Project Milestone 2

Analysis of Amazon Health and Personal Care Customer Reviews for Customer Satisfaction Improvement

BA820 - B1 Group 5

Jyun-Ru Huang, Pin-Hao Pan, Ruo-Rong Wang, Zehui Wang

3 March 2025

Proposal - Updated to capture new ideas

Our investigation focuses on extracting actionable insights from Amazon's Health and Personal Care product reviews. Our findings will provide insights into product performance, improve customer satisfaction, and enable businesses to understand consumer sentiment across platforms beyond Amazon. We aim to:

1. **Dive into the Sentiment Trend for 3-Star Reviews**: investigate whether 3-star ratings lean positive or negative using sentiment analysis techniques.
2. **Identify Themes of Reviews Based on Product Subcategories**: we want to extract patterns of satisfaction and dissatisfaction from customer reviews (average rating, common complaints, etc.) of each product subcategory (defined through clustering or topic modeling) for more domain-specific insights.
3. **Develop a Sentiment Prediction Model**: with our labeled review data, we want to build a machine learning model to predict sentiment from textual reviews, which could be useful for sellers to quantify customer sentiment with external data sources (e.g., Twitter reviews), even in the absence of explicit star ratings.

Exploratory Data Analysis (EDA) & Preprocessing

Our datasets are Amazon's product review data and product metadata of the Health and Personal Care category, collected by McAuley Lab (Computer Science Department, UC San Diego) in 2023. The user reviews subset includes 494,121 entries across 10 columns, while the item metadata subset has 60,293 entries across 14 columns. The two datasets can be connected via the 'parent_asin' column, which is essentially the product id. In M1, we dropped some columns due to their irrelevance or uninformativeness. We transformed User_Review['timestamp'] into separated columns of time, and preprocessed textual data as our investigation focused primarily on text. These steps from M1, including text preprocess and tokenizing methods, remain unchanged. During our M1 EDA, we found that certain review text is extremely long (outlier by word count), which may pose challenges to our clustering steps and sentiment analysis. In M2, we have made the following progress:

Topic Modeling for Product Subcategory Defining: we realized that, since our dataset does not contain a column for the subcategories of products, we would need to define this by ourselves. We applied both clustering and topic modeling methods on the metadata's ['title'] column, hoping to classify products based on keywords in their titles. We first tried K-Means and Hierarchical Clustering. Unfortunately, they result in very low silhouette scores (0.02 - 0.08 for both methods), which indicates ineffective clustering (Supplemental Material 1).

Since clustering did not work well on our data, we tried topic modeling methods like LDA, NMF, and BERTopic. During this, we found some top words seem irrelevant to our domain:

```
--------------------------------------------------
Topic 5: party, blue, large, set, cap, gold, black, nose, size, silicone, kit, birthday, banner, christmas, pieces, kids, balloons, shoe, cover, pairs, bath_bombs, whit
e, wedding, or, color, water, replacement, pure_pharmaceutical_grade_safety, usp_kosher, glycerin_vegetable_quart_non
```

"Party", "Birthday", "Christmas", and "Wedding" seem irrelevant, so we decided to remove all products containing these keywords in their title to keep our analysis focus on the health domain. We started with LDA, but soon realized there is a tradeoff between LDA Coherence Score and the interpretability of topics. We eventually chose a set of hyperparameters that resulted in a low score (0.4152) but gave more interpretable top words (Supplement Material 2), but there was still some overlaps between topics. LDA also took a relatively long time to run. We then tried NMF

topic modeling. With TF-IDF vectorizing, this achieved a more interpretable top words (Supplement Material 3), as well as a higher coherence score (0.5447). We also did BERTopic, but it struggled to return a balanced distribution of products between topics when we reduced topics to 10 (Supplement Material 4).

We eventually chose the TF-IDF + NMF method; please refer to our topic modeling notebook for more details. After finalizing topic modeling, we summarized each topic based on its top words and added a new column to label each product. For example, Topic 10, with top words including *"pill, powder, supplement, electrolyte, kosher, vitamin, capsule, container, serving, organizer, flavor, box, gmo, protein, drink, non, sugar, day, hydration, vegetable"*, was summarized as "Nutritional Supplement." With this new column added, we were able to conduct further analysis on each subcategory and extract more specific insights into different product types.

Vectorization Method Finalization: To make sure we process the reviews properly, we tried different vectorization methods on our textual data. Considering that we are handling a large amount of unstructured data with limited computational resources, we eventually finalized with TF-IDF for its relatively fast processing time (compared to Word2Vec) and its ability to weigh more unique words with higher importance.

Analysis & Experiments

Analysis 1: Should a rating of 3 be classified as positive or negative?

Amazon customer reviews follow a 5-point rating scale. We can categorize 1-star and 2-star reviews as negative and 4-star and 5-star reviews as positive. This raises an important question: What about 3-star reviews? Should they be considered positive or negative? To resolve this ambiguity, we decided to quantify review sentiment and compare their cosine similarity. We used cosine similarity because it measures the directional similarity between text vectors regardless of magnitude. This is useful because we want to compare the semantic closeness of reviews of different rating groups.

We used TF-IDF to vectorize our review data, because it is efficient and weighs important words while reducing the impact of frequently occurring, non-informative words. We used a maximum feature limit of 150 words of our TF-IDF vectorizer to balance expressiveness and computational efficiency. For all the vectorized reviews of 5-star rating (most positive) and 1-star rating (most negative), we used the average of their vectors to calculate a centroid, representing the most representative positive and negative review sentiment. Then, for each 3-star review, we computed the cosine similarity of that review with the centroid of 1-and-5-star-review. This gives us an idea of each individual 3-star rating's closeness to 1 (negative) and 5 (positive), and we were able to classify every 3-star review based on their similarity. Through this process, we quantified sentiment polarity of 3-star reviews, identifying if they leaned more positive or negative. We found that 73.11% of 3-star reviews are closer to 5-star review's centroid point and the rest 26.89% are closer to 1-star's centroid. Since the majority of 3-star reviews are closer to the 5-star centroid, **we classified 3-star reviews as positive reviews.** This classification helped us understand the behavior of 3-star Amazon reviews, and also set the foundation for our Logistic Regression in later sections. To validate our findings, we randomly selected 10 data points from each rating category and used PCA to visualize these selected data points in a 3D plot. See Supplement Material 5 for our PCA plot.

Analysis 2: Update on Logistics Regression from M1

In M1, we used logistic regression for sentiment analysis and compared the results by applying three different vectorization methods in preprocessing. Since our previous analysis classified 3-star reviews as positive, which differed from our assumption in M1, we decided to reattempt the classification to see whether we could improve predictive performance. This time, we also incorporated n-grams (tri-grams) into our vectorization process, as n-grams enable us to capture contextual relationships between words.

See Supplement Material 6 for our updated Logistics Regression results.

Analysis 3: BoW & POS Trigrams: Analyzing Top 20 Key Customer Reviews Across Product Categories

One of the primary aims of this project is to uncover the most common positive and negative keywords in certain categories of product reviews. By isolating and ranking these terms, we seek to understand the major pain points (negative feedback) and the strengths (positive feedback) that customers highlighted. Focusing on the language customers use in reviews allows us to glean meaningful insights for product improvement, customer satisfaction, and marketing strategies.

Based on our result from topic modeling of product titles (introduced in EDA), we labeled our dataset with a new column, "subcategory", with the 10 topics we extracted from topic modeling:

Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
Hair Accessories	Fish Oil	Fitness Accessories	Joint Support	Portable Comfort Products
Topic 6	**Topic 7**	**Topic 8**	**Topic 9**	**Topic 10**
Odor & Germ Elimination	Beauty & Cleansing Tools	Health-Related IoT	Sleep Accessories	Nutritional Supplements

See Supplement Material 7 for a detailed composition of the ten subcategories.

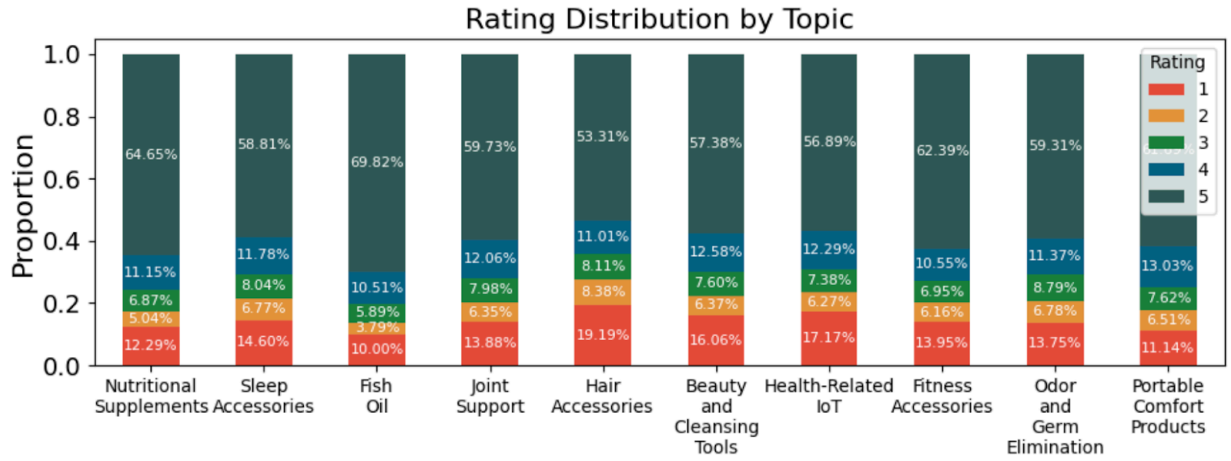
From the subcategories produced by topic modeling, we classified the remaining reviews into positive (3, 4 or 5 stars) and negative (1 or 2 stars) groups. This allowed us to isolate both the compliments and complaints customers shared, thereby offering a well-rounded perspective on how different products perform in the eyes of real users.

To capture sentiment-laden expressions, we first use BoW to identify high-frequency words and manually divide them into positive and negative sentiment groups. We defined two sets of keywords: allowed_positive (e.g., "good," "excellent," "amazing," "recommend") and allowed_negative (e.g., "bad," "waste," "terrible," "toxic"). We then utilized NLTK's pos_tag to assign part-of-speech labels to each word in a review. Based on these labels, we apply POS Trigrams to extract three-word phrases based on grammatical structure. For this part, we initially considered unigrams or bigrams for context extraction, but these were often too broad or lacked sufficient context to capture the nuance of user complaints. For instance, single words like "bad" or two-word phrases like "bad product", which did not convey enough detail about the nature of the problem, while "bad allergic reaction" is more informative. Consequently, we shifted our approach to using three-word patterns (trigrams), which struck a better balance between capturing contextual detail and avoiding overly sparse data. In this case, we can capture contextually meaningful phrases. By grouping the extracted phrases by product subcategory, we can focus on more specific terms and identify the most common positive and negative opinions

within each category. This approach helps us gain deeper insights into customer sentiment for different types of products.

Findings and Interpretations

Rating Distribution of 10 Subcategories:



In our defined 10 subcategories, we attempt to understand the components of ratings. Interestingly, the differences are significant: Nutritional Supplements and Fish Oil, which do not provide instant effects, tend to receive less negative feedback, whereas Hair Accessories and Beauty & Cleansing Tools, which deliver immediate results, are more prone to get negative feedback. Given that most reviews are positive, the posting of negative feedback typically serves as a way for frustrated or dissatisfied customers to express their discontent.

Top Review Phrases from 10 Subcategories:

According to the rating distribution bar chart, we can tell that the Fish Oil category has the highest proportion of positive sentiment. Using BoW & POS Trigrams analysis, keywords like "good_quality_oil," "great_quality_product," and "good_quality_brand" suggest that consumers generally perceive these products as high quality. Additionally, customer service also performs well. Keywords such as "excellent_customer_service," "great_customer_service," and "good_customer_service" indicate that consumers have a favorable view of the support they receive when purchasing these products.

The Nutritional Supplements category holds the second-highest proportion of positive reviews. Mentions of "good_price_point" suggest that affordability and perceived value for money contribute to the positive feedback. Additionally, "super_fast_shipping" underscores the importance of efficient delivery services, while "good_size_pill" and "good_pill_organizer" highlight the significance of product design and convenience.

On the other hand, the category “Hair Accessories” has the highest proportion of negative reviews. One of the most notable issues is product effectiveness and quality, as seen in keywords like "s_total_waste," "ineffective_sticky_stuff," and "Horrible_quality_product." These terms indicate that consumers feel some products do not perform as promised, leading to frustration and disappointment. Mentions of "pretty_darn_toxic" and "un_rash_horrible" may raise concerns about product safety and formulation (Supplement Material 8 for details).

In the Health-Related IoT category, consumer feedback highlights several key areas of dissatisfaction, primarily centered around product quality, durability, safety, and pricing. Keywords such as "poor_quality_thread," "low_quality_material," and "cheap_flimsy_plastic suggest that consumers perceive certain health-related IoT devices as being made from subpar materials. Additionally, complaints about "low_handle_height" and "flimsy_little_arm" suggest that some devices may not be ergonomically designed or user-friendly.

In conclusion, these findings highlight the importance of aligning product features with consumer needs and expectations. Addressing concerns related to comfort, functionality, and product safety can help businesses enhance customer satisfaction, improve brand perception, and ultimately drive product success in the market. By leveraging consumer insights, companies can refine their product development strategies and create more reliable, high-quality offerings that better serve their target audience. From Amazon's perspective, monitoring review data across millions of products helps maintain a high level of trust and satisfaction among shoppers. Amazon wants customers to have a smooth buying experience, and consistent negative feedback—like "unsafe" or "expired"—can hurt Amazon's reputation if not addressed. By spotting these common complaints early, Amazon can refine its product listings, highlight important safety information, or even remove problematic items from the marketplace.

Challenges, Dead Ends & Adjustments

One of the biggest challenges in our project was handling a large, messy, and unstructured dataset while working within limited time and computational resources. Initially, we applied clustering techniques (K-Means, hierarchical clustering) to both product titles and reviews, hoping to uncover meaningful patterns. However, despite extensive efforts, these methods failed to produce useful clusters. Additionally, we found that some Amazon sellers miscategorized products, such as birthday party supplies under Health & Personal Care, introducing further noise. To address this, we pivoted from clustering to topic modeling and implemented additional filtering criteria to focus on relevant products. From this process, we learned that clustering is not always the best method for text-based analysis—especially when the goal is to extract interpretable insights rather than just grouping data points.

Another major challenge was text cleaning and preprocessing. Despite multiple iterations, unexpected, irrelevant, or misspelled words continued to appear in top-word extractions, distorting topic modeling results. Unlike structured data, where missing values can be handled with statistical imputation, text data has no clear metric to evaluate cleaning quality. Also, text preprocessing lacks a direct evaluation method, making it difficult to determine whether an approach improved or worsened data quality. To address this, we developed an iterative process of cleaning → modeling → analyzing, where we manually reviewed top extracted words and adjusted preprocessing accordingly. Our key takeaways from this process was that text cleaning has no one-size-fits-all approach, and we had to refine our process through trial-and-error.

We also encountered an unexpected failure with incorporating n-gram methods leading to a decline in all logistic regression benchmarks, presenting a challenge for our analysis. We suspect this was due to switching from Word2Vec embeddings (used in M1) to TF-IDF, which was necessary to apply N-grams but may have lost semantic context captured by Word2Vec. This shift highlighted a key limitation of TF-IDF—while it improves phrase-level representation, it lacks the ability to capture word meaning and relationships, ultimately leading to weaker model performance.

Appendix

Contribution:

Name	Task Done before M2	Task Done for M2
Jyun-Ru Huang	EDA: review text lengths/helpful votes related antalysis; review ratings/helpful votes related antalysis M1 preliminary results and analysis	Clustering (K-Means and Hierchical) of reviews, Analysis 1 (Cosine similarity), Part of Analysis 2 (NMF Topic Modeling), Organizing the Notebook File
Pin-Hao Pan	EDA: user_id/ helpful votes related antalysis; text vectorization method; M1 preliminary results and analysis	Clustering (K-Means and Hierarchical), M2 Problem Statement Refinement, Part 3 Analysis (BoW & POS Trigrams), Findings and Interpretations
Ruo-Rong Wang	EDA: active_user/ helpful votes related analysis; Text vectorization method; Data handling	Clustering (K-Means and Hierarchical),Sentiment Analysis by Logistic Regression and Random Forest, Part 3 Analysis (BoW & POS Trigrams), Findings and Interpretations
Zehui Wang	EDA: time-trend related analysis; text preprocessing, text tokenization, M1 composing of Challenges, EDA, Preprocessing of dataset, and Analysis Plan	Clustering (K-Means and Hierchical) and Topic Modeling (LDA, NMF, BERTopic) for product titles to identify product sub-groups, M2 composing of Proposal, EDA, Challenge, and overall M2 refinement

- **GitHub Project Repository Link:** https://github.com/pinhaopan/Analysis-of-Amazon-Health-and-Personal-Care-Customer-Reviews-for-Customer-Satisfaction-Improvement

- **References**:

https://chatgpt.com/share/67c3c4bb-0c04-8012-bc30-d82862a5c766

https://chatgpt.com/share/67b6aed5-6794-8012-b069-3446a89ef5b3
https://chatgpt.com/share/67c67f75-9238-8006-8de0-3b45d61e2561
https://chatgpt.com/share/67c683c3-7e70-8010-8f79-5461ef40ae82
https://chatgpt.com/share/67c68740-4b58-8005-b19a-65e6f55a2b5b

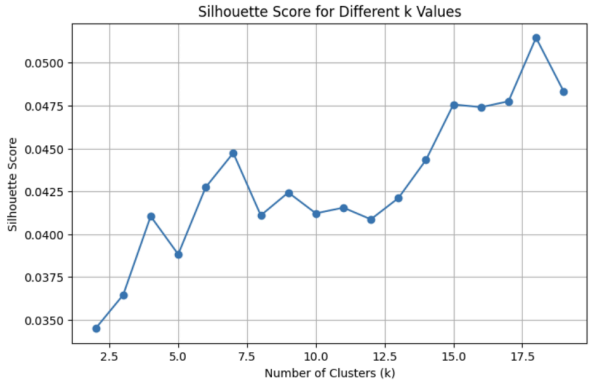
We ask ChatGPT to summarize text preprocessing techniques to help us understand their purpose and determine which ones to apply. Additionally, it helps fine-tune our wording to ensure our ideas are clearly articulated. Moreover, it clarifies every coding step when we are confused.
When we try to use the Word2Vec function, the dataset is too large, causing computational issues. To resolve this, we seek assistance from ChatGPT.
We also asked ChatGPT for many concept explanations and clarifications to help us better understand and compare different methods.

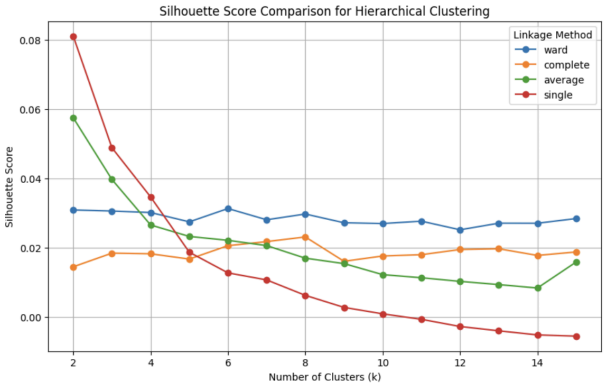
- **Timeline:** Provide a rough timeline for completing tasks.

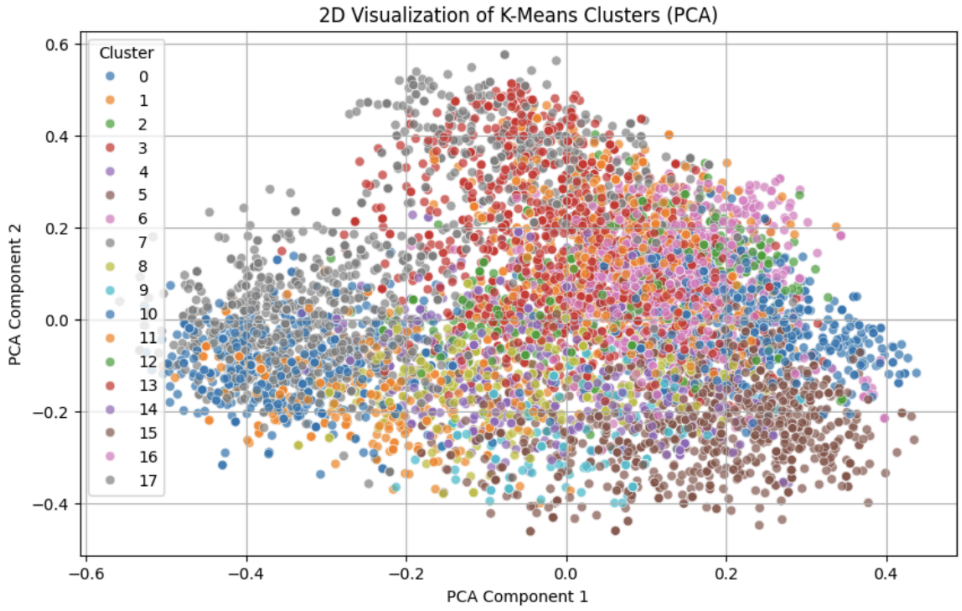
Tasks	**Start Date**	**End Date**	**Duration**
Data Cleaning	2/6	2/9	4days
Exploratory Data Analysis	2/9	2/12	4days
Feature Engineering	2/12	2/14	3days
Text Vectorization	2/15	2/17	2days
Analysis 1	2/18	2/20	3days
Analysis 2	2/21	2/25	5days
Analysis 3	2/26	3/1	4days
Documentation	3/1	3/3	3days

Supplemental Data and Results

Supplemental Material 1: Silhouette Score and 2D PCA for Product Title Clustering







Conclusion: Ineffective clustering - not applicable to our dataset

Supplement Material 2: Top 20 Words Extracted from LDA Topic Modeling

```
Topic 0: brush, facial, body, shower, sponge, bath, glass, clean, scrubber, head, replacement, electric, steel, stainless, face, handle, wipe, wo
man, beard, clipper
Topic 1: foot, massage, massager, roller, remover, file, callus, spa, skin, pedicure, solution, care, clean, muscle, deep, dry, tool, heat, valu
e, clear
Topic 2: color, light, ultra, red, free, water, medical, nail, filter, bottle, supply, original, paper, probiotic, repair, fit, tablet, patch, ba
by, cup
Topic 3: hair, ear, iron, noise, sleep, ceramic, flat, machine, plug, air, hair_straightener, portable, sound, snore, straighten, plantar, fascii
tis, brush, fast, scale
Topic 4: supplement, powder, capsule, support, organic, vitamin, natural, health, energy, liquid, fish_oil, vegan, formula, hair, strength, flavo
r, joint, nongmo, extract, free
Topic 5: clean, protein, mat, natural, cleaning, bar, organic, towel, cloth, fitness, bag, safe, restore, equipment, reusable, yoga, aroma, gear,
residue, pad
Topic 6: toothbrush, non, air, head, muscle, cap, reduce, technology, odor, base, gmo, uvc, sanitizer, pro, guardian, ggpk, germsfreshen, germgua
rdian, petssmokemoldcooke, laundrypack
Topic 7: mask, compression, glass, sock, holder, strap, bandage, woman, adjustable, support, adult, man, ankle, eyeglass, sleeve, cotton, face_ma
sk, tissue, nail_file, knee
Topic 8: oil, tooth, essential, organic, natural, water, spray, whiten, pure, scale, cold, shoe, digital, repellent, electrolyte, peppermint, bat
hroom, moisturizing, dropper, powder
Topic 9: pill, box, design, case, battery, travel, glove, organizer, ounce, portable, make, pressure, day, reduce, plastic, usa, comfort, pump, c
ushion, free
Topic 10: pad, woman, man, neck, pain_relief, support, pain, brace, adjustable, eyelash, shoe, shoulder, foot, heel, gel, electric, insole, pair,
posture, eye
```

Coherence Score = 0.4152

Conclusion: Takes time to run; Some topics have overlapping products

Supplement Material 3: Top 20 Words Extracted from TF-IDF + NMF Topic Modeling

```
Topic 1: sponge, silicone, makeup, brush, body, shower, facial, bath, scrubber, skin, loofah, exfoliate, massager, scrub, cosmetic, latex, care, handle, soft, lotion
Topic 2: clean, mat, natural, cleaning, towel, restore, mindful, refreshe, deepcleanse, asutra, wmicrofiber, slippery, lemongrass, residue, yoga, fitness, come, arom
a, gym, gear
Topic 3: pain, relief, support, neck, brace, posture, shoulder, corrector, adjustable, pad, upper, provide, therapy, knee, clavicle, compression, heating, cervical, m
uscle, traction
Topic 4: oil, organic, pure, essential, powder, cap, usda, dropper, supplement, certify, peppermint, available, wimprove, variation, fish, natural, capsule, liquid, v
itamin, pill
Topic 5: hair, iron, straightener, ceramic, curler, brush, straighten, trimmer, flat, curl, styling, comb, shaver, straight, fast, electric, tool, eyelash, heating, s
alon
Topic 6: design, cushion, pressure, comfort, chair, grid, gaming, ultimate, usa, reduce, travel, relieve, seat, charcoal, backnobber, positive, tooth, foam, ergonomi
c, water
Topic 7: foot, file, remover, callus, heel, pedicure, shoe, electronic, nail, tool, insole, skin, spa, massage, pair, massager, pad, dead, crack, roller
Topic 8: air, odor, guardian, uvc, sanitizer, deodorizerkill, petssmokemoldcooke, germsfreshen, germguardian, laundrypack, ggpk, pluggable, technology, reduce, scale,
spray, freshener, bathroom, digital, eliminator
Topic 9: mask, face, glass, ear, adjustable, box, sleep, adult, reusable, pill, case, cotton, cover, cloth, comfortable, washable, holder, scale, strap, buttonsmith
Topic 10: toothbrush, head, replacement, electric, brush, compatible, rechargeable, tooth, water, oral, shaver, sonic, mode, waterproof, battery, dental, cordless, fl
osser, sonicare, usb
```

Coherence Score = 0.5447

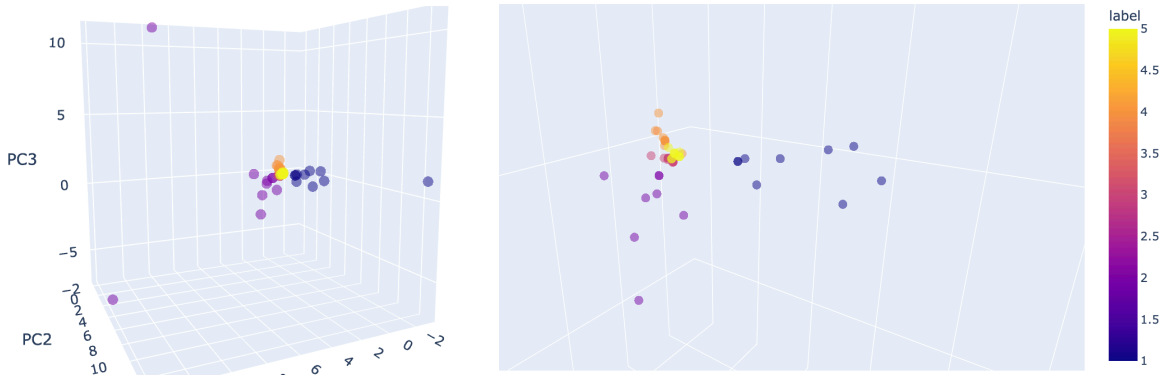
Conclusion: We think this set of top words makes more sense for interpretation. Also, this method achieve a higher score and took shorter time to run. We eventually chose this method.

Supplement Material 4: Distribution of BERTopic after Topic Number Reduced to 10

Topic	Count
-1	1617
0	6928
1	270
2	192
3	93
4	88
5	47
6	38
7	13
8	12

Conclusion: We find this distribution of topic not reasonable, therefore this method is not suitable for our task here.

Supplement Material 5: PCA Plot of 10 Randomly Selected Product Review Vectors from Each Rating Categories



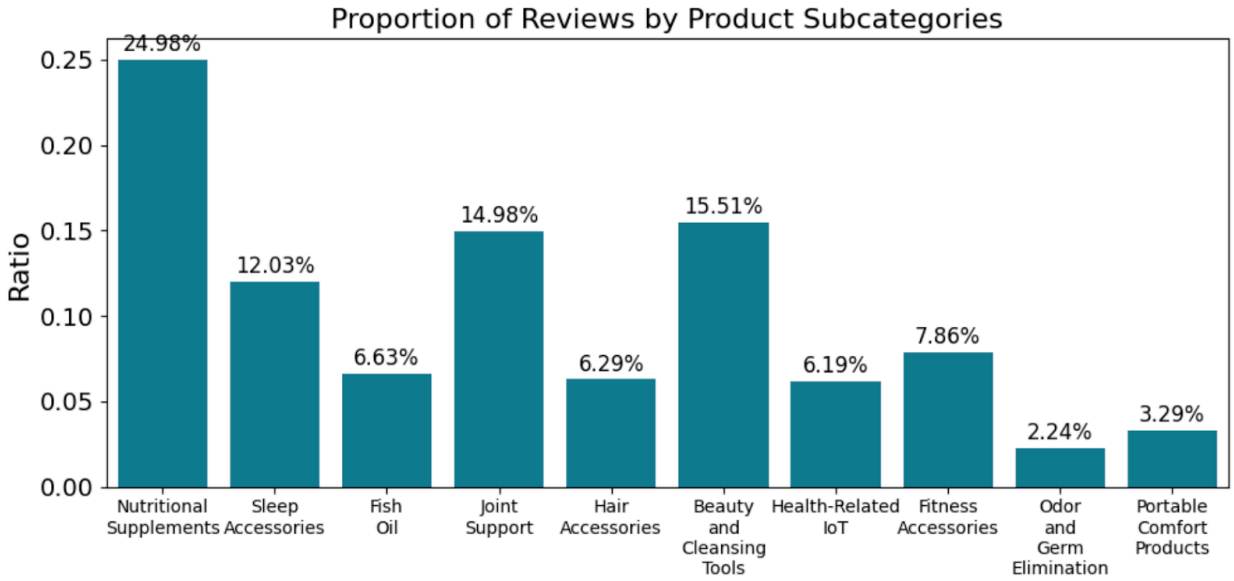
Conclusion: The data points for 5-star (yellow), 4-star (Orange), and 3-star (Red) reviews are clustered together, while the data points for 2-star and 1-star reviews are more spread out, exhibiting different behavior. This result solidifies our finding that 3-star review tend to lean toward 4 and 5 stars reviews.

Supplement Material 6: Sentiment Analysis by Logistic Regression

		Text Vectorization Method		
		M2 TF-IDF + N-gram Random Forest	M2 TF-IDF + N-gram Logistic Regression	M1 Best Result Embedding + Word2vec Logistic Regression
Accuracy		0.81	0.82	0.87
Positive	Precision	0.83	0.82	0.89
	Recall	0.98	0.99	0.95
	F1 Score	0.89	0.90	0.92
Negative	Precision	0.64	0.74	0.73
	Recall	0.17	0.16	0.51
	F1 Score	0.27	0.27	0.60

After incorporating the N-gram method, the results did not meet our expectations: all regression benchmarks appeared to decrease. We suspect this was due to the change in the vectorization method. In M1, we used Word2Vec embedding, but we switched to TF-IDF to enable the use of the N-gram method, which is not supported by Word2Vec or GloVe.

Supplement Material 7:Topic Modeling Results Details



Topic modeling doesn't consider the balance of amounts between subcategories; it only groups by category based on the meaning of the product title. In reality, it is natural for different types of products to be purchased in varying amounts, leading us to view this uneven distribution as normal.

And here is the table showing the breakdown of positive and negative reviews.

	Proportion of Positive and Negative Reviews in Each Topic (%)									
	1	2	3	4	5	6	7	8	9	10
Positive	72.43	86.22	79.89	79.77	82.35	79.47	77.56	76.56	78.63	82.68
Negative	24.57	13.78	20.11	20.23	17.65	20.53	22.44	23.44	21.37	17.32

Supplement Material 8: BoW & POS Trigrams — Insights from the Top 20 Key Customer Reviews Keywords for Fish Oil and Hair Accessories

	Fish Oil	Hair Accessories
1	good_quality_oil	'_t_waste
2	great_<_br	pretty_darn_toxic
3	good_quality_brand	un_rash_horrible
4	cellent_customer_service	s_total_waste
5	Great_quality_product	ineffective_sticky_stuff
6	great_customer_service	expensive_flat_iron
7	good_beard_oil	Terrible_tool_..
8	good_customer_service	Terrible_cheap_product
9	good_muscle_relaxer	Horrible_quality_product
10	good_ear_health	
11	good_overall_health	
12	good_bug_spray	
13	good_clean_cotton	
14	good_clean_flavor	
15	good_cotton_wick	
16	good_oil_emulsion	
17	good_energy_boost	
18	good_natural_remedy	
19	good_energy_drink	
20	good_night_sleep	

From these findings, fish oil products' positive feedback, highlighting "good quality," "great customer service," and notable health benefits. In contrast, the hair accessories category's negative reviews, highlighting customers complaining about "ineffectiveness," "poor quality," and "high prices." These observations underscore the distinct differences in customer focus and satisfaction across product categories.