Text Mining Report: Amazon Fine Foods Reviews

Topic Modeling and Sentiment Analysis

HUANG SHAN
0430-37-9985
Department of Economist
Kyoto University

February 1, 2026

Contents

1 Introduction

1.1 Background

In recent years, online consumer reviews have become an important source of information for both customers and businesses. Compared with structured numerical data, textual reviews contain richer semantic details, personal experiences, and emotional expressions that cannot be fully captured by rating scores alone. The rapid growth of e-commerce platforms has therefore created large volumes of unstructured text data, making text mining and natural language processing essential tools for extracting meaningful insights.

In the food retail domain, reviews are very informative because purchasing decisions are often influenced by taste preferences, perceived quality, and brand familiarity. Analyzing large-scale food review data allows researchers to observe not only what consumers purchase, but also how they describe their experiences and how their attitudes evolve over time. Such analysis provides a broader understanding of consumer behavior beyond simple numerical evaluations.

1.2 Research Aim

The aim of this study is to explore consumer review behavior and preference patterns using large-scale textual data from online food reviews.

Through the combination of topic modeling, sentiment analysis, and temporal trend analysis, this study aims to provide a comprehensive view of how consumers express opinions and preferences in online food review communities.

2 Data Description

The dataset used in this study is the Web Data: Amazon Fine Foods Reviews from the Stanford SNAP project.[1] It contains large-scale consumer review records related to food products on Amazon.

The full dataset consists of 568,454 reviews, contributed by 256,059 users and covering 74,258 products. The reviews span a long temporal period from October 1999 to October 2012.

Each review entry contains both structured and unstructured information. Key fields include product identifiers, user identifiers, helpfulness votes, numerical rating scores, timestamps, short summaries, and full plaintext review content. Specifically, the main attributes used in this study include:

productId (unique identifier of the reviewed product), **userId** (unique identifier of the reviewer), **profileName** (display name of the user), **helpful_yes / helpful_total** (community helpfulness votes), **score** (numerical rating ranging from 1 to 5), **time_unix / time** (timestamp of the review), **summary** (short review headline), and **text** (full review content).

[1] `https://snap.stanford.edu/data/web-FineFoods.html`

3 Data & Preprocessing

3.1 Data Preparation

The original dataset was provided in a raw `.txt` format where each review record contained multiple key–value pairs, including product identifiers, user information, rating scores, timestamps, and plaintext review content. For the convenience of further analysis, the text file was first parsed into a structured tabular format `.csv`.

A custom parsing function was implemented in Python to extract fields such as *productId*, *userId*, *score*, *time*, *summary*, and *text*. Special handling was applied to multi-line review texts to ensure that long comments were not truncated. The cleaned records were then exported into a CSV file for the following processing.

Due to the large scale of the original dataset, this study primarily employs a random sample of 50,000 reviews for text mining, topic modeling, and sentiment analysis. To ensure reproducibility, a fixed random seed was applied during the sampling process.

3.2 Text Cleaning

The following cleaning steps were applied to each review text. First, all words were converted to lowercase to keep the format consistent. HTML-like tags and special symbols (many of which were consideredto be emojis) were removed, mainly to reduce noise and because emoji sentiment was not the focus of this study. Punctuation and non-alphanumeric characters were deleted, and extra spaces were compressed. The text was then tokenized by splitting on whitespace.

Short tokens (fewer than three characters) were removed since they usually act as grammatical fillers rather than meaningful words. Reviews with missing or empty text fields were also excluded from later analysis.

3.3 Stopword Strategy

Stopword filtering was conducted in two stages. First, a standard English stopword list provided by the NLTK library was applied to remove high-frequency functional words such as *the*, *and*, and *is*, which contribute little semantic value in topic modeling.

Second, a domain-specific stopword set was constructed based on frequency inspection and sentiment overlap analysis. Words such as *one*, *first*, *time*, and *would* appeared frequently across both positive and negative reviews but did not convey meaningful topical information. These terms were therefore added to an extended stopword list to improve topic clarity.

Notably, strongly sentiment-bearing words such as *good*, *great*, and *love* were initially retained to preserve emotional signals during early sentiment exploration. Decisions regarding their removal were deferred until later modeling stages to avoid prematurely discarding potentially informative features.

3.4 Vocabulary Filtering

After tokenization, a dictionary was constructed using the Gensim library. Extremely rare words (fewer than 10 documents) and overly common words (more than 50% of documents)

were filtered out.

The resulting vocabulary provided a balanced representation of meaningful terms suitable for TF–IDF weighting and LDA topic extraction.

4 Exploratory Analysis

4.1 Corpus Overview

After preprocessing and random sampling, a total of 50,000 review documents were used for exploratory analysis. The cleaned corpus contained approximately 1,954,069 tokens, providing sufficient textual volume for following semantic modeling.

Before vocabulary filtering, the dictionary contained 40,298 unique tokens. After removing rare terms and overly common words (using frequency thresholds), the effective vocabulary was reduced to 8,947 tokens. This filtering step improves signal-to-noise ratio and stabilizes later TF-IDF and topic modeling results.

Table 1: Corpus Statistics After Preprocessing (Random Sample of 50,000 Reviews)

Statistic	**Value**
Number of documents	50,000
Total tokens	1,954,069
Vocabulary size (before filtering)	40,298
Vocabulary size (after filtering)	8,947

4.2 High-Frequency Words

Table 2 reports the most frequent tokens in the sampled corpus. Overall, high-frequency words reflect a mixture of (i) *food/product categories* (e.g., *coffee, tea, food*), (ii) *sensory attributes* (e.g., *taste, flavor*), and (iii) *evaluation or shopping terms* (e.g., *good, great, price, buy, amazon*).

Table 2: Selected High-Frequency Words in the Sampled Review Corpus (Top 15)

Word	**Freq**	**Word**	**Freq**	**Word**	**Freq**
like	22,363	good	17,836	one	15,745
taste	15,153	coffee	14,718	great	14,699
product	13,373	flavor	12,976	tea	12,455
food	11,497	love	11,263	would	10,787
get	9,457	really	8,867	amazon	8,001

4.3 Sentiment Distribution

Sentiment labels were derived from rating scores: reviews with score ≥ 4 were in this research set as *positive*, score ≤ 2 as *negative*, and score $= 3$ as *neutral*. The dataset exhibits a clear

positive trend: approximately 78.07% of reviews are positive, while 14.43% are negative and 7.50% are neutral.

This imbalance suggests that users are more likely to leave reviews after satisfied experiences. It also implies that observed language patterns and topics in the corpus may be disproportionately shaped by positive consumption narratives.

Table 3: Sentiment Distribution Based on Rating-Derived Labels (Full Dataset)

Sentiment	**Count**	**Proportion**
Positive	443,777	0.7807
Negative	82,037	0.1443
Neutral	42,640	0.0750

4.4 Positive vs. Negative Language Patterns

A comparison of word usage between positive and negative reviews shows both overlap and divergence. High-frequency tokens such as *taste*, *coffee*, and *tea* appear in both sentiment groups, implying that the same product categories are discussed under both satisfaction and dissatisfaction. However, negative reviews tend to include more failure-oriented language (e.g., packaging and delivery complaints such as *box*, or explicit dissatisfaction such as *bad*), while positive reviews more often emphasize enjoyment and recommendation behavior.

5 Topic Modeling

Latent Dirichlet Allocation (LDA) model was applied on the TF–IDF weighted document-term matrix. After several experiments, the number of topics was set to 8.

5.1 Topic Keywords

Each topic is represented by its top weighted keywords. Table 4 summarizes the dominant terms for all identified topics.

Table 4: Top Keywords of LDA Topics

Topic	**Representative Keywords**
0	pill, matcha, toddler, pomegranate, walmart
1	coffee, tea, cup, flavor, drink, strong
2	hair, muffins, cider, shampoo, wrap
3	food, product, sauce, chicken, cat
4	chips, amazon, order, price, box
5	taste, sugar, sweet, chocolate, bars
6	dog, treats, chew, teeth, food
7	dressing, mustard, plant, season

From the keyword distributions, it can be observed that several topics naturally align with intuitive product categories, such as beverages, sweets, pet food, and shopping experience. Other topics capture smaller or mixed niches, indicating that user discussions are not evenly distributed but rather concentrated around several dominant food themes.

5.2 Topic Distribution

The topic proportion analysis shows that most reviews are concentrated in a few major categories, while only a small portion belongs to niche or brand-specific themes. This indicates that user discussions mainly focus on common and frequently purchased food products rather than rare or specialized items.

6 Topic–Sentiment Analysis

A sentiment cross-analysis was conducted to examine how user emotions vary across food categories. This section focuses on interpretability and behavioral insights.

6.1 Sentiment Distribution by Topic

Table 5 presents the sentiment proportions within each discovered topic. The table reveals that positive reviews dominate across all food categories, although the strength of positivity and degree of controversy vary between topics.

Table 5: Sentiment Proportion by Topic

Topic	Negative	Neutral	Positive
0	0.107	0.040	0.853
1	0.110	0.081	0.809
2	0.203	0.136	0.661
3	0.167	0.079	0.753
4	0.180	0.065	0.755
5	0.111	0.077	0.812
6	0.121	0.073	0.806
7	0.091	0.034	0.875

Across almost all topics, positive reviews dominate the sentiment distribution. However, the degree of positivity and controversy differs.

6.2 Topic Interpretations

Topics 0 and 7 mainly represent niche or brand-specific discussions. Although their review volume is small, they show very high positive ratios, suggesting loyal but limited user groups rather than mainstream food categories.

Topic 1 – Coffee / Tea / Drinks focuses on beverages such as coffee and tea. Reviews are mostly positive, indicating stable taste preferences and strong brand loyalty. Negative comments are usually about flavor strength rather than product failure.

Topic 2 – Mixed / Product Quality contains scattered keywords and fewer clear food anchors. Its relatively higher negative ratio suggests dissatisfaction often comes from inconsistent quality or mismatched expectations rather than a specific food type.

Topic 3 – General Food / Cats / Sauce reflects broad food discussions with some overlap into pet-related items. Sentiment is generally positive but less stable, showing mixed usage contexts between human and pet consumption.

Topic 4 – Shopping / Amazon / Price is more about purchasing experience than food itself. Words like *order*, *price*, and *box* indicate logistics or packaging concerns, and negative feedback is often linked to delivery or pricing issues.

Topic 5 – Sweet / Chocolate / Sugar centers on desserts and snacks. The strong positive sentiment aligns with the naturally enjoyable and emotional appeal of sweet foods.

Topic 6 – Pet Food / Treats shows a stable positive trend. Keywords such as *dog*, *treats*, and *chew* suggest repetitive purchasing behavior and consistent satisfaction among pet owners.

7 Topic × Time Analysis

This section examines how user interest across different food-related topics evolves over time.

7.1 Temporal Volume Trends

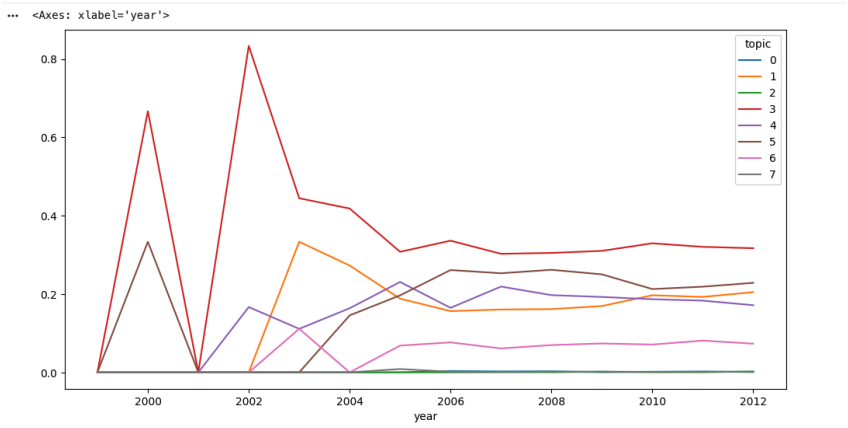
Starting around 2006, the total number of reviews increases substantially year by year. This growth reflects the rapid expansion of the Amazon platform as well as increasing user engagement in online food reviews. Major topics such as beverages, general food products, sweets, shopping experience, and pet food all show rising absolute counts over time.

Importantly, this increase is largely driven by overall platform growth instead of sudden popularity shifts in individual food categories.

7.2 Topic Proportion Stability

To control for platform expansion, topic proportions were calculated for each year by normalizing topic counts within annual review totals. Figure 1 illustrates the evolution of topic proportions from 1999 to 2012.

Figure 1: Topic Proportions over Time (1999–2012)



Despite strong growth in total review volume, the relative distribution of topics remains remarkably stable. Topics related to general food consumption, beverages, sweet snacks, and shopping experience consistently occupy the largest shares across years. This stability suggests that consumer interest in food categories is driven more by routine consumption habits than by short-lived trends.

7.3 Behavioral Interpretation

Topic 3 (General Food) consistently holds the largest proportion, indicating that everyday food discussions dominate long-term user behavior. Topic 1 (Coffee / Tea / Drinks) maintains a steady presence across all years, which aligns with habitual beverage consumption and strong brand loyalty in drink-related products.

Similarly, Topic 5 (Sweet / Chocolate / Sugar) remains stable, reflecting the emotional and repetitive appeal of snack consumption. Topic 4 (Shopping / Amazon / Price) fluctuates slightly but remains persistently relevant, suggesting that logistics and purchasing experience are ongoing concerns rather than temporary issues.

In contrast, smaller topics such as Topic 0, Topic 2, and Topic 7 remain at consistently low proportions throughout the timeline. These topics likely represent niche, brand-specific, or semantically mixed discussions rather than mainstream food consumption behavior.

Overall, the Topic $\times$ Time analysis reveals that while user participation expands rapidly, the underlying structure of food-related discussions remains stable over more than a decade. This pattern highlights the habitual and stable nature of food consumption and evaluation behavior in online reviews.

8 Key Insights

- Beverage and pet food categories show strong long-term stability, indicating habitual consumption patterns and higher brand loyalty.
- Sweet and snack-related products consistently receive the highest positive sentiment, suggesting emotional satisfaction plays a major role in these purchases.

- Negative reviews are more frequently associated with shopping logistics and pricing concerns rather than taste or food quality.
- Although the overall number of reviews increases over time, the relative topic distribution remains stable, implying that consumer food interests are persistent rather than trend-driven.

9 Limitations

Although this study reveals meaningful patterns, several limitations should be noted.

First, the analysis is based on a random sample of 50,000 reviews instead of the full dataset. While the sample is large enough to observe general trends, rare products or minor topic variations may be underrepresented.

Second, the dataset does not include geographic information, which prevents analysis of regional preferences or cultural differences in food consumption.

Third, non-textual signals such as emojis and special symbols were excluded from the analysis. These elements may contain additional emotional cues that word-based models cannot fully capture.

Overall, future work with larger samples, location metadata, and richer sentiment features could provide more comprehensive insights.

10 Future Work

Several directions could further extend this research.

First, incorporating location-based Named Entity Recognition (NER) would enable regional and cultural analysis of food preferences.

Second, integrating emoji and symbol-level sentiment detection could capture emotional nuances beyond plain text.

Third, applying brand-level clustering or entity recognition could reveal finer-grained insights into brand loyalty, market competition, and product differentiation.

References

Stanford Network Analysis Project (SNAP). *Amazon Fine Food Reviews Dataset.* Available at: `https://snap.stanford.edu/data/web-FineFoods.html`