# SHEIN Women Clothes Review Analysis

Yahan Yang

2024-12-10

## 1. Introduction

In today's digital age, e-commerce platforms have become essential for shopping, offering consumers the convenience of browsing and purchasing a vast range of products online. This project examines a dataset of women's clothing reviews from SHEIN, encompassing attributes like review content, rating, age, product title, and department classification for different categories of clothes. The primary research question driving this analysis is: *How do factors such as product category, and rating impact the sentiment and language of customer reviews? Do they have interaction?* These questions are quite compelling because it looks into consumer behavior and sentiment—a crucial area for online retail as it seeks to personalize and enhance the shopping experience. By analyzing these textual reviews, this study aims to uncover insights into customer preferences and expectations, providing valuable information that can guide product recommendations and improve customer satisfaction. Understanding these dynamics is essential in a competitive market where consumer insights can make a substantial difference in product positioning and customer loyalty.

## 2. Data

The dataset for this project consists of customer reviews for women's clothing on an e-commerce platform called SHEIN. It is downloaded from *HuggingFace* It includes 23,486 entries with 12 columns, providing various attributes about each review and product. After filtering all the columns where reviews are not blank, there are a total of 22641 observations left. Here's an overview of the dataset columns:

1. **Clothing ID**: Unique identifier for each clothing item.
2. **Age**: Age of the reviewer.
3. **Title**: Title of the review, giving a brief description or opinion.
4. **Review Text**: Detailed text of the customer review.
5. **Rating**: Rating given by the customer, ranging from 1 to 5.
6. **Recommended IND**: Indicator if the reviewer recommends the product (1 for yes, 0 for no).
7. **Positive Feedback Count**: Count of positive feedback received for the review.
8. **Division Name**: High-level division in which the item is categorized (e.g., General, Initmates).
9. **Department Name**: More specific department name under the division (e.g., Dresses, Tops).
10. **Class Name**: Product category within the department (e.g., Blouses, Dresses, Pants).

In this experiment, the *Review Text* column was used for text tokenization and further analysis, along with *Department Name*, as we want to look into a basic trend of the data. Other variables such as *age*, *Rating* are also of interest.

Table 1: Statistics Summary of Departments

| | binary_rating | Total Words | Review Count | Avg Words per Review |
|---|---|---|---|---|
| 0 | high | 1033159 | 17448 | 59.213606 |
| 1 | low | 324327 | 5193 | 62.454650 |

# 3. Methods

## 3.1 Choice of Methods and Reasons

This project uses frequency analysis, sentiment analysis and Principal Component Analysis for clustering. Before implementing any analysis, we feature engineer a new column called "binary rating", which takes column "rating" from 1 to 3 as "low", 4 to 5 as "high". As the previous coffee break experiment shows us there is no significant difference between each category, we instead focus on the binary rating.

**Frequency Analysis**: Frequency analysis was chosen to identify the most commonly used words in customer reviews for each clothing category. This is very useful for exploring lexical differences and thematic focus across product categories. The method provides a straightforward way to understand the key topics or features that customers frequently mention, such as "fit," "size," or "style." By comparing the top words across ratings, the project hopes to understand some interesting features (e.g., positive adjectives for high ratings and negative words for low ratings) and category-specific language features (e.g., "dress" and "look" in dresses, or "jacket" and "warm" in jackets).

**Sentiment Analysis**: Sentiment analysis was used to explore how customers feel about different categories of clothing. This is important for understanding the nuances of language use, because words with similar frequencies may carry vastly different sentiments depending on the context. This method helps discover the emotional tone of reviews, such as positive sentiments about quality or negative sentiments about poor fit. By combining sentiment analysis with frequency analysis, we can assess not only what customers are talking about but also how they feel about ratings hopefully. For example, we can understand better about the tone for high and low ratings.

**PCA Cluster Analysis**: Clustering allows for an unsupervised exploration of patterns in the text to discover natural groupings that may not align strictly with predefined categories (Gries and Newman (2011)). For instance, reviews about dresses and tops may share similar language around "style," forming a cluster. Clustering helps to identify underlying themes in customer feedback and uncover shared or unique linguistic patterns that could inform product-specific marketing strategies or improvements. Here, we adopt the method to investigate the difference between clothes categories.

# 4. Results

## 4.1 Frequency Analysis

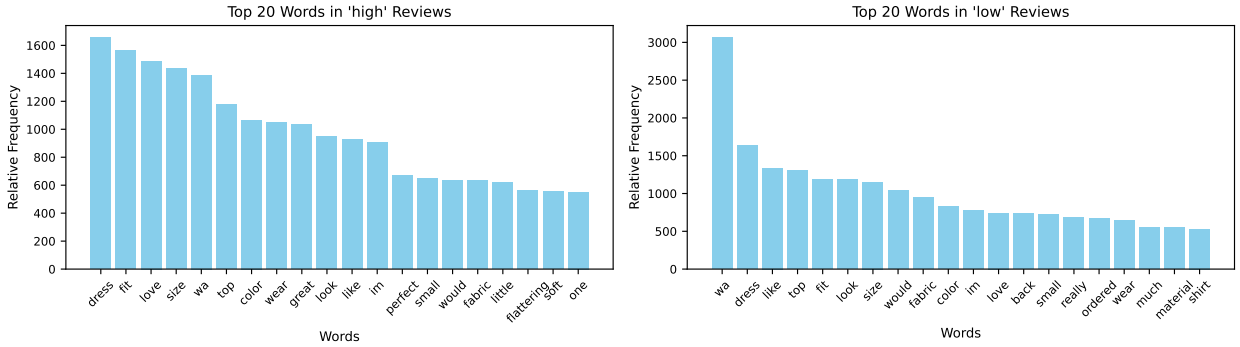


Figure 1: Token Frequency Display of Binary Rating

Looking at the top 20 tokens of both "high" and "low" ratings, we spot the need to understand the word appearing frequency. Hence, a wordcloud graph is used to visualize the words in both rating categories.


Figure 2: Word Cloud of Binary Rating

## 4.2 Interaction between binary rating and clothes department

After categorizing the ratings into high and low, we are interested in how they are distributed in each clothes department. The following statistics summary shows that within each category, high ratings take up around 80% for bottoms, dresses, intimate, jackets and tops. As for trend, the high rating percentage is lower than other 5 categories (65%). Using Chi-squared test, we calculate the p value to be less than 0.001, which suggests the distribution of binary ratings differs significantly between each clothes category.

Table 2: Binary Rating Varies Across Departments

| binary_rating<br>Department Name | high | low | low_ratio | high_ratio |
|---|---|---|---|---|
| Bottoms | 3058 | 741 | 0.20 | 0.80 |
| Dresses | 4792 | 1527 | 0.24 | 0.76 |
| Intimate | 1404 | 331 | 0.19 | 0.81 |
| Jackets | 832 | 200 | 0.19 | 0.81 |
| Tops | 8030 | 2438 | 0.23 | 0.77 |
| Trend | 78 | 41 | 0.34 | 0.65 |

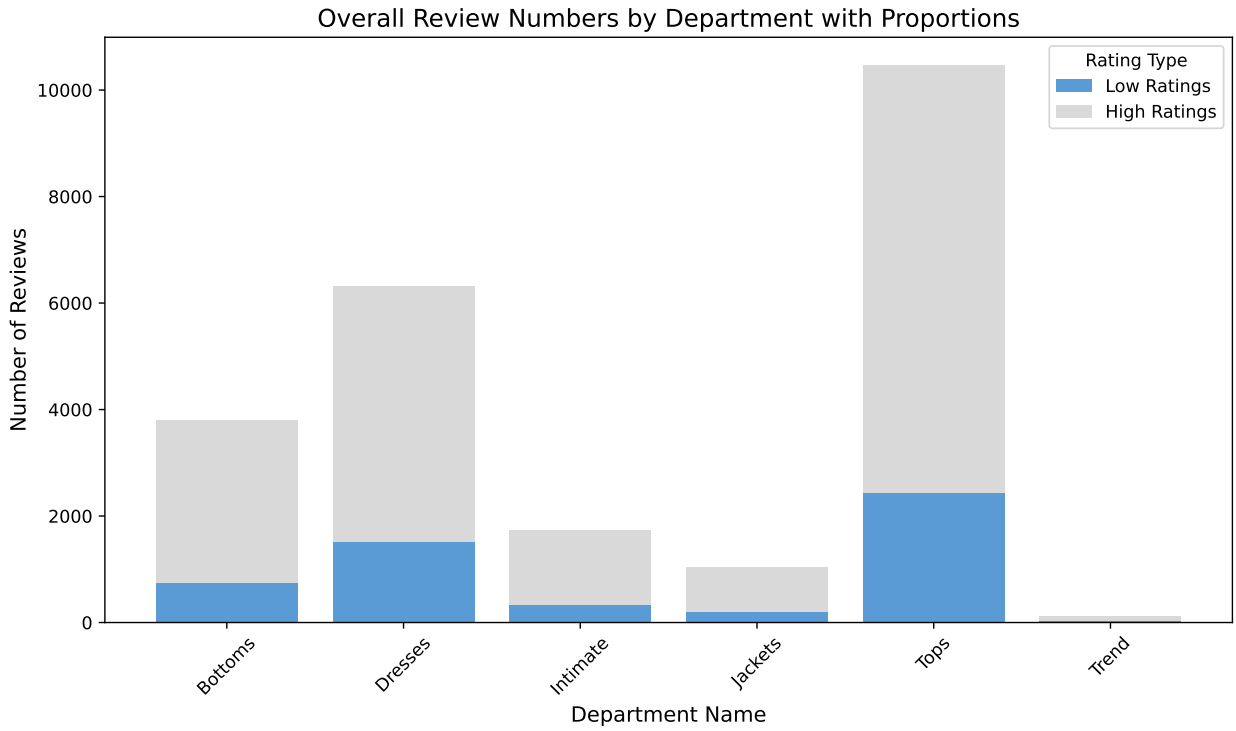


Figure 3: Binary Rating within Each Department

### 4.3 Sentiment Analysis

Proceeding to sentiment analysis, we found that two binary ratings differ in their sentiment scores: High rating has higher positive score while low rating has higher neutral score.

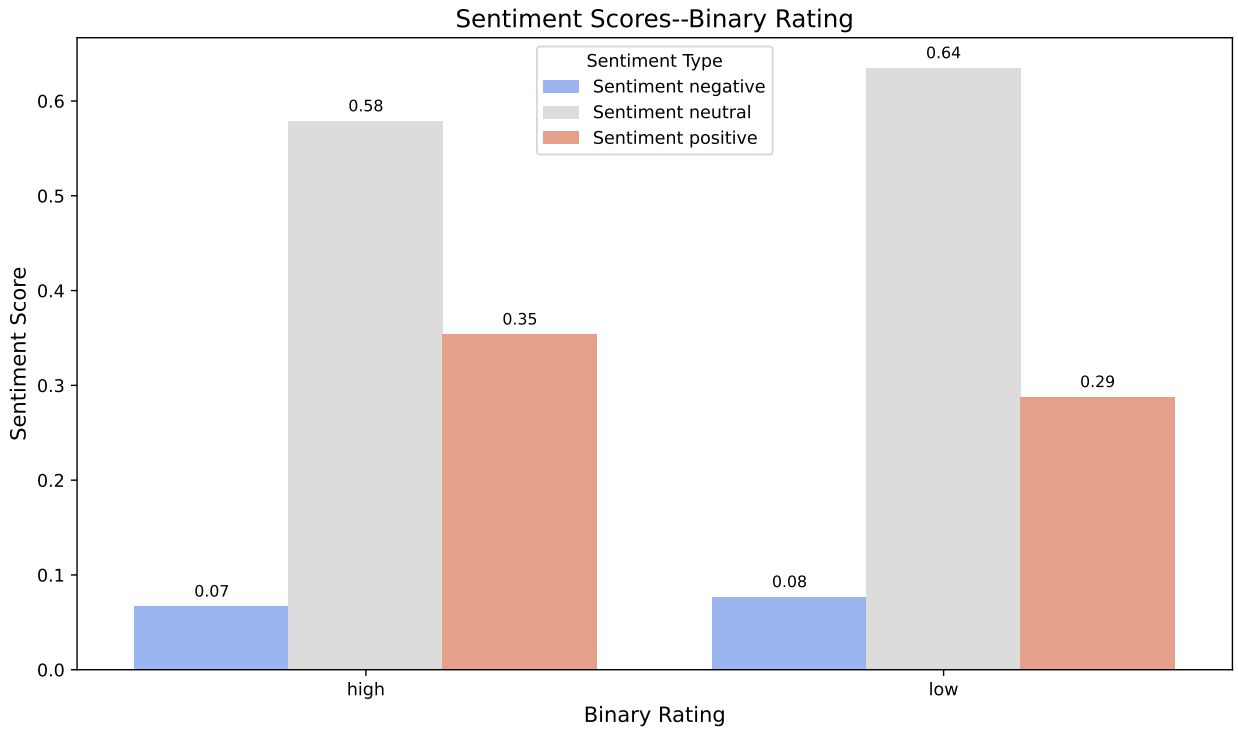


Figure 4: Sentiment Score by Binary Rating

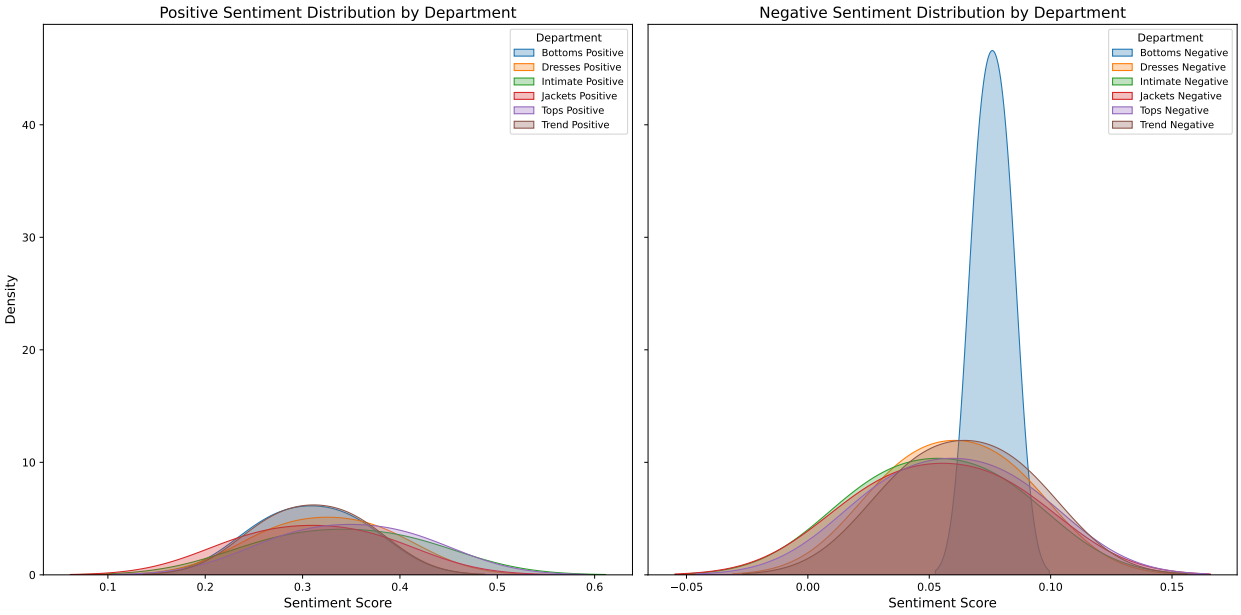


Figure 5: Sentiment Distribution

## 4.4 Principal Component Analysis

With the numerical variables in the given dataset, we use PCA analysis to cluster the departments. Here, we consider all variables inluding age, rating, recommend IND (binary), and positive feedback count in an attempt to give a more extensive analysis of the data set.

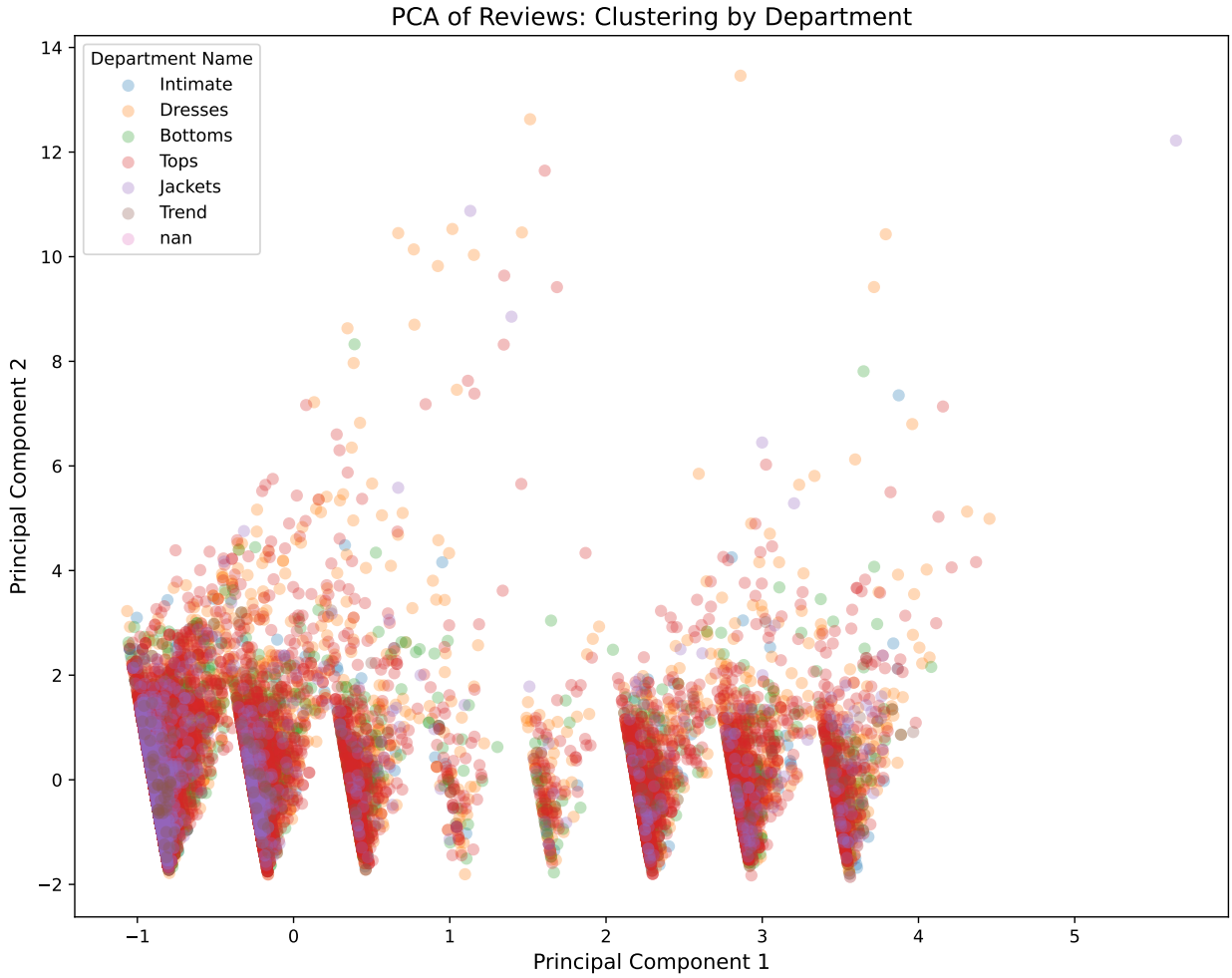


Figure 6: PCA Cluster Analysis

# 5. Discussion

## 5.1 Conclusion

### 5.1.1 Frequency Analysis and Interaction

The frequency analysis graph of words across two ratings shows the key terms customers frequently use in their reviews. For example, words like "dress" and "top" appear the most frequently in both high and low ratings, reflecting their importance in clothing reviews. However, there are also unique terms within each category. For instance, for high rating, the top words would mention extreme adjectives such as "love" "great""perfect", while low rating places more importance in descriptive words. We also observe the "love" word appearing, but this should not be concerning as we categorize rating = 3 as low, when in reality it might representive a neutral perspective.

However, one thing to notice is that many of the describing words are overlapping despite the preassumptions that they could differ in terms of rating. Even though words with strong emotions appear at the same time, their frequency are very different. This suggests that low rating does not necessarily lead to bad words, deviating from what we anticipated.

We were hoping to distinguish the interaction between clothes departments and rating classification, but Figure 2 shows the distribution is generally the same across all categories except for trend. Intuitively, trend

is a relatively fashionable type of clothing, which may lead to a more polar review collection.

#### 5.1.2 Sentiment Analysis Result

Sentiment analysis provides valuable insights into the emotional tone of customer reviews for different clothing categories. By examining the distribution of positive and negative sentiment, we can better understand customers' perceptions of specific products. This dual-plot visualization separates positive and negative sentiment distributions for the purpose of a clearer comparison of customer feedback across departments.

The left plot consists of positive sentiment distributions, which show significant variation across departments but generally peak within the mid-range of sentiment scores. Categories like "Tops" and "Dresses" have broader peaks in the positive sentiment distribution, indicating higher overall satisfaction and a diverse range of positive experiences among reviewers. On the contrary, the "Trend" category shows a relatively lower density for positive sentiment, meaning that it might not meet customer expectations as effectively as other departments.

The right plot focuses on negative sentiment distributions, which reveal a generally low density across all departments. However, the "Trend" category exhibits a slightly higher peak in negative sentiment scores compared to others, highlighting a potential area for improvement. This distribution aligns with the idea that customers in this category might be more vocal about dissatisfaction. Departments like "Intimate" and "Jackets" show minimal negative sentiment, which also reflects consistency in customer satisfaction.

Initially, the goal was to uncover significant disparities in sentiment between departments, with the expectation that certain types of products might attract more critical feedback. However, the data suggests that reviews tend to lean towards a positive or neutral tone overall, possibly due to the tendency of reviewers to share feedback primarily when they are satisfied with a product. For future research, a deeper dive into subcategories within each department might uncover more granular trends. Besides, we can also considering the role of specific product attributes, such as fit or material quality, as they could also provide actionable insights for enhancing customer satisfaction across all departments.

#### 5.1.3 PCA Analysis Result

The PCA visualization reveals how customer reviews are distributed across departments after reducing the data to two principal components. The dense clustering of points suggests that numerical features like Age, Rating, and Positive Feedback Count do not carry enough variance to separate departments meaningfully. As most of the variables are categorical, it makes sense to observe such similar clustering around integers Interestingly, some departments like "Intimate" and "Trend" appear more isolated, potentially indicating distinct feedback patterns. However, more features could be included for better separation. For example, incorporating text-based attributes, such as sentiment analysis scores or word embeddings derived from Review Text, could enhance the ability to distinguish departments. The vertical spread of some clusters further suggests the presence of noise or uninformative features that might dilute the effectiveness of PCA.

Moreover, the explained variance of the first two principal components should be analyzed to determine how much information they capture. If these components fail to represent a significant portion of the variance, additional dimensions may be required to fully understand the data.

### 5.2 Improvements and Next step

The binary rating of manual operation does not seem to add any features that are worth deep diving. The sentiment analysis aligned with what we expected, but other analyses do not reveal prominent trends. Further steps can be taken to check for the keyness potentially between each department and ratings, or to dive deep into the interactions.

## 6. Work Cited

(1) Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

(2) Biber, D. (1988). Variation across Speech and Writing. Cambridge: Cambridge University Press.

Gries, Stefan Th., and John Newman. 2011. "N -Grams and the Clustering of Registers." In. https://api.semanticscholar.org/CorpusID:163160063.