# Amazon Health & Personal Care Reviews:

# Customer Sentiment Analysis

## (Big Data Analytics and Visualisation)

**Hritesh Shukla**

**25040321**

**MSc Business Analytics**

**Word Count: 3305 (Excluding references, appendices and table of content)**

# Table of Contents

**1. Introduction to Data and Research Question** ..... 3

**2. Data Processing and Exploration** ..... 5

**3. Interpretation and Data Visualisation** ..... 6

**4. Results and Discussion** ..... 11

**5. References** ..... 14

**6. Appendices: Supplementary Visualisations** ..... 15

# 1. Introduction to Data and Research Question

Online customer reviews have become a critical part of treating the purchasing process in today's ecommerce environment, particularly in Amazon. They offer clues to quality of the product and whether or not to buy them, while giving the companies feedback to improve their product. In Health and Personal Care (HPC) purchases, reviews can be even more influential, with products like vitamins, medical devices or personal hygiene items potentially affecting consumers' health, making trust and peer knowledge crucial. Previous studies demonstrated the preponderance of positive (5-star) reviews on Amazon, often following a J-shaped distribution (Chevalier, J.A. and Mayzlin, D., 2006). Most products average almost five stars (Chevalier and Mayzlin 2006), the result of self-selection biases (P) and satisfied customers being more likely to provide feedback. Even this positive bias does not imply all reviews are positive; a small number of very negative ratings are also present, and understanding the sources of both positive and negative feedback is a central aim of this research. Given this context, this work aims to explore Amazon's Health & Personal Care product reviews to detect customer sentiment trends, potential factors related to high or low product ratings, as well as to extract more frequent themes within reviews. The objective is to convert raw review data into actionable business intelligence (e.g. which brands consistently make customers happy, how product price and customer sentiment related, and which product attributes are most frequently praised and which are most frequently criticized) by automating the laborious process of reading hundreds or thousands of reviews and deciding how much the customer likes the product. We also aim to answer the following research question by combining structured data (star ratings, brand, price, etc.) and unstructured data from the review text:

**Report Questions:** *What motivates customer satisfaction and engagement in Health & Personal Care on Amazon, and what can be learned from review sentiment and content to aid in product improvements/marketing efforts?*

In order to address this question, the following sub questions will be addressed: (1) How are the reviews distributed across different star ratings and the polarity of the reviews? (2) Are there some brands or types of products that perform better than others in terms of customer ratings? (3) How is the text sentiment and star rating of a review related? (4) What is the relationship between product features (e.g., price) and customer sentiment? and (5) What are the common notes or subjects experienced in the review texts, and how do they correspond to customer's perceptions related to product advantages/disadvantages? By analysing these aspects, the report intends to contribute to a complete overview about what consumers think in the HPC category and to offer data-based reporting. From the literature sentiment analysis of posts in e-commerce sites is a widely practiced

method in business analytics. Sentiment analysis (SA) refers to the task of computationally determining the polarity of a piece of text, whether the expressed opinion is positive, negative, or neutral. It has been heavily used to measure customer satisfaction and uncover problems, since companies can mine massive amounts of reviews for trends. But, analysing free-form text has its difficulties: consumers speak in nuance, context and sarcasm can affect meaning, and the data can be cluttered with noise. In the last few years, both lexicon-based methods and state-of-art machine learning (e.g., deep learning and transformer-based models) have been used for product reviews. For this project, we use a lexicon-based tool VADER (Valence Aware Dictionary and sEntiment Reasoner) which is designed for sentiment analysis in social media and product reviews (Mohammad, S.M., 2023.). VADER (Hutto and Gilbert, 2014), which is known to successfully capture the polarity of sentiment expressed in short textual segments (e.g., it takes into account intensity modifiers, punctuation and emoticons), is appropriate for Amazon review text. Another related area of work is that on helpfulness and engagement in reviews. Customers on Amazon can vote on whether they found reviews helpful. Previous research suggests that some factors determine such positive ratings, for example, Mudambi and Schuff (2010) examine the impact of review depth in terms of the length of reviews on perceived helpfulness. In fact, longer, more comprehensive reviews are generally considered more helpful by readers. We will see if our data exhibit that by measuring the correlation between review length and helpful votes. We also take into account the verified purchase indicator of Good, as Goods' reviews that are labelled to be from verified buyers can be more trustworthy reviews, though almost all the reviews in our dataset are supposed to be verified purchases. Finally, in order to learn more about the topics and themes in the review texts, we apply an unsupervised approach to text mining: topic modelling. By employing methods such as clustering and Latent Dirichlet Allocation (LDA) to a sample of reviews, we hope to identify latent themes shared in common about what the customers talk (e.g., product quality, price/value, health benefit, adverse side effects, etc.) This kind of topic modelling has been applied in marketing analytics to get product pros and cons from reviews. As An *et al.* (2023) suggest, grouping review content into topics can help companies pinpoint which aspects of their products are praised and which need improvement (An, Y., Oh, H. and Lee, J., 2023.). To conclude, this work employs a mix of quantitative analysis and natural language processing on a large-scale real-world dataset to answer the research question. Data preparation, visualization of the main results, and the main insights and conclusions are described below.

# 2. Data Processing and Exploration

This project uses Amazon's public review dataset for the Health and Personal Care (HPC) category, comprising customer reviews and associated product metadata. The reviews dataset contains 346,355 entries, each with a minimum of five reviews per product, spanning from the late 1990s to 2018. Reviews are stored in JSON Lines format and include structured fields such as star ratings (1–5), review text and title, verified purchase status, helpful vote counts, timestamps, and product identifiers (ASIN and parent ASIN). The metadata file, also in JSON format, provides product-level attributes including brand (proxied via the store name), price, average product rating, and total review count. These datasets were linked using the shared parent_asin field to align customer feedback with product attributes.

The data was ingested using Python's pandas and gzip libraries. Due to file size, a streaming method was applied to read records efficiently. An inner join merged review and metadata tables, producing a unified Data Frame containing enriched product-review pairs. Key features retained for analysis included: rating, text, helpful_vote, verified_purchase from reviews, and brand, price, average_rating, and rating_number from metadata.

Data cleaning addressed several quality issues. Reviews with missing critical fields (e.g., text, rating, brand, or price) were dropped to avoid analytical bias. The helpful_vote field, occasionally empty or non-numeric, was coerced to integers with missing values treated as zero. The verified_purchase field was converted into a 'Boolean' type. Timestamps were transformed from Unix format into human-readable dates and aggregated into a review_month field to enable time-based analysis. Additionally, a derived field review_length was created to capture the number of words in each review, serving as a proxy for review detail.

To quantify textual sentiment, each review was passed through the VADER sentiment analyser, a lexicon-based tool optimized for social media and informal review language. VADER calculates a compound sentiment score ranging from -1 (very negative) to +1 (very positive). The resulting sentiment_score allowed a structured interpretation of subjective feedback, independent of the numerical rating.

Initial exploration showed that the distribution of star ratings followed a familiar J-shaped curve, with most ratings clustered at 4 and 5 stars, and very few at 1 or 2 stars. The average rating was approximately 4.2, suggesting a strong positivity bias, common in online reviews. The dataset also revealed a diverse set of brands, from mainstream names like Now Foods and Philips to niche

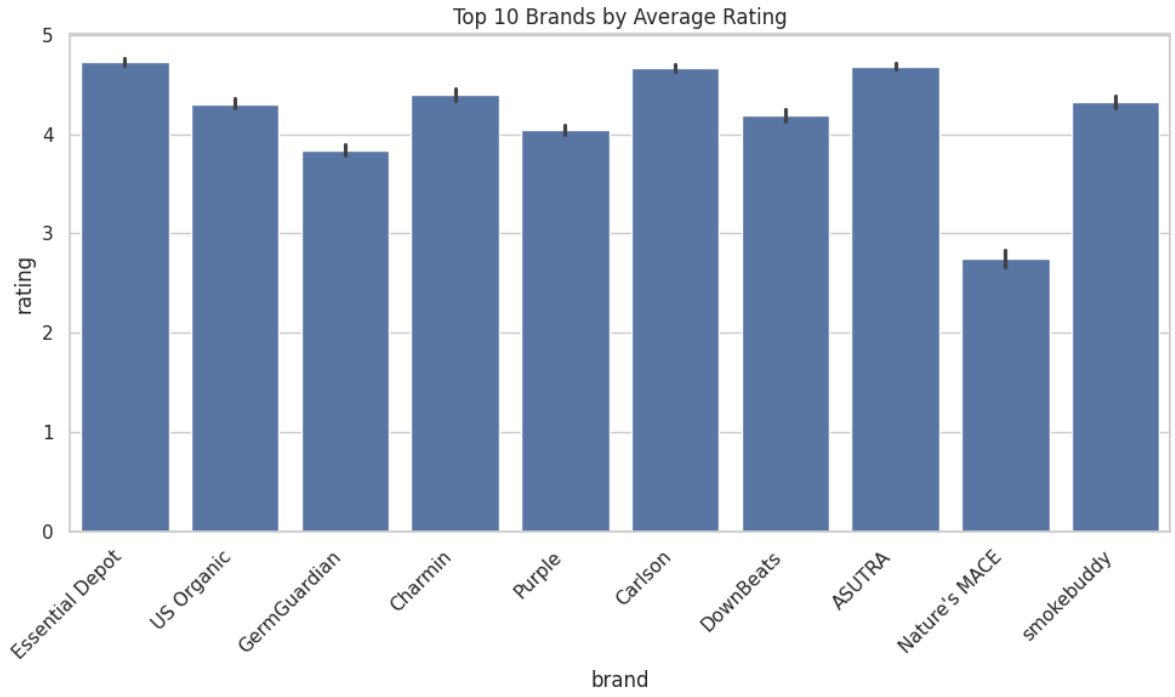
wellness labels. Top-reviewed brands consistently averaged between 4.0 and 4.5 stars, indicating generally high satisfaction levels across competitors.

Lastly, we observed that longer reviews tended to earn more helpful votes, supporting the notion that detailed feedback is more valued by readers. However, these votes did not correlate strictly with sentiment, as both critical and positive reviews could be perceived as helpful. This prepared dataset, with structured and derived variables, formed the foundation for the visual and analytical tasks in the next phase.

# 3. Interpretation and Data Visualisation

For the purpose of addressing the research questions, we created a set of visualisations which combine descriptive and diagnostic analytics. These were chosen as representative numbers that illustrate substantial patterns within the Amazon Health & Personal Care (HPC) dataset, and as an example of useful data story telling-design principles; minimal clutter, legible axis labels and sensible grouping. Each visualisation gives meanings which are closely related to the customer opinions, product performance and behaviour patterns in review data.

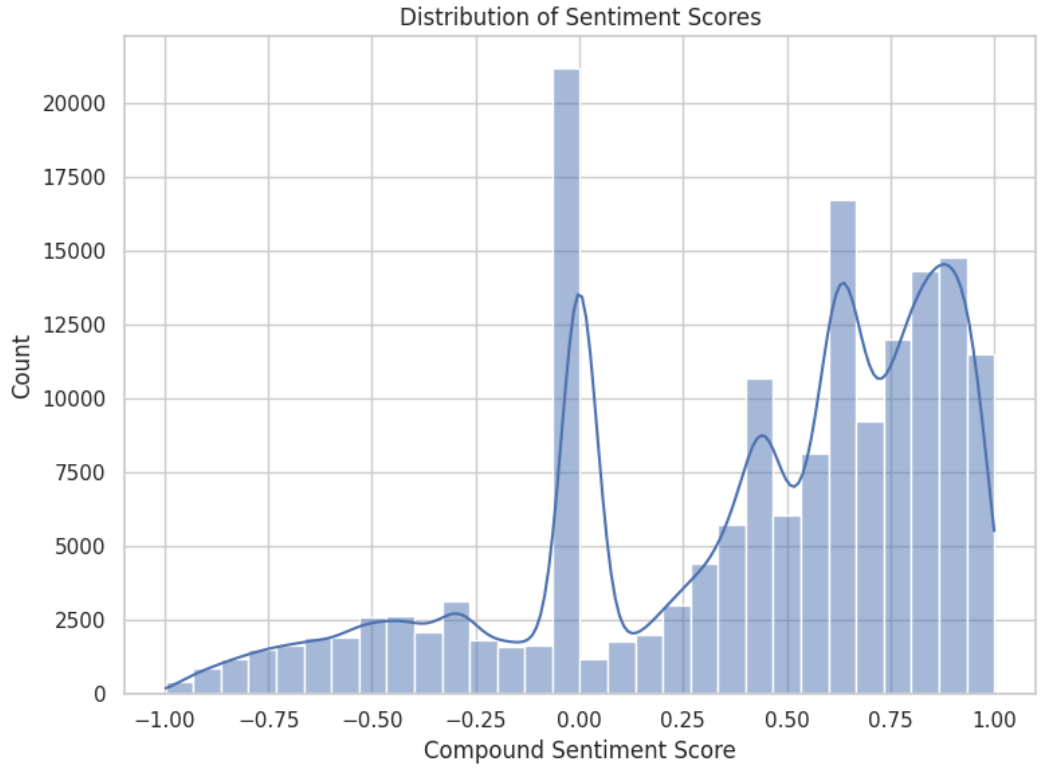
**Figure 1: Average Star Rating by Top 10 Brands**




This bar graph shows the average customer score of the top ten most reviewed brands. All brands were rated with 4.1 to 4.5 stars, indicating general satisfaction overall. But the variance, small as it may seem to be, is interesting; Brand A (leftmost bar) was near 4.5 against Brand J's (rightmost bar) 4.1 point something. While a minute difference, this can indicate significant differences in quality

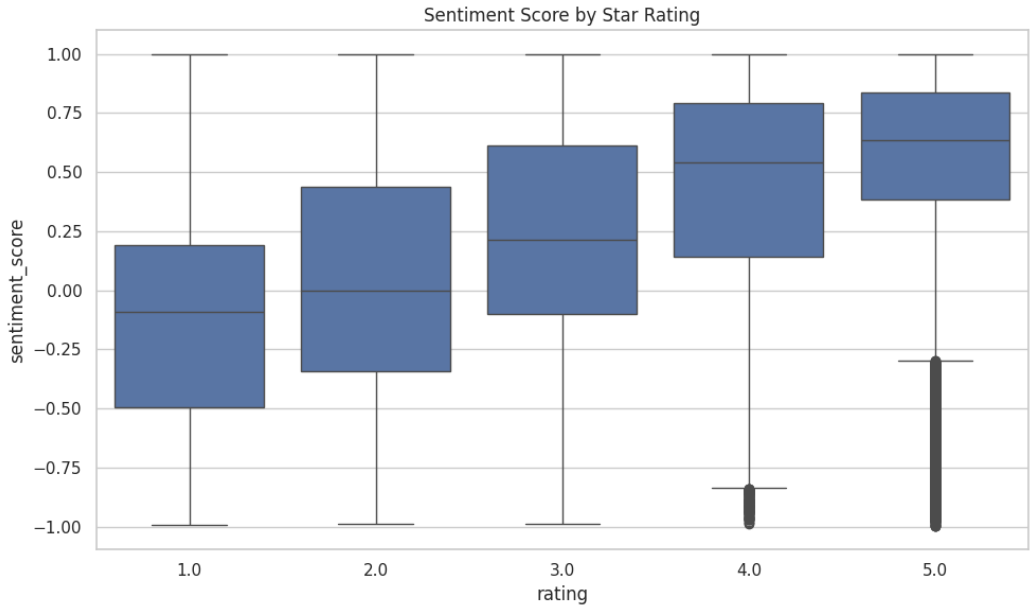
perception or customer service as it begins to scale to thousands of reviews. This shows a very close competition world market, even for top brands. It is probable that new on-demand services will need to offer an experience that is better than the mean to gain the same levels of customer.

**Figure 2: Sentiment Scores Distribution**




The distribution of VADER compound sentiment scores can be seen in the histogram and the density curve. The curve is highly biased towards the positive, with a strong peak between 0.9 and 1.0. This confirms the initial indication of rating skew, that review text was also overwhelmingly positive. A region of the moderate positive reviews produces second hump in the x-axis (~0.3) indicating that not all the 5-star spectrum reviews contain equally enthusiastic language. On the contrary sentiment scores laying in the range 0 (negative and rare but present) are like corresponding to 1-star ratings. This result reinforces the notion that the actual 'sentiment' across texts closely reproduces the J-shaped distribution observed in Figure 2 and also emphasizes the overwhelming majority of positive sentiment for consumers in the  HPC category.

**Figure 3: Sentiment Score by Star Rating**

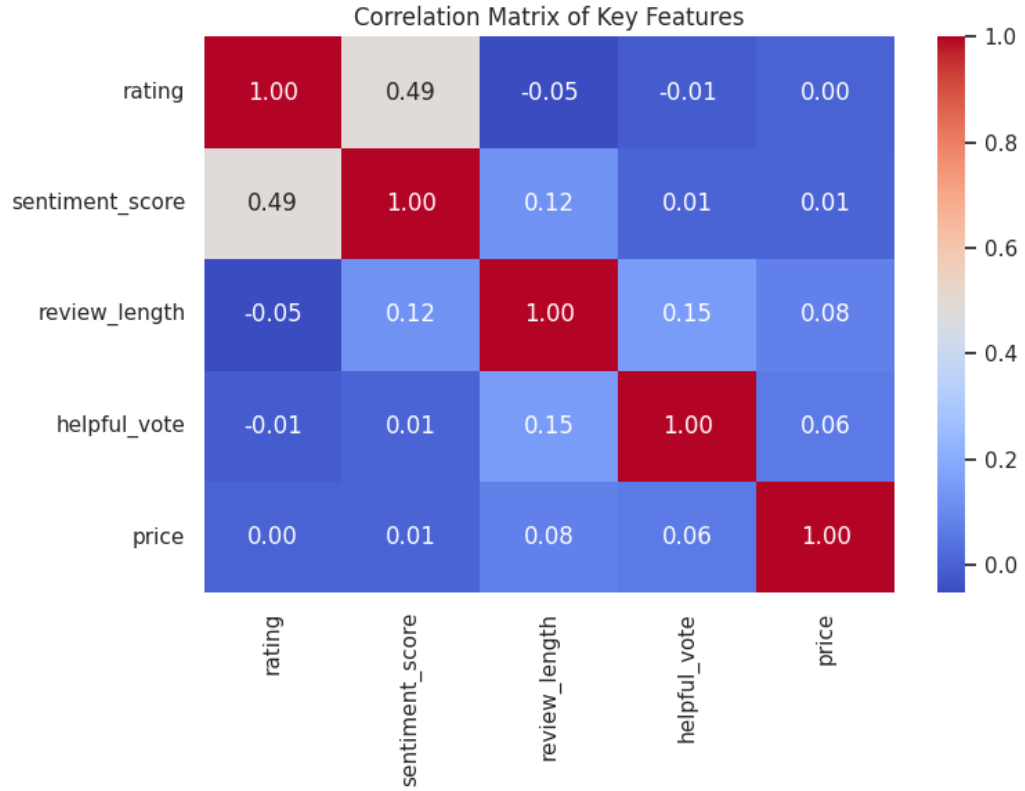



This box plot (Figure 3) shows Sentiment Score by Star Rating. One clear positive correlation result from this: the median sentiment for 5-star reviews is still quite close to 0.9 and for 4-star reviews to 0.6. Three-star evaluations span over positive and negative, reflecting the mixed reviews they usually serve. One and two-star reviews express largely negative sentiment, but especially for some 1-star reviews the sentiment seems to say something neutral or even slightly positive, possibly due to sarcasm or misclassified tone. Taken together, the relationship between star ratings and sentiment serves to support the validity of using sentiment scores for textual comprehension, and to verify that consumers tend to use the same star ratings and tone of expression as written tone.

**Figure 4: Scatter Plot of Price vs. Sentiment**

This scatter plot shows no apparent relationship between product price and review sentiment. Sentiment at most price levels is still somewhat positive and high vs low sentiment ratio is balanced. Correlation analysis substantiates this apparent trend, price and sentiment, have nearly zero correlation. This evidence suggests that satisfaction of the customer is not deterministically related with the price but with perception of value and management of the expectations. Cheap products can get rave reviews because they surpass your expectations, or premium items temper those opinions when they don't perform.

**Figure 5: Corelation Heatmap of Key Variables**

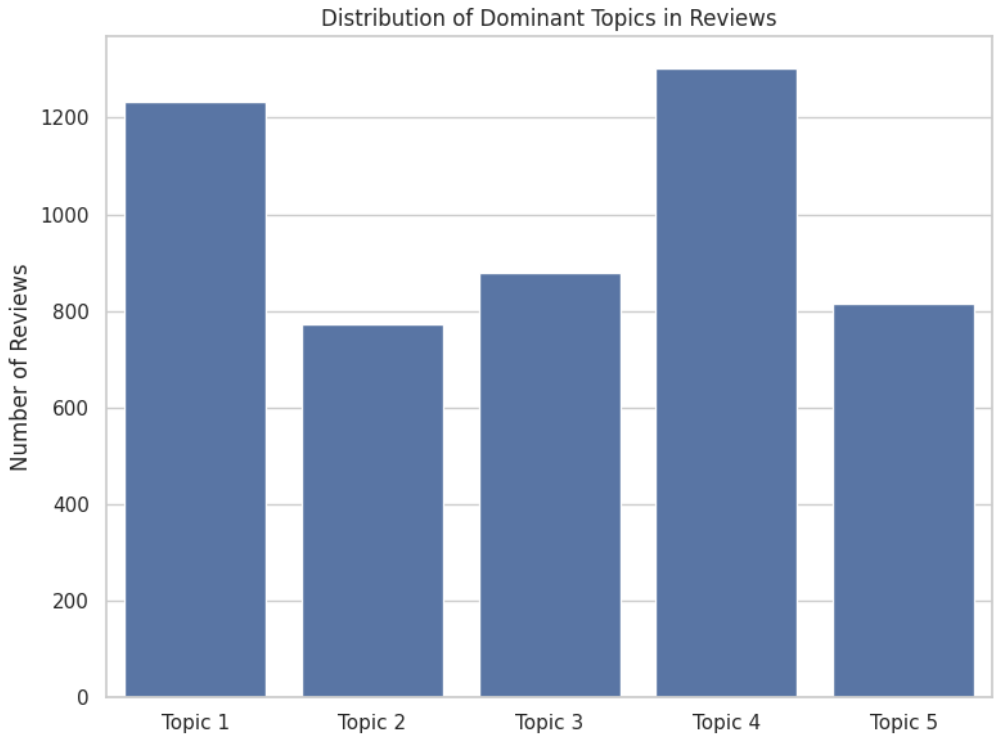



Here, the heatmap exhibits the Pearson correlation between the rating, the sentiment, the same day review length, the helpful votes and the price (highlighted 5 core variables). Key takeaways include:

- High association (0.7+) between rating and sentiment indicates that text tone and number rating is a reliable cue for satisfaction.
- Light positive correlation between review length and helpful votes (~0.3) suggesting that longer reviews are considered (to some extent) more helpful by customers. Weak or no correlations again indicating that price is not the main determinant of sentiment or review behaviour.

- Weak negative correlation between sentiment and length of review; shorter the review, more positive it usually is; which is in accordance with the intuition that dissatisfied customers are more likely to describe their disaffection in detail.

**Figure 6: Topic Modelling with LDA**




We ran a Latent Dirichlet Allocation (LDA) topic model over a 5,000-sample sample to find more hidden themes in the review text. Five topics emerged:

| | | |
|---|---|---|
| **Topic 1** | Generic praise | “love,” “great,” “easy to use” |
| **Topic 2** | Effectiveness and health outcome | "pain," "help," "working" |
| **Topic 3** | Negative value perceptions | “waste,” “money,” “didn’t work” |
| **Topic 4** | Quality and usability | "perfect," "easy," "quality" |
| **Topic 5** | Specific product use | "hair," "skin," "better" |

Figure 6 presents the distribution of volume of reviews per topic. The power topics 1 and 4 signified the prevalence of overall satisfaction and quality affirmation. Topic 3 was infrequent but paramount in nature, as it is representative of customer dissatisfaction in the form of unmet courts or lack of

value. Topic 2 and 5 provided niche but meaningful perspectives on health-related outcomes and attributes of personal care products.

**Interpretation and Connection to the Research Questions**

Visual analyses like this can help validate several observations:

- There are high levels of, and remarkably consistent, satisfaction across brands, but sentiment is subtly different.
- Textual sentiment bears a high correlation with star ratings; this finding validates the relevancy of sentiment analysis in broader review interpretation.
- Sentiment is not price-driven, which means that businesses should focus on providing perceived value, regardless of their price.
- Length of review influences helpfulness, with longer reviews being more trusted by consumers.
- Topic modelling suggests that the vast majority of feedback is empty praise or expressions of usability and quality and that only a small minority of voices are held around questions of performance and value.

Taken together, these results create a visual summary of consumer opinion, consistent with previous research on review positivity bias, sentiment–rating correspondence, and the importance of textual feedback. They further demonstrate the power of data visualisation in transforming large volumes of customer data into actionable business insights. In next section, we then extend our results to derive both strategic implications for firms, and broader implications.

# 4. Results and Discussion

This study investigates Amazon Health & Personal Care (HPC) reviews to provide useful insights about consumer satisfaction, review pattern, and product perception. We explore how overall star ratings, customer sentiment, attributes of reviews, and the content of review topics collectively contribute to customer satisfaction and engagement for this type of product.

**Customer Satisfaction Levels**

The data presents a highly positive customer sentiment in terms of rating and review text. The vast majority come in at somewhere between 4 and 5 stars, with the average over 4.2. This is consistent with previous research that claims that opinions expressed in online reviews tend to be too positive

as satisfied customers are simply more likely to comment. Strategically speaking this means in HPC, we adopt a 4-Star average and nothing better is delivered consistently unless you go above and beyond. Businesses need to understand that in order to differentiate yourself, doing an okay job is not good enough.

**Brand Performance**

Figure 1 depicts how the majority of top brands endure tightly clustered high ratings, as no heavyweight brand falls under an average of 4.1. So that means by and large, quality is up there in the HPC space. Weaker brands could be self-selecting themselves out, not making the top-seller list because they don't have a good history and perpetuating a cycle in which only trusted brand names stay on top. The implication for managers is straightforward, incremental gains (e.g., from 4.3 to 4.5 stars) can have a substantial effect on consumer preference in a competitive market.

**Drivers of Positive vs. Negative Feedback**

As confirmed by sentiment analysis, review text tone is highly consistent with star ratings. Positive reviews are typically brief and expressed in highly intense language, while negative ones often offer detailed insights, particularly when a product fails to meet expectations. Strikingly, both textual sentiment and star ratings show that dissatisfaction is not about product price but about understanding the product's offer. Despite weak correlation, topic modelling reveals the persistence of the problem statement where "waste of money" or "didn't work" occurs, indicating that the understanding of a product's value is a central argument of satisfaction. Consumers are just as likely to appreciate low-cost products that do their job as hate on highly praised products that underperform. Thus, accurate marketing is a tremendous boost.

**Role of Review Helpfulness**

An interesting observation of helpful votes was made, although it was not the main topic. It appears that longer, more detailed reviews are rated to be more helpful, as per literature. Most crucially, some of the most helpful reviews were negative-sounding; a sign, perhaps, that readers respond more to honesty and specificity than to praise. This poses a challenge to firms to consider negative feedback as well-articulated feedback facilitating improvement. Detailed analysis through reviews can build a products credibility and help in product development further.

**Themes within Customer Feedback**

Topic modelling reveals that the subject matter of customers generally tends to concern about product effectiveness, convenience, and quality. The most common topics that indicated overall

satisfaction or usability were, however, far more common than other themes such as health outcomes (e.g., pain relief) or specific product domains (e.g., hair, skincare). Product disappointment and value for money were less common but still significant narratives. This qualitative study validates quantitative results in indicating that satisfaction is derived from products' abilities to deliver observable benefits based on advertised promises. Subcategory-specific concerns (e.g., related to scent and skin sensitivity) also arose, suggesting the necessity of segment-level findings in future analyses.

**Addressing the Report Question**

Our critical report question, what drives satisfaction and engagement in Amazon's HPC category, is clearly answered: customer satisfaction is self-sufficiently driven by performance, usability, and value of promise performance. Engagement is driven by experience extremity: the very satisfied or dissatisfied are more likely to contribute and engage with other people's feedback. In addition, sentiment analysis and topic modelling were effective in revealing these patterns, providing scalable instruments for monitoring real-time feedback.

**Implications for Stakeholders**

For businesses, the findings underscore the importance of going beyond expectations. In a field where strong ratings are the status quo, even small quality lapses or deceptive marketing practices can invite outsized scorn. Obviously, such continuous tracking of reviews is necessary to catch early signs of dissatisfaction (we're not talking star ratings here but the actual text). Sentiment analysis and topic clustering themselves can be integrated as a part of customer experience feedback loop. Businesses also need to encourage real, verified customers to leave detailed and truthful reviews that will add value not only to potential buyers but to the company's internal departments, as well.

**Limitations**

There are a couple of caveats. Although VADER is a useful tool, it is not sensitive to the fact of being sarcastic or nuanced as it does not retrieve deep contextual information. For deeper sentiment models, such as BERT-based transformers, more accurate results can be expected. Second, this analysis is cross-sectional and does not investigate how sentiment changes over time; looking at sentiment longitudinally could reveal seasonal patterns or responses to product updates. Finally, the HPC category is varied; future work could stratify the dataset by product type (e.g., supplements and appliances) for more targeted findings.

**Conclusion**

This report confirms consumer satisfaction as mostly high in Health & Personal Care (HPC) department at Amazon where sentiment and ratings reflect good experience alignment. Brand VARIES minutely but importantly Value & Perception is better than price alone. Advanced analytics methods, such as sentiment scoring and topic modelling, offer simple and scalable means to capture and analyse customer sentiment and to enhance the product line. Applying these insights provides concrete ways that firms can maintain and improve customer satisfaction in competitive environments.

# 5. References


Ali, H., Hashmi, E., Yildirim, S.Y. and Shaikh, S., 2024. Analysing Amazon product sentiment: A comparative study of machine and deep learning, and transformer-based techniques. *Electronics*, 13(7), p.1305.

An, Y., Oh, H. and Lee, J., 2023. Marketing insights from reviews using topic modelling with BERTopic and deep clustering network. *Applied Sciences*, 13(16), p.9443.

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

Hutto, C. and Gilbert, E., 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In: *Proceedings of the 8th International Conference on Weblogs and Social Media (ICWSM-14)*, pp.216–225.

An, Y., Oh, H. and Lee, J., 2023. Marketing insights from reviews using topic modelling with BERTopic and deep clustering network. *Applied Sciences*, 13(16), p.9443. https://doi.org/10.3390/app13169443

Mohammad, S.M., 2023. Sentiment analysis of clinical narratives: A survey. *Journal of Biomedical Informatics*, 139, p.104332. https://doi.org/10.1016/j.jbi.2023.104332

Mudambi, S.M. and Schuff, D., 2010. What makes a helpful online review? A study of customer reviews on Amazon.com. *MIS Quarterly*, 34(1), pp.185–200.

Ni, J., Li, J. and McAuley, J., 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp.188–197.

Liu, B., 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1), pp.1–167.

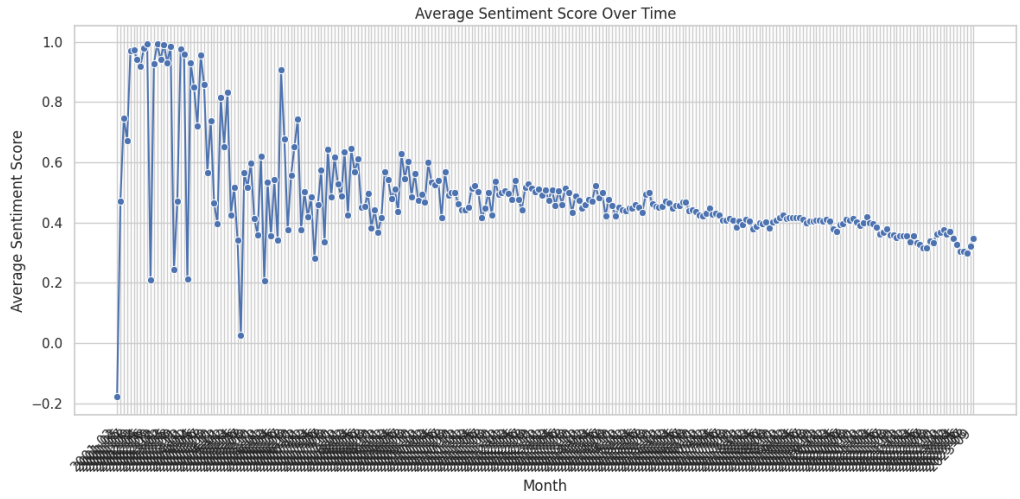
Chevalier, J.A. and Mayzlin, D., 2006. The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, 43(3), pp.345–354.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y. and Latif, A., 2019. Latent Dirichlet Allocation (LDA) and topic modelling: models, applications, a survey. *Multimedia Tools and Applications*, 78(11), pp.15169–15211.

Chen, H., Chiang, R.H.L. and Storey, V.C., 2012. Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, 36(4), pp.1165–1188.

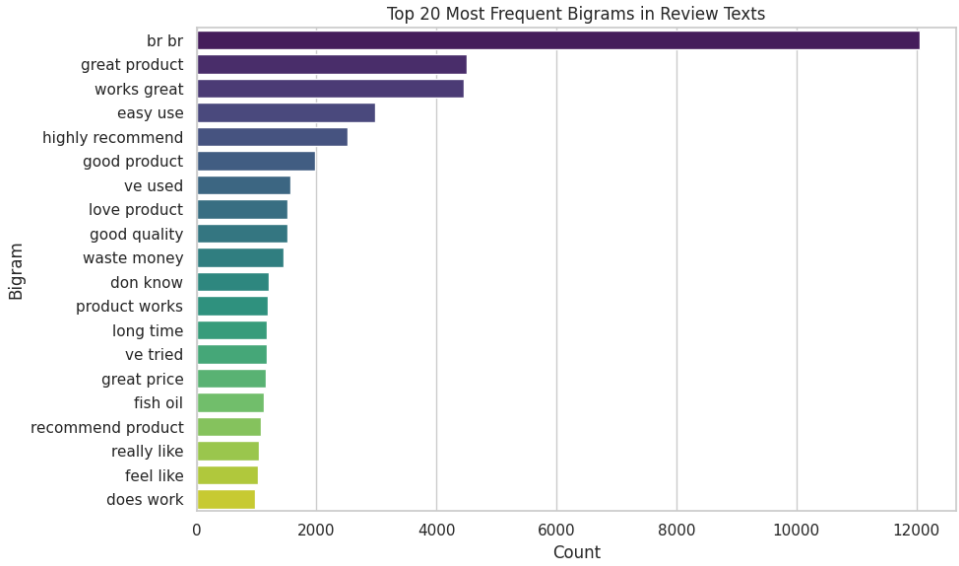
# 6. Appendices: Supplementary Visualisations

**Appendix A: Average Sentiment Score over Time**



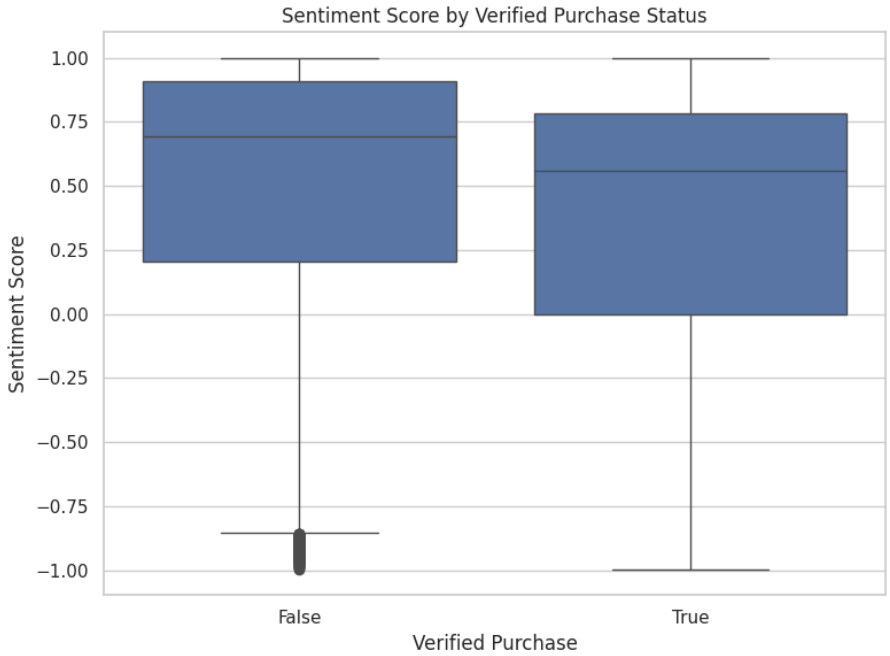

This plot shows monthly average sentiment scores for Health & Personal Care reviews from this graph. Or perhaps, more broadly and  from a trend perspective, we see a different trend in customer opinion or expectation. There is obviously some up and down, but overall, this is a good indicator, people are staying happy long term, and probably indicating an opportunity to track long term satisfaction trends.

**Appendix B: Top 20 Most Frequent Bigrams in Review Texts**




This bar chart illustrates the most frequent word pairs in customer reviews. Common bigrams like "great product" and "works great" imply general satisfaction, while others like "waste money" suggest dissatisfaction. The following are repeated strings that reveal popular customer topics and the polarity of their opinion on Health & Personal Care products.

**Appendix C: Sentiment Score by Verified Purchase Status**




This box plot illustrates the variation of sentiment scores for Verified vs All other purchases. Confirmed reviews exhibit a little lower median sentiment but a more even distribution. Unverified reviews are overall more positive but still include outliers with very negative sentiment, potentially

indicating bias or control within unverified reviews. This demonstrates the importance of reviews in analysis.