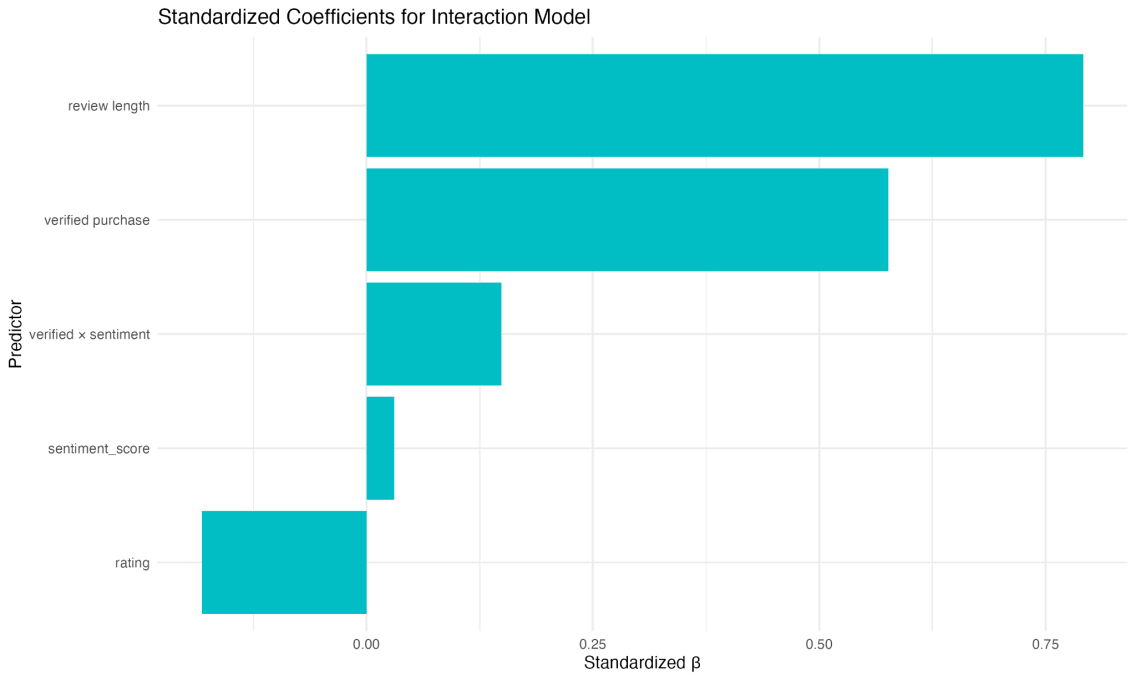


Figure2 Standardized coefficients from the Negative Binomial model with an interaction between verified purchase and sentiment score. Review length and verified status were the strongest variables. The interaction term shows that sentiment matters more when the review is from a verified source.

This pattern is also visible in the standardized coefficient plot (Figure 2), where the interaction term contributes a moderate positive association, while verified purchase and review length remain the most influential predictors

To evaluate the impact of emotional tone, sentiment analysis was performed using three lexicons: AFINN, BING, and NRC. Using the AFINN lexicon, sentiment score showed a positive and statistically significant association with helpfulness (Model 1: β = 0.085, $p < 0.001$). While the effect size is moderate, it supports the idea that emotional tone contributes to helpfulness, particularly in verified reviews (Model 2 β = 0.147, $p < 0.001$). The BING lexicon, which classifies words as either positive or negative, revealed a significant negative association between negative sentiment and helpfulness (Model 1: β = –0.036, $p < 0.001$). This effect was slightly stronger in the interaction model (Model 2: β = –0.079, $p < 0.001$). Notably, the interaction term between verified status and BING score was positive and significant (β = 0.192, $p < 0.001$), indicating that the adverse effect of negative sentiment was mitigated in

verified reviews. These findings suggest that trust cues can buffer the negative relation of emotionally negative language on helpfulness (see Model 2.)

The NRC emotion lexicon provided further granularity by mapping reviews to specific emotional categories. Expressions of trust were positively associated with helpfulness ($\beta = 0.022$, $p < 0.001$), while joy showed a small but significant negative association ($\beta = -0.012$, $p = 0.005$). Negative emotions such as anger ($\beta = -0.014$, $p < 0.001$) and disgust ($\beta = -0.014$, $p = 0.001$) were also negatively related to helpfulness. Notably, the interaction between sadness and review length was positive and statistically significant ($\beta = 0.030$, $p < 0.001$) when it is indicating that emotional expressions were more relational when conveyed in longer, more detailed reviews (See Table B3. Standardized Coefficients from NRC Emotion-Enriched Negative Binomial Model).

In addition to emotional tone and source credibility, textual characteristics also played a significant role in shaping helpfulness perceptions. Review length showed a consistently strong positive association with helpfulness across all models ($\beta = 0.778$, $p < 0.001$), underscoring that detailed reviews are perceived as more informative and trustworthy. Review age also had a positive, though comparatively smaller, effect ($\beta = 0.388$, $p < 0.001$), suggesting that older reviews still retain value in influencing decisions. Interestingly, higher product ratings (e.g., five-star reviews) were negatively associated with helpfulness ($\beta = -0.185$, $p < 0.001$). One possible explanation is that extremely positive reviews may be seen as less objective or lacking in critical insight, reducing their perceived usefulness. Price was also found to be statistically significantly associated with helpfulness. Reviews of higher-priced products were linked to slightly more helpful votes ($\beta = 0.050$, $p < 0.001$), suggesting that consumers are more attentive to detailed and credible feedback when making higher stakes purchasing decisions. This highlights the value of including control variables to disentangle the nuanced relationship between review characteristics and perceived helpfulness (See Model 2: Interaction).

A brief look across product categories showed that the factors linked to review helpfulness can vary depending on the type of item. For example, verified purchase status seemed especially important in more subjective categories like Beauty and

Personal Care, while review length stood out in Fashion. In Electronics, emotional tone appeared to play a bigger role, and in fast-changing categories like Software, the age of the review seemed to matter more. These differences give some insight into how different features may carry more weight depending on the context.

Model comparison supported the robustness of the results. The baseline Negative Binomial model (AIC = 397708.6) confirmed verified purchase status as the most influential factor. The interaction model (AIC = 397233.4) highlighted how sentiment had a stronger effect when expressed by verified users, with the interaction term yielding $\beta = 0.147$ ($p < 0.001$).

The Zero-Inflated Negative Binomial (ZINB) model (AIC = 410609.3) offered additional insight by identifying a structural "always-zero" group, estimated to comprise 26.5% of the reviews. These reviews were mostly unverified, emotionally flat or negative, brief, older, and often related to lower-priced products. Rather than a lack of quality, their invisibility and perceived lack of credibility may explain why they failed to attract engagement. This result underscores the added value of using ZINB in online review research, as it reveals systematic barriers to engagement that standard models may overlook (See Model Zinb).

The interaction between verified purchase status and sentiment score was positive and statistically significant ($\beta = 0.147$, $p < 0.001$), suggesting that emotional tone had a stronger effect on helpfulness when expressed by verified users. Additional analyses also hinted that emotions like sadness became more influential when included in longer, more detailed reviews, indicating that elaboration may amplify emotional impact.

While these findings are robust, several limitations should be noted. First, sentiment lexicons such as AFINN and BING may miss contextual cues like sarcasm or irony, limiting the depth of emotional interpretation. Second, the average age of reviews in the dataset was 7.4 years, which may not reflect current user behavior or algorithmic ranking mechanisms. Lastly, the large share of reviews with zero helpful votes (over 68%) required specialized models, which, although appropriate for the data structure, may complicate the interpretation and generalizability of the effects.

In summary, verified purchase status was found to be the most relational factor review helpfulness. Relation of this was further enhanced by emotional expressiveness, review elaboration, pricing context, and alignment with product types. These findings suggest that platforms could improve user experience by prioritizing reviews that are verified, emotionally rich, and content-dense, especially for higher-priced products.

4. DISCUSSION

The findings of this study align closely with the theoretical expectations outlined in the introduction and supported by the literature. Verified purchase status was consistently and strongly associated with higher helpfulness scores, aligning with the expectations of Source Credibility Theory (Hovland et al., 1953). This finding supports the idea that information tends to be seen as more useful when it comes from sources perceived as knowledgeable and trustworthy. This credibility signal proved especially impactful across all product categories and remained stable across different statistical models. These findings are consistent with earlier research (e.g., Chevalier & Mayzlin, 2006), which reports that verified reviews are generally perceived as more persuasive.

Furthermore, prior studies have shown that reviews including emotionally rich content such as expressions of regret or excitement tend to be more engaging and are more likely to be considered helpful. This aligns with narrative persuasion theory, which suggests that messages become more effective when they include immersive and emotionally expressive storytelling (Yin et al., 2014; Green & Brock, 2000). Emotional language from credible sources may enhance relatability and engagement. In contrast, similar emotional expressions from unverified users were often associated with lower helpfulness or showed no significant relationship, suggesting that such content may be viewed as less credible in the absence of verification.

The interaction between emotional tone and verified status further demonstrates that these variables should not be analyzed in isolation. Emotions such as trust and sadness, especially in longer reviews, appeared more impactful, suggesting that elaboration can enhance how diagnostic and useful a message is perceived to be. This reinforces the idea that the helpfulness of a review is shaped not only by emotional content or source credibility but also by their interplay.

The influence of control variables also deserves attention. For example, review length consistently increased helpfulness, which echoes earlier findings on the value of detail (Mudambi & Schuff, 2010). Star ratings and review age had more nuanced relationships: overly positive ratings were associated with fewer helpful votes, possibly due to perceived bias, while older reviews were viewed as less relevant, particularly when unverified. Price also played a role, suggesting that consumers seek more credible

and informative reviews when making higher-stakes purchasing decisions which means a pattern in line with previous research on context-aware review processing (Filieri, 2015).

Product category comparisons further highlighted that review processing is context-dependent. As previous studies suggest, consumers assess reviews differently depending on the product type (Forman et al., 2008). For instance, in categories like Electronics and Software, trust-based language had greater influence, whereas in Fashion or Beauty, review length and subjective descriptions were more important. These findings confirm that review helpfulness is shaped not only by content or reviewer traits but also by broader consumption contexts.

From a practical standpoint, platforms such as Amazon could improve user experience by prioritizing verified, emotionally expressive, and content-rich reviews in their ranking algorithms. Features like “Show Verified Reviews with High Sentiment” or writing assistance tools could guide users toward writing more helpful content by encouraging emotional and credible narratives.

The application of a Zero-Inflated Negative Binomial model also helped illuminate why some reviews receive no helpful votes. Many of these reviews were short, unverified, and emotionally neutral or negative, suggesting that lack of visibility or trust might explain their lower engagement. This raises broader concerns about how feedback systems operate and whether current algorithms might overlook potentially valuable reviews.

In sum, perceived review helpfulness appears to be shaped by the interaction of emotional expression, credibility, textual richness, and contextual factors. These insights contribute to the growing literature on digital trust and offer actionable recommendations for platforms aiming to improve review visibility, personalization, and overall consumer confidence in online settings.

4.1 Limitation

While the results appear consistent, several limitations should be taken into account. A notable constraint is the reliance on lexicon-based sentiment analysis, which may struggle to capture the full nuance and contextual richness of natural language.

Lexicons such as AFINN, BING, and NRC do not effectively detect elements like sarcasm, irony, negation, or culturally specific expressions. As a result, estimates of emotional tone may involve some degree of measurement error.

Second, the dataset consists of historical reviews with an average age exceeding seven years. This temporal skew may affect the generalizability of results, as user behaviors, platform algorithms, and consumer expectations have evolved over time. Future readers might interpret older reviews differently or assign different importance to verification cues due to increased awareness of review authenticity mechanisms.

Third, a considerable proportion of reviews (over 68%) received zero helpful votes. While the Zero-Inflated Negative Binomial model addressed this analytically, it remains unclear whether these zeros reflect a lack of quality, visibility, or user engagement. The platform's internal ranking system that based on recency, upvotes, or personalization may influence which reviews are even seen, introducing a visibility bias that this study could not fully control.

Fourth, the models did not incorporate reviewer-level metadata such as prior review activity, expertise, or review frequency, which could influence perceived helpfulness. A reviewer's reputation or review history might interact with verification or sentiment in shaping credibility perceptions.

Fifth, the study is limited to English-language reviews and a fixed set of eight product categories, which may restrict the findings' cross-linguistic and cross-sectoral applicability. Sentiment interpretation and the role of verification may vary in multilingual or culturally diverse consumer populations. The role of culture in shaping the perception of emotions such as how trust, anger, or joy are expressed and interpreted warrants deeper exploration.

Finally, although 160,000 reviews were analyzed, the sampling approach could have been elaborated further. Reviews were selected to balance verified and unverified status across categories such as Beauty, Electronics, Fashion, and Books. However, the exact distribution and filtering criteria (e.g., language, minimum text length, exclusion of duplicates) could be more clearly stated to improve replicability.

4.2 Future Work

First of all future research should explore the use of deep learning and transformer-based models to better capture the semantic nuances and context of emotional expressions. These models can address the rigidity of lexicon based approaches and provide richer sentiment representations.

Additionally, future studies could incorporate visibility metrics such as number of views, review position, or user scrolling behavior to disentangle helpfulness from exposure. Understanding the mechanics of review visibility will help clarify whether a review's low helpfulness is due to content or lack of attention.

Cross-linguistic and cross-platform comparisons would also offer valuable insight. Platforms like Ebay, TripAdvisor, or Alibaba may exhibit different patterns in the use of verification badges and emotional tone. Similarly, exploring the same models in different languages would test the universality of the findings and reveal cultural differences in emotional expression and credibility evaluation.

Moreover, future work should consider panel or longitudinal datasets to observe how helpfulness ratings evolve over time. This would help distinguish early perceptions from cumulative reputation effects and better assess the temporal dynamics of credibility and emotional impact.

Finally, incorporating qualitative methods such as user interviews, surveys, or experimental designs may complement the quantitative analysis and offer deeper insights into the psychological and cognitive processes involved in how consumers evaluate reviews. This mixed-methods approach would enable a more comprehensive exploration of how emotion and credibility jointly shape digital trust.

4.3 Conclusion

This research set out to explore what makes certain product reviews more helpful than others in an online shopping environment. Analyzing a large dataset of 160,000 Amazon reviews, the results clearly indicated that verified purchase status is the most consistent and influential factor associated with helpfulness. Reviews written by verified users were more frequently associated with higher helpfulness ratings from other consumers. In addition, factors such as emotional tone, review length, product

category, and contextual elements like price and star rating showed significant associations with how helpful a review was perceived to be. Most notably, reviews that combined emotional expression with credibility that signaled through verification were especially persuasive, highlighting the importance of trust in user-generated content. By combining sentiment analysis with advanced count-based modeling, the study contributes both theoretically and practically. Platforms could improve review sorting systems by prioritizing verified reviews that are emotionally informative and textually rich. This approach would not only enhance user trust but also help customers make more informed decisions.

All code, datasets, and analysis scripts used in this study are publicly accessible.[1]

[1] Available at: https://github.com/ekinkizildas/TFM

5. BIBLIOGRAPHY

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

Filieri, R. (2015). What makes online reviews helpful? A diagnosticity-adoption framework to explain informational and normative influences in e-WOM. *Journal of Business Research*, 68(6), 1261–1270. https://doi.org/10.1016/j.jbusres.2014.11.006

Forman, C., Ghose, A., & Wiesenfeld, B. (2008). Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Information Systems Research*, 19(3), 291–313. https://doi.org/10.1287/isre.1080.0193

Ghose, A., & Ipeirotis, P. G. (2011). Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Transactions on Knowledge and Data Engineering*, 23(10), 1498–1512. https://doi.org/10.1109/TKDE.2010.188

Green, M. C., & Brock, T. C. (2000). The role of transportation in the persuasiveness of public narratives. *Journal of Personality and Social Psychology*, 79(5), 701–721. https://doi.org/10.1037/0022-3514.79.5.701

Hoffman, D. L., & Fodor, M. (2010). *Can you measure the ROI of your social media marketing?* MIT Sloan Management Review, 52(1), 41–49.

Hou, Y., Li, J., He, Z., Yan, A., Chen, X., & McAuley, J. (2024). *Bridging language and items for retrieval and recommendation*. arXiv preprint arXiv:2403.03952.

Hovland, C. I., Janis, I. L., & Kelley, H. H. (1953). *Communication and persuasion: Psychological studies of opinion change*. Yale University Press.

Ismagilova, E., Slade, E., Rana, N. P., & Dwivedi, Y. K. (2020). *The effect of characteristics of source credibility on consumer behavior: A meta-analysis.* Journal of Retailing and Consumer Services, 53, 101736. https://doi.org/10.1016/j.jretconser.2019.01.005

Liu, Y., Huang, X., An, A., & Yu, X. (2008). Modeling and predicting the helpfulness of online reviews. In *2008 Eighth IEEE International Conference on Data Mining* (pp. 443–452). IEEE. https://doi.org/10.1109/ICDM.2008.55

Luca, M. (2016). Reviews, reputation, and revenue: The case of Yelp.com. H*arvard Business School NOM Unit Working Paper*, (12-016). https://doi.org/10.2139/ssrn.1928601

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://ssrn.com/abstract=2175066

Sen, S., & Lerman, D. (2007). *Why are you telling me this? An examination into negative consumer reviews on the web.* Journal of Interactive Marketing, 21(4), 76–94. https://doi.org/10.1002/dir.20090

Yin, D., Bond, S. D., & Zhang, H. (2014). Anxious or angry? Effects of discrete emotions on the perceived helpfulness of online reviews. *MIS Quarterly*, 38(2), 539–560.

Zhang, Z., Ye, Q., Law, R., & Li, Y. (2010). *The impact of e-word-of-mouth on the online popularity of restaurants: A comparison of consumer reviews and editor reviews.* International Journal of Hospitality Management, 29(4), 694–700. https://doi.org/10.1016/j.ijhm.2010.02.002

Zhu, F., & Zhang, X. (2010). Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of Marketing*, 74(2), 133–148. https://doi.org/10.1509/jmkg.74.2.133

6. APPENDIX

Appendix A: Descriptive Figures

Figure A1. Distribution of Emotions Detected by NRC Lexicon

This bar plot displays the relative frequencies of eight primary emotions (e.g., trust, joy, anger) identified in review texts. Trust and joy are the most common emotions, suggesting a generally positive affective tone in the dataset.

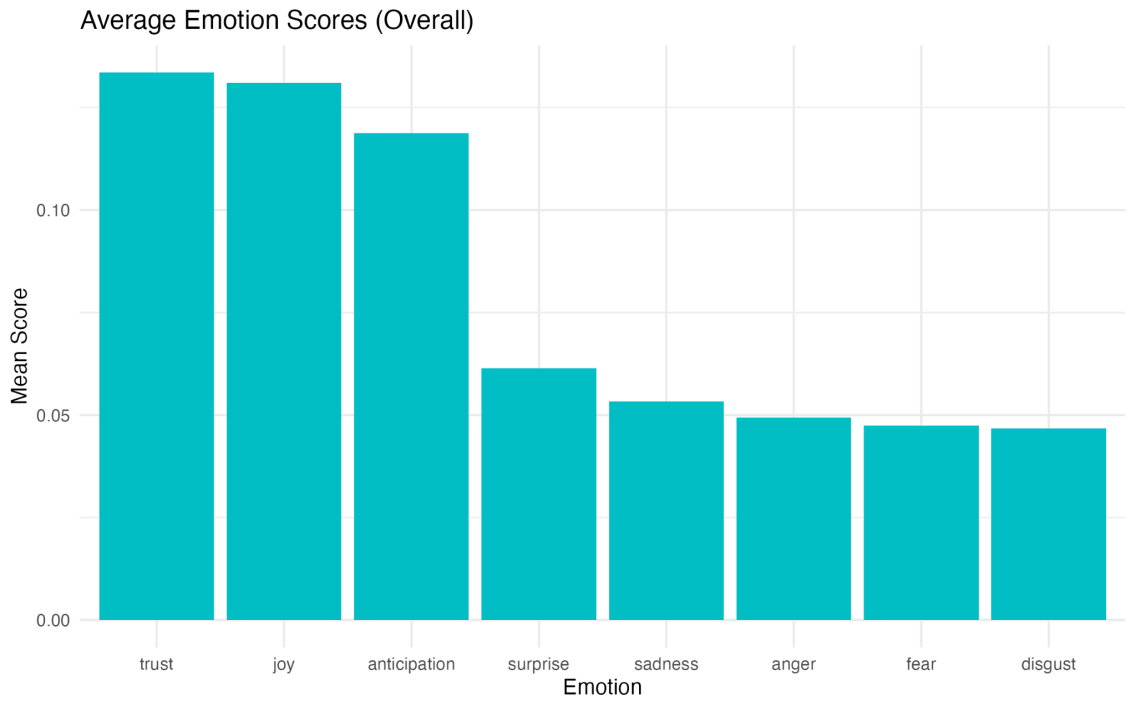


Figure A2. Distribution of Star Ratings

This histogram shows that the majority of reviews have 5-star ratings, indicating a strong positivity bias in product evaluation on the platform.

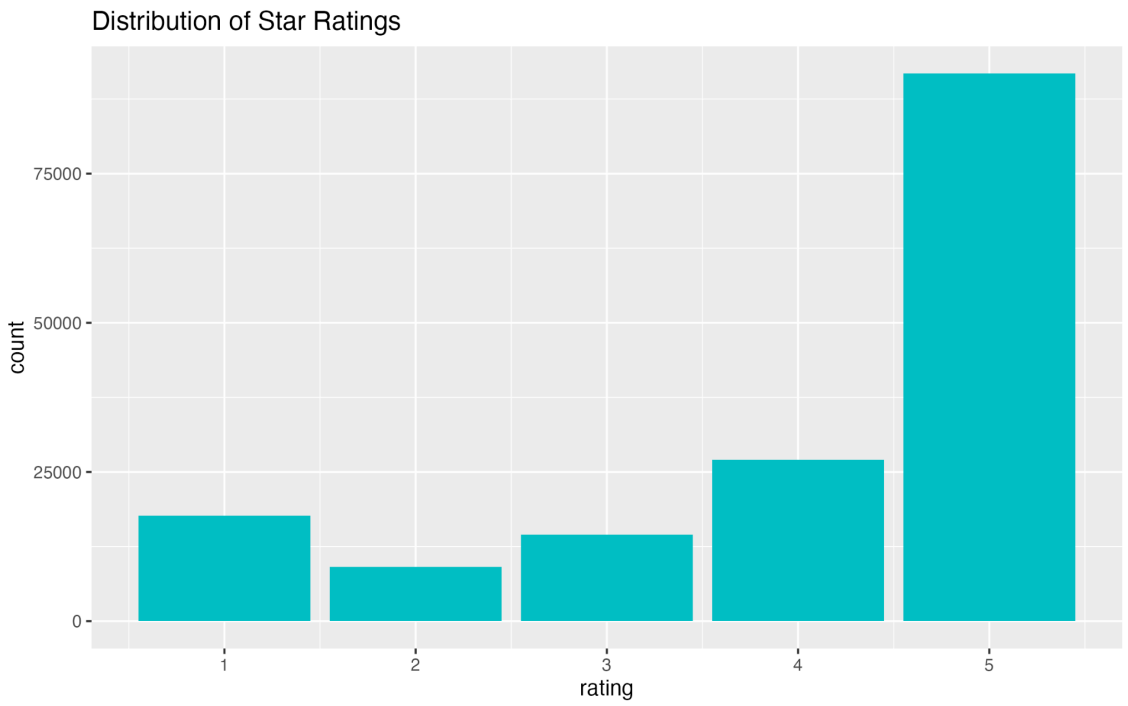


Figure A4. Average Helpful Votes by Star Rating

This bar plot illustrates the relationship between star ratings and helpfulness votes, regardless of verification status. Reviews with extreme ratings (1-star and 5-star) are less helpful than moderate ones.

Figure A5. AFINN Sentiment Scores by Review Length

This scatterplot shows that longer reviews tend to have more extreme AFINN sentiment values, both positive and negative.

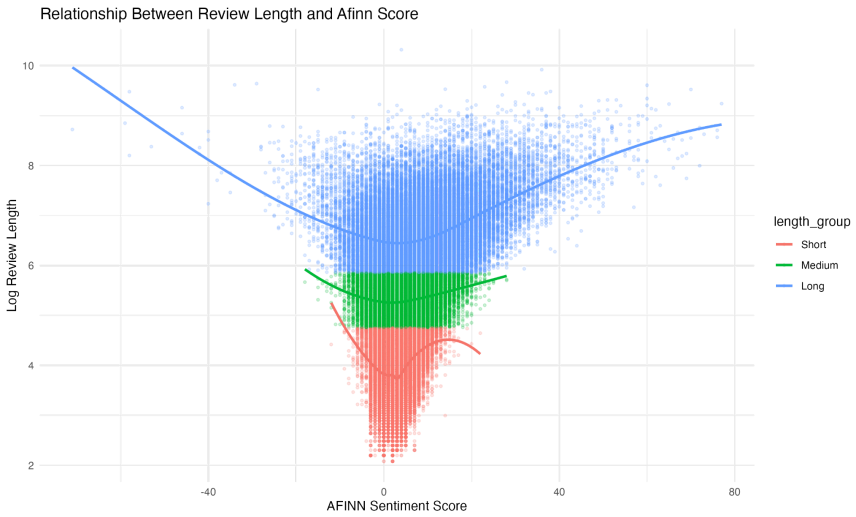


Figure A6. BING Sentiment Categories by Review Length

This plot indicates how BING-labeled positive, negative, and neutral reviews are distributed across review lengths. Longer reviews tend to be more polarized in sentiment expression.

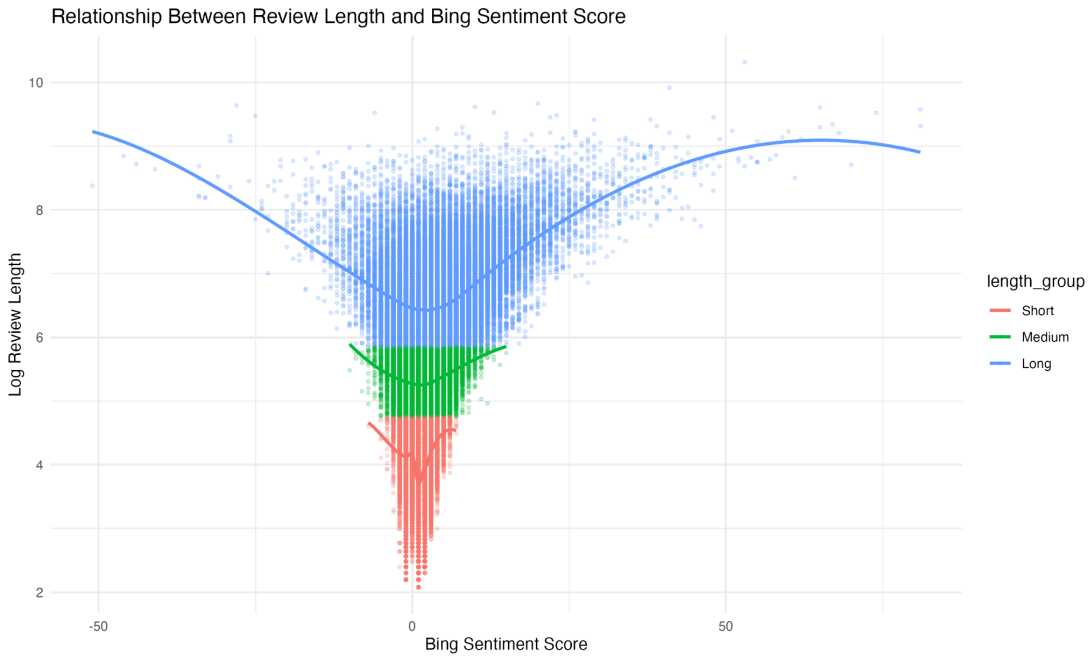


Figure A7. Emotional Expression by Review Length

This line chart maps the relative frequency of each emotion across short, medium, and long reviews, revealing how review length affects emotional richness.

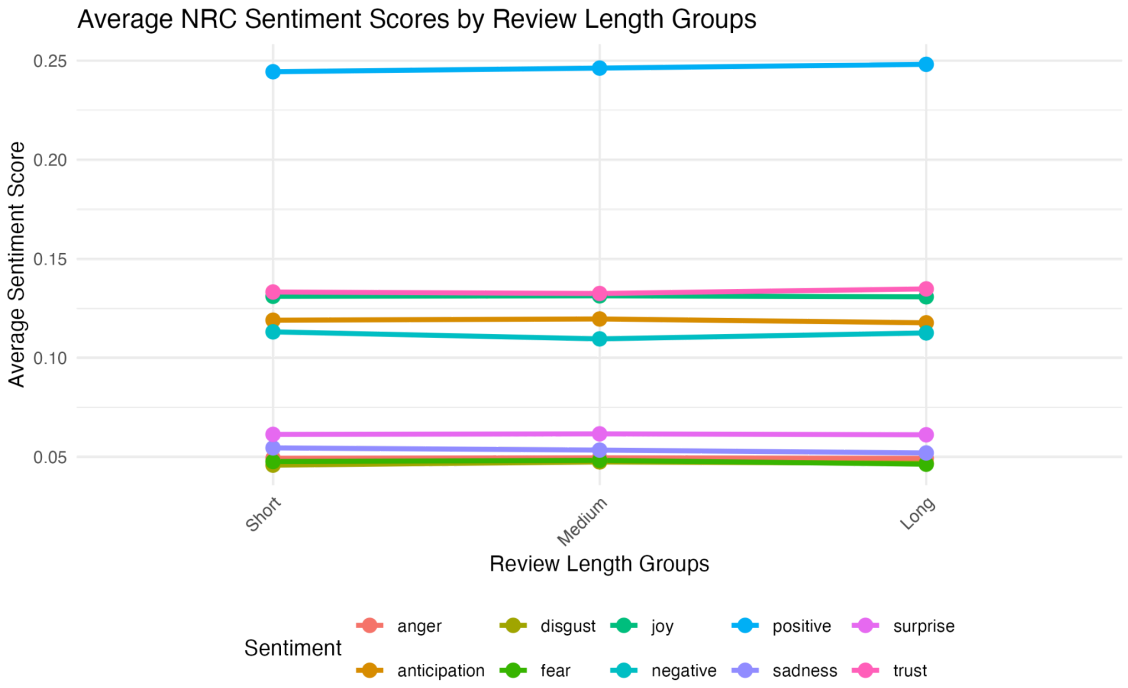


Figure A8. Helpfulness Vote Distribution Across Product Categories

This violin plot shows the distribution of helpful votes by category on a log scale. Electronics, Software, and Beauty reviews show higher variance in helpfulness.

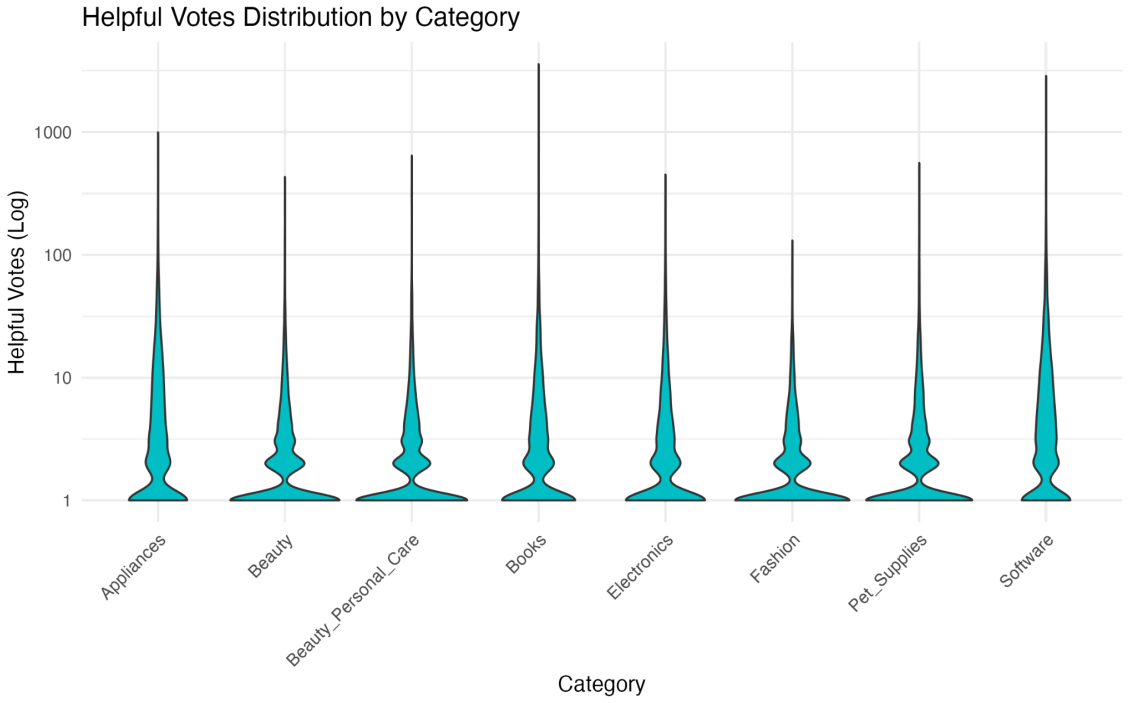


Figure A9. Distribution of Review Age

This histogram illustrates the distribution of review ages (in days). Most reviews are from the last 2,500 days, indicating recency in the dataset and potential visibility bias.

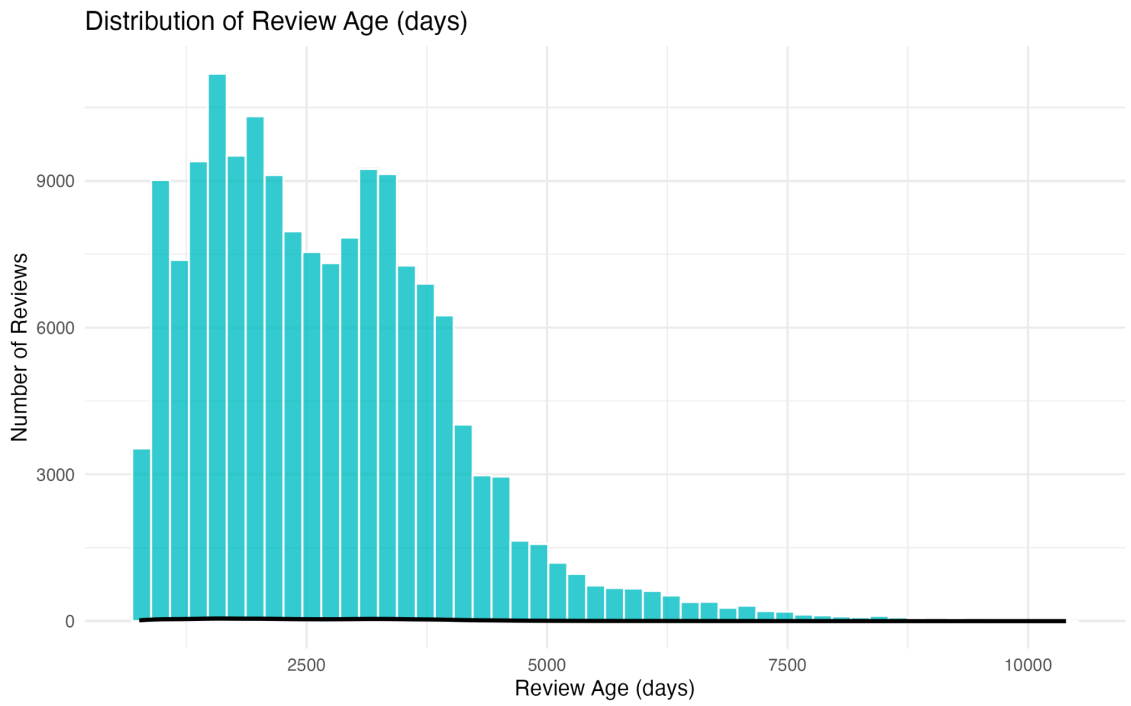


Figure A10. Review Age vs. Helpful Votes

This scatterplot illustrates the relationship between review age and helpfulness (log-scaled). Older reviews tend to have accumulated more helpful votes, suggesting temporal visibility advantages.

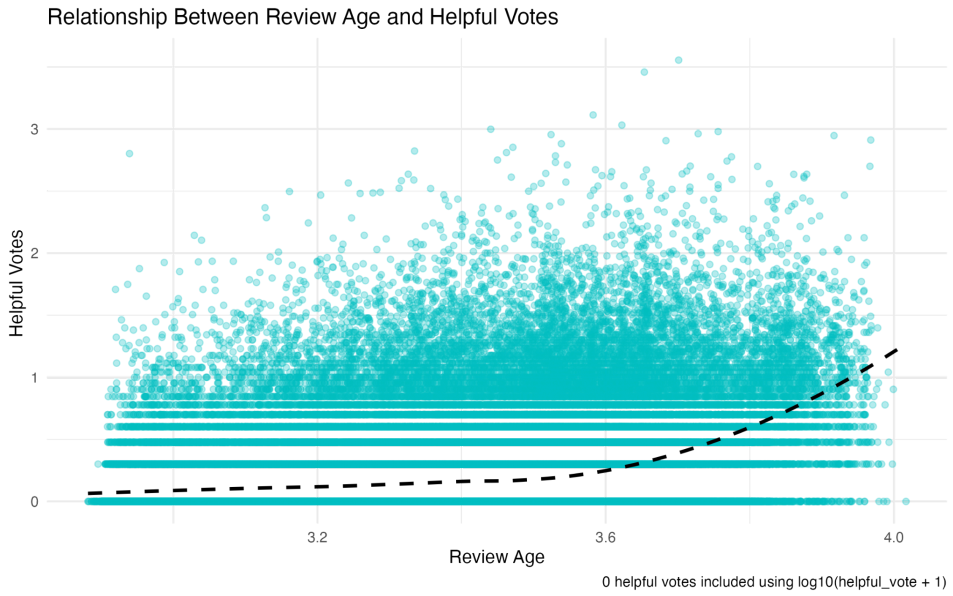


Figure A11. AFINN Sentiment vs. Helpfulness

This scatterplot with a trendline shows that helpfulness votes increase slightly with higher AFINN sentiment scores.

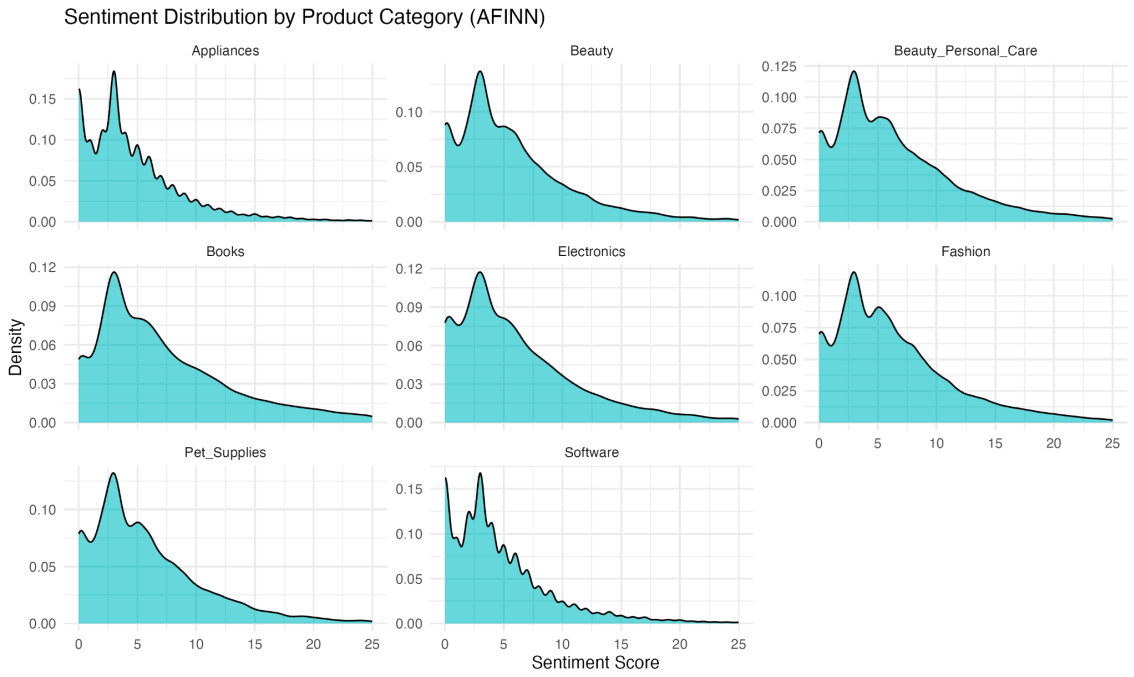


Figure A12. Distribution of BING Sentiment Scores by Product Category
This series of density plots shows the distribution of BING sentiment scores across different product categories. Reviews in categories like Software and Books exhibit wider sentiment ranges, indicating higher emotional polarization. In contrast, categories such as Fashion or Appliances show more concentrated sentiment patterns. These variations suggest that the emotional tone of reviews may depend on the nature of the product being evaluated.

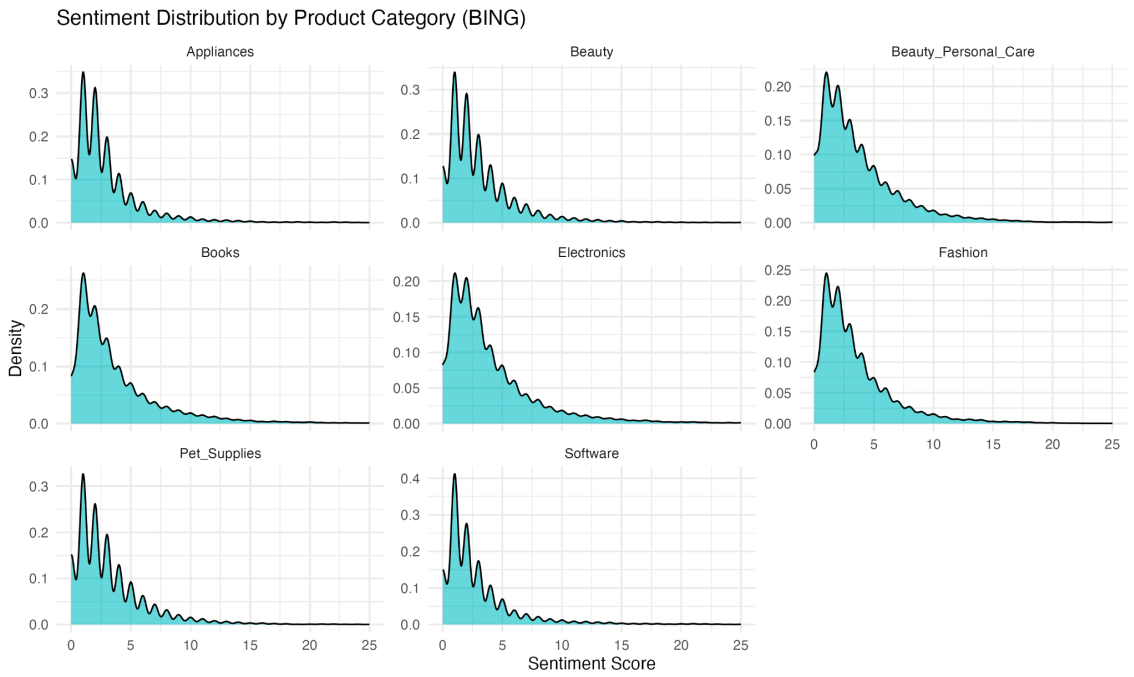


Figure A13. Summary Table of Review Characteristics by Product Category and Verification Status

This table summarizes key review characteristics across different product categories, separated by verification status. It includes average helpful votes, review length, sentiment scores (AFINN), and BING sentiment proportions (positive and negative). Verified reviews generally tend to be longer, more emotionally polarized, and receive more helpful votes, especially in categories such as Software and Books. This summary complements earlier visualizations by offering a consolidated numerical comparison.

Category	Verified	N	Avg. Helpful	Avg. Length	Avg. Rating	AFINN (Mean)	BING Negative (%)	BING_Neutral (%)	BING Positive (%)
Appliances	FALSE	10,000	4.0101	603.4338	3.5373	4.1143	22.6	11.8	65.6
Appliances	TRUE	10,000	1.0633	182.2477	4.3709	2.9848	12.0	16.9	71.1
Beauty	FALSE	10,000	1.1168	407.0899	4.0934	6.4621	10.6	9.4	79.9
Beauty	TRUE	10,000	0.9438	173.3011	3.9459	3.6838	12.3	15.9	71.7
Beauty_Personal_Care	FALSE	10,000	0.9665	547.6680	4.2207	8.0643	7.3	7.6	85.1
Beauty_Personal_Care	TRUE	10,000	1.4629	205.3649	4.1837	4.0799	10.2	15.0	74.8
Books	FALSE	10,000	3.4181	1,011.0256	4.2368	8.9436	15.9	7.3	76.8
Books	TRUE	10,000	1.8300	258.6933	4.5323	5.1852	7.1	12.2	80.6
Electronics	FALSE	10,000	2.2149	720.9100	4.0539	7.0550	11.0	7.5	81.5
Electronics	TRUE	10,000	1.1367	274.7588	4.1737	3.8549	10.3	12.7	77.0
Fashion	FALSE	10,000	0.5423	385.5604	4.0454	7.7787	7.8	8.5	83.7
Fashion	TRUE	10,000	0.6527	156.2549	3.9543	4.3936	9.5	14.6	75.9
Pet_Supplies	FALSE	10,000	1.2027	543.4696	4.1049	6.3143	13.5	10.3	76.2
Pet_Supplies	TRUE	10,000	1.2369	229.5851	4.1526	3.6489	13.3	15.6	71.1
Software	FALSE	10,000	5.4032	743.7768	3.1550	3.9447	22.0	11.8	66.1
Software	TRUE	10,000	4.2113	150.0952	3.8596	2.9209	11.2	19.6	69.2

Appendix B: Model Outputs

Table B1. Standardized Coefficients from Models

This table presents the standardized results from three core models used in the analysis. Model 1 includes the main predictors like verified status, sentiment, and review characteristics. Model 2 adds interaction terms to explore how verified status might change the impact of emotional tone.

	Model 1: NB	Model 2: Interaction
(Intercept)	0.070***	0.063***
	(0.019)	(0.019)

verified_purchaseTRUE	0.555***	0.570***
	(0.014)	(0.014)
sentiment_score	0.086***	0.031**
	(0.010)	(0.012)
bing_score	-0.036***	-0.079***
	(0.010)	(0.012)
review_length	0.765***	0.778***
	(0.007)	(0.007)
rating	-0.184***	-0.184***
	(0.007)	(0.007)
review_age	0.394***	0.388***
	(0.007)	(0.007)
price	0.057***	0.055***
	(0.006)	(0.006)
categoryBeauty	-0.385***	-0.362***
	(0.026)	(0.026)
categoryBeauty_Personal_Care	-0.361***	-0.339***
	(0.026)	(0.025)
categoryBooks	-0.421***	-0.412***
	(0.026)	(0.026)
categoryElectronics	-0.520***	-0.492***
	(0.025)	(0.025)
categoryFashion	-0.808***	-0.781***
	(0.027)	(0.027)
categoryPet_Supplies	-0.377***	-0.342***
	(0.025)	(0.025)
categorySoftware	0.867***	0.909***
	(0.025)	(0.025)
verified_purchaseTRUE × sentiment_score		0.147***
		(0.022)
verified_purchaseTRUE × bing_score		0.192***

		(0.024)
Num.Obs.	152213	152213
AIC	397695.5	397221.3
BIC	397854.5	397400.1
Log.Lik.	-198831.768	-198592.647
F	2419.450	2153.717
RMSE	1.7e+13	1.9e+13

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Table B2. Extended Model Comparison: Coefficient Estimates Across Eight Models
This figure presents standardized coefficient estimates from eight regression models, showcasing the influence of various predictors and interaction effects. The broader comparison helps illustrate how emotional tone, verification, and structural characteristics interact to shape helpfulness perceptions.

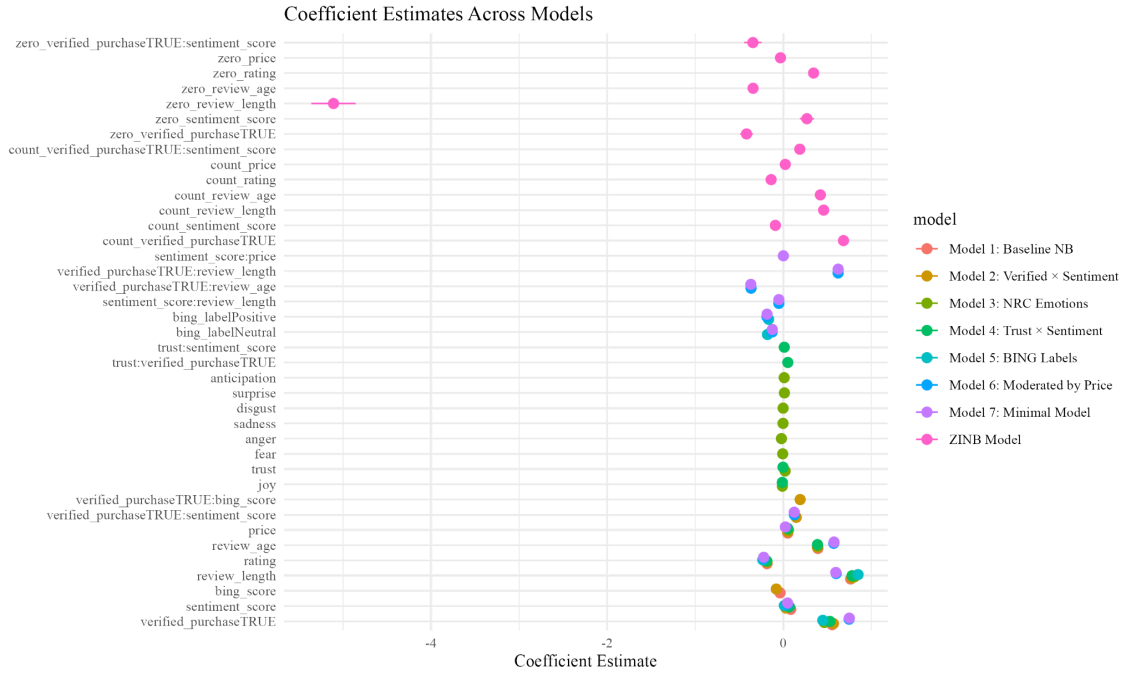


Table B3. Standardized Coefficients from NRC Emotion-Enriched Negative Binomial Model
This model highlights how individual emotions impact helpfulness. Trust increases helpful votes, while anger and disgust reduce them. Interaction effects show that longer reviews amplify the influence of emotions like sadness.

	(1)
(Intercept)	0.121***
	(0.019)
joy	-0.010
	(0.007)
trust	0.022**
	(0.007)
fear	-0.006
	(0.007)
anger	-0.020**
	(0.007)
sadness	-0.002
	(0.007)
disgust	-0.001
	(0.007)
surprise	0.014*
	(0.007)
anticipation	0.011
	(0.007)
verified_purchaseTRUE	0.466***
	(0.013)
review_length	0.808***
	(0.006)
review_age	0.392***
	(0.007)
price	0.059***
	(0.006)
categoryBeauty	-0.406***
	(0.025)
categoryBeauty_Personal_Care	-0.385***
	(0.025)
categoryBooks	-0.445***
	(0.025)
categoryElectronics	-0.555***
	(0.025)
categoryFashion	-0.815***
	(0.026)
categoryPet_Supplies	-0.416***
	(0.025)

	(1)
categorySoftware	0.921***
	(0.025)
Num.Obs.	160000
AIC	412221.2
BIC	412430.8
Log.Lik.	-206089.577
F	1813.789
RMSE	1.9e+14

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Table B4. Standardized Coefficients from ZINB Model
Summary of predictors influencing both helpful vote counts and the probability of a review receiving no helpful votes. Verified status, emotional tone, and length are key differentiators in visibility and engagement.

	(1)
(Intercept)	0.070***
	(0.019)
verified_purchaseTRUE	0.555***
	(0.014)
sentiment_score	0.086***
	(0.010)
bing_score	-0.036***
	(0.010)
review_length	0.765***
	(0.007)
rating	-0.184***
	(0.007)
review_age	0.394***
	(0.007)
price	0.057***
	(0.006)
categoryBeauty	-0.385***
	(0.026)

	(1)
categoryBeauty_Personal_Care	-0.361***
	(0.026)
categoryBooks	-0.421***
	(0.026)
categoryElectronics	-0.520***
	(0.025)
categoryFashion	-0.808***
	(0.027)
categoryPet_Supplies	-0.377***
	(0.025)
categorySoftware	0.867***
	(0.025)
Num.Obs.	152213
AIC	397695.5
BIC	397854.5
Log.Lik.	-198831.768
F	2419.450
RMSE	1.7e+13

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001