Beyond the Stars: Sentiment, Emotion, and Topic Analysis of E-Commerce Customer Reviews Using Pre-Trained Transformer Models

AUTHOR: **YLG**

September 20, 2025

This document provides an excerpt from the original project, presenting the research question, analytical approach, and key findings.

Abstract

Online customer reviews play a key role in e-commerce, influencing purchasing decisions, brand reputation, and product success. While star ratings offer a quick measure of satisfaction, they lack the depth to explain why customers rate products as they do. Prior research often concentrates narrowly on sentiment polarity, neglecting the role of emotions and thematic features.

This study analyses Amazon grocery and gourmet reviews to investigate how sentiment, emotions, and thematic content relate to satisfaction, measured by star ratings. Pre-trained transformer models extract sentiment scores, emotion and topic probabilities from unstructured text, creating a multidimensional view of customer feedback.

Regression analysis identifies sentiment intensity as the strongest predictor of ratings. At the same time, emotions such as joy, fear, and disgust, along with topics related to quality, freshness, unmet expectations, damaged items, and expired products, are also associated with satisfaction. Clustering with UMAP and HDBSCAN uncovers four feedback types: "Surprised Reviews with Disconfirmed Expectations," where expectation gaps lead to mixed evaluations; "Disappointed by Product Condition," characterised by sadness over damaged products and delivery issues; "Emotionally Intense Complaints," dominated by disgust linked to failures in taste and quality; and "Neutral to Joyful Experiences," centred on perceived value, freshness, and quality of products.

The findings demonstrate that combining sentiment, emotion, and topic features provides a deeper understanding of customer satisfaction than sentiment analysis alone. The study presents a scalable and interpretable framework for review analysis and offers practical implications for managing expectations, improving reliability, and mitigating reputational risks associated with emotionally intense complaints.

Table of Contents

List of Figures

List of Tables

List of Appendices

1. Introduction

1.1 Background

In the digital marketplace, numerical ratings are a widely recognised measure of customer satisfaction, and research shows their direct influence on consumer behaviour. For example, increasing product ratings by half a star can lead to a 5% rise in demand, driven by higher sales (Fülöp, 2022). However, ratings alone lack the context needed to explain why customers form particular opinions. As one of the most influential forms of user-generated content, online customer reviews (OCRs) provide this context by shaping consumer perceptions, brand reputation, and purchase decisions.

On e-commerce platforms such as Amazon, thousands of reviews are posted daily, offering businesses a valuable yet complex source of insights into customer satisfaction, preferences, and behaviour. With an estimated 321 million users and 9.7 million sellers worldwide in 2025 (Michael, 2025), reviews have evolved from being a peripheral marketing consideration to a decisive factor in purchase decisions, brand image, market trends, competitor performance, and product development strategies, all of which are essential for strategic decision-making (Hidayatullah et al., 2024). Notably, 97% of the customer needs identified in traditional market research interviews also appear in Amazon reviews, which additionally reveal eight more needs, representing nearly 10% of the total, that traditional focus group interviews do not mention (Timoshenko and Hauser, 2019).

Recent survey findings by Rachmiani, Oktadinna, and Fauzan (2024) report that 85% of respondents always read reviews before making a purchase, with 60% stating that reviews have a strong influence on their buying decisions. 65% of respondents report decreased interest in a product after reading negative reviews. In a digital environment where physical inspection of products is impossible, reviews serve as a trusted and authentic source of information, often regarded as more reliable than seller-provided descriptions. In the UK, around 90% of shoppers rely on reviews, which influence an estimated £23 billion in annual spending (Wood, 2025).

The abundance of this feedback, however, presents a challenge: how can businesses efficiently extract meaning from such vast and diverse textual data? As Malik and Bilal (2024) show in

Figure 1, the analysis of online reviews has become a key research domain within natural language processing (NLP), with applications in marketing, product innovation, and customer experience management. NLP techniques such as sentiment analysis, emotion detection, and topic classification enable the processing of large volumes of feedback at scale.

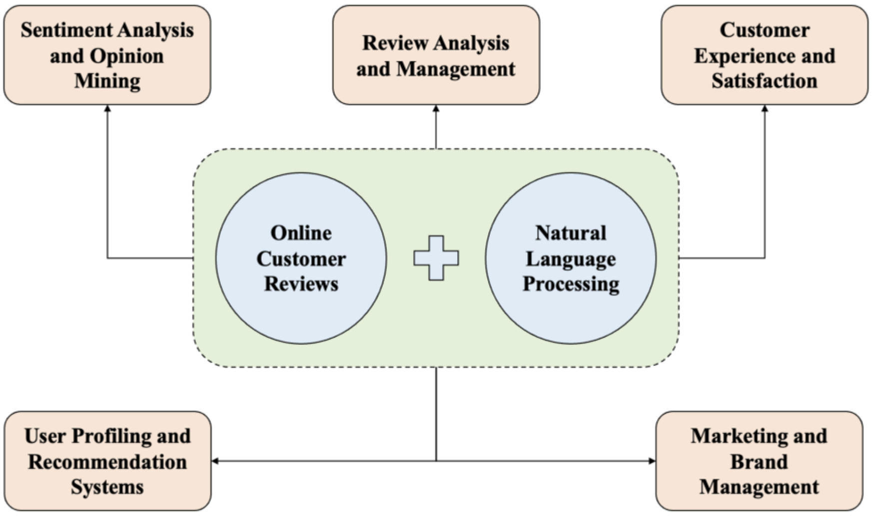


Figure 1: Taxonomy of applications of NLP in online customer reviews (Malik and Bilal, 2024, fig.1)

Recent advances in transfer learning have made it easier to apply NLP in real-world scenarios. Pre-trained transformer models available through open-source platforms, such as Hugging Face (2025), allow the extraction of sentiment, emotion, and topic insights without requiring labelled datasets or training from scratch. This enables businesses and researchers to apply advanced language models directly to raw customer reviews, generating consistent and scalable insights even in data-rich but unlabelled environments. Their ability to process large and complex datasets facilitates accurate sentiment analysis, which can improve service quality, foster customer loyalty, and boost company revenue by up to 15% (Rahman and Maryani, 2024).

Although some recent studies analyse customer reviews using generative large language models (LLMs) such as ChatGPT (McCloskey, LaCasse and Cox, 2024), these approaches face challenges in stability and reproducibility, as outputs vary across sessions and API versions (Jang and Lukasiewicz, 2023). By contrast, this project applies text classification models within a single transformer-based framework, ensuring consistent, stable, and reproducible outputs across the dataset.

Despite these opportunities, most studies address analytical tasks in isolation, often focusing on aspect-based sentiment analysis (ABSA), which concentrates on specific product features while neglecting the holistic evaluative tone of reviews (Meng, Gao, and Bao, 2023). Classification accuracy is also frequently treated as an end goal (Punitha, Raja Shree and Cruz Antony, 2025), rather than as a means of supporting downstream analyses such as predictive modelling or customer feedback segmentation. While sentiment analysis, emotion detection, and topic classification are each well-established individually, existing research typically addresses only one or two dimensions, such as sentiment and topic, rather than integrating all three (Zhang et al., 2022). This fragmented approach limits the depth of insight into customer experience and the underlying factors associated with satisfaction or dissatisfaction.

Scholars also emphasise the importance of accessible analysis tools in business settings and the need for multidimensional analytical frameworks that integrate numerical ratings with textual review data (Davis and Tabrizi, 2021; Kyriakidis and Tsafarakis, 2024). Although NLP has substantially improved the ability to process and interpret large volumes of OCRs, gaps remain in research that (i) captures sentiment spectrum, emotional tone, and thematic framing of entire reviews; (ii) integrates multiple analytical tasks within a single workflow; and (iii) applies outputs directly to business-relevant predictive and exploratory analyses.

This project addresses these gaps by contributing both methodological and managerial insights. It aims to provide a more comprehensive view of customer satisfaction by linking quantitative ratings with the qualitative nuances embedded in written feedback.

1.2 Research Objectives

The study aims to analyse customer reviews from an e-commerce platform by examining their sentiment, emotional tone, and thematic framing to understand how these dimensions shape the overall customer experience. To achieve this, the study employs feature extraction with state-of-the-art, off-the-shelf NLP models to produce structured variables for sentiment, emotions, and topics from unstructured reviews. The study then examines which features are most strongly associated with star ratings, identifying the key factors that explain customer satisfaction and dissatisfaction. Ultimately, it aims to use these features to uncover common patterns in customer feedback through unsupervised learning, thereby providing a more nuanced understanding of the diversity of customer experiences.

Although the focus lies on grocery and gourmet food reviews from Amazon, the methodology can be extended to other product categories and e-commerce platforms. The central research question is:

Which sentiment, emotion, and topic-based features extracted from online customer reviews using pre-trained transformer-based models best explain variations in star ratings, and which of these features contribute to distinguishing different types of customer feedback?

By addressing this question, the study aims to generate actionable insights that enable businesses to enhance their product offerings and refine customer engagement strategies.

2. Methodology

2.1 Research Approach

The study adopts a primarily quantitative approach, consistent with its objective of identifying features associated with customer satisfaction and uncovering patterns in e-commerce reviews using data-driven methods. Structured variables are derived from unstructured text to examine the relationships between sentiment, emotion, topic features, and customer satisfaction ratings. The approach also supports unsupervised learning to identify natural groupings within customer feedback — patterns that would be difficult to detect through qualitative methods alone.

Although quantitative in focus, the research incorporates a targeted qualitative element to validate and contextualise results where predicted sentiment scores diverge significantly from star ratings. In such cases, the original review text is examined to identify possible explanations, such as sarcasm, mixed sentiment, or emphasis on aspects unrelated to the overall rating. This qualitative validation ensures that quantitative findings are interpreted with greater nuance.

2.2 Data

2.2.1 Data Collection

The study uses the grocery and gourmet food category of the large-scale Amazon Reviews dataset compiled by Hou et al. (2024). This category represents products with high consumer awareness, and interpreting reviews does not require specialist domain knowledge, making the findings broadly accessible and relevant.

The researcher merges the reviews with the corresponding product metadata to produce a unified data frame. As the study focuses on ratings and review text, it retains only a subset of variables to provide contextual information.

2.2.2 Data Pre-Processing

Pre-processing ensures that the dataset is clean, consistent, and suitable for analysis. The study follows a multi-stage process of transformation, cleaning, and sampling. Transformation standardises formats and ensures data consistency, while cleaning removes irrelevant or low-quality entries. Stratified sampling ensures balanced representation across satisfaction levels (Stephan, 1941).

2.2.3 Feature Extraction

Feature extraction transforms unstructured review text into structured variables suitable for regression and clustering. The study employs the *transformers* package in Python to generate three categories of features.

Sentiment Scores

nlptown/bert-base-multilingual-uncased-sentiment (NLP Town, 2023), a model fine-tuned for sentiment analysis on product reviews, outputs a five-point sentiment classification, recorded as a categorical score. This granularity provides more nuanced insights than a three-class polarity classification, supporting personalised marketing and customer service strategies (Cherukuri, 2024).

Emotion Probability Scores

j-hartmann/emotion-english-distilroberta-base (Hartmann, 2022) model generates continuous probability scores for Ekman's six basic emotions (Ekman, 1992), labelled by the model as joy (which Ekman refers to as enjoyment), anger, fear, disgust, sadness, and surprise, plus a neutral category. Each review's emotional profile is represented by probabilities that sum to 1 across the seven categories. This probabilistic output captures the full range of variation in emotional signals, offering more nuanced insights and improving predictive and clustering performance (Dawson and Weiss, 2012).

Topic Probability Scores

The researcher defines candidate topic labels using insights from the literature review and text mining techniques, including TF-IDF, frequent bigrams, trigrams, and word clouds (Li, Liu and Chen, 2024). This approach ensures that topics are both conceptually relevant and empirically grounded in the dataset. Zero-shot classification with the NLI model *facebook/bart-large-mnli* (Lewis et al., 2019) then assigns probability scores to each topic for every review through a multi-label classification approach.

For models with token limits, truncation allows longer reviews to fit within the maximum sequence length.

2.3 Data Analysis

2.3.1 Exploratory Data Analysis (EDA)

EDA examines the structure and variability of the dataset to inform model specification. Descriptive statistics for star ratings and reviews include distribution plots, boxplots, a word cloud, and frequency analysis of common words, bigrams, trigrams, and TF-IDF scores by rating level. Together, these methods provide insights into the vocabulary characteristics of the reviews.

2.3.2 Regression Analysis

This study employs Ordinary Least Squares (OLS) regression to examine the relationship between extracted features and customer satisfaction. OLS regression offers simplicity and interpretability, making it a common method in customer review analysis (Routaray and Chitra, 2024). The dataset is split into training (80%) and testing (20%) subsets to assess the model's generalisability.

To enable meaningful comparison of effect sizes across predictors measured on different scales, the analysis standardises all independent variables as z-scores. The reported coefficients then represent the expected change in customer rating associated with a one-standard-deviation increase in the predictor, holding all other variables constant. This standardisation supports clearer interpretation of the relative importance of predictors (Hunter and Hamilton, 2002). Marginal effect plots illustrate the influence of statistically significant predictors on ratings.

For benchmarking, the study fits a Random Forest regressor with 100 estimators and compares its performance to the OLS model using mean squared error (*MSE*), root mean squared error (*RMSE*), and R^2. This comparison evaluates whether the linear model provides predictive accuracy comparable to a more flexible machine learning approach.

The researcher considers the inclusion of interaction terms, as they can capture complex relationships between features (Mikucka, Sarracino and Dubrow, 2015). However, guided by the principle of parsimony, the writer retains the baseline model to preserve explanatory clarity and statistical validity (Dunstan, Crowne and Drew, 2022).

2.3.3 Clustering Analysis

The study employs unsupervised learning to reveal latent structures within the review data. It standardises numerical variables, excluding *Rating*, and reduces dimensionality using UMAP with two components, which maintains both local and global structures in high-dimensional data while supporting effective visualisation (Herrmann et al., 2023). The reduced feature set is then clustered with HDBSCAN, chosen for its ability to detect clusters of varying density and shape, handle noise points, and avoid the need to predefine the number of clusters. Previous research analysing customer reviews indicates that the UMAP and HDBSCAN pipeline produces coherent and interpretable clusters (Lakatos et al., 2024).

The researcher fine-tunes HDBSCAN parameters using a grid search. The *min_cluster_size* parameter, which determines the minimum cluster size, ranges from 5 to 40 in increments of 5. The *min_samples* parameter, which defines how strictly points are classified as core or noise, ranges from 5 to 35 in increments of 5. From an exploratory perspective, the choice of cluster granularity depends on the analytical goals rather than solely on statistical stability metrics.

Cluster quality is assessed using multiple internal validity indices to capture different dimensions of performance. The silhouette score (Rousseeuw, 1987) evaluates cohesion and separation between clusters, the Davies–Bouldin Index (Davies and Bouldin, 1979) measures cluster compactness relative to separation, and the Calinski–Harabasz Index (Caliński and Harabasz, 1974) considers the ratio of between- to within-cluster dispersion. To test whether feature distributions differ significantly across clusters, the analysis applies the Kruskal–Wallis test (Kruskal and Wallis, 1952) with false discovery rate (FDR) correction to ensure more reliable conclusions (Posch, Zehetmayer and Bauer, 2009). Stability is examined using the Adjusted Rand Index (Rand, 1971) and the Jaccard Index (Jaccard, 1912), which assess the consistency of clustering solutions under resampling. Visualisations, including a UMAP projection, a heatmap and a radar chart of mean feature values, support interpretation, while Pearson correlation matrices for each cluster further contextualise relationships between features.

This framework evaluates clustering solutions not only on statistical performance but also on interpretability, ensuring that the final selections reported in the Results chapter are both meaningful and robust.

2.4 Tools and Environment

The study employs a combination of open-source programming tools. Python, run through Jupyter, serves as the primary language for data pre-processing, feature extraction, regression, and clustering. R, executed in RStudio, supports additional pre-processing, data visualisation, and statistical testing. Pre-trained transformer models are implemented through the *transformers* package in Python. Other packages include *scikit-learn* and *statsmodels* for regression and evaluation, *umap* for dimensionality reduction, and *hdbscan* for clustering.

Visualisations are created using *ggplot2*, *matplotlib*, *seaborn*, and Tableau, which collectively support exploratory analysis and clear presentation of results. Using these tools guarantees transparency, reproducibility, and efficiency in the research process.

2.5 Ethical Considerations

The methodology employed in this study adheres to established ethical guidelines for data science and social data research. All data is publicly available and anonymised, ensuring no risk of breaching confidentiality or privacy (Ting, 2022). Nonetheless, working with user-generated content presents specific ethical challenges. Reviews may occasionally contain sensitive opinions, and the automated analysis of sentiment and emotion raises broader issues of transparency and the potential misuse of insights if applied to influence consumer behaviour in commercial contexts (Oğuz, 2024).

Representation and bias remain important concerns. Customer reviews are inherently subjective and may reflect cultural, linguistic, and demographic biases (Skotis and Livas, 2024). Transformer models trained on large-scale corpora may reproduce or amplify such biases, leading to skewed or unfair representations of customer experiences (Yang et al., 2023). Variability in linguistic structures and cultural expressions can also reduce accuracy and reliability, emphasising the importance of cautious interpretation and recognising that both training data and linguistic context shape outcomes.

Algorithmic fairness and interpretability add further considerations. While the models used in this study produce stable and reproducible outputs, their transparency varies. Linear regression provides interpretable coefficients that directly measure relationships between variables, whereas transformer models, UMAP, and HDBSCAN function more as black-box systems. Their internal decision-making processes remain opaque, making it difficult for non-technical stakeholders to understand how classifications are made or to assess model behaviour in edge cases (Adak, Pradhan and Shukla, 2022).

3. Results

This chapter addresses the central research question: Which sentiment, emotion, and topic-based features from online reviews best explain the variation in customer satisfaction and distinguish feedback types? While star ratings capture satisfaction levels, they lack context about why customers evaluate products as they do. Text-derived features provide this depth and underpin the analyses presented in this chapter.

The results unfold in four stages: data explanation outlines the data pre-processing steps and provides an overview of the data used for exploratory analysis and feature extraction. Exploratory analysis describes review characteristics and feature distributions. Regression analysis identifies key predictors of star ratings. Clustering reveals distinct feedback types and shows how sentiment, emotional and thematic signals shape satisfaction profiles.

Together, these findings provide the empirical foundation for interpreting customer satisfaction and dissatisfaction, setting the stage for the discussion in the following chapter.

3.1 Data Explanation

Upon merging the grocery and gourmet food reviews with product metadata, a unified data frame contains more than 50,000 rows and 23 columns, including fields such as product images and descriptions.

The data cleaning process involves removing duplicate rows, handling missing values, and retaining only English-language reviews from verified purchases with a minimum length of ten words, thereby ensuring meaningful feedback. It also removes textual noise, such as excessive whitespace, HTML tags, and emojis (Raza et al., 2024).

Although transformer-based models used in this study for the feature extraction process raw text directly using their built-in tokenisers, the researcher selectively applies traditional NLP pre-processing steps, such as stop-word removal and manual tokenisation, during exploratory analyses (e.g., n-grams and TF-IDF) to highlight distinctive vocabulary patterns across rating levels.

The researcher standardises data types, converts star ratings into a numeric format, and simplifies field names for readability. Merging the review title and body into a single text field provides richer and more contextually complete inputs for NLP analysis. The researcher removes duplicate text from the title where it appears verbatim at the beginning of the body. Sentence count, derived as a variable from the merged reviews, serves as an indicator of review length and potential content richness, which can be associated with sentiment expression and feature extraction in subsequent analyses.

To achieve balanced representation across satisfaction levels, the researcher draws a stratified random sample of 500 reviews for each star rating (on a five-point scale, where one represents strongly negative and five strongly positive). This approach satisfies the minimum sample size requirements for multiple correlation analysis (Green, 1991).

The sampled dataset, comprising 2,500 reviews, retains five variables relevant to the research objectives: full review with title and body merged, number of sentences in the review, product title, retailer name, and star rating. Table 1 presents characteristics of the dataset used for exploratory analysis.

Table 1: Overview of the stratified sample

Field	**Count**	**Count Distinct**	**Type**	**Explanation**
Rating	2500	5	Numeric	Product rating on a five-star scale (1 = lowest, 5 = highest)
Review	2500	2500	Text	Review text comprising the title merged with the body
Sentence Count	2500	37	Numeric	Number of sentences in the merged review
Product Title	2500	2200	Text	Name of the reviewed product
Retailer	2500	1399	Text	Retailer or store associated with the product

3.. Exploratory Analysis

The exploratory analysis begins with descriptive statistics and distributions to establish a baseline understanding of the dataset. It then examines sentiment scores, emotion probabilities, and topic probabilities, all extracted by transformer models, to assess their distributions, variation across rating levels, and correlations. This step provides insight into the richness of the feature set and its suitability for downstream modelling.

3.2.1 Data Overview

This balanced sampling produces a mean rating of 3.0 with a standard deviation of 1.41. The structure minimises bias towards positive or negative reviews and allows for more reliable comparisons of linguistic features across the satisfaction spectrum. Figure 2 and Figure 3 confirm that no single product or retailer dominates the dataset. The most frequent products and retailers each account for less than 3% of total reviews, ensuring diversity and reducing the risk of bias toward individual brands.

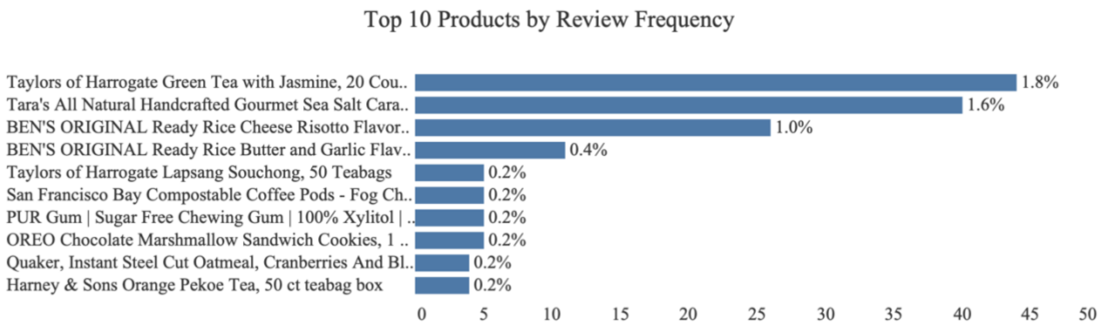


Figure 2: 10 most reviewed products

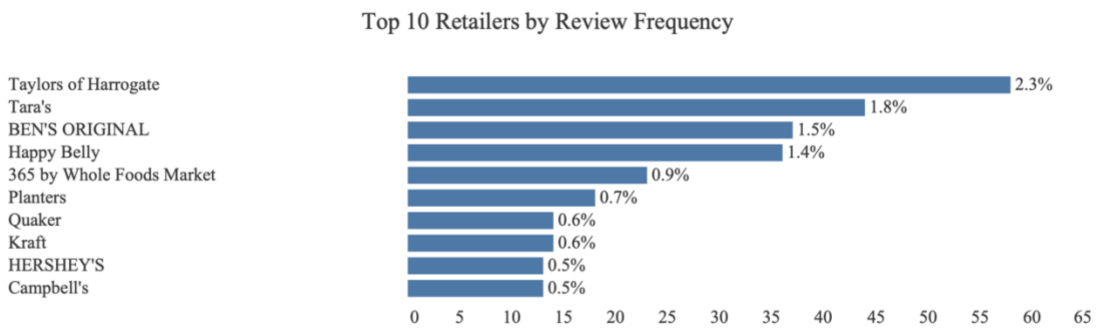


Figure 3: 10 retailers with the largest review counts

Review Length Characteristics

The distribution of review length, measured in sentence count, shows a highly skewed pattern (Figure 4). Most reviews are short, with the mode concentrated at two to five sentences. Counts decline sharply after ten sentences, and reviews longer than 20 sentences are rare. This long-tail distribution suggests that consumers typically provide concise feedback, while only a minority invest effort in writing substantially longer reviews.

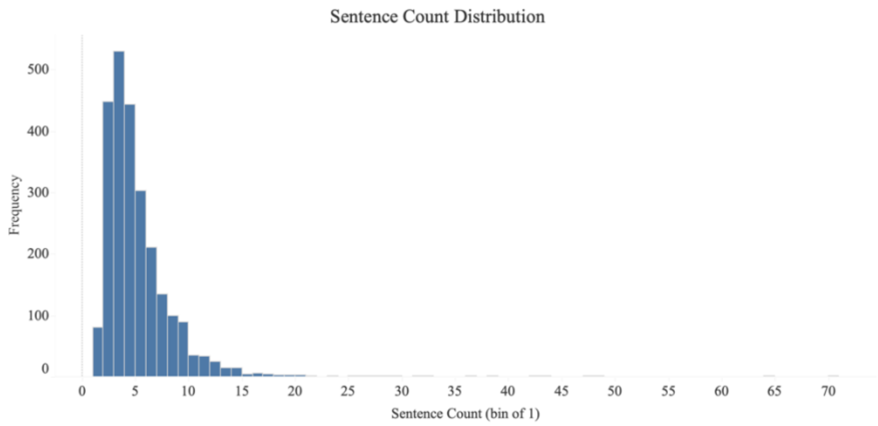


Figure 4: Distribution of sentence counts

Table 2 presents summary statistics for sentence counts. These values confirm the skewed distribution, with a low median and mode but a much higher maximum driven by extreme outliers.

Table 2: Summary statistics of review sentence counts

mean	sd	mode	min	25%	50%	75%	max	range	skew	kurtosis
4.86	4.08	3	1	3	4	6	70	69	6.02	65.17

Disaggregating review length by rating (Figure 5) shows that this pattern is consistent across categories. Median length remains low at three to five sentences, although each category contains outliers. Some reviews, particularly in the two-star and four-star groups, extend beyond 50 sentences. This indicates that moderately negative or moderately positive experiences sometimes prompt detailed narratives.

To complement the sentence-level analysis, review length is also measured in words (Figure 6). The distribution again shows a skew, with most reviews under 100 words and a long tail of very lengthy entries. Median lengths remain consistently low across ratings, but upper ranges are higher for three- and four-star reviews. Outliers exceeding 700 words appear in the mid-range categories, suggesting that more nuanced experiences tend to generate longer narratives. In contrast, extreme satisfaction or dissatisfaction is often expressed briefly.

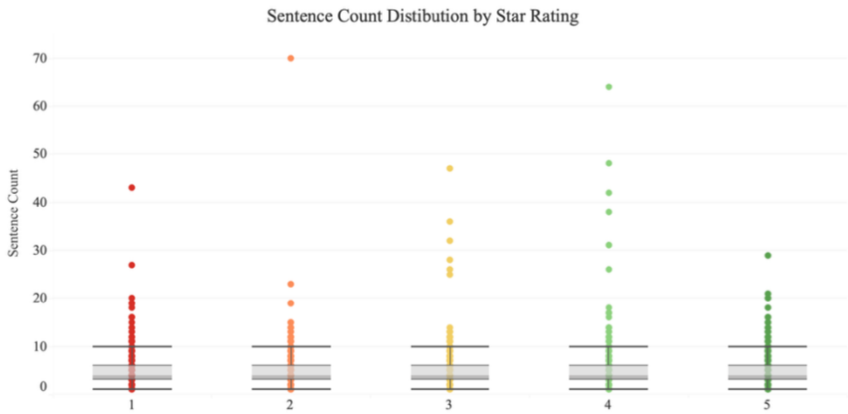


Figure 5: Distribution of sentence counts by rating

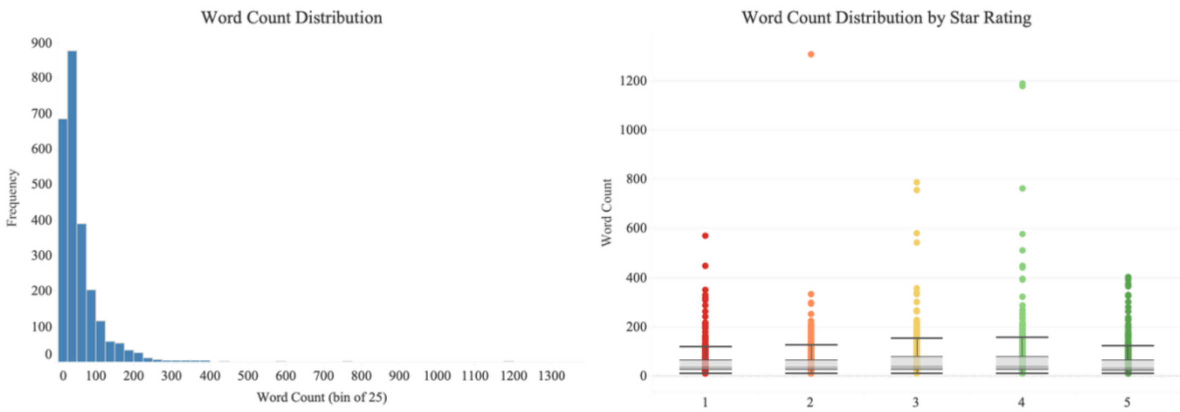


Figure 6: Distributions of word counts

The analysis of review length highlights two behavioural patterns. First, brevity is the norm across all ratings, with both highly positive and highly negative experiences often expressed in short statements. Second, although infrequent, longer reviews provide disproportionately rich qualitative insights. They appear most often in mid-range ratings, where feedback is more balanced and nuanced, and they may be especially valuable for identifying themes and opportunities for product improvement.

Vocabulary Exploration

The study explores vocabulary and linguistic patterns across satisfaction levels to gain insight into the thematic structure of customer feedback. TF-IDF plots identify distinctive words associated with each rating, while the word cloud provides a visual overview of frequently used terms. Most frequent words, bigrams, and trigrams are also examined by rating, revealing common multi-word expressions and illustrating how language use varies across different satisfaction levels. Together, these exploratory techniques offer an initial view of thematic and stylistic variation before feature extraction.

The overall word cloud in Figure 7 highlights that the most frequent terms relate directly to the product experience. Words like *taste, flavor, good, bad, love, expensive*, and *price* dominate the lexicon, suggesting that evaluations are anchored in sensory quality, hedonic evaluation, and perceived value.

Figure 7: Word cloud of most frequent terms in reviews

Figure 8 displays the most common words in reviews by rating, highlighting clear shifts in tone across satisfaction levels. Words like *taste*, *flavor*, and *product* appear across all ratings, but negative reviews (1–2 stars) are characterised by terms such as *disappointed*, *bad*, *bag*, and *box*, reflecting dissatisfaction and complaints. Mid-range reviews (3 stars) show a lexical mix, with both critical and positive terms like *bad* and *love*, indicating more neutral or ambivalent feedback. In contrast, four- and five-star reviews feature more positive language, including *love*, *delicious*, *nice*, and *price*, reinforcing the central role of quality, perceived value, and expectations in shaping customer sentiment.

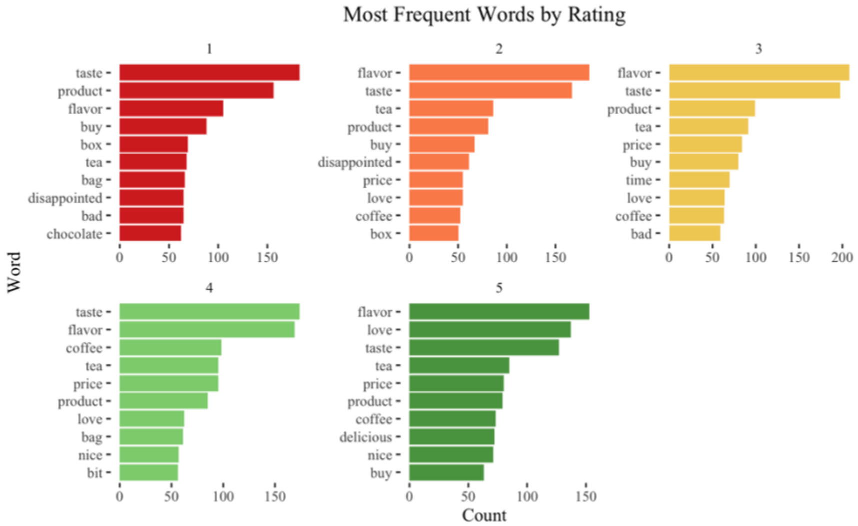


Figure 8: Most frequent words by rating

The TF-IDF analysis by rating (Figure 9) further uncovers distinctive vocabulary. Negative reviews (1–2 stars) frequently use strongly emotive descriptors such as *rotten, disgusting,* or *awful,* alongside references to issues like *expiration date* or *stale.* Mid-range reviews (3 stars) contain more functional words, e.g., *repurchase*, reflecting a less polarised tone. Positive reviews (4–5 stars) often include specific product descriptors such as *cardamom, lactic, melon,* and terms with positive connotations such as *excellent.*

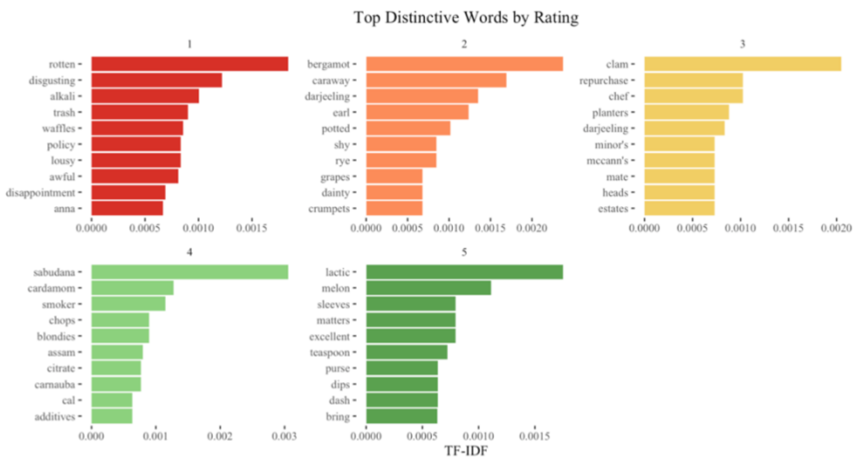


Figure 9: Most distinctive words by rating (TF-IDF)

Analysis of bigrams (Figure 10) and trigrams (Figure 11) confirms these observations. Lower-rated reviews emphasise product faults through expressions such as *bad batch* and *didn't taste* *fresh.* Higher-rated reviews include recommendation-oriented phrases like *perfect size, absolutely delicious,* and *coffee highly recommend.* This contrast suggests that dissatisfied customers highlight quality failures, whereas satisfied customers emphasise endorsement and product distinctiveness.

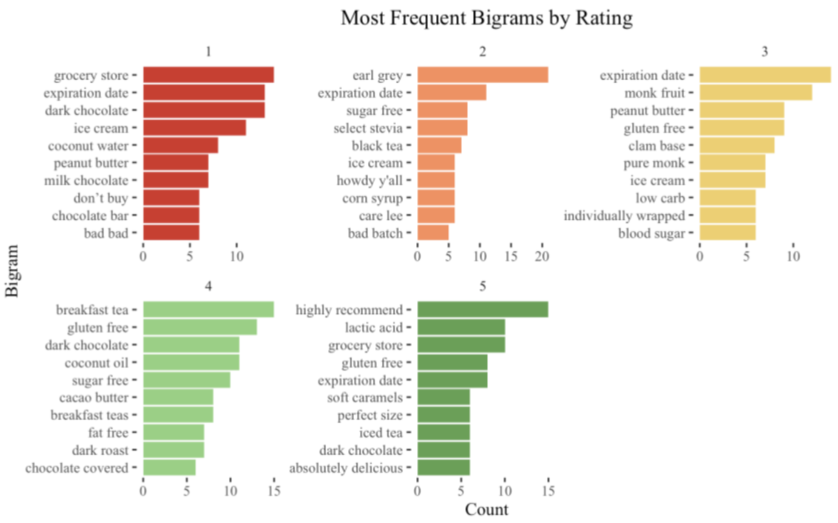


Figure 10: Most frequent bigrams by rating

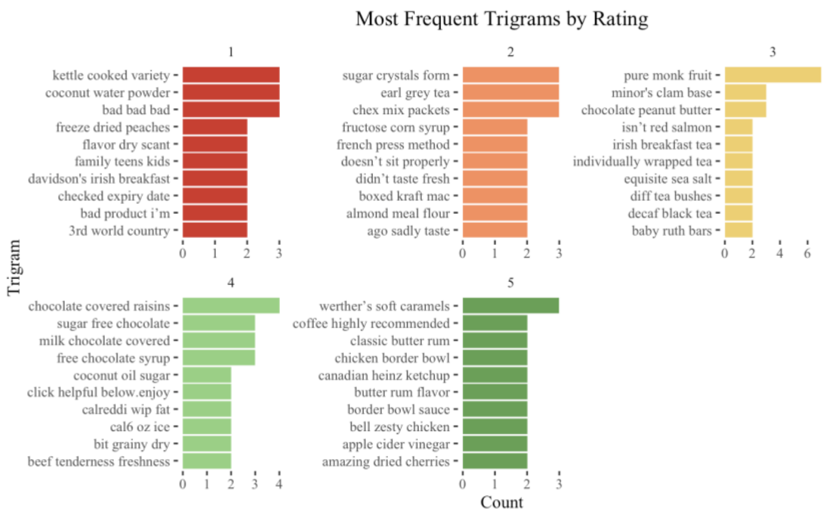


Figure 11: Most frequent trigrams by rating

Vocabulary analysis also detects terms that appear in both positive and negative contexts, most notably *expiration date.* In negative reviews, the phrase refers to product defects such as *bad batch* or dissatisfaction with shelf life. In contrast, positive reviews use *expiration date* as reassurance of product quality and freshness. This dual use shows that certain attributes can act as both risk factors and differentiators: failure generates dissatisfaction, while meeting or exceeding expectations enhances positive evaluations. This finding highlights the importance of context-sensitive analysis, as identical terms may carry opposite meanings depending on the framing of the review.

Vocabulary exploration shows systematic variation in language across rating levels. Negative reviews use emotionally charged and problem-focused terms, mid-range reviews adopt a more neutral descriptive tone, and positive reviews emphasise product qualities and endorsements. These findings support further analysis and provide linguistic cues that inform regression and clustering, where such terms act as predictive features of customer satisfaction.

This initial analysis confirms that the dataset provides a balanced foundation for further modelling. It combines coverage across rating levels with the diversity and noise typical of real-world reviews, making it suitable for examining how sentiment, emotion, and topic features vary across satisfaction levels, as well as for applying predictive and clustering techniques.

3.2.2 Sentiment Analysis

The sentiment analysis model produces scores on a five-point scale, which are compared with star ratings used as proxy labels for ground-truth sentiment to assess their alignment. Performance evaluation relies on standard classification metrics. Accuracy represents the proportion of correct predictions among all predictions, providing an overall indication of the model's performance. Precision measures the proportion of correctly predicted items within a class relative to all items assigned to that class. Recall indicates the proportion of correctly predicted items within a class relative to all items that truly belong to that class. The *F1*-score combines precision and recall through their harmonic mean, providing a balanced indicator of model performance. Support indicates the number of true instances in each class, while macro averages report the unweighted mean of precision, recall, and *F1*-score across all classes.

The classification report in Table 3 shows an overall accuracy of 0.5680, indicating moderate alignment between predicted sentiment and user-assigned ratings. While this accuracy reflects the complexity of mapping free-text sentiment to discrete star ratings, it provides sufficient reliability to examine patterns and inconsistencies.

Table 3: Classification report of sentiment model performance with ratings as proxy labels

Star Rating	Precision	Recall	F1-score	Support
1	0.6257	0.6620	0.6433	500
2	0.4351	0.5300	0.4779	500
3	0.4957	0.4580	0.4761	500
4	0.5747	0.4540	0.5073	500
5	0.7287	0.7360	0.7323	500
Accuracy			0.5680	2500
Macro avg	0.5720	0.5680	0.5674	2500

Performance varies across classes. The model performs best on extreme ratings. Five-star reviews achieve the highest *F1*-score of 0.73, supported by both high precision (0.73) and recall (0.74). One-star reviews also perform reasonably well (*F1* = 0.64). Mid-range ratings are more difficult to classify. Two- and three-star reviews both score an *F1* of 0.48, suggesting the ambiguity and mixed sentiment typical of these reviews. The four-star category shows reduced recall (0.45), suggesting frequent misclassification into neighbouring classes.

The confusion matrix in Figure 12 highlights these tendencies. The one-star and five-star categories show strong diagonal concentration, while mid-range reviews display higher dispersion. Many four-star reviews are misclassified as five-star, and a large share of two-star reviews are misclassified as either one- or three-star. This pattern shows the “off-by-one” nature of most errors. When applying a tolerance of ±1 star, accuracy increases sharply to 0.928, suggesting that the model typically predicts the sentiment score within one category of the rating.

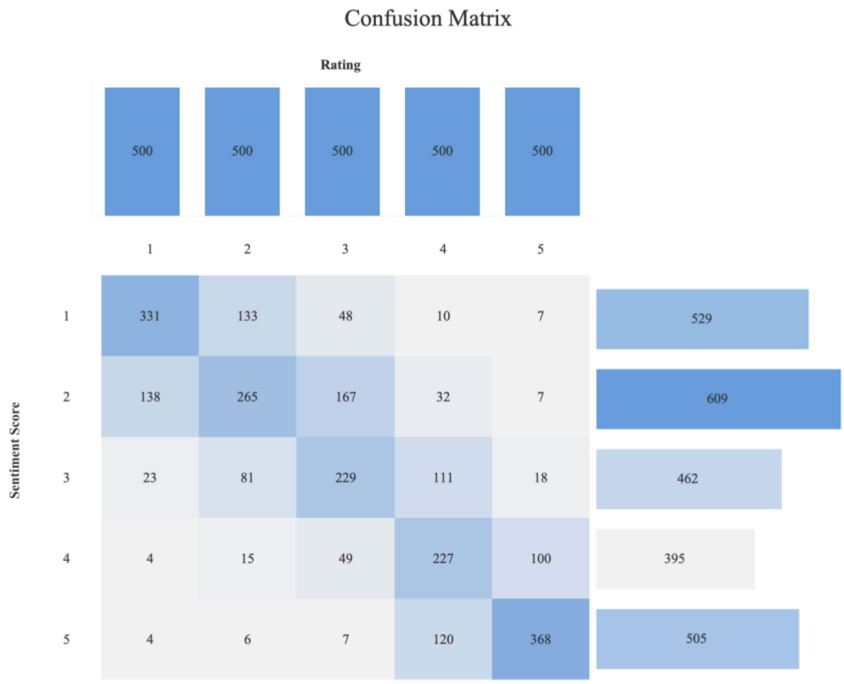


Figure 12: Confusion matrix of predicted sentiment scores versus actual ratings

Robustness checks confirm these results. Both the Wilcoxon signed-rank test ($p < 0.001$) and the chi-square test ($p < 0.001$) reject the null hypothesis of no association between predictions and ratings. The mean absolute error (*MAE*) of 0.52, compared to the *Rating* standard deviation of 1.41, indicates that prediction errors are small in magnitude. This is unsurprising given that user-generated ratings may not always reflect consistent or objective sentiment.

Figure 13 visualises individual review-level mismatches between *Sentiment Score* and *Rating*. While most reviews align closely, some show notable divergence, such as positive sentiment paired with low ratings or negative sentiment paired with high ratings. These cases may arise when textual content emphasises specific issues that do not drive the overall rating. They may also reflect sarcasm, rating biases, or the model's failure to detect sentiment correctly.

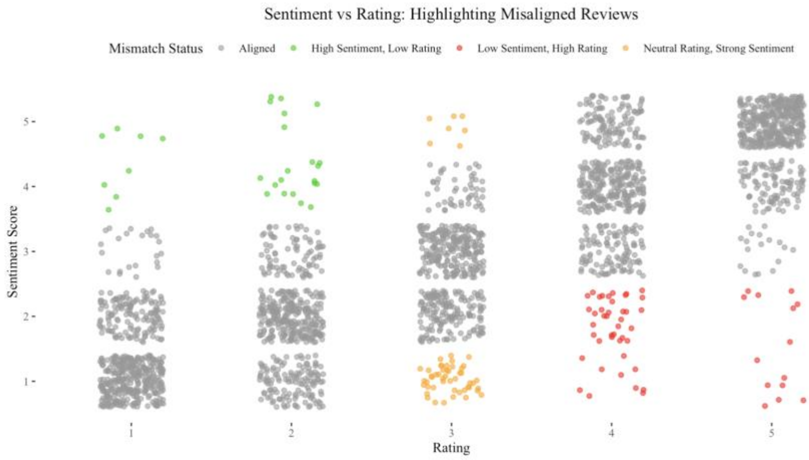


Figure 13: Scatterplot of reviews with misaligned sentiment scores and star ratings

Table 4 illustrates three examples of mismatches between *Rating* and *Sentiment Score*. In the first, a review describing shipping damage still carries a four-star rating. In the second, a sarcastic comment receives a low sentiment score despite a five-star rating. In the third, a positive review about ease of use is paired with a two-star rating, likely due to concerns about price. These examples highlight that ratings do not always align with expressed sentiment and that automated models struggle with sarcasm, contextual nuance, and value judgements.

Table 4: Illustrative examples of rating-sentiment mismatches

No.	**Star Rating**	**Sentiment Score**	**Review**
1	4	1	"Exploded in shipping. Waited awhile for these… BUT I do not see a place to contact seller as two of the ginger ones (the ones I really wanted) exploded in my box."
2	5	1	"Made a beautiful burger out of these. And my dumb brain didn't even realize it was vegetarian."
3	2	5	"Easy to use. Last year I brined a turkey with a home made bri ne. Had to boil water and the sugar and salt. Then wait for it to cool. It's time consuming. This product is more expensive tha n making your own, but the trade off is no prep. Mix with wat er and you're done. No boiling or waiting to cook."

Overall, the sentiment model demonstrates strong performance in identifying clear positive and negative extremes but weaker performance in mid-range categories. Despite these limitations, the off-by-one proximity indicates that the model provides a reliable basis for subsequent regression and clustering analyses.

3.2.3 Emotion Analysis

Emotion analysis complements sentiment classification by providing contextual insight into the linguistic and affective patterns in customer feedback. The model assigns probabilities to each category, with the dominant emotion defined as the one with the highest probability.

The distribution in Figure 14 shows that most reviews are classified as *Neutral*, followed by *Disgust* and *Joy*. *Anger* and *Fear* appear less frequently, suggesting that dissatisfaction is often expressed through milder emotions rather than strong negative affect. Figure 15 illustrates a clear alignment between dominant emotional tone and satisfaction: *Joy* is associated with higher ratings, while *Disgust*, *Sadness*, *Anger*, and *Fear* are more prevalent in lower ratings. *Neutral* and *Surprise* occur across all satisfaction levels.

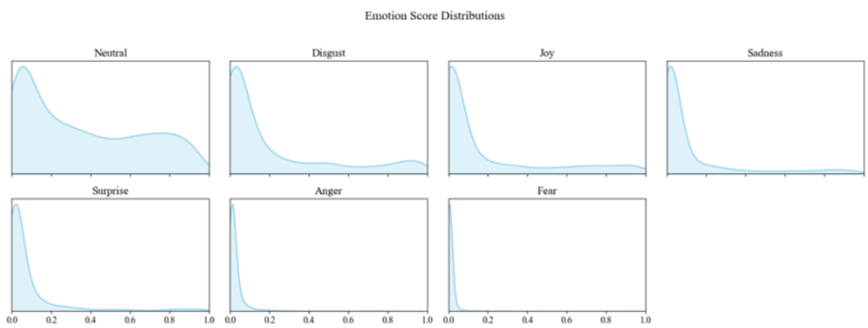


Figure 14: Distribution of emotion probabilities across reviews

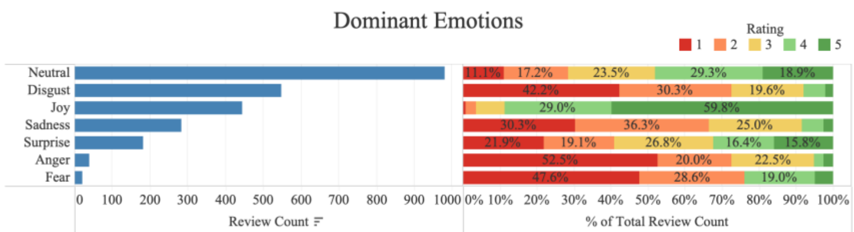


Figure 15: Dominant emotions and their distribution by rating

A chi-squared test confirms a statistically significant association between dominant emotion and sentiment score ($\chi^2 = 1381.44$, $p < 0.001$). The effect size, measured by Cramér's V (0.369), indicates a relatively strong relationship between emotion categories and sentiment scores, supporting the model's validity in capturing meaningful emotional signals.

Qualitative inspection further illustrates the role of emotions. Review No.1 in Table 5 shows that neutrality in emotion does not necessarily imply neutrality in judgement. Although the text lacks strong affective language, it reflects dissatisfaction with product quantity despite a high rating. Review No.2 demonstrates that emotion detection can sometimes outperform sentiment scoring by correctly capturing dissatisfaction in cases where sentiment predictions are misleading.

Table 5: Illustrative examples of rating-sentiment-emotion mismatches

No.	**Star Rating**	**Sentiment Score**	**Emotion**	**Review**
1	5	1	Neutral	"Quantity not what I ordered. The order I placed was for a pack of 6 jars. The box contained only one 13 oz jar."
2	1	5	Disgust	"Tastes like bubble gum with a hint of tropical flavor… These taste like classic bubble gum (yuck)… whereas the Tropical Beat actually tastes like what I would expect."

These examples highlight that neither model is flawless, yet emotion detection may capture nuances overlooked by sentiment scores. It is particularly effective in cases involving mixed evaluations, contextual phrasing, or subtle dissatisfaction. Together, sentiment and emotion analysis provide complementary perspectives that enrich the interpretation of customer feedback.

3.2.4 Topic Analysis

Topic analysis provides insights into the aspects linked to customer satisfaction and dissatisfaction. The researcher defines topic labels based on the literature review and vocabulary analysis. These include: *Not as Expected*, *Taste and Flavour*, *Price*, *Value for Money*, *Damaged*, *Quality*, *Freshness*, *Returns*, *Ingredients*, *Delivery*, *Size and Quantity*, *Packaging*, *Easy to Use*, and *Expiration Date*.

Assuming reviews may cover multiple aspects, the model assigns probability scores for each topic within a review. Figure 16 shows these distributions on a scale from 0 to 1, with higher values denoting stronger alignment with the corresponding topic. The observed distributions illustrate distinct patterns across topics, underscoring variation in the degree to which specific aspects are emphasised within the corpus. For example, *Not as Expected* and *Taste and Flavour* display sharp peaks near 1.0, indicating that when present, they dominate the review narrative. *Packaging*, *Delivery*, and *Size and Quantity* exhibit broader, flatter distributions, suggesting that customers mention them as secondary attributes. *Damaged, Quality* and *Value for Money* display bimodal distributions, reflecting that reviews either strongly emphasise or ignore these attributes. *Easy to Use* and *Expiration Date* are rarely mentioned, with distributions skewed toward zero, although *Expiration Date* shows a thicker right tail. These patterns suggest that some topics are more strongly associated with satisfaction or dissatisfaction, while others function as supporting contextual factors.

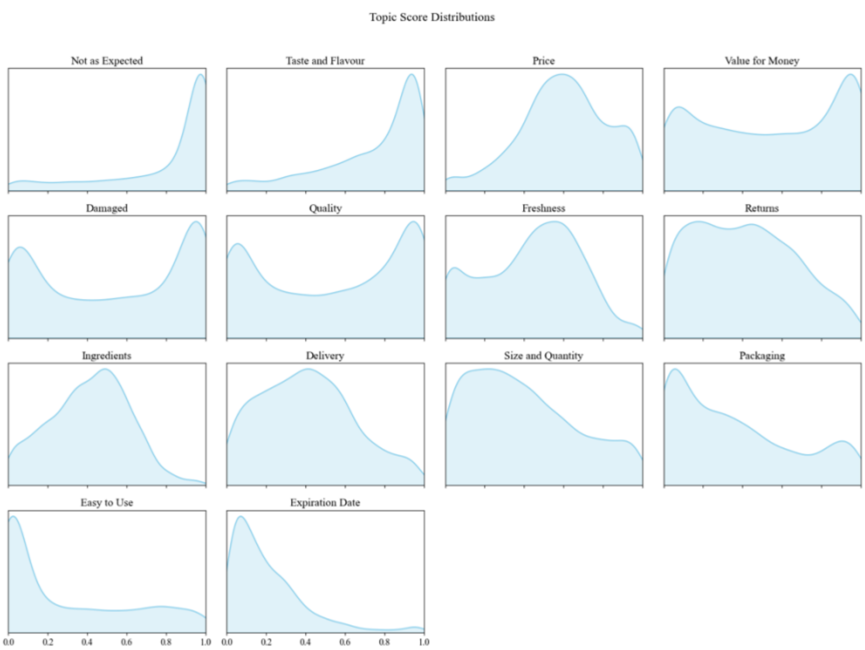


Figure 16: Distribution of topic probabilities across reviews

Analysis of dominant topics, defined as the label with the highest probability (Figure 17), highlights *Not as Expected* as the single most frequent feature, appearing in more than one-third of all reviews. Although this topic spans all ratings, it is concentrated in one- to three-star reviews, underscoring unmet expectations as a leading cause of dissatisfaction. Other negative attributes include *Damaged* and *Returns*, both of which are heavily concentrated in low ratings, with more than half of the *Damaged* reviews assigned a one-star rating. By contrast, *Quality*, *Taste and Flavour*, and *Value for Money* are more evenly distributed, with higher representation in four- and five-star reviews, suggesting that meeting or exceeding these expectations contributes to satisfaction. Less frequent dominant topics also provide meaningful insights. *Freshness* is most often associated with higher ratings, while *Expiration Date* is more frequently linked to critical reviews. Only nine reviews fall into the *Other* residual group, where all topic probabilities are below 0.5, indicating that the identified topics capture most of the review content and provide clear insights into the factors that shape satisfaction and dissatisfaction.

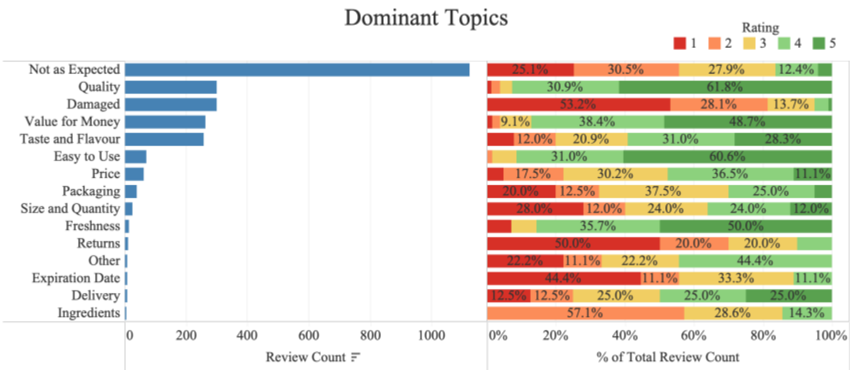


Figure 17: Dominant topics and their distribution by rating

3.2.5 Analysis of Dominant Topic–Emotion Combinations

The analysis of dominant topic-emotion combinations provides an integrated view of how customer concerns and emotional responses interact to shape evaluations. By mapping these combinations against mean sentiment scores and star ratings, Figure 18 identifies not only which topics are associated with satisfaction or dissatisfaction but also how emotions mediate these effects. The quadrant structure highlights alignment and misalignment between predicted sentiment and assigned ratings, offering insights that would not emerge from analysing topics, emotions, or sentiment independently. Bubble size further contextualises the prevalence of each combination, helping to prioritise issues by both impact and frequency. This integrated approach strengthens the interpretability of customer feedback and supports targeted, evidence-based decision-making. Clear patterns emerge across the four quadrants.

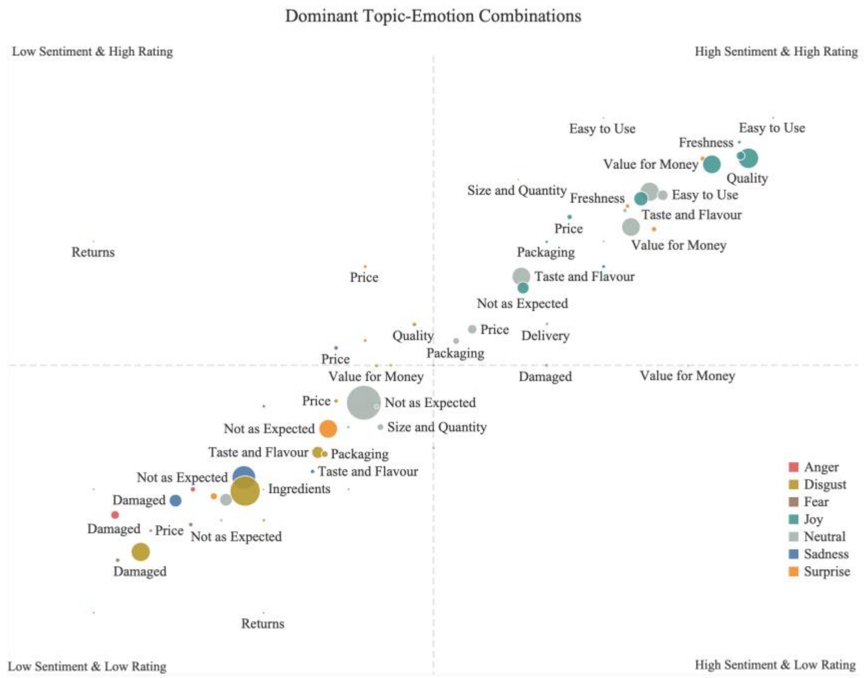


Figure 18: Dominant topic-emotion combinations by mean sentiment score and rating

In the high sentiment and high rating quadrant, combinations such as *Quality* + *Joy*, *Easy to Use* + *Sadness*, and *Value for Money* + *Neutral* dominate. These reflect positive experiences where emotional tone aligns with the evaluation. In this context, *Sadness* may signal effort or relief rather than dissatisfaction.

In the low sentiment and low rating quadrant, combinations such as *Damaged* + *Disgust*, *Returns* + *Sadness*, and *Not as Expected* + *Fear* appear. These represent emotionally charged complaints, with *Disgust* strongly associated with product failure and low satisfaction, confirming its role as a marker of intense negative feedback.

The low sentiment and high rating quadrant is sparsely populated. It includes combinations such as *Returns* + *Surprise*, suggesting that some reviewers award generous star ratings despite neutral or mixed emotional tones, possibly due to brand loyalty or isolated issues.

The high sentiment and low rating quadrant contains rare or anomalous combinations, such as *Damaged* + *Joy*. These may represent sarcasm, mixed signals, or model misclassification.

Overall, the analysis shows that topic–emotion combinations shape the alignment between sentiment and rating. *Joy*, combined with *Quality*, signals satisfaction, while *Disgust*, combined with condition-related topics, reliably indicates dissatisfaction. Topics such as *Not as Expected* and *Returns* recur across emotional tones, highlighting their broad relevance to customer perception.

Collectively, these findings demonstrate that customer satisfaction is associated with both functional reliability and sensory and value-based qualities, underscoring the multidimensional nature of consumer evaluations.

3.2.6 Correlation Analysis

After examining how sentiment, emotions, and topics interact to shape customer evaluations, correlation analysis establishes the internal coherence of the feature set and explores how the extracted features relate both to one another and to star ratings. Table A1 in Appendix 1 summarises all variables retained for further analysis, while the feature correlation matrix in Figure 19 presents their interrelationships. Statistical significance is indicated by asterisks at the levels of $p < 0.05$ (*), $p < 0.01$ (**), and $p < 0.001$ (***).

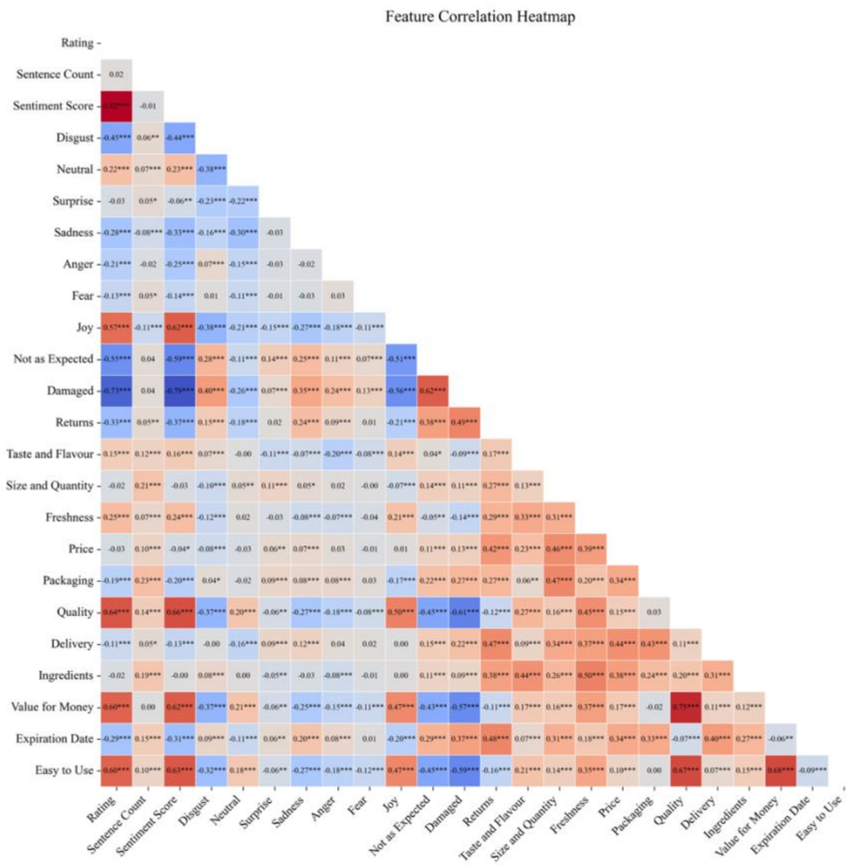


Figure 19: Correlation matrix of features

As expected, *Rating* correlates strongly and positively with *Sentiment Score* ($r = 0.82$, $p < 0.001$), confirming alignment between the user-assigned evaluations and the predicted sentiment. Among the emotions, *Joy* correlates positively with *Rating* ($r = 0.57$, $p < 0.001$), while *Disgust* correlates negatively with it ($r = –0.45$, $p < 0.001$). *Neutral* shows a weaker but significant positive correlation with *Rating* ($r = 0.22$, $p < 0.001$), consistent with its prevalence in less polarised reviews.

Topic features also show meaningful patterns. Negative themes such as *Damaged* ($r = –0.73$, $p < 0.001$) and *Not as Expected* ($r = –0.55$, $p < 0.001$) display strong negative correlations with star ratings. In contrast, *Value for Money* ($r = 0.60$, $p < 0.001$), *Easy to Use* ($r = 0.60$, $p < 0.001$), and *Quality* (r = 0.64, $p < 0.001$) correlate positively, indicating their role as factors significantly associated with satisfaction. *Taste and Flavour* shows only a modest correlation ($r = 0.15$, $p < 0.001$), suggesting that sensory descriptions occur across both positive and negative reviews and do not, by themselves, determine satisfaction.

Sentence Count is weakly correlated with ratings ($r = 0.02$) and is not statistically significant, confirming earlier observations that review length is not a primary determinant of satisfaction.

Beyond their associations with ratings, several internal relationships are evident among the extracted features. *Sentiment Score* correlates positively with *Joy* ($r = 0.62$, $p < 0.001$) and negatively with *Disgust* ($r = –0.44$, $p < 0.001$), *Damaged* ($r = –0.79$, $p < 0.001$), and *Not as Expected* ($r = –0.59$, $p < 0.001$), indicating that the sentiment model aligns with emotion and topic signals. *Damaged* correlates positively with *Returns* ($r = 0.49$, $p < 0.001$) and *Not as Expected* ($r = 0.62$, $p < 0.001$), and negatively with *Quality* ($r = –0.61$, $p < 0.001$), *Value for Money* ($r = –0.57$, $p < 0.001$), and *Easy to Use* ($r = –0.59$, $p < 0.001$). These associations show that product condition issues often overlap with dissatisfaction about expectations, quality, and returns. *Quality* correlates positively with *Value for Money* ($r = 0.75$, $p < 0.001$) and *Easy to Use* ($r = 0.67$, $p < 0.001$), reflecting that these attributes complement each other in shaping positive evaluations.

These results highlight the internal coherence of the feature set. Predicted sentiment aligns with star ratings, emotions align with sentiment polarity, and topics cluster into groups reflecting satisfaction (quality, freshness, value for money) or dissatisfaction (damaged, returns, not as expected). This structure provides a strong conceptual basis for including these features in subsequent regression analysis.

3.3 Regression Analysis

The OLS regression model (Figure 20) explains a substantial proportion of the variation in star ratings, with an R^2 of 0.717 and an adjusted R^2 of 0.714. This indicates that approximately 71% of the variance in ratings is accounted for by the predictors. The close alignment of R^2 and adjusted R^2 suggests that the relatively large set of predictors (22 variables, excluding *Neutral* to address multicollinearity) does not result in excessive overfitting and that most predictors contribute meaningfully to the model.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Rating   R-squared:                       0.717
Model:                            OLS   Adj. R-squared:                  0.714
Method:                 Least Squares   F-statistic:                     359.8
Date:                Sun, 31 Aug 2025   Prob (F-statistic):               0.00
Time:                        18:15:09   Log-Likelihood:                -2275.6
No. Observations:                2000   AIC:                             4597.
Df Residuals:                    1977   BIC:                             4726.
Df Model:                          22
Covariance Type:                  HC3
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                 2.9955      0.017    175.155      0.000       2.962       3.029
Sentence Count        0.0603      0.025      2.379      0.017       0.011       0.110
Sentiment Score       0.6771      0.045     14.950      0.000       0.588       0.766
Disgust              -0.1198      0.026     -4.675      0.000      -0.170      -0.070
Surprise              0.0141      0.020      0.697      0.486      -0.026       0.054
Sadness              -0.0028      0.024     -0.117      0.907      -0.049       0.044
Anger                 0.0081      0.021      0.383      0.701      -0.033       0.050
Fear                 -0.0459      0.022     -2.120      0.034      -0.088      -0.003
Joy                   0.0482      0.022      2.176      0.030       0.005       0.092
Not as Expected      -0.0743      0.023     -3.193      0.001      -0.120      -0.029
Damaged              -0.1600      0.036     -4.389      0.000      -0.232      -0.089
Returns               0.0054      0.025      0.215      0.829      -0.044       0.055
Taste and Flavour     0.0299      0.021      1.399      0.162      -0.012       0.072
Size and Quantity    -0.0080      0.023     -0.350      0.726      -0.053       0.037
Freshness             0.1022      0.022      4.688      0.000       0.059       0.145
Price                 0.0161      0.022      0.717      0.474      -0.028       0.060
Packaging            -0.0313      0.021     -1.458      0.145      -0.073       0.011
Quality               0.1073      0.034      3.120      0.002       0.040       0.175
Delivery             -0.0094      0.024     -0.392      0.695      -0.057       0.038
Ingredients          -0.0981      0.022     -4.507      0.000      -0.141      -0.055
Value for Money       0.0883      0.030      2.912      0.004       0.029       0.148
Expiration Date      -0.0925      0.025     -3.775      0.000      -0.141      -0.044
Easy to Use           0.0513      0.028      1.838      0.066      -0.003       0.106
==============================================================================
Omnibus:                       32.653   Durbin-Watson:                   1.930
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               61.181
Skew:                          -0.031   Prob(JB):                     5.18e-14
Kurtosis:                       3.855   Cond. No.                         5.37
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)
```

Figure 20: Regression model summary with standardised coefficients

The overall model fit is statistically significant ($F(22, 1977) = 359.8$, $p < 0.001$). On the test set, the model shows similar performance ($MSE = 0.583$, $RMSE = 0.763$, $R^2 = 0.700$), indicating that it generalises well beyond the training sample. While the Random Forest achieved slightly higher predictive accuracy **(**$MSE = 0.556$, $RMSE = 0.745$, $R^2 = 0.714$), the OLS model offers greater interpretability, allowing for clear identification of feature-level effects and their relative importance.

The regression analysis focuses on standardised coefficients to allow direct comparison of predictor importance. For completeness, Figure A1 in Appendix 2 provides the regression summary with non-standardised coefficients. These represent the expected change in *Rating* for a one-unit change in each predictor in its original scale. Including both sets of coefficients ensures transparency: the standardised results facilitate comparison across predictors, while the non-standardised values allow real-world interpretation.

3.3.1 Model Diagnostics

The regression diagnostics indicate broad consistency with the key assumptions of linear regression. The Durbin–Watson statistic (1.93) shows no evidence of autocorrelation, and the condition number (5.37) suggests that multicollinearity is not a concern. Variance inflation factors (VIFs) in Figure 21 confirm this finding: the highest values are observed for *Damaged* (3.93), *Sentiment Score* (3.83), and *Quality* (3.34), all of which are well below the conservative threshold of 5. These results indicate that multicollinearity among predictors does not compromise coefficient stability.

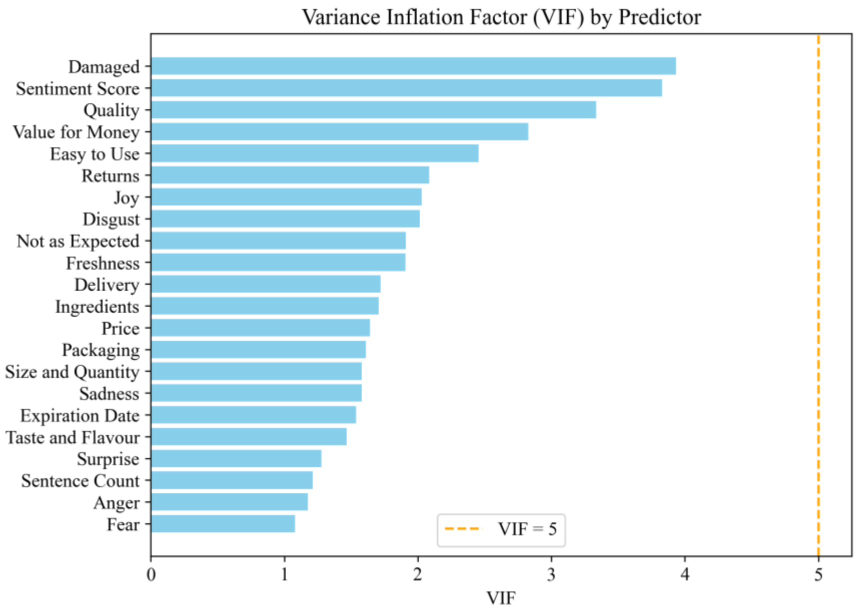


Figure 21: Variance Inflation Factor (VIF) of predictors

The Jarque–Bera test is significant ($p < 0.001$), indicating some deviation from normality in the residuals. However, skewness is close to zero and kurtosis (3.85) indicate only mild leptokurtosis. Residual diagnostics (Figure A2 and Figure A3 in Appendix 3) confirm that residuals are approximately normally distributed, with only minor deviations in the tails. Given the large training sample size (n = 2,000), these departures are unlikely to bias estimates. Heteroscedasticity is present, as indicated by the Breusch–Pagan test (LM = 82.40, $p < 0.001$;

$F = 3.86$, $p < 0.001$), but it is addressed through HC3 heteroscedasticity-robust standard errors, ensuring valid inference.

Taken together, the diagnostic checks confirm that the assumptions of linear regression are sufficiently satisfied, supporting a reliable interpretation of the model results.

3.3.2 Key Predictors of Rating

The regression results indicate that sentiment, emotions, and topic-specific attributes collectively exhibit a significant association with ratings. Standardised coefficients in Figure 22 identify the strongest predictors.

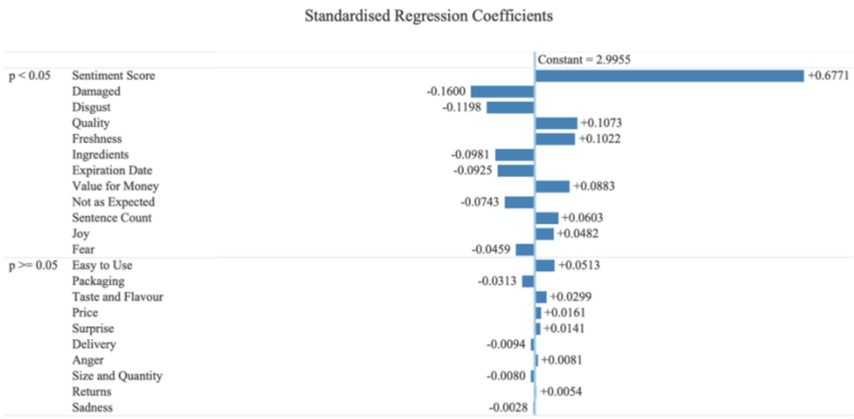


Figure 22: Standardised regression coefficients for rating predictors

The *Sentiment Score* emerges as the dominant predictor, with a relatively large positive coefficient ($\beta = 0.677$, $p < 0.001$), indicating that the overall tone of reviews has a greater association with ratings than any individual attribute. Positive features, such as *Freshness* ($\beta = 0.102$, $p < 0.001$), *Quality* ($\beta = 0.107$, $p = 0.002$), and *Value for Money* ($\beta = 0.088$, $p = 0.004$), are also positively associated with higher ratings, albeit with more modest effects.

Negative experiences exert stronger effects. Mentions of *Damaged* items ($\beta = –0.160$, $p < 0.001$), *Expiration Date* concerns ($\beta = –0.093$, $p < 0.001$), and unmet expectations ($\beta = –0.074$, $p = 0.001$) are all strongly associated with lower ratings. Negative emotions, particularly *Disgust* ($\beta = –0.120$, $p < 0.001$) and *Fear* ($\beta = –0.046$, $p = 0.034$), also demonstrate decreasing effects. The greater magnitude of these negative coefficients underscores the asymmetry in customer evaluations, where product failures and negative emotions carry more weight than positive attributes.

Several predictors, including *Surprise*, *Sadness*, *Anger*, *Returns*, *Taste and Flavour*, *Size* and *Quantity*, *Price*, *Packaging*, *Delivery*, and *Ease of Use* are not statistically significant. This suggests that, once other factors are controlled for, these attributes do not exhibit an independent statistical association with ratings, although they may still be relevant in specific contexts or in interaction with other variables.

The intercept of approximately 3.0 represents the expected star rating for an average review with all predictors at their mean values, closely matching the sample mean of the dependent variable.

3.3.3 Marginal Effects

Figure 23 illustrates the effect of each statistically significant predictor on ratings, holding other variables constant. *Sentiment Score*, *Freshness*, and *Value for Money* all show positive slopes, indicating that higher values consistently translate into higher ratings. *Damaged*, *Disgust*, *Fear*, and *Expiration Date* concerns display negative slopes, confirming that these factors systematically reduce ratings.

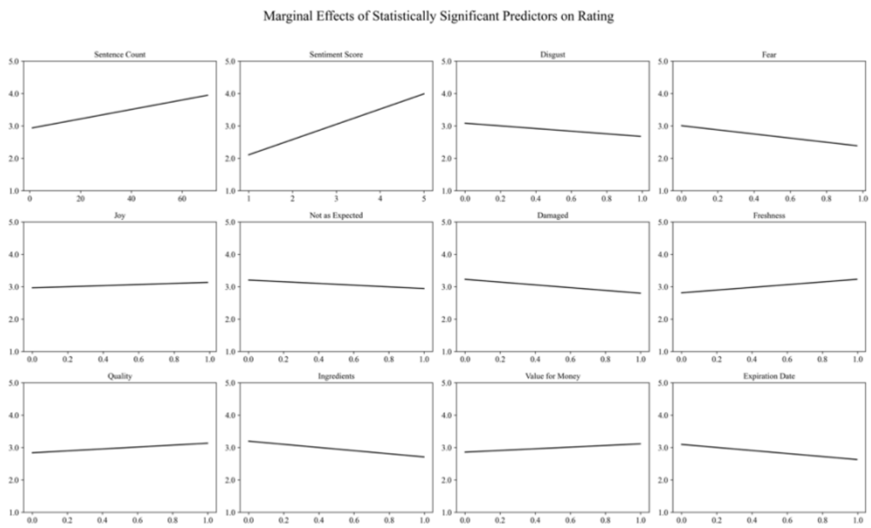


Figure 23: Marginal effects of statistically significant predictors on rating

The differences in gradient are notable. The steep slope for *Sentiment Score* reflects its dominant association with evaluations. *Freshness* and *Value for Money* have more moderate positive slopes, suggesting incremental gains in satisfaction. By contrast, *Fear* and *Expiration Date* concerns show sharper negative slopes, reinforcing the conclusion that negative experiences exert stronger marginal effects than positive ones.

The regression results indicate that sentiment, emotions, and content-specific attributes are collectively associated with customer satisfaction. Sentiment serves as a broad evaluative signal, capturing the overall tone of reviews, while discrete emotions add explanatory value by reflecting the intensity and nature of customer reactions. Positive attributes and emotions contribute to incremental increases in ratings, whereas negative product experiences and adverse emotions exert disproportionately strong effects. This asymmetry underscores the importance of product reliability and quality control in maintaining satisfaction and minimising dissatisfaction.

3.4 Clustering Analysis

Clustering analysis reveals patterns in customer feedback that do not emerge from regression or feature-by-feature analysis. Instead of focusing on how individual variables predict ratings, clustering groups reviews based on their overall sentiment, emotion, and topic features. This approach helps identify natural customer experience profiles and directly addresses the second part of the research question: understanding how textual features distinguish different types of feedback.

3.4.1 Cluster Structure

Using UMAP for dimensionality reduction and HDBSCAN for clustering, the researcher selects a configuration that prioritises interpretability and meaningful insights over statistical robustness among the alternatives generated by the grid search. The chosen settings require clusters to contain at least 35 reviews and apply a comparable threshold for density. This configuration identifies four distinct clusters (Figure 24), with 15.9% of reviews (n = 397) classified as noise (Cluster 0). Figure 25 shows that cluster sizes vary considerably. Cluster 4 is the largest, accounting for 51.6% of the reviews (n = 1,290), while Cluster 1 is the smallest, comprising only 5.9% (n = 147).

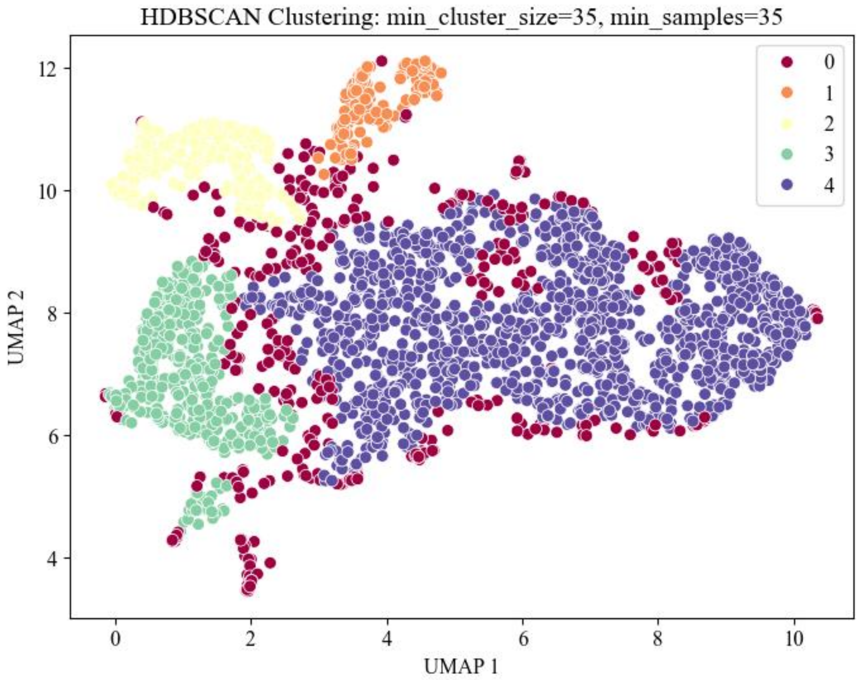


Figure 24: UMAP projection of reviews clustered with HDBSCAN

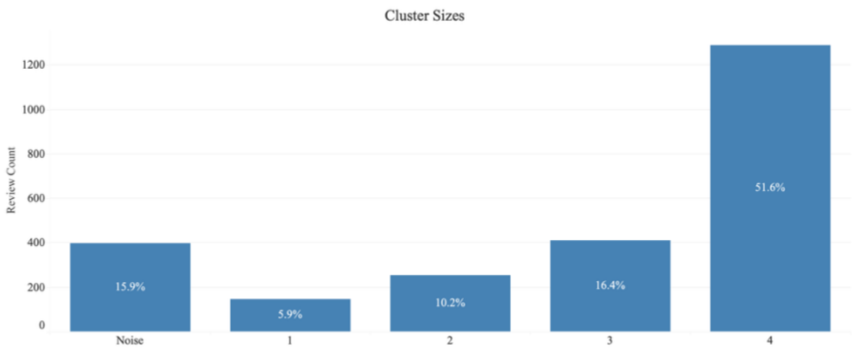


Figure 25: Cluster sizes and proportion of reviews

3.4.2 Cluster Diagnostics

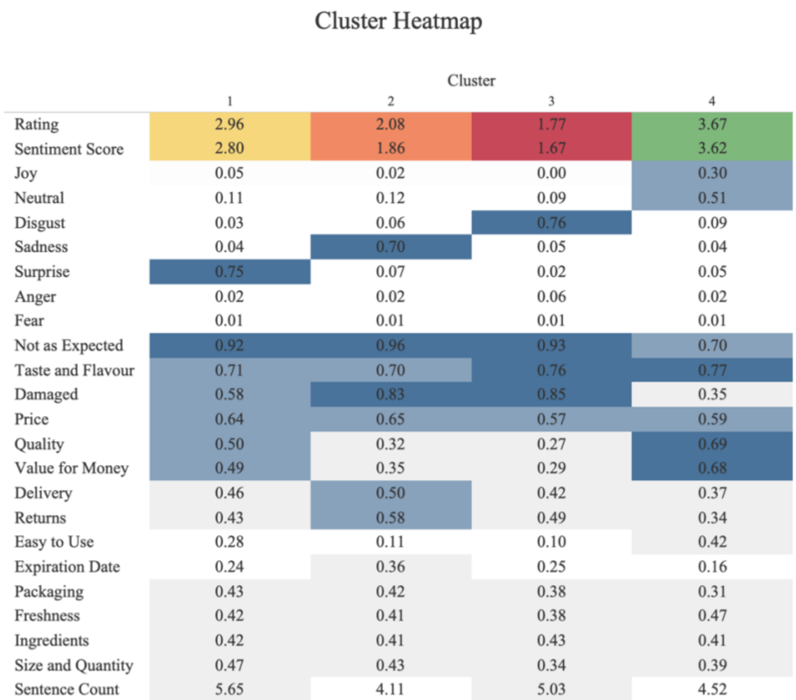
Although the clustering structure shows moderate separation (silhouette score = 0.406) and meaningful between-cluster variance (Calinski–Harabasz Index = 1375.3), the Davies–Bouldin Index (0.59) further indicates relatively distinct clusters. Stability metrics, however, indicate considerable variability in individual assignments (ARI = 0.083 ± 0.022, Jaccard Index = 0.439), suggesting that cluster membership is not highly reproducible. Kruskal–Wallis tests confirm significant differences across clusters for most features after FDR correction (p <

0.05), supporting the interpretability of the results. *Joy*, *Disgust*, *Neutral*, *Sentiment Score*, and *Damaged* demonstrate strong affects (all $H > 750$, $p < 0.001$), as well as *Quality* ($H = 538$, $p < 0.001$), *Value for Money* ($H = 507$, $p < 0.001$), and negative or functional aspects such as *Not as Expected*, *Expiration Date*, and *Returns* (all $H > 225$, $p < 0.001$). *Ingredients* is the only feature without significant variation ($p = 0.184$). For completeness, Table A2 in Appendix 4 provides the Kruskal–Wallis H test results.

Taken together, these results indicate that while individual cluster assignments should be treated cautiously, the overall structure is sufficiently consistent to identify broad patterns and profile customer feedback, rather than for rigid segmentation.

3.4.3 Cluster Characteristics

Cluster characteristics draw on the heatmap of mean feature values (Figure 26), the radar chart with normalised *Rating* and *Sentiment Score* for clearer comparison (Figure 27), and supplementary correlation matrices for each cluster (Figures A4-A7 in Appendix 5). *Taste and Flavour* and *Not as Expected* appear consistently across all clusters, confirming their central role in customer reviews.



Cluster Heatmap

	Cluster			
	1	2	3	4
Rating	2.96	2.08	1.77	3.67
Sentiment Score	2.80	1.86	1.67	3.62
Joy	0.05	0.02	0.00	0.30
Neutral	0.11	0.12	0.09	0.51
Disgust	0.03	0.06	0.76	0.09
Sadness	0.04	0.70	0.05	0.04
Surprise	0.75	0.07	0.02	0.05
Anger	0.02	0.02	0.06	0.02
Fear	0.01	0.01	0.01	0.01
Not as Expected	0.92	0.96	0.93	0.70
Taste and Flavour	0.71	0.70	0.76	0.77
Damaged	0.58	0.83	0.85	0.35
Price	0.64	0.65	0.57	0.59
Quality	0.50	0.32	0.27	0.69
Value for Money	0.49	0.35	0.29	0.68
Delivery	0.46	0.50	0.42	0.37
Returns	0.43	0.58	0.49	0.34
Easy to Use	0.28	0.11	0.10	0.42
Expiration Date	0.24	0.36	0.25	0.16
Packaging	0.43	0.42	0.38	0.31
Freshness	0.42	0.41	0.38	0.47
Ingredients	0.42	0.41	0.43	0.41
Size and Quantity	0.47	0.43	0.34	0.39
Sentence Count	5.65	4.11	5.03	4.52

Figure 26: Cluster characteristics based on mean feature values

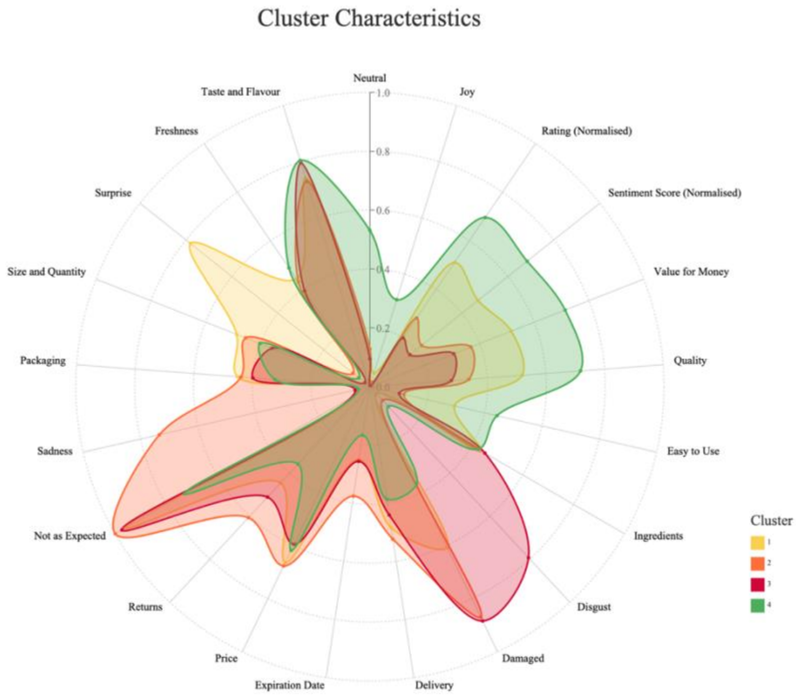


Figure 27: Radar chart of cluster characteristics

Cluster 1 (*n* = 147, mean *Rating* = 2.96) is characterised by a moderate mean sentiment score (2.80) and a high prominence of *Not as Expected* (mean = 0.92). Reviews frequently express surprise (mean = 0.75) and include references to *Price* (mean = 0.64) and *Damaged* (mean = 0.58), indicating a gap between customer expectations and actual experience. The correlation matrix (Figure A4 in Appendix 4) shows strong positive associations between *Sentiment Score* and *Rating* ($r = 0.79$, $p < 0.001$), *Quality* ($r = 0.59$, $p < 0.001$), *Value for Money* ($r = 0.66$, $p < 0.001$), and *Easy to Use* ($r = 0.64$, $p < 0.001$), suggesting that these topic attributes are scarce in reviews with low sentiment. In contrast, references to *Damaged* items correlate negatively with *Sentiment Score* ($r = -0.69$, $p < 0.001$) and evaluative features, reinforcing the view that product condition undermines perceived quality and satisfaction. *Surprise* also shows a moderate negative correlation with *Neutral* emotion ($r = -0.70$, $p < 0.001$), which implies heightened emotional responses when expectations are not met.

Cluster 2 (*n* = 255, mean *Rating* = 2.08) is strongly associated with negative topics, including *Damaged* (mean = 0.83), *Not as Expected* (mean = 0.96), *Returns* (mean = 0.58), *Delivery* (mean = 0.50), and *Expiration Date* (mean = 0.36). The mean *Sentiment Score* is low (1.86) but still correlates positively with *Rating* ($r = 0.52$, $p < 0.001$). *Sadness* dominates as the primary emotion (mean = 0.70) and displays a strong negative correlation with *Neutral* (r = -0.84, $p < 0.001$), showing that dissatisfaction with product condition and logistical problems plays a significant role in shaping the cluster.

Cluster 3 (n = 411, mean *Rating* = 1.77) has the lowest *Sentiment Score* (mean = 1.67). Reviews in this group focus heavily on *Damaged* items (mean = 0.85), unmet expectations (mean = 0.93), and *Taste and Flavour* (mean = 0.76). *Disgust* is the dominant emotion (mean = 0.76), and the correlation structure mirrors that of Cluster 2, though with even stronger negative associations between *Damaged* items and *Sentiment Score* (r = -0.61, $p < 0.001$). This suggests that negative reviews are highly likely to be emotionally intense and mention issues with product condition.

Cluster 4 (n = 1290, mean *Rating* = 3.67) contains reviews with the highest *Sentiment Score* (mean = 3.62). Customer feedback here is primarily neutral (mean = 0.51) or joyful (mean = 0.30), with emphasis on *Quality* (mean = 0.69), *Value for Money* (mean = 0.68), and *Taste and Flavour* (mean = 0.77). Mentions of negative issues are limited, with relatively low mean probability scores for *Damaged* items (0.35) and *Expiration Date* (0.16). *Not as Expected* records the lowest mean probability among clusters, although still high at 0.70. Within this group, star ratings correlate positively with sentiment (r = 0.77, $p < 0.001$) and *Joy* (r = 0.50, $p < 0.001$), while showing negative correlations with *Sadness* (r = -0.37, $p < 0.001$), *Damaged* items (r = -0.67, $p < 0.001$), and unmet expectations (r = -0.52, $p < 0.001$) (Figure A7 in Appendix 4), indicating that these features are rare in the reviews from this group.

Appendix 6 presents summary statistics for each cluster, offering an overview of the central tendencies and distributions to complement the results in this chapter.

4. Discussion

This study aims to examine the relationships between sentiment, emotional tone, and thematic content in online customer reviews and different levels of satisfaction, as measured by star ratings. It also aims to identify the features that distinguish various types of customer feedback. Using a stratified sample of 2,500 Amazon reviews from the grocery and gourmet food category, the researcher employs pre-trained transformer models to extract sentiment, emotion, and topic features from unstructured reviews. The study applies regression analysis to identify the features that explain variation in customer satisfaction, and uses clustering analysis to reveal distinct types of customer feedback.

The findings demonstrate that examining reviews through an integrated lens of sentiment, emotion, and topics offers a richer and more nuanced understanding of satisfaction than ratings or sentiment alone. The combination of descriptive and predictive methods allows the analysis to move beyond isolated variable associations and to reveal broader patterns in customer feedback.

4.1 Summary of Key Findings

The exploratory analysis shows that shorter reviews are more frequent, while mid-range ratings tend to include longer and more detailed reviews. Vocabulary patterns highlight the role of taste, flavour, price, value, unmet expectations, and product condition issues, such as expired or damaged items, in distinguishing reviews across satisfaction levels.

The pre-trained transformer models successfully capture systematic variation in sentiment scores, emotional tones, and thematic content across rating levels. Consistent with prior studies, the classification report confirms that the sentiment model performs well at distinguishing polarity (positive versus negative) but faces challenges with more nuanced evaluations (Wang and Ester, 2014). The analysis also reveals some inconsistencies: in some cases, the predicted sentiment or emotion diverges from the assigned star rating, particularly in reviews that contain sarcasm, irony, or positive language paired with low ratings. These findings reflect the complexity of subjective, user-generated content and support prior research, demonstrating that implicit sentiment and mixed polarity present significant challenges for accurate classification (Li et al., 2021).

The correlation analysis provides further insights into relationships between features. Sentiment scores are most strongly correlated with star ratings, while emotions such as disgust and joy, as well as several topic features, also show significant but more moderate associations. This confirms the multidimensional nature of the extracted features and their relevance for explaining variation in customer ratings (Wang, Du and Wang, 2023).

The regression analysis demonstrates the predictive power of sentiment, emotion, and topic features, which collectively explain over 70% of the variation in ratings. The findings indicate that the sentiment score is the strongest overall predictor of star ratings, reaffirming its value in capturing evaluative tone. However, emotions and topics also contribute explanatory power, aligning with prior studies that indicate sentiment analysis alone does not capture the full complexity of customer feedback (AL-Barrak and Al-Alawi, 2024). Disgust and fear are associated with lower ratings, while joy best explains higher ratings. Disgust emerges as the most significant negative factor, consistent with prior findings that this emotion, along with sadness, strongly reduces consumer preference for food products (Motoki, Motoki and Sugiura, 2018). In contrast, reviews expressing joy are positively associated with higher ratings, signalling quality and satisfaction (Wang et al., 2019). Topic-level analysis reveals that negative product-related themes such as damaged items, unmet expectations, ingredients, and expiration dates predict lower ratings, underscoring the importance of quality control and effective return policies (Barutçu and Basak, 2018). Conversely, aspects such as freshness, quality, and value for money are consistently associated with higher ratings, reinforcing their importance for customer satisfaction.

Clustering analysis supports these findings by identifying four distinct types of customer feedback. Emotional tone serves as the primary differentiator across clusters, with joy standing out as the most distinctive variable, while sadness dominates the reviews, albeit not as a statistically significant predictor of star ratings. Dissatisfaction often arises from unmet expectations and product failures, consistent with expectation-disconfirmation theory, which suggests that satisfaction reflects the gap between expected and actual experiences (Park et al., 2021).

Cluster 1, described as “Surprised Reviews with Disconfirmed Expectations,” reflects mean sentiment scores and ratings but high levels of surprise, highlighting the influence of expectation gaps (Vanhamme, 2003).

Cluster 2, "Disappointed by Product Condition," includes reviews with lower mean ratings and expressions of sadness related to damaged items, returns, and delivery problems, consistent with previous evidence that logistical issues weaken future purchase intentions (Tolulope and Owoseni, 2024). Although reviews expressing sadness often indicate dissatisfaction, research indicates that such reviews are seen as less helpful, as they may lack the emotional intensity needed to strongly resonate with other consumers (Ren and Hong, 2019). This implies that while such reviews still signal reputational risk, their persuasive influence on future buyers might be weaker.

Cluster 3, "Emotionally Intense Complaints," includes the most negative reviews, mainly characterised by disgust and associated with failures in taste, quality, and product integrity. Studies indicate that expressions of disgust diminish purchase intentions and trigger contagion effects across related products (Morales and Fitzsimons, 2007; Guido, Pino and Peluso, 2017).

In contrast, Cluster 4, "Neutral to Joyful Experiences," represents broadly satisfied customers, emphasising hedonic and functional features such as taste, value, freshness, and quality. This aligns with the concept of customer delight, where joy arises when experiences exceed expectations (Barnes and Krallman, 2019).

While not a primary focus of the analysis, sentence count provides structural context for analysis. It shows a moderate positive association with ratings in regression analysis but is relatively weak in clustering results. This finding aligns with prior research, which suggests that longer reviews may reflect higher engagement or more complex experiences, but have limited predictive power compared to sentiment, emotions, and topics (Ghasemaghaei et al., 2018).

Taken together, the exploratory, regression, and clustering results confirm the multidimensional nature of customer satisfaction. Feedback varies systematically in sentiment, emotional tone, and thematic focus, underscoring the heterogeneity of consumer experiences and the value of integrated analysis (Kyriakidis and Tsafarakis, 2024).

4.2 Business Implications

The findings suggest several actionable strategies for enhancing the customer experience.

Expectation management is central to shaping satisfaction (Ofir and Simonson, 2007). The frequent occurrence of unmet expectations and their significant impact on ratings suggest that businesses should invest in clearer product descriptions and more accurate visual representations to minimise post-purchase disappointment.

Product condition and fulfilment emerge as critical vulnerabilities. Clusters 2 and 3 indicate that damaged products have a strong negative association with satisfaction. With quality ranking as the second strongest positive predictor of ratings, addressing these tangible issues can yield immediate improvements in customer experience (Renko, Petljak and Naletina, 2019).

Emotional intensity amplifies dissatisfaction. The prevalence of disgust in low-rated reviews, along with the negative association between fear and ratings, demonstrates that emotionally charged failures have lasting effects. Since disgust is one of the strongest negative predictors of ratings, businesses should prioritise these cases through rapid response systems, proactive customer service interventions, and compensation policies to mitigate reputational risks (Langaro et al., 2022).

Positive reinforcement also plays a key role. Attributes such as value for money, freshness, and quality are consistently linked to higher ratings and the group with the most positive reviews. Reinforcing these perceptions through transparent pricing, accurate product descriptions, and strong branding can strengthen satisfaction and encourage repeat purchases. In e-commerce, sales promotions and personalised shopping experiences have been shown to boost positive emotions, which in turn increase purchase intentions and consumer loyalty (Cuong, 2024). Leveraging these strategies can foster customer satisfaction and drive long-term growth.

4.3 Limitations

While this study provides valuable insights into customer feedback, several limitations must be acknowledged.

The analysis relies on a dataset restricted to a single product category and platform, which may limit generalisability to other domains. It focuses on document-level features and does not incorporate aspect-based sentiment analysis, reducing precision in identifying causal links between features and satisfaction.

Online customer reviews are inherently noisy and variable in quality, often containing informal language, typographical errors, and ambiguous phrasing (Tribe, 2023). Although transformer-based models improve contextual understanding, the findings reveal cases when the models used in the study do not reliably capture implicit sentiment, mixed polarity, and sarcasm. Prior research emphasises that sarcasm, irony, and context-dependent phrasing can invert or obscure sentiment polarity (Potamias, Siolas and Stafylopatis, 2020). Consequently, some regression results and cluster profiles may only partially reflect the underlying affective content.

The results are also model-dependent. Alternative transformer architectures or variants fine-tuned on different datasets may produce different outputs for sentiment, emotions, or topics. While the topic labels used in this study are grounded in the literature review and exploratory analysis, their selection involves the researcher's judgement, and alternative label sets could yield different outcomes.

Similarly, feature selection and regression results depend on methodological decisions, such as the choice of predictive algorithms and inclusion criteria, which influence estimates of feature importance. Clustering outcomes are sensitive to model design, parameter settings, and choices of dimensionality reduction and clustering algorithm.

Although the models and settings used in this study prioritise interpretability, alternative approaches may achieve higher accuracy at the expense of transparency. Taken together, these limitations emphasise that the presented patterns and insights depend on the specific modelling and methodological choices made in this study.

References

Achar, C., So, J., Agrawal, N. and Duhachek, A., 2016. What we feel and why we buy: the influence of emotions on consumer decision-making. *Current Opinion in Psychology* [Online], 10(10), pp.166–170. Available from: https://doi.org/10.1016/J.COPSYC.2016.01.009.

Adak, A., Pradhan, B. and Shukla, N., 2022. Sentiment Analysis of Customer Reviews of Food Delivery Services Using Deep Learning and Explainable Artificial Intelligence: Systematic Review. *Foods* [Online], 11(10), p.1500. Available from: https://doi.org/10.3390/foods11101500.

Agarwal, T., Jangid, J. and Kumar, G., 2023.Transformer and Natural language processing; A recent development. *Tuijin Jishu/Journal of Propulsion Technology* [Online], 44(1), pp.140–143. Available from: https://doi.org/10.52783/tjjpt.v44.i1.2225.

AL-Barrak, M.A. and Al-Alawi, A.I., 2024.Sentiment Analysis on Customer Feedback for Improved Decision Making: A Literature Review. *2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS)* [Online], 28-29 January 2024, Manama, Bahrain, IEEE, pp.207–212. Available from: https://doi.org/10.1109/icetsis61505.2024.10459452.

Angelis, J.N., Murthy, R.S., Beaulieu, T. and Miller, J.C., 2022. Better Angry Than Afraid: The Case of Post Data Breach Emotions on Customer Engagement. *IEEE Transactions on Engineering Management* [Online], 71, pp.2593–2605. Available from: https://doi.org/10.1109/tem.2022.3189599.

Archak, N., Ghose, A. and Ipeirotis, P.G., 2011. Deriving the pricing power of product features by mining consumer reviews. *Management Science* [Online], 57(8), pp.1485-1509. Available from: https://doi.org/10.1287/mnsc.1110.1370.

Arora, S., Mahapatra, S., Jadav, A., Barla, M. and Mallick, N., 2024. Temporal and Sentimental Analysis of Customer Reviews. *2024 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* [Online], 18-19 January 2024, Noida, India, IEEE, pp.520–525. Available from: https://doi.org/10.1109/confluence60223.2024.10463473.

Bar-Haim, R., Eden, L., Kantor, Y., Friedman, R. and Slonim, N., 2021. Every Bite Is an Experience: Key Point Analysis of Business Reviews. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* [Online], Association for Computational Linguistics, pp.3376–3386. Available from: https://doi.org/10.18653/V1/2021.ACL-LONG.262.

Barnes, D.C. and Krallman, A., 2019. Customer Delight: A Review and Agenda for Research. *The Journal of Marketing Theory and Practice* [Online], 27(2), pp.174–195. Available from: https://doi.org/10.1080/10696679.2019.1577686.

Barutçu, M.T. and Basak, B., 2018. Customer Complaints about E-Commerce Sites: Content Analysis. *The Eurasia Proceedings of Educational & Social Sciences (EPESS)* [Online], 10, pp.238–243. Available from: https://dergipark.org.tr/tr/download/article-file/536232 [Accessed 7 September 2025].

Bashath, S., Perera, N., Tripathi, S., Manjang, K., Dehmer, M. and Emmert Streib, F., 2022. A data-centric review of deep transfer learning with applications to text data. *Information Sciences* [Online], 585, pp.498–528. Available from: https://doi.org/10.1016/j.ins.2021.11.061.

Birjali, M., Kasri, M. and Beni-Hssane, A., 2021. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems* [Online], 226, p.107134. Available from: https://doi.org/10.1016/J.KNOSYS.2021.107134.

Böhm, M., 2022. The interpretation of topic models for scholarly analysis: An evaluation and critique of current practice. *Digital Scholarship in the Humanities* [Online]. Available from: https://doi.org/10.1093/llc/fqac075.

Caliński, T. and Harabasz, J., 1974. A dendrite method for cluster analysis. *Communications in Statistics* [Online], 3(1), pp.1–27. Available from: https://doi.org/10.1080/03610927408827101.

Changchit, C. and Klaus, T., 2019. Determinants and Impact of Online Reviews on Product Satisfaction. *Journal of Internet Commerce* [Online], 19(1), pp.82-102. Available from: https://doi.org/10.1080/15332861.2019.1672135.

ChatGPT, 2025. [Online] Available from: https://chat.openai.com/auth/login. [Accessed 4 August 2025].

Chefer, H., Gur, S. and Wolf, L., 2021. Transformer Interpretability Beyond Attention Visualization. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* [Online], 20-25 June 2021, Nashville, TN, USA, IEEE, pp.782–791. Available from: https://doi.org/10.1109/CVPR46437.2021.00084.

Cherukuri, M., 2024.Exploring Multi-Dimensional Sentiment Analysis: A Study on Emotion Representation Structures and Prediction Models. *REST Journal on Data Analytics and Artificial Intelligence* [Online]. 3(3), pp.55–76. Available from: https://doi.org/10.46632/jdaai/3/3/7.

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

Cho, H.S., Sosa, M.E. and Hasija, S., 2022. Reading Between the Stars: Understanding the Effects of Online Customer Reviews on Product Demand. *Manufacturing & Service Operations Management* [Online], 24(4), pp.1977–1996. Available from: https://doi.org/10.1287/msom.2021.1048.

Churchill, R. and Singh, L., 2022. The Evolution of Topic Modeling. *ACM Computing Surveys* [Online], 54(10s), pp.1-35. Available from: https://doi.org/10.1145/3507900.

Cui, C., Wei, M., Che, L., Wu, S. and Wang, E., 2022. Hotel recommendation algorithms based on online reviews and probabilistic linguistic term sets. *Expert Systems with Applications* [Online], 210, p.118503. Available from: https://doi.org/10.1016/j.eswa.2022.118503.

Cuong, D.T., 2024. Positive emotions influencing consumer shopping behavior on e-commerce platforms. *Management & Marketing* [Online], 19(1), pp.15–31. Available from: https://doi.org/10.2478/mmcks-2024-0002.

Davies, D.L. and Bouldin, D.W., 1979. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence* [Online], 1(2), pp.224–227. Available from: https://doi.org/10.1109/TPAMI.1979.4766909.

Davis, S. and Tabrizi, N., 2021. Customer Review Analysis: A Systematic Review. *2021 IEEE/ACIS 6th International Conference on Big Data, Cloud Computing, and Data Science (BCD)* [Online], 13-15 September 2021, Zhuhai, China, IEEE, pp.91–97. Available from: https://doi.org/10.1109/BCD51206.2021.9581965.

Dawson, N.V. and Weiss, R.E., 2012. Dichotomizing Continuous Variables in Statistical Analysis: A Practice to Avoid. *Medical Decision Making* [Online], 32(2), pp.225–226. Available from: https://doi.org/10.1177/0272989X12437605.

Deng, J. and Ren, F., 2023. A Survey of Textual Emotion Recognition and Its Challenges. *IEEE Transactions on Affective Computing* [Online], 14(1), pp.49–67. Available from: https://doi.org/10.1109/TAFFC.2021.3053275.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *North American Chapter of the Association for Computational Linguistics* [Online]. Available from: https://api.semanticscholar.org/CorpusID:52967399 [Accessed 5 August 2025].

Dunstan, D., Crowne, J.E. and Drew, A.J., 2022. Easy computation of the Bayes factor to fully quantify Occam's razor in least-squares fitting and to guide actions. *Dental science reports* [Online], 12(1). Available from: https://doi.org/10.1038/s41598-021-04694-7.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion* [Online], 6(3-4), pp.169–200. https://doi.org/10.1080/02699939208411068.

Fülöp, F., 2022. Unboxing the Causal Effect of Ratings on Product Demand: Evidence from Wayfair.com. *Journal of Industrial Economics* [Online]. Available from: https://doi.org/10.1111/joie.12302.

Ghasemaghaei, M., Eslami, S.P., Deal, K. and Hassanein, K., 2018. Reviews' length and sentiment as correlates of online reviews' ratings. *Internet Research* [Online], 28(3), pp.544–563. Available from: https://doi.org/10.1108/INTR-12-2016-0394.

Gorai, J. and Shaw, D.K., 2024. Multi-Modal Sentiment Analysis of Product Reviews. *2024 International Conference on Computer, Electronics, Electrical Engineering & their Applications (IC2E3)* [Online], 06-07 June 2024, Srinagar Garhwal, Uttarakhand, India, IEEE, pp.1–4. Available from: https://doi.org/10.1109/ic2e362166.2024.10827296.

Green, S.B., 1991. How Many Subjects Does It Take To Do A Regression Analysis. *Multivariate Behavioral Research* [Online], 26(3), pp.499–510. Available from: https://doi.org/10.1207/S15327906MBR2603_7.

Guido, G., Pino, G. and Peluso, A.M., 2017. The impact of disgust on consumers' purchase intentions: An empirical assessment. *Journal of Consumer Marketing* [Online], 35(1), pp.105–115. Available from: https://doi.org/10.1108/JCM-04-2016-1786.

Hartmann, J., 2022. *Emotion English DistilRoBERTa-base* [Online]. Available from: https://huggingface.co/j-hartmann/emotion-english-distilroberta-base [Accessed 15 August 2025].

Hashmi, E. and Yayilgan, S.Y., 2024. A robust hybrid approach with product context-aware learning and explainable AI for sentiment analysis in Amazon user reviews. *Electronic Commerce Research* [Online]. Available from: https://doi.org/10.1007/s10660-024-09896-5.

Herqutanto, M.F., Zatari, R.P. and Sutoyo, R., 2023. Topic Modeling Using LDA-Based and Machine Learning for Aspect Sentiment Analysis. *2023 International Conference on Informatics, Multimedia, Cyber and Informations System (ICIMCIS)* [Online], 07-08 November 2023, Jakarta Selatan, Indonesia, IEEE, pp.142–148. Available from: https://doi.org/10.1109/icimcis60089.2023.10349056.

Herrmann, M., Kazempour, D., Scheipl, F. and Kröger, P., 2023. Enhancing cluster analysis via topological manifold learning. *Data Mining and Knowledge Discovery* [Online], 38, pp. 840-887. Available from: https://doi.org/10.1007/s10618-023-00980-2.

Hidayatullah, A.F., Kalinaki, K., Gul, H., Zakari, R.Y. and Shafik, W., 2024. Leveraging Natural Language Processing for Enhanced Text Analysis in Business Intelligence. *Advances in computational intelligence and robotics book series* [Online], pp.151–182. Available from: https://doi.org/10.4018/979-8-3693-5288-5.ch006.

Hou, Y., Li, J., He, Z., Yan, A., Chen, X. and McAuley, J., 2024. Bridging Language and Items for Retrieval and Recommendation. *arXiv* [Online]. Available from: https://doi.org/10.48550/arxiv.2403.03952.

Huang, B., Guo, R., Zhu, Y., Fang, Z., Zeng, G., Liu, J., Wang, Y., Fujita, H. and Shi, Z., 2022. Aspect-level sentiment analysis with aspect-specific context position information. *Knowledge-Based Systems* [Online], 243, p.108473. Available from: https://doi.org/10.1016/j.knosys.2022.108473.

Hugging Face, 2025. [Online] Available from: https://huggingface.co/ [Accessed 10 August 2025].

Hunter, J.E. and Hamilton, M., 2002. The Advantages of Using Standardized Scores in Causal Analysis. *Human Communication Research* [Online], 28(4), pp.552–561. Available from: https://doi.org/10.1111/J.1468-2958.2002.TB00823.X.

Ismail, W.S., Ghareeb, M.M. and Youssry, H., 2024. Enhancing Customer Experience through Sentiment Analysis and Natural Language Processing in E-commerce. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications* [Online], 15(3), pp.60–72. Available from: https://doi.org/10.58346/jowua.2024.i3.005.

Jaccard, P., 1912. The Distribution of the Flora in the Alpine Zone. *The New Phytologist* [Online], 11, pp.37–50. Available from: https://doi.org/10.1111/j.1469-8137.1912.tb05611.x.

Jang, M. and Lukasiewicz, T., 2023. Consistency Analysis of ChatGPT. *arXiv* [Online]. Available from: https://doi.org/10.48550/arXiv.2303.06273.

Jia, M., Kim, H.-S. and Tao, S., 2023. B&B Customer Experience and Satisfaction: Evidence from Online Customer Reviews. *Service Science* [Online], 16(1), pp. 42-54. Available from: https://doi.org/10.1287/serv.2022.0080.

Kanwal, I., Wahid, F., Ali, S., Rehman, A.-U., Alkhayyat, A. and Al-Radaei, A., 2023. Sentiment Analysis Using Hybrid Model of Stacked Auto-Encoder-Based Feature Extraction and Long Short Term Memory-Based Classification Approach. *IEEE Access* [Online], 11, pp.124181–124197. Available from: https://doi.org/10.1109/ACCESS.2023.3313189.

Kemper, J., 2017. The Power of Online Customer Reviews in Fashion E-Commerce - An Empirical Analysis Across Categories and Brands. *25th European Conference on Information Systems* [Online], June 5-10, 2017, Guimarães, Portugal. Available from: https://aisel.aisnet.org/ecis2017_rp/29/ [Accessed 8 September 2025].

Kim, J.H., 2019. Multicollinearity and misleading statistical results. *Korean Journal of Anesthesiology* [Online], 72(6), pp.558–569. Available from: https://doi.org/10.4097/KJA.19087.

Kranzbühler, A.M., Zerres, A., Kleijnen, M.H.P. and Verlegh, P.W.J., 2020. Beyond valence: a meta-analysis of discrete emotions in firm-customer encounters. *Journal of the Academy of Marketing Science* [Online], 48(3), pp.478–498. Available from: https://doi.org/10.1007/S11747-019-00707-0.

Kruskal, W.H. and Wallis, W.A., 1952. Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association* [Online], 47(260), pp.583–621. Available from: https://doi.org/10.1080/01621459.1952.10483441.

Kyriakidis, A. and Tsafarakis, S., 2024a. Extracting knowledge from customer reviews: an integrated framework for digital platform analytics. *International Transactions in Operational Research* [Online]. Available from: https://doi.org/10.1111/itor.13537.

Kyritsis, K., Spatiotis, N., Perikos, I. and Paraskevas, M., 2023. *A Comparative Performance Evaluation of Algorithms for the Analysis and Recognition of Emotional Content* [Online]. IntechOpen. Available from: https://doi.org/10.5772/intechopen.112627.

Lakatos, R., Bogacsovics, G., Harangi, B., Lakatos, I., Tiba, A., Toth, J., Szabó, M. and Hajdu, A., 2024. A Machine Learning-Based Pipeline for the Extraction of Insights from Customer Reviews. *Big data and cognitive computing* [Online]. Available from: https://doi.org/10.3390/bdcc8030020.

Langaro, D., Loureiro, S., Schivinski, B. and Neves, H., 2022. In the eye of the (fire)storm: better safe or sorry? Crisis communication strategies for managing virality of online negative brand-related content. *Journal of Marketing Communications* [Online], 30(3), pp. 301–317. Available from: https://doi.org/10.1080/13527266.2022.2109056.

Lei, Z., Yin, D., Mitra, S. and Zhang, H., 2022. Swayed by the reviews: Disentangling the effects of mean ratings and individual reviews in online word-of-mouth. *Production and Operations Management* [Online], 31(6), pp.2393–2411. Available from: https://doi.org/10.1111/poms.13695.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V. and Zettlemoyer, L., 2019. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *arXiv* [Online]. Available from: https://doi.org/10.48550/arxiv.1910.13461.

Li, N., Liu, Y. and Chen, Z., 2024. Unlocking insights: integrated text mining and interpretive structural modeling for enhanced user review analysis. *PeerJ Computer Science* [Online], 10, p.e2541. Available from: https://doi.org/10.7717/peerj-cs.2541.

Li, Z., Zou, Y., Zhang, C., Zhang, Q. and Wei, Z., 2021. Learning Implicit Sentiment in Aspect-based Sentiment Analysis with Supervised Contrastive Pre-Training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* [Online], pp.246–256. Available from: https://doi.org/10.18653/v1/2021.emnlp-main.22.

Lien, Y.-C., Zhang, R., Harper, F.M., Murdock, V. and Lee, C.-J., 2022.Leveraging Customer Reviews for E-commerce Query Generation. *Lecture notes in computer science* [Online], 13186, pp.190–198. Available from: https://doi.org/10.1007/978-3-030-99739-7_22.

Liu, R., Shi, Y., Ji, C. and Jia, M., 2019. A Survey of Sentiment Analysis Based on Transfer Learning. *IEEE Access* [Online], 7, pp.85401–85412. Available from: https://doi.org/10.1109/ACCESS.2019.2925059.

Lourdusamy, R., Thangavel, P. and Johnbosco, S., 2024. Sentiments Unleashed: Pioneering the Frontier of Sentiment Analysis through Cutting-Edge Applications and Methodologies. *International journal of scientific research in computer science, engineering and information technology* [Online], 10(5), pp.205–220. Available from: https://doi.org/10.32628/cseit24105105.

Lupulescu, M.G.G., Dincă, V.M., Taranu, S.-D. and Blănuţă, B.A., 2024. Data-Driven Insights from 10,000 Reviews: Fostering Sustainability through Rapid Adaptation to Guest Feedback. *Sustainability* [Online], 16(7), p.2759. Available from: https://doi.org/10.3390/su16072759.

Ma, Q., 2023. Product Reviews: Analyzing Sentiment, Identifying Trends, and Designing Marketing Strategies. In *Managing Product Reviews: A Comprehensive Guide for Brands and Businesses* [Online]. CSMFL Publications eBooks, pp.74–92. Available from: https://doi.org/10.46679/978819573226506.

Malik, N. and Bilal, M., 2024. Natural language processing for analyzing online customer reviews: a survey, taxonomy, and open research challenges. *PeerJ Computer Science* [Online], 10, p.e2203. Available from: https://doi.org/10.7717/peerj-cs.2203.

McCloskey, B.J., LaCasse, P.M. and Cox, B.A., 2024. Natural language processing analysis of online reviews for small business: extracting insight from small corpora. *Annals of Operations Research* [Online], 341, pp.295–312. Available from: https://doi.org/10.1007/s10479-023-05816-2.

Meng, F., Gao, J. and Bao, S., 2023.Aspect-Level Sentiment Analysis Review: Recent Advances and Trends. *2023 IEEE 15th International Conference on Computational Intelligence and Communication Networks (CICN)* [Online], 22-23 December 2023, Bangkok, Thailand, IEEE. Available from: https://doi.org/10.1109/cicn59264.2023.10402225.

Michael, P., 2025. *Amazon Statistics: A 2025 Look at Users, Sellers, Revenue, and Marketplace Trends.Analyzer.tools* [Online]. Available from: https://www.analyzer.tools/amazon-statistics/ [Accessed 11 August 2025].

Mikucka, M., Sarracino, F. and Dubrow, J.K., 2015. Costs and Benefits of Including or Omitting Interaction Terms: A Monte Carlo Simulation. *Sociological Methods & Research* [Online]. Available from: https://api.semanticscholar.org/CorpusID:124121327 [Accessed 8 September 2025].

Morales, A.C. and Fitzsimons, G.J., 2007. Product Contagion: Changing Consumer Evaluations Through Physical Contact with “Disgusting” Products. *Journal of Marketing Research* [Online], 44(2), pp.272–283. Available from: https://doi.org/10.1509/JMKR.44.2.272.

Motoki, K., Motoki, K. and Sugiura, M., 2018. Disgust, Sadness, and Appraisal: Disgusted Consumers Dislike Food More Than Sad Ones. *Frontiers in Psychology* [Online], 9, p.76. Available from: https://doi.org/10.3389/FPSYG.2018.00076.

Mylonas, N., Mollas, I. and Tsoumakas, G., 2022. An attention matrix for every decision: faithfulness-based arbitration among multiple attention-based interpretations of transformers in text classification. *Data Mining and Knowledge Discovery* [Online], 38, pp. 128–153. Available from: https://doi.org/10.1007/s10618-023-00962-4.

Narejo, K.R., 2024. EEBERT: An Emoji-Enhanced BERT Fine-Tuning on Amazon Product Reviews for Text Sentiment Classification. *IEEE Access* [Online], 12, pp.131954–131967. Available from: https://doi.org/10.1109/ACCESS.2024.3456039.

NLP Town, 2023. *BERT-base-multilingual-uncased-sentiment (Revision edd66ab)* [Online]. Hugging Face. Available at: https://doi.org/10.57967/hf/1515.

Nwangburuka, C., Ijomah, M. and Nwakuya, M., 2023. Heteroscedasticity of unknown form: a comparison of five heteroscedasticity-consistent covariance matrix (hccm) estimators. *Global Journal of Pure and Applied Sciences* [Online], 29(1), pp.83–90. Available from: https://doi.org/10.4314/gjpas.v29i1.10.

Obiedat, R., Qaddoura, R., Al-Zoubi, A.M., Al-Qaisi, L., Harfoushi, O., Alrefai, M. and Faris, H., 2022. Sentiment Analysis of Customers' Reviews Using a Hybrid Evolutionary SVM-Based Approach in an Imbalanced Data Distribution. *IEEE Access* [Online], 10, pp.22260–22273. Available from: https://doi.org/10.1109/ACCESS.2022.3149482.

Oğuz, A., 2024. Consumer Behavior in the Era of AI-Driven Marketing. *Human computer interaction* [Online], 8(1), p.147. Available from: https://doi.org/10.62802/h9frxh42.

Ofir, C. and Simonson, I., 2007. The Effect of Stating Expectations on Customer Satisfaction and Shopping Experience. *Journal of Marketing Research* [Online], 44(1), pp.164–174. Available from: https://journals.sagepub.com/doi/abs/10.1509/jmkr.44.1.164 [Accessed 8 September 2025].

Park, Y.J., Joo, J., Polpanumas, C. and Yoon, Y., 2021. "Worse Than What I Read?" The External Effect of Review Ratings on the Online Review Generation Process: An Empirical Analysis of Multiple Product Categories Using Amazon.com Review Data. *Sustainability* [Online], 13(19), p.10912. Available from: https://doi.org/10.3390/SU131910912.

Park, K., Park, S. and Joung, J., 2024. Contextual Meaning-Based Approach to Fine-Grained Online Product Review Analysis for Product Design. *IEEE Access* [Online], 12, pp.4225–4238. Available from: https://doi.org/10.1109/ACCESS.2023.3343501.

Posch, M., Zehetmayer, S. and Bauer, P., 2009. Hunting for Significance With the False Discovery Rate. *Journal of the American Statistical Association* [Online], 104(486), pp.832–840. Available from: https://doi.org/10.1198/JASA.2009.0137.

Potamias, R.A., Siolas, G. and Stafylopatis, A., 2020. A transformer-based approach to irony and sarcasm detection. *Neural Computing and Applications* [Online], 32(23), pp.17309–17320. Available from: https://doi.org/10.1007/S00521-020-05102-3.

Punitha, K., Raja Shree, S. and Cruz Antony, J., 2025. A Systematic Review of Sentiment Analysis Approaches and Techniques. *2025 6th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI)* [Online], 07-08 January 2025, Goathgaun, Nepal, IEEE, pp.603–607. Available from: https://doi.org/10.1109/ICMCSI64620.2025.10883219.

Rachmiani, R., Oktadinna, N. and Fauzan, T., 2024. The Impact of Online Reviews and Ratings on Consumer Purchasing Decisions on E-commerce Platforms. *International Journal of Management Science and Information Technology* [Online], 4(2), pp.504–515. Available from: https://doi.org/10.35870/ijmsit.v4i2.3373.

Rahman, B. and Maryani, M., 2024. Optimizing Customer Satisfaction through Sentiment Analysis: A BERT-based Machine Learning Approach to Extract Insights. *IEEE Access* [Online], 2, pp. 151476-151489. Available from: https://doi.org/10.1109/access.2024.3478835.

Rand, W.M., 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* [Online], 66(336), pp.846–850. Available from: https://doi.org/10.2307/2284239.

Ray, A., Bala, P.K. and Jain, R., 2020. Utilizing emotion scores for improving classifier performance for predicting customer's intended ratings from social media posts. *Benchmarking: An International Journal* [Online], 28(2), pp.438–464. Available from: https://doi.org/10.1108/BIJ-01-2020-0004.

Raza, S., Garg, M., Reji, D.J., Bashir, S.R. and Ding, C., 2024. Nbias: A natural language processing framework for BIAS identification in text. *Expert Systems with Applications* [Online], 237, p.121542. Available from: https://doi.org/10.1016/j.eswa.2023.121542.

Reddy, P., Indrani, P., Janaki, P., Gayathri, P., Chandrahasini, P. and Apparao, G., 2024. Product Review Sentiment Analysis. *International Journal For Multidisciplinary Research* [Online], 6(3). Available from: https://doi.org/10.36948/ijfmr.2024.v06i03.20551.

Ren, G. and Hong, T., 2019. Examining the relationship between specific negative emotions and the perceived helpfulness of online reviews. *Information Processing and Management* [Online], 56(4), pp.1425–1438. Available from: https://doi.org/10.1016/J.IPM.2018.04.003.

Renko, S., Petljak, K. and Naletina, D., 2019.Food integrity throughout the chain: the case of good distribution practice. *Logforum* [Online], 15(1), pp.53–69. Available from: https://doi.org/10.17270/J.LOG.2019.318.

Rousseeuw, Jp.J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* [Online], 20, pp.53–65. Available from: https://doi.org/10.1016/0377-0427(87)90125-7.

Routaray, A.P. and Chitra, K., 2024. Sentiment Analysis and Rating Predicting for Hotel Review. *International Journal of Advanced Research in Science, Communication and Technology* [Online], 4(6), pp.49–51. Available from: https://doi.org/10.48175/ijarsct-22511.

Selçuk, B., Dursun, B.I. and Şerif, T., 2024.AI Assisted Customer Review Sentiment Analysis and Department Classification Tool. *2024 9th International Conference on Computer Science and Engineering (UBMK)* [Online], 26-28 October 2024, Antalya, Turkiye, IEEE, pp.365–370. Available from: https://doi.org/10.1109/ubmk63289.2024.10773394.

Sharma, N. and Verma, B., 2024. Recent Advances in Transfer Learning for Natural Language Processing (NLP). In *A Handbook of Computational Linguistics: Artificial Intelligence in Natural Language Processing* [Online]. BENTHAM SCIENCE PUBLISHERS, pp.228–254. Available from: https://doi.org/10.2174/9789815238488124020014.

Skotis, A. and Livas, C., 2024.Forms of Bias in Online Reviews and Their Implications for Management of Customer Knowledge: A Literature Review. In: Alper Erturk, et al., ed. *Convergence of Digitalization, Innovation, and Sustainable Development in Business* [Online]. IGI Global, pp.206–236. Available from: https://doi.org/10.4018/979-8-3693-0798-4.ch010.

Stephan, F.F., 1941. Stratification in Representative Sampling. *Journal of Marketing* [Online], 6(1), pp.38–46. Available from: https://doi.org/10.1177/002224294100600107.

Suits, D.B., 1957. Use of Dummy Variables in Regression Equations. *Journal of the American Statistical Association* [Online], 52(280), pp.548–551. Available from: https://doi.org/10.1080/01621459.1957.10501412.

Sun, Y., 2024. The evolution of transformer models from unidirectional to bidirectional in Natural Language Processing. *Applied and Computational Engineering* [Online], 42(1), pp.281–289. Available from: https://doi.org/10.54254/2755-2721/42/20230794.

Sykora, M., Elayan, S., Hodkinson, I., Jackson, T. and West, A., 2022. The power of emotions: Leveraging user generated content for customer experience management. *Journal of business research* [Online], 144, pp.997-1006. Available from: https://doi.org/10.1016/j.jbusres.2022.02.048.

Timoshenko, A. and Hauser, J.R., 2019. Identifying Customer Needs from User-Generated Content. *Marketing Science* [Online], 38(1), pp.1–20. Available from: https://doi.org/10.1287/mksc.2018.1123.

Ting, C.-K., 2022. Quest for the Balance of AI and Privacy [Editor's Remarks]. *IEEE Computational Intelligence Magazine* [Online], 17(3), p.2. Available from: https://doi.org/10.1109/mci.2022.3180649.

Tolulope, A.I. and Owoseni, T., 2024.Enhancing Customer Experience through E-commerce Review Analysis: Using Topic Modeling and Rule Induction for Understanding User Perception. *2024 International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG)* [Online], 2-4 April 2024, Omu-Aran, Nigeria, IEEE, pp.1–11. Available from: https://doi.org/10.1109/seb4sdg60871.2024.10630067.

Tribe, D., 2023. Presence of informal language, such as emoticons, hashtags, and slang, impact the performance of sentiment analysis models on social media text? *arXiv* [Online]. Available from: https://doi.org/10.48550/arxiv.2301.12303.

Vairamani, A.D. and Nayyar, A., 2024. Decoding product sentiments: Unraveling reviews with explainable analysis using Hugging-Face transformer. In: F. Al-Turjman, A. Nayyar, M. Naved, A.K. Singh and M. Bilal, eds. *XAI Based Intelligent Systems for Society 5.0* [Online]. Elsevier, pp.173–199. Available from: https://doi.org/10.1016/B978-0-323-95315-3.00003-6.

Vanhamme, J., 2003. *Surprise... Surprise..., An Empirical Investigation on How Surprise is Connected to Customer Satisfaction* [Online]. Rotterdam: Erasmus Research Institute of Management. Available from: https://repub.eur.nl/pub/273/erimrs20030211172951.pdf [Accessed 8 September 2025].

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. Advances in Neural Information Processing Systems *arXiv* [Online]. Available from: https://doi.org/10.48550/arXiv.1706.03762.

Venugopalan, M., Nalayini, G., Radhakrishnan, G. and Gupta, D., 2018. Rating prediction model for reviews using a novel weighted textual feature method. In: Sa, P., Bakshi, S., Hatzilygeroudis, I., Sahoo, M. eds. *Recent Findings in Intelligent Computing Techniques . Advances in Intelligent Systems and Computing*, Singapore, Springer [Online], 709, pp.177-190. Available from: https://doi.org/10.1007/978-981-10-8633-5_19.

Wang, F., Du, Z. and Wang, S., 2023. Information multidimensionality in online customer reviews. *Journal of Business Research* [Online], 159, p.113727. Available from: https://doi.org/10.1016/j.jbusres.2023.113727.

Wang, H. and Ester, M., 2014. A Sentiment-aligned Topic Model for Product Aspect Rating Prediction. *Empirical Methods in Natural Language Processing* [Online], pp.1192–1202. Available from: https://doi.org/10.3115/V1/D14-1126.

Wang, X., Guo, J., Wu, Y. and Liu, N., 2019. Emotion as signal of product quality. *Internet Research* [Online], 30(2), pp.463–485. Available from: https://doi.org/10.1108/INTR-09-2018-0415.

Weng, D., Yang, Y. and Zhao, J., 2022. Positive emotions help rank negative reviews for sellers and producers in e-commerce. *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)* [Online], 13-16 October 2022, Shenzhen, China, IEEE, pp.1–10. Available from: https://doi.org/10.1109/DSAA54385.2022.10032403.

Wood, Z., 2025. Amazon promises fake reviews crackdown after investigation by UK watchdog. *The Guardian* [Online]. Available from: https://www.theguardian.com/technology/2025/jun/06/amazon-promises-fake-reviews-crackdown-after-investigation-by-uk-watchdog [Accessed 10 August 2025].

Wu, D., Guo, X., Wang, Y. and Chen, G., 2023. A Warning Approach to Mitigating Bandwagon Bias in Online Ratings: Theoretical Analysis and Experimental Investigations.

Journal of the Association for Information Systems [Online], 24(4), pp.1132–1161. Available from: https://doi.org/10.17705/1jais.00817.

Wu, H., Guo, G., Yang, E., Luo, Y., Chu, Y., Jiang, L. and Wang, X., 2024. PESI: Personalized Explanation recommendation with Sentiment Inconsistency between ratings and reviews. *Knowledge-Based Systems* [Online], 283. Available from: https://doi.org/10.1016/j.knosys.2023.111133.

Yadav, K., 2021. A comprehensive survey on aspect-based sentiment analysis. *arXiv* [Online]. Available from: https:// doi.org/10.48550/arxiv.2006.04611.

Yang, Y., Duan, H., Abbasi, A., Lalor, J.P. and Tam, K.Y., 2023. Bias A-head? Analyzing Bias in Transformer-Based Language Model Attention Heads. *arXiv* [Online]. Available from: https://doi.org/10.48550/arxiv.2311.10395.

Yin, D., Bond, S.D. and Zhang, H., 2021. Anger in Consumer Reviews: Unhelpful but Persuasive? *Management Information Systems Quarterly* [Online], 45(3), pp.1059–1086. Available from: https://doi.org/10.25300/MISQ/2021/15363.

Yin, W., Hay, J. and Roth, D., 2019. Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach. *ArXiv* [Online], abs/1909.00161. Available from: https://api.semanticscholar.org/CorpusID:202540839 [Accessed 8 September 2025].

Zhang, Y., Tiwari, P., Rong, L., Chen, R., Alnajem, N.A. and Hossain, M.S., 2022.Affective Interaction: Attentive Representation Learning for Multi-Modal Sentiment Classification. *ACM Transactions on Multimedia Computing, Communications, and Applications* [Online], 18(3s), pp. 1-23. Association for Computing Machinery. Available from: https://doi.org/10.1145/3527175.

Zhang, S., Zhang, D., Zhong, H. and Wang, G., 2020. A Multiclassification Model of Sentiment for E-Commerce Reviews. *IEEE Access* [Online], 8, pp.189513–189526. Available from: https://doi.org/10.1109/ACCESS.2020.3031588.

Zhu, Y., Zhou, R., Chen, G. and Zhang, B., 2024. Enhancing sentiment analysis of online comments: a novel approach integrating topic modeling and deep learning. *PeerJ Computer Science* [Online], 10, p.e2542. Available from: https://doi.org/10.7717/peerj-cs.2542.

Appendices

Appendix 1. Summary Statistics

Table A1: Summary statistics of the dataset

	vars	n	mean	sd	min	25%	50%	75%	max	skew	kurtosis
Rating	1	2500	3	1.41	1	2	3	4	5	0	-1.3
Sentence Count	2	2500	4.86	4.08	1	3	4	6	70	6.02	65.17
Sentiment Score	3	2500	2.9	1.43	1	2	3	4	5	0.16	-1.31
Disgust	4	2500	0.21	0.29	0	0.01	0.06	0.30	0.99	1.46	0.76
Neutral	5	2500	0.35	0.3	0	0.07	0.28	0.62	0.96	0.49	-1.16
Surprise	6	2500	0.09	0.19	0	0.01	0.02	0.07	0.98	3.15	9.69
Sadness	7	2500	0.12	0.23	0	0.01	0.02	0.09	0.99	2.42	4.83
Anger	8	2500	0.03	0.08	0	0.01	0.01	0.02	0.92	6.23	46.4
Fear	9	2500	0.01	0.07	0	0.00	0.00	0.01	0.97	9.59	104.18
Joy	10	2500	0.18	0.28	0	0.00	0.02	0.22	0.99	1.62	1.2
Not as Expected	11	2500	0.8	0.28	0	0.70	0.95	0.99	1	-1.5	1.01
Damaged	12	2500	0.54	0.37	0	0.14	0.59	0.93	1	-0.15	-1.59
Returns	13	2500	0.41	0.26	0	0.19	0.40	0.61	1	0.28	-0.89
Taste and Flavour	14	2500	0.74	0.26	0	0.61	0.84	0.94	1	-1.18	0.46
Size and Quantity	15	2500	0.4	0.27	0	0.17	0.35	0.58	0.99	0.5	-0.76
Freshness	16	2500	0.44	0.25	0	0.24	0.46	0.62	1	-0.13	-0.81
Price	17	2500	0.59	0.22	0	0.46	0.60	0.75	1	-0.34	-0.22
Packaging	18	2500	0.36	0.31	0	0.09	0.29	0.58	1	0.61	-0.88
Quality	19	2500	0.54	0.36	0	0.15	0.58	0.91	1	-0.17	-1.54
Delivery	20	2500	0.41	0.24	0	0.22	0.40	0.57	0.99	0.32	-0.59
Ingredients	21	2500	0.41	0.2	0	0.27	0.42	0.55	1	-0.03	-0.46
Value for Money	22	2500	0.54	0.34	0	0.21	0.56	0.89	1	-0.12	-1.45
Expiration Date	23	2500	0.21	0.2	0	0.07	0.15	0.29	0.99	1.71	3.26
Easy to Use	24	2500	0.3	0.33	0	0.01	0.13	0.57	1	0.77	-0.9

Appendix 2. Non-Standardised Regression Model Summary

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Rating   R-squared:                       0.717
Model:                            OLS   Adj. R-squared:                  0.714
Method:                 Least Squares   F-statistic:                     359.8
Date:                Sun, 31 Aug 2025   Prob (F-statistic):               0.00
Time:                        18:15:08   Log-Likelihood:                -2275.6
No. Observations:                2000   AIC:                             4597.
Df Residuals:                    1977   BIC:                             4726.
Df Model:                          22
Covariance Type:                  HC3
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                 1.7651      0.157     11.264      0.000       1.458       2.072
Sentence Count        0.0146      0.006      2.379      0.017       0.003       0.027
Sentiment Score       0.4705      0.031     14.950      0.000       0.409       0.532
Disgust              -0.4087      0.087     -4.675      0.000      -0.580      -0.237
Surprise              0.0754      0.108      0.697      0.486      -0.137       0.288
Sadness              -0.0124      0.106     -0.117      0.907      -0.221       0.196
Anger                 0.0943      0.246      0.383      0.701      -0.388       0.576
Fear                 -0.6388      0.301     -2.120      0.034      -1.230      -0.048
Joy                   0.1690      0.078      2.176      0.030       0.017       0.321
Not as Expected      -0.2631      0.082     -3.193      0.001      -0.425      -0.102
Damaged              -0.4282      0.098     -4.389      0.000      -0.620      -0.237
Returns               0.0212      0.098      0.215      0.829      -0.172       0.214
Taste and Flavour     0.1162      0.083      1.399      0.162      -0.047       0.279
Size and Quantity    -0.0293      0.084     -0.350      0.726      -0.193       0.135
Freshness             0.4188      0.089      4.688      0.000       0.244       0.594
Price                 0.0719      0.100      0.717      0.474      -0.125       0.269
Packaging            -0.1021      0.070     -1.458      0.145      -0.239       0.035
Quality               0.2951      0.095      3.120      0.002       0.110       0.481
Delivery             -0.0395      0.101     -0.392      0.695      -0.237       0.158
Ingredients          -0.4868      0.108     -4.507      0.000      -0.699      -0.275
Value for Money       0.2557      0.088      2.912      0.004       0.084       0.428
Expiration Date      -0.4681      0.124     -3.775      0.000      -0.711      -0.225
Easy to Use           0.1562      0.085      1.838      0.066      -0.010       0.323
==============================================================================
Omnibus:                       32.653   Durbin-Watson:                   1.930
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               61.181
Skew:                          -0.031   Prob(JB):                     5.18e-14
Kurtosis:                       3.855   Cond. No.                         105.
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)
```

Figure A1: Regression model summary with non-standardised coefficients

Appendix 3. Regression Model: Residuals Diagnostics

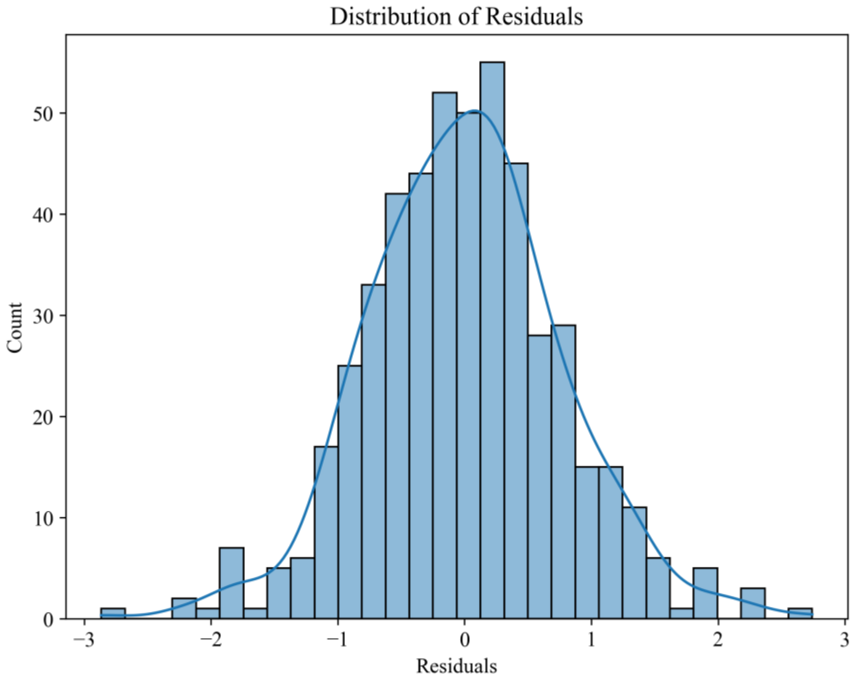


Figure A2: Distribution of regression model residuals

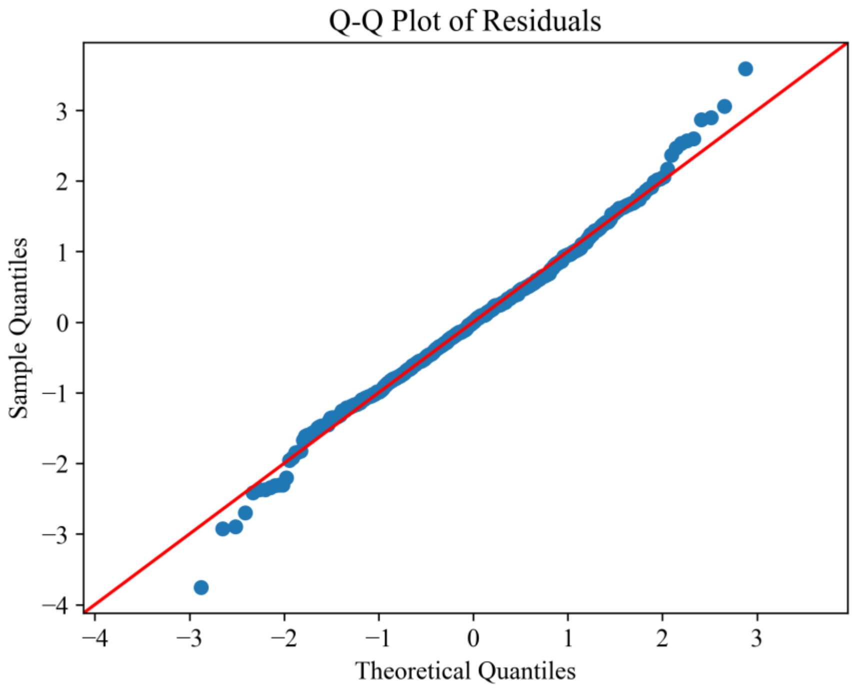


Figure A3: Q-Q plot of regression model residuals

Appendix 4. Kruskal–Wallis H test results

Table A2: Kruskal–Wallis H test results

	H Statistic	**p-value**	**FDR-corrected p-value**
Joy	1048.13	0.000	0.000
Disgust	975.54	0.000	0.000
Neutral	911.53	0.000	0.000
Sentiment Score	775.81	0.000	0.000
Damaged	766.10	0.000	0.000
Sadness	682.09	0.000	0.000
Surprise	674.30	0.000	0.000
Quality	538.11	0.000	0.000
Value for Money	507.00	0.000	0.000
Fear	468.50	0.000	0.000
Not as Expected	460.75	0.000	0.000
Easy to Use	455.00	0.000	0.000
Expiration Date	240.00	0.000	0.000
Returns	226.35	0.000	0.000
Anger	156.75	0.000	0.000
Delivery	79.97	0.000	0.000
Packaging	61.85	0.000	0.000
Freshness	43.74	0.000	0.000
Size and Quantity	33.93	0.000	0.000
Price	32.04	0.000	0.000
Sentence Count	28.68	0.000	0.000
Taste and Flavour	22.93	0.000	0.000
Ingredients	4.84	0.184	0.184

Appendix 5. Correlation Matrices per Cluster

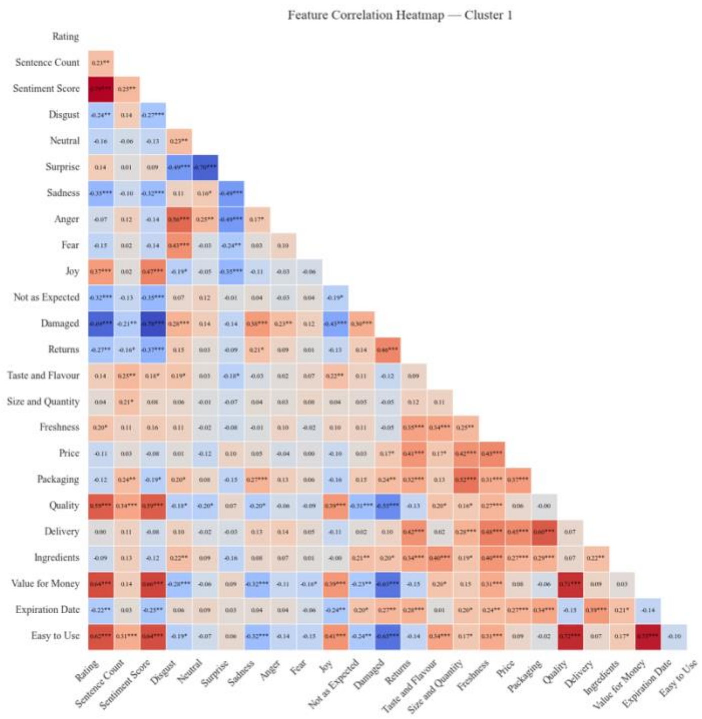


Figure A4: Correlation matrix of features in Cluster 1

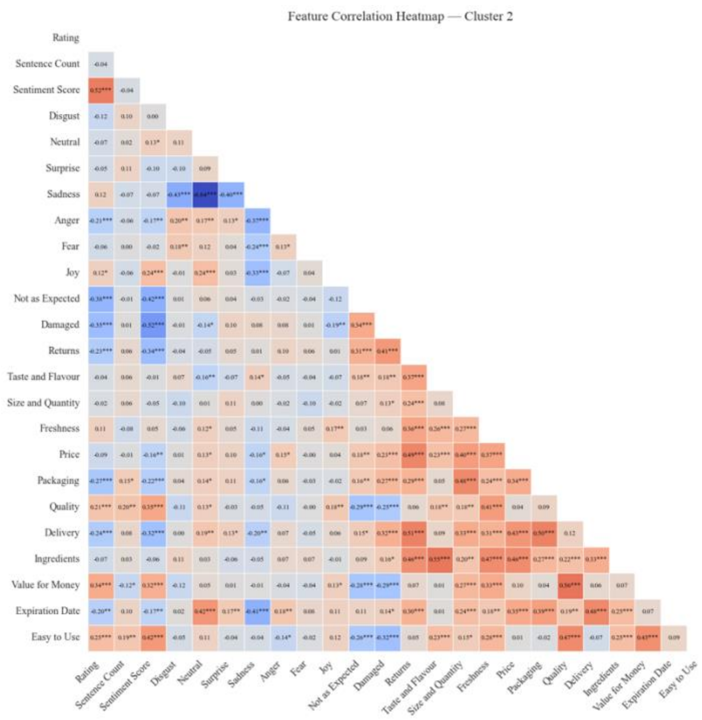


Figure A5: Correlation matrix of features in Cluster 2

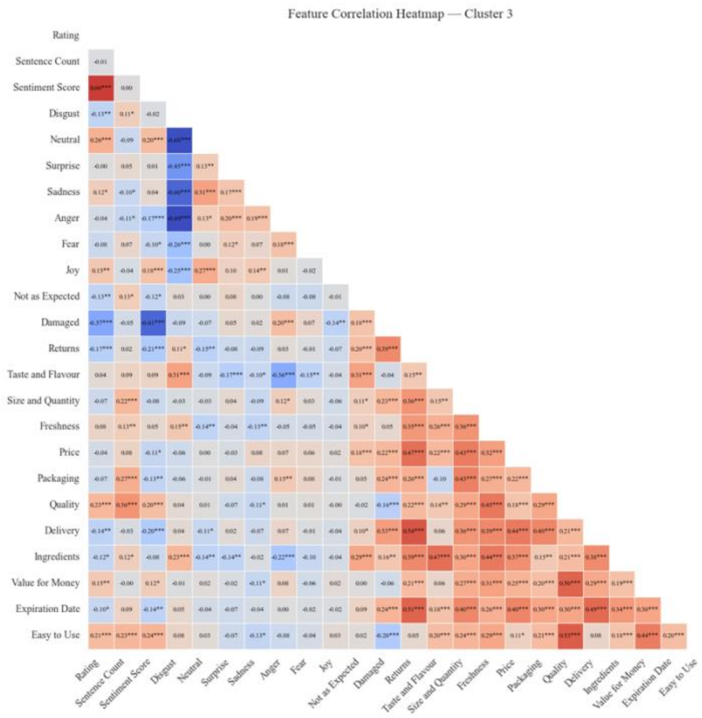


Figure A6: Correlation matrix of features in Cluster 3

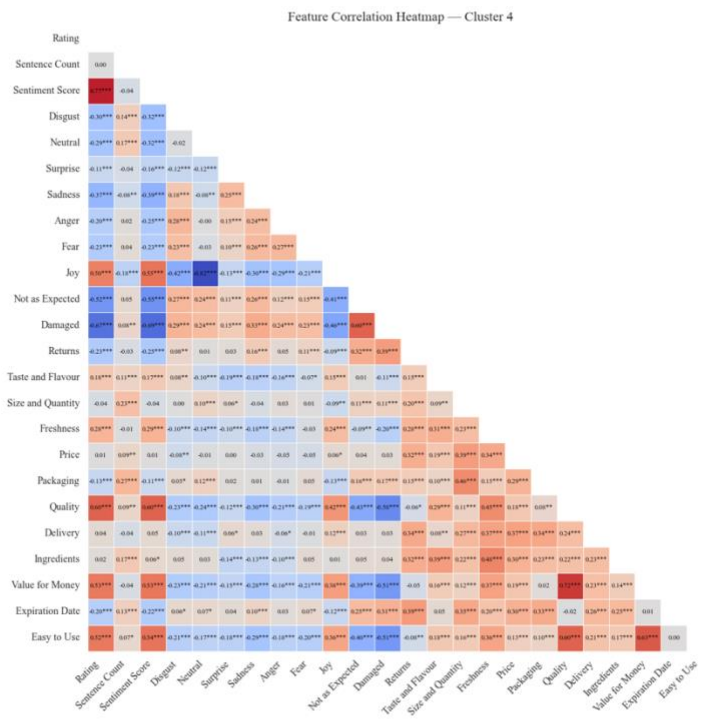


Figure A7: Correlation matrix of features in Cluster 4

Appendix 6. Summary Statistics per Cluster

Table A3: Summary statistics of Cluster 1

	count	mean	std	min	25%	50%	75%	max
Rating	147	2.96	1.37	1	2	3	4	5
Sentence Count	147	5.65	3.49	1	3	5	8	17
Sentiment Score	147	2.80	1.40	1	2	2	4	5
Disgust	147	0.03	0.05	0.00	0.00	0.01	0.03	0.28
Neutral	147	0.11	0.10	0.00	0.03	0.08	0.15	0.48
Surprise	147	0.75	0.17	0.34	0.60	0.80	0.90	0.98
Sadness	147	0.04	0.07	0.00	0.00	0.01	0.05	0.35
Anger	147	0.02	0.02	0.00	0.00	0.01	0.02	0.11
Fear	147	0.01	0.03	0.00	0.00	0.00	0.01	0.17
Joy	147	0.05	0.09	0.00	0.00	0.01	0.04	0.44
Not as Expected	147	0.92	0.19	0.01	0.95	0.99	0.99	1.00
Damaged	147	0.58	0.34	0.00	0.23	0.64	0.90	1.00
Returns	147	0.43	0.26	0.00	0.22	0.44	0.60	0.98
Taste and Flavour	147	0.71	0.25	0.01	0.54	0.80	0.93	0.99
Size and Quantity	147	0.47	0.27	0.02	0.24	0.44	0.63	0.99
Freshness	147	0.42	0.23	0.00	0.23	0.45	0.59	0.98
Price	147	0.64	0.22	0.00	0.51	0.66	0.79	0.99
Packaging	147	0.43	0.29	0.00	0.17	0.39	0.66	1.00
Quality	147	0.50	0.36	0.00	0.15	0.53	0.86	1.00
Delivery	147	0.46	0.25	0.00	0.27	0.47	0.63	0.98
Ingredients	147	0.42	0.19	0.00	0.28	0.44	0.56	0.96
Value for Money	147	0.49	0.34	0.00	0.18	0.47	0.86	1.00
Expiration Date	147	0.24	0.18	0.00	0.11	0.22	0.32	0.99
Easy to Use	147	0.28	0.33	0.00	0.01	0.08	0.55	0.99

Table A4: Summary statistics of Cluster 2

	count	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
Rating	255	2.08	1.02	1	1	2	3	5
Sentence Count	255	4.11	2.34	1	3	4	5	14
Sentiment Score	255	1.86	0.82	1	1	2	2	5
Disgust	255	0.06	0.08	0.00	0.01	0.03	0.09	0.38
Neutral	255	0.12	0.15	0.00	0.03	0.06	0.14	0.88
Surprise	255	0.07	0.07	0.00	0.02	0.04	0.09	0.46
Sadness	255	0.70	0.21	0.01	0.59	0.75	0.86	0.99
Anger	255	0.02	0.03	0.00	0.01	0.01	0.02	0.23
Fear	255	0.01	0.01	0.00	0.00	0.00	0.01	0.11
Joy	255	0.02	0.03	0.00	0.00	0.01	0.01	0.34
Not as Expected	255	0.96	0.09	0.34	0.97	0.99	1.00	1.00
Damaged	255	0.83	0.24	0.01	0.80	0.95	0.99	1.00
Returns	255	0.58	0.26	0.00	0.39	0.63	0.80	0.99
Taste and Flavour	255	0.70	0.27	0.01	0.55	0.79	0.92	0.99
Size and Quantity	255	0.43	0.25	0.00	0.23	0.40	0.60	0.99
Freshness	255	0.41	0.24	0.00	0.21	0.46	0.61	0.99
Price	255	0.65	0.20	0.01	0.51	0.67	0.79	1.00
Packaging	255	0.42	0.29	0.00	0.19	0.37	0.63	0.99
Quality	255	0.32	0.30	0.00	0.04	0.22	0.58	0.99
Delivery	255	0.50	0.25	0.01	0.33	0.50	0.69	0.99
Ingredients	255	0.41	0.20	0.00	0.27	0.44	0.56	0.92
Value for Money	255	0.35	0.28	0.00	0.10	0.29	0.55	0.99
Expiration Date	255	0.36	0.26	0.01	0.16	0.30	0.51	0.99
Easy to Use	255	0.11	0.18	0.00	0.00	0.02	0.15	0.95

Table A5: Summary statistics of Cluster 3

	count	mean	std	min	25%	50%	75%	max
Rating	411	1.77	0.89	1	1	2	2	5
Sentence Count	411	5.03	3.14	1	3	4	6	27
Sentiment Score	411	1.67	0.82	1	1	1	2	5
Disgust	411	0.76	0.21	0.03	0.65	0.83	0.93	0.99
Neutral	411	0.09	0.10	0.00	0.02	0.05	0.14	0.46
Surprise	411	0.02	0.05	0.00	0.00	0.01	0.02	0.48
Sadness	411	0.05	0.07	0.00	0.01	0.02	0.07	0.40
Anger	411	0.06	0.11	0.00	0.01	0.02	0.05	0.60
Fear	411	0.01	0.03	0.00	0.00	0.01	0.01	0.25
Joy	411	0.00	0.01	0.00	0.00	0.00	0.00	0.13
Not as Expected	411	0.93	0.10	0.27	0.92	0.97	0.99	1.00
Damaged	411	0.85	0.19	0.02	0.79	0.92	0.98	1.00
Returns	411	0.49	0.25	0.00	0.28	0.50	0.67	1.00
Taste and Flavour	411	0.76	0.25	0.01	0.64	0.88	0.94	1.00
Size and Quantity	411	0.34	0.23	0.00	0.15	0.31	0.49	0.97
Freshness	411	0.38	0.24	0.00	0.16	0.41	0.57	0.98
Price	411	0.57	0.20	0.00	0.45	0.57	0.71	1.00
Packaging	411	0.38	0.31	0.00	0.11	0.33	0.61	1.00
Quality	411	0.27	0.27	0.00	0.04	0.17	0.45	0.98
Delivery	411	0.42	0.22	0.00	0.26	0.42	0.57	0.99
Ingredients	411	0.43	0.20	0.01	0.29	0.45	0.57	0.98
Value for Money	411	0.29	0.26	0.00	0.07	0.23	0.46	0.99
Expiration Date	411	0.25	0.19	0.00	0.10	0.21	0.35	0.95
Easy to Use	411	0.10	0.17	0.00	0.00	0.01	0.12	0.96

Table A6: Summary statistics of Cluster 4

	count	mean	std	min	25%	50%	75%	max
Rating	1290	3.67	1.24	1	3	4	5	5
Sentence Count	1290	4.52	2.71	1	3	4	6	23
Sentiment Score	1290	3.62	1.25	1	3	4	5	5
Disgust	1290	0.09	0.13	0.00	0.01	0.03	0.11	0.96
Neutral	1290	0.51	0.29	0.00	0.26	0.55	0.76	0.96
Surprise	1290	0.05	0.07	0.00	0.01	0.02	0.05	0.63
Sadness	1290	0.04	0.07	0.00	0.01	0.01	0.03	0.54
Anger	1290	0.02	0.03	0.00	0.01	0.01	0.02	0.34
Fear	1290	0.01	0.01	0.00	0.00	0.00	0.00	0.18
Joy	1290	0.30	0.33	0.00	0.02	0.13	0.59	0.99
Not as Expected	1290	0.70	0.32	0.00	0.47	0.84	0.97	1.00
Damaged	1290	0.35	0.33	0.00	0.06	0.24	0.62	1.00
Returns	1290	0.34	0.23	0.00	0.16	0.32	0.50	0.98
Taste and Flavour	1290	0.77	0.23	0.00	0.66	0.86	0.95	1.00
Size and Quantity	1290	0.39	0.27	0.00	0.16	0.33	0.57	0.99
Freshness	1290	0.47	0.24	0.00	0.30	0.49	0.64	1.00
Price	1290	0.59	0.22	0.00	0.45	0.59	0.75	0.99
Packaging	1290	0.31	0.29	0.00	0.07	0.22	0.47	1.00
Quality	1290	0.69	0.32	0.00	0.46	0.82	0.97	1.00
Delivery	1290	0.37	0.22	0.00	0.20	0.36	0.52	0.98
Ingredients	1290	0.41	0.19	0.00	0.28	0.42	0.54	1.00
Value for Money	1290	0.68	0.31	0.00	0.46	0.79	0.96	1.00
Expiration Date	1290	0.16	0.15	0.00	0.05	0.12	0.22	0.99
Easy to Use	1290	0.42	0.34	0.00	0.06	0.40	0.75	1.00