Probabilistic Topic Modeling for E-commerce Reviews using LDA

Samudyata Sudarshan Jagirdar
Arizona State University
ASU ID: 1233568442
sjagird1@asu.edu

Ramanuja Magadi Anantha Krishna
Arizona State University
ASU ID: 1233054006
rkrish79@asu.edu

Shalin Nilesh Bhavsar
Arizona State University
ASU ID: 1233281896
sbhavsa8@asu.edu

Nikhil Srikant Kulkarni
Arizona State University
ASU ID: 1235817442
nkulka37@asu.edu

Abstract

The project uses Latent Dirichlet Allocation (LDA) to analyze women's clothing reviews, extracting key topics like quality, fit, and style to help businesses understand consumer preferences and improve products.

1 Introduction

In modern e-commerce, user-generated content (UGC) such as customer reviews plays an increasingly important role in shaping consumer perceptions and business decisions. These reviews – often generated in large volumes for segments like women's fashion retail – constitute free-text feedback that is inherently unstructured. The abundance of such unstructured opinion data presents a challenge: manual analysis of thousands of comments is infeasible. To address this, researchers have turned to natural language processing (NLP) and machine learning (ML) techniques to automatically organize and interpret review content. In particular, unsupervised methods like topic modeling (e.g., Latent Dirichlet Allocation) combined with clustering can group reviews into coherent themes, summarizing each group by its most salient keywords. For example, previous studies have applied LDA topic modeling and K-means clustering to consumer reviews, yielding clusters whose top keywords reveal the main areas of customer feedback. Others have analyzed large datasets of women's clothing e-commerce reviews using these techniques, uncovering key topics and concerns in the fashion domain. By combining topic modeling and clustering, these approaches convert large, unstructured review corpora into structured thematic insights, thereby enabling retailers to identify prevalent issues and guide product improvements.

1.1 Motivation and Importance

Online reviews are a cornerstone of e-commerce, shaping consumer trust and business strategies. For women's clothing, where fit, quality, and style are critical, reviews provide a wealth of unstructured feedback that is challenging to analyze manually. Extracting meaningful patterns from this data can help retailers address customer pain points, improve product offerings, and tailor marketing campaigns. Our project tackles this challenge by applying advanced natural language processing (NLP) to uncover hidden themes and segment customers, offering a scalable solution for retail analytics.

The goal of this project is to demonstrate how unsupervised NLP techniques can be applied to large-scale retail feedback to uncover structure and actionable knowledge. We specifically examine a public dataset of women's fashion product reviews and apply Latent Dirichlet Allocation (LDA) topic modeling and K-means clustering. The key contributions include: (1) converting thousands of unstructured text reviews into a set of interpretable topics (e.g., themes like fit or quality) without

manual labeling, (2) segmenting the review data (or customers) into clusters based on those topics or embedding patterns, thereby revealing distinct groups of feedback, and (3) evaluating the coherence and practical relevance of these discovered topics and segments. Our work illustrates a pipeline that can help retailers automatically "listen" to customer voices at scale – bridging the gap between raw text and structured customer insights.

1.2 Project Description

The goal is to use K-means clustering to group customers based on feedback patterns and Latent Dirichlet Allocation (LDA) to identify latent topics in the Women's Clothing E-Commerce Reviews dataset (`https://www.kaggle.com/datasets/nicapotato/womens-ecommerce-clothing-reviews/data`). Specific objectives include preprocessing review text to ensure consistency and quality; extracting topics (e.g., fit, quality) to understand customer priorities; clustering reviews to identify distinct customer segments; visualizing results to provide actionable insights for e-commerce businesses. This approach bridges NLP and machine learning to transform raw feedback into strategic recommendations.

2 Methodology

2.1 Dataset and Text Preprocessing

We conducted our experiments on the Women's Clothing E-Commerce Reviews dataset, which contains 23,486 customer reviews from a women's apparel retailer. Each review includes a free-text comment and structured fields such as a review title, a star rating (1–5), a binary recommendation flag, the reviewer's age, and product category labels (division, department, and class). All personal or brand identifiers were removed to ensure anonymity (company references were replaced with the token "retailer"). This dataset's combination of unstructured text and categorical product data makes it ideal for multifaceted analysis. In our work, however, we focus on the review text for NLP modeling and use the other fields (like star ratings) only to help interpret the results. Before modeling, we applied a standard text preprocessing pipeline to clean and normalize the review texts. We tokenized each review into words and converted all text to lowercase. Punctuation, numeric digits, and stop words (common function words like "and" or "the") were removed, and any very short tokens (1–2 characters) were dropped to reduce noise. Next, we applied lemmatization (using WordNet) to reduce words to their base forms (for example, "dresses" becomes "dress" and "running" becomes "run"), ensuring that different inflections of the same word are treated uniformly. After preprocessing, the review corpus consists of tokenized, lowercased, lemmatized texts containing only content-bearing terms. We built a vocabulary from this cleaned corpus and represented the data as a document-term matrix for topic modeling. We also removed 845 reviews ($\tilde{3}$.6 percent of the data) that had no review text, leaving 22,641 reviews for the final analysis. These preprocessing steps ensure that the analysis is not skewed by irrelevant artifacts or noise, and that the core content of each review is preserved in a consistent, machine-readable form.

2.2 Feature Engineering

We engineered features to enhance modeling:

- **Text-based**: Word counts (`rev_word_count`), unique word counts (`unique_word_count`), and sentiment scores computed with VADER, which quantifies positive, neutral, and negative tones (e.g., "Love this dress!" scores high positivity).
- **Categorical**: Binary encoding of clothing categories (e.g., 1 for dresses, 0 otherwise) to reduce dimensionality.
- **Numerical**: Normalized ratings (1–5 scaled to 0–1) and age for clustering compatibility.

Rare clothing IDs were bucketed into an "other" category to mitigate sparsity. Detailed statistics are provided in Appendix A.1.

2.3 LDA Topic Modeling

LDA assumes reviews are mixtures of topics, with each topic a distribution over words. The probability of a word w in a document d is:

$$P(w|d) = \sum_{k=1}^{K} P(w|k)P(k|d),$$

where K is the number of topics, $P(w|k)$ is the word distribution for topic k, and $P(k|d)$ is the topic distribution for d. Using Gensim, we:

1. Created a bag-of-words representation and document-term matrix.
2. Selected $K = 4$ based on manual inspection of topic interpretability.
3. Tuned hyperparameters alpha (document-topic density) and beta (topic-word density) to balance topic clarity and diversity.

The model was trained on preprocessed text, emphasizing nouns and adjectives for interpretability.

2.4 Clustering

K-means clustering grouped reviews by engineered features, minimizing within-cluster variance:

$$J = \sum_{i=1}^{n} \sum_{k=1}^{K} r_{ik} \|x_i - \mu_k\|^2,$$

where x_i is a data point, μ_k is the cluster centroid, and r_{ik} is the cluster assignment. We:

1. Applied PCA to reduce feature dimensions, retaining 95% variance.
2. Used t-SNE for 2D visualization.
3. Evaluated $K = 3$ to $K = 11$ using the elbow method, plotting within-cluster sum of squares (WCSS) to identify the optimal K. Silhouette scores were computed to assess cluster separation:
$$S = \frac{b - a}{\max(a, b)},$$
where a is intra-cluster distance and b is nearest-cluster distance.
4. Selected $K = 3$ based on the elbow method and highest silhouette score (0.5301), testing initialization methods (k-means++, random) for robustness.

2.5 Pipeline

The pipeline (Figure 1) integrates preprocessing, feature engineering, LDA, and clustering, enabling end-to-end analysis.

3 Experiments and Results

3.1 Experimental Setup and Optimization Strategy

All experiments were conducted on Google Colab using common Python libraries: NLTK for text preprocessing, Gensim for topic modeling, and scikit-learn for machine learning tasks.

Topic modeling was performed using Latent Dirichlet Allocation (LDA) with four topics. This number was chosen based on manual inspection of semantic coherence and interpretability, as resource constraints prevented automated hyperparameter tuning.

K-means clustering was applied with K=3, validated externally using the elbow method and silhouette score analysis. These evaluations informed the final choice of K to balance simplicity and performance.

To manage Colab's memory limits, optimizations included saving intermediate outputs to Google Drive and precomputing t-SNE projections for visualizing high-dimensional data. These strategies improved execution efficiency and resource utilization.

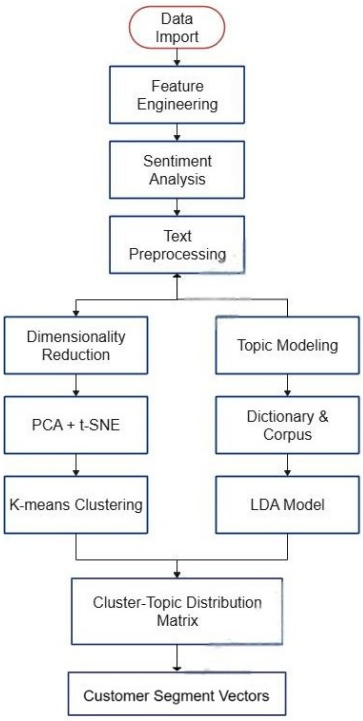


Figure 1: Review Analysis Pipeline

3.2 Data Exploration

The initial dataset consisted of 23,486 customer reviews from an online women's clothing retailer. After removing 845 entries (approximately 3.6 perccent) due to missing review text, a cleaned dataset of 22,641 valid reviews was obtained for further natural language processing and analysis. Reviewer ages ranged from 18 to 99 years, with a mean age of 43.28 and a median of 41. This broad age distribution indicates a diverse customer base, which enhances the generalizability and relevance of the insights derived from the analysis.

In addition to review text and reviewer age, the dataset included numerical ratings on a scale of 1 to 5, representing customer satisfaction levels. It also categorized reviewed products into various clothing types such as dresses, tops, bottoms, and other apparel categories. These features—age, rating, and product category—were incorporated as contextual metadata to support downstream tasks such as clustering and topic modeling, providing meaningful dimensions for segmenting the data and extracting interpretable themes from the reviews. Additional visualizations (e.g., age and rating distributions) are included in Appendix A.2.

3.3 Feature Engineering Results

Feature engineering yielded

- **Word Counts**: Average review length was 59.35 words (`rev_word_count`, min 2, max 115), with 33.17 unique words (`unique_word_count`, min 2, max 62).
- **Sentiment Scores**: VADER compound scores averaged 0.673 (min -0.9735, max 0.9923), reflecting generally positive sentiment. Positive, negative, and neutral scores were also computed.

Detailed statistics are provided in Appendix A.1.

3.4 Clustering Results

K-means clustered 22,641 reviews into three balanced segments (Table 1), visualized via t-SNE (Figure 3):

- **Cluster 0**: 33.06% (7,484 reviews), mixed sentiment, focused on fit issues (e.g., "Dress too tight").
- **Cluster 1**: 34.29% (7,763 reviews), positive sentiment, style-driven (e.g., "Stylish and comfortable").
- **Cluster 2**: 32.66% (7,394 reviews), neutral sentiment, quality concerns (e.g., "Material faded quickly").

The elbow method confirmed $K = 3$ as optimal, with the WCSS plot (Figure 3) showing a clear elbow at $K = 3$. Silhouette scores (Table 2) indicated good cluster separation for $K = 3$ (0.5301), with higher K values yielding lower scores, suggesting reduced cohesion. Additional cluster analysis visualizations are provided in Appendix A.2.

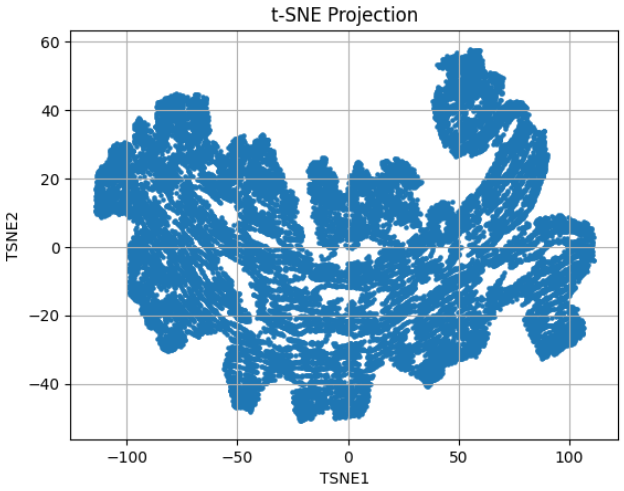
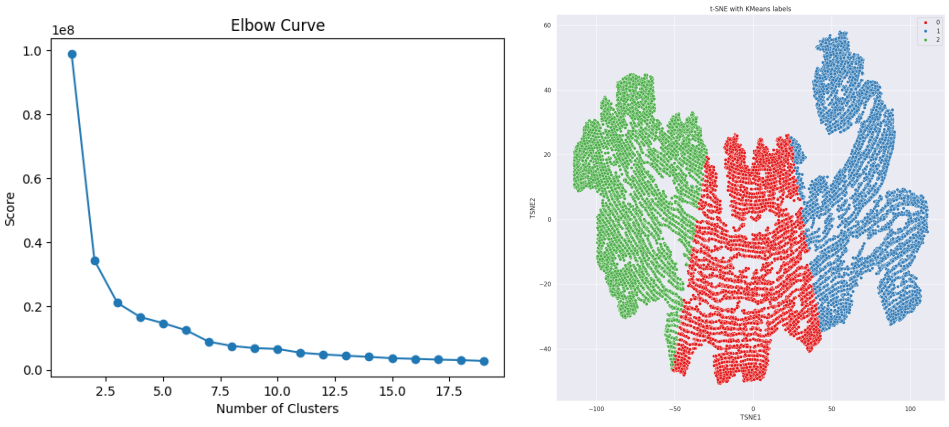


Figure 2: Unlabeled t-SNE



(a)

Figure 3: (a) Elbow Method Plot for K-means Clustering (b) t-SNE Visualization with K-means Clusters

3.5 Topic Modeling Results

The LDA model identified four interpretable topics, derived from the review text using Gensim:

1. **Dresses and Skirts**: Words like "dress," "size," "skirt," "fabric," "petite," "waist," "beautiful" (e.g., "The dress fits perfectly but is a bit short for petites").
2. **Jeans and Pants**: Words like "jean," "pant," "size," "great," "jacket," "comfortable," "pair" (e.g., "These jeans are super comfortable and fit well").
3. **Sweaters and Colors**: Words like "color," "great," "sweater," "soft," "love," "nice," "black" (e.g., "Love the soft sweater, especially in black").
4. **Sizing and Shirts**: Words like "size," "small," "fabric," "large," "medium," "shirt," "didnt," "cute" (e.g., "The shirt runs small, but the fabric is cute").

3.6 Evaluation

LDA performance was assessed with:

- **Perplexity**: 69.56, indicating reasonable model fit to the data, though lower values would suggest better generalization.
- **Jensen-Shannon Divergence (JSD)**: Pairwise JSD between topics ranged from 0.2284 to 0.3116, showing moderate topic distinctiveness, computed as:

$$JSD(P||Q) = \frac{1}{2}(KL(P||M) + KL(Q||M)),$$

  where $M = \frac{P+Q}{2}$. Specific values include:
  - Topic 0 vs. 1: 0.2692
  - Topic 0 vs. 2: 0.2747
  - Topic 0 vs. 3: 0.2284
  - Topic 1 vs. 2: 0.2814
  - Topic 1 vs. 3: 0.3116
  - Topic 2 vs. 3: 0.2974

Clustering performance was evaluated with:

- **Silhouette Score**: 0.5301 for $K = 3$, indicating good cluster separation, with scores for higher K values (Table 2) showing reduced cohesion.

Qualitative analysis via manual review sampling and pyLDAvis visualizations confirmed topic relevance (e.g., "Dresses and Skirts" topic aligned with petite sizing complaints). No coherence metrics (e.g., C_v) were computed due to computational constraints, relying on manual inspection for topic selection.

3.7 Analysis

The clustering analysis identified three distinct customer segments: Cluster 0 (fit-focused) emphasized garment sizing and comfort, Cluster 1 (style-driven) focused on design and trends, and Cluster 2 (quality-concerned) highlighted fabric durability and construction. These segments inform targeted marketing strategies, such as promoting fit accuracy to Cluster 0 or emphasizing aesthetics to Cluster 1.

Topic modeling with LDA uncovered key themes: sizing issues in dresses and skirts (especially for petite customers), comfort concerns for jeans and pants, visual and tactile preferences for sweaters, and inconsistent sizing for shirts. These insights guide product improvements, like refining size guides and material descriptions.

Sentiment analysis using VADER revealed a positive overall sentiment (mean score: 0.673), reflecting general customer satisfaction.

Analytical challenges included noisy text, which was resolved through NLTK-based preprocessing and stopword removal. Topic clarity was enhanced by focusing on nouns and adjectives during tokenization. Computational constraints were mitigated using checkpointing and sentence embeddings to improve efficiency.

4 Conclusion

The project applied Latent Dirichlet Allocation (LDA) topic modeling and K-means clustering to 22,641 women's clothing reviews to extract meaningful insights. LDA uncovered four dominant themes aligned with product categories and customer concerns: dresses and skirts; jeans and pants; sweaters and color preferences; and sizing issues with shirts. Meanwhile, K-means grouped reviews into three distinct customer segments by primary concern: fit-focused, style-driven, and quality-conscious shoppers. These complementary results linked what customers discuss with who is discussing it, offering a comprehensive view of customer feedback. The analysis surpassed its initial objectives by delivering robust findings and visualizations beyond preliminary goals. The approach is also modular and scalable, enabling updates with new data and adjustments to the number of topics or clusters as needed. Overall, the study demonstrates that advanced text mining techniques like LDA and K-means yield actionable insights from unstructured reviews, providing a roadmap for data-driven improvements in product design, sizing policies, and targeted marketing in the women's apparel industry, ultimately enhancing customer satisfaction and business outcomes.

5 References

1. Dataset: https://www.kaggle.com/datasets/nicapotato/womens-ecommerce-clothing-reviews/data
2. LDA for Sentiment Analysis: https://datascience.stackexchange.com/questions/53271/lda-for-sentiment-analysis
3. LDA in NLP: https://mohamedbakrey094.medium.com/all-about-latent-dirichlet-allocation-lda-in-nlp-6cfa7825034e
4. Understanding LDA: https://medium.com/@pinakdatta/understanding-lda-unveiling-hidden-topics-in-text-data-9bbbd25ae162
5. Visualize Document Clusters: https://www.mathworks.com/help/textanalytics/ug/visualize-document-clusters-using-lda-model.html
6. Clustering with LDA: https://stats.stackexchange.com/questions/292281/clustering-with-latent-dirichlet-allocation-lda-distance-measure
7. Clustering and Topic Modeling Evaluation: https://pmc.ncbi.nlm.nih.gov/articles/PMC9040385/
8. Age-wise Sentiment Analysis: https://www.kaggle.com/code/adhok93/understanding-age-wise-sentiments-using-k-means
9. COVID-19 Literature Clustering: https://www.kaggle.com/code/maksimeren/covid-19-literature-clustering/notebook
10. Topic Modeling Tutorial: https://github.com/adashofdata/nlp-in-python-tutorial/blob/master/4-Topic-Modeling.ipynb
11. t-SNE Guide: https://distill.pub/2016/misread-tsne/
12. Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis. ICWSM-14, Ann Arbor, MI.
13. Lossio-Ventura, J.A., et al. (2021). Evaluation of Clustering and Topic Modeling Methods. Artificial Intelligence in Medicine, 117:102096. https://doi.org/10.1016/j.artmed.2021.102096
14. ChatGPT Prompts:
    - Create LaTeX table template for dataset summary.
    - Provide Python code for LDA topic modeling.
    - https://chatgpt.com/share/67da5ad2-b7f4-800d-8751-bd44a5c92b87

- https://chatgpt.com/share/6803110c-a440-800d-8f3d-d7249d6d0ed5
- https://chatgpt.com/share/6803112a-edb0-800d-be61-91f218d71557
- Explain LDA and topic modeling in detail.
- List potential challenges in NLP projects.
- Debug Python code for LDA implementation.

A Supplementary Material

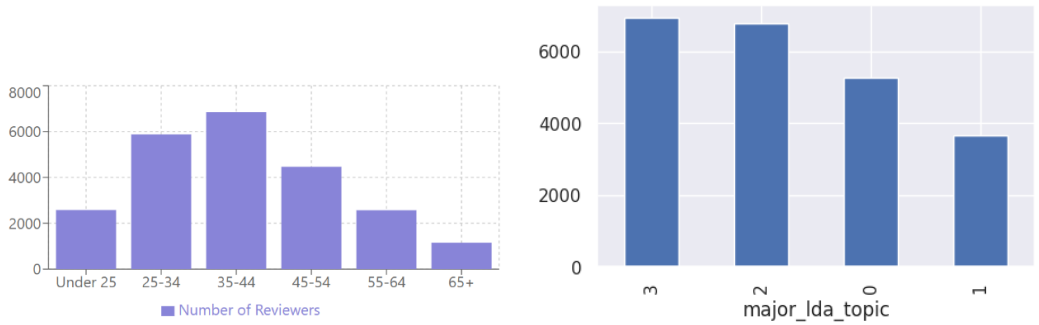
This appendix provides additional figures and feature engineering details to support the main report.

A.1 Feature Engineering Details

Table 3 summarizes the statistics for engineered features, including word counts and VADER sentiment scores, computed for the 22,641 reviews.

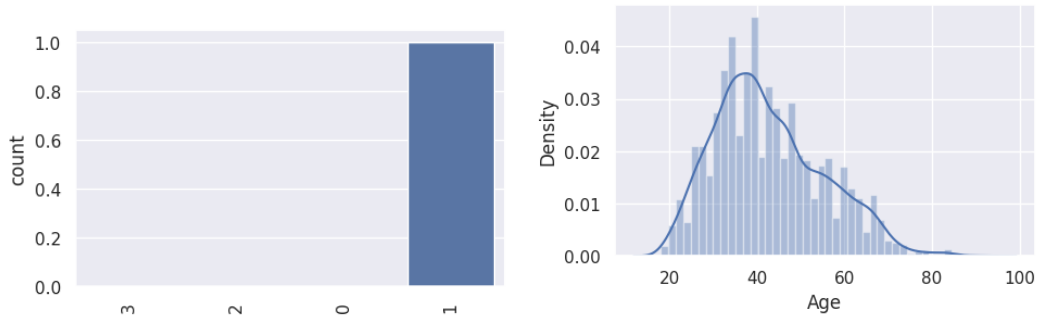
A.2 Additional Figures

This section includes supplementary visualizations for data exploration, clustering, and topic modeling.



(a) (b)

Figure 4: (a) Age Distribution of Reviewers (b) LDA Topic Value Counts



(a) (b)

Figure 5: (a) Categorical Feature Distribution for Cluster 0 (b) Numerical Feature Distribution for Cluster 0

Table 1: Cluster Distribution Analysis ($K = 3$)

Cluster	Number of Points	Percentage (%)
Cluster 0	7,484	33.06
Cluster 1	7,763	34.29
Cluster 2	7,394	32.66

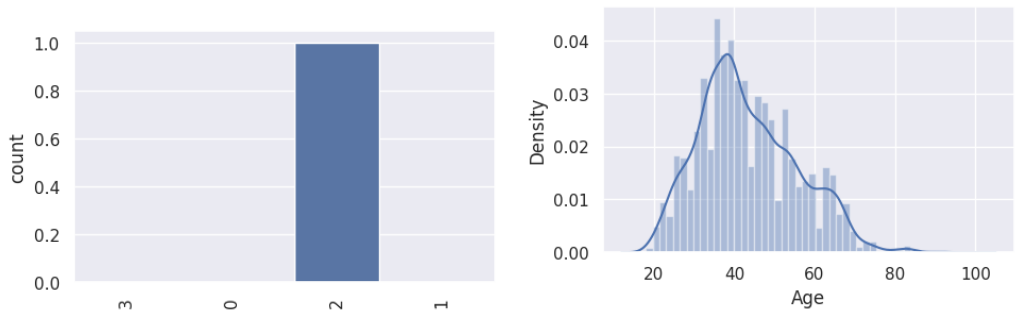
Total reviews analyzed: 22,641

Table 2: Silhouette Scores for Different K Values

K	Silhouette Score
3	0.5301
4	0.4545
5	0.3890
6	0.4019
7	0.4120
8	0.4084
9	0.4236
10	0.4241
11	0.4089

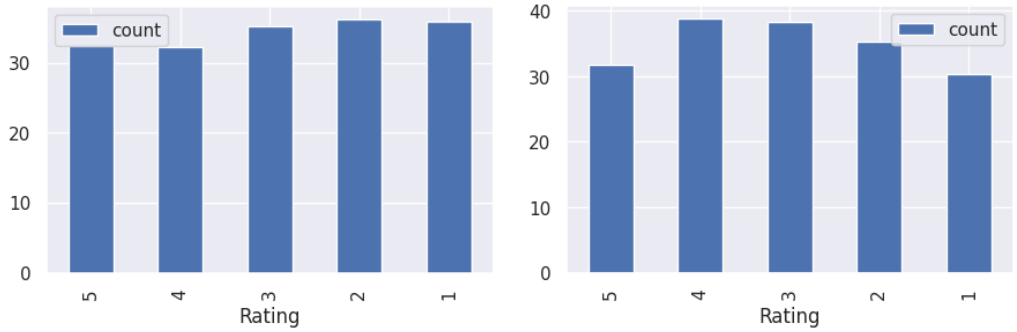
Table 3: Feature Engineering Statistics

Feature	Mean	Std	Min	Median	Max
`rev_word_count`	59.35	28.28	2	57	115
`unique_word_count`	33.17	14.08	2	32	62
Compound Sentiment	0.673	0.390	-0.9735	0.8481	0.9923
Positive Sentiment	0.279	0.165	0.0	0.260	1.0
Negative Sentiment	0.048	0.089	0.0	0.0	0.737
Neutral Sentiment	0.673	0.180	0.0	0.677	1.0



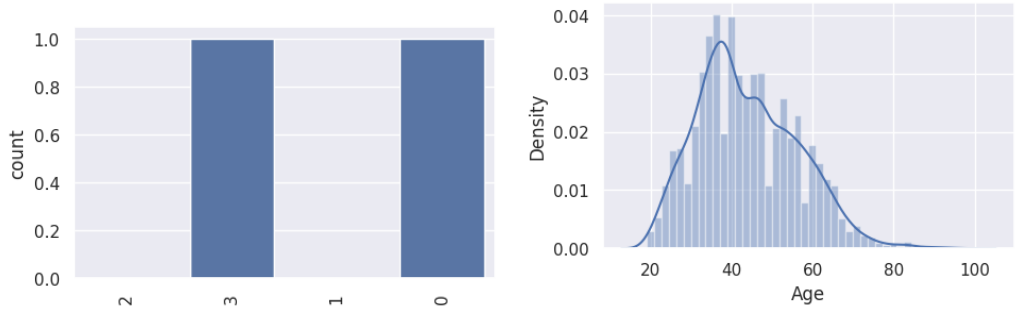
(a) (b)

Figure 6: (a) Categorical Feature Distribution for Cluster 1 (b) Numerical Feature Distribution for Cluster 1



(a) (b)

Figure 7: (a) Rating Distribution for Cluster 0 (b) Rating Distribution for Cluster 1



(a) (b)

Figure 8: (a) Categorical Feature Distribution for Cluster 2 (b) Numerical Feature Distribution for Cluster 2

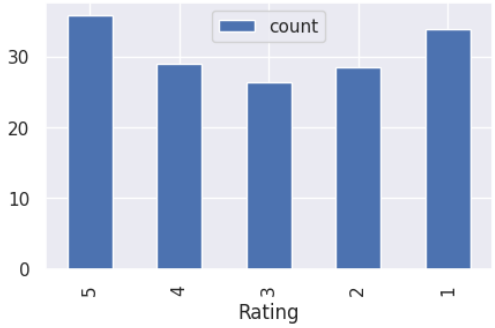


Figure 9: Rating Distribution for Cluster 2

Topic	Keyword	Probability
0	dress	0.1312
	size	0.0353
	true	0.0060
1	jean	0.0397
	pant	0.0359
	work	0.0077
2	color	0.0458
	great	0.0301
	little	0.0076
3	size	0.0390
	small	0.0333
	side	0.0067

Table 4: Top 3 Keywords per Topic with Probabilities