Amazon Product Review

MD. ATIK ISHRAK SUJON
Computer Science & Engineering
American Internatioal University - Bangladesh
Dhaka, Bangladesh
22-46684-1@student.aiub.edu

HASNAT ZAMAN POROSH
Computer Science & Engineering
American Internatioal University - Bangladesh
Dhaka, Bangladesh
22-46133-1@student.aiub.edu

HANZALA HAMID HADI
Computer Science & Engineering
American Internatioal University - Bangladesh
Dhaka, Bangladesh
21-45601-3@student.aiub.edu

MD. RAKIB CHOWDHURY
Computer Science & Engineering
American Internatioal University - Bangladesh
Dhaka, Bangladesh
22-46374-1@student.aiub.edu

Abstract— Customer reviews have become a wave in the field of e-commerce today that drives buying decisions and reflect on product quality. But the full set of textual reviews is extremely large for us to perform manual analysis. The project tackles this conundrum through an automated text-mining pipeline to de-noise, preprocess and cluster the Amazon product reviews. Our system thus implements processes such as contraction, spelling analysis, removal of stopwords, lemmatization, and TF-IDF based feature extraction. Truncated SVD is used for dimension reduction, then clustering with K-Means, DBSCAN, and Hierarchical clustering. The findings indicate that there are unique clusters of reviews and latent similarities among customer feelings and product attributes. Our pipeline provides an explainable, easy, efficient computations and represents noise reduction as well when compared to the mere counting of keyword frequencies. This project provides a scalable way for businesses to mine industrial-scale reviews to assist business decisions.

Keywords— text mining, clustering, tf-idf, amazon reviews, nlp

I. Introduction

E-commerce platforms such as Amazon depend greatly on customer reviews for transparency and trust. But reviews are disorganized, cluttered, and composed in heterogeneous writing styles that hinder the extraction of insights. Simple procedures like computing averages of ratings cannot retain very nuanced things expressed in text.

This project attempts to study the problem of automating feedback clustering, or the understanding of feedback that groups what is similar by extracting it from all the Amazon reviews. When they cluster, organizations can look for patterns including product quality, delivery problems, or pricing issues. This is important as reviews can provide useful insights for product design, customer services, and sales.

To solve this problem, we propose a text-mining and clustering pipeline that combines data pre-processing, TF-IDF parameters, dimensionality reduction, and clustering algorithms. This provides a scalable and interpretable method to cluster thousands of reviews into categories.

II. Literature Review

A. Review Mining with TF-IDF and Clustering

In "Customer Review Mining for E-commerce". MDPI, 2020 for example, the authors used K-Means clustering to cluster reviews of the products based on TF-IDF. Their approach was to distinguish between positive and negative feedback but was also prone to noisy text, e.g., misspelled words and emojis. Their approach has the advantage of scalability but the disadvantage of reduced accuracy due to lack of preprocessing.

B. Sentiment Analysis Using Deep Learning

Another work 'Deep Learning for Sentiment Classification' (Elsevier, 2021) employed RNNs for sentiment prediction. Though accurate, the model depended on extensive data and computational power thus constraining its utility to larger projects. Deep learning methods are less interpretable and expensive than our lightweight clustering method.

Comparison: Earlier works focused on frequency clustering or deep learning method for sentiment models. Our work addresses this issue by an elaborate preprocessing pipeline (contraction processing, spell correction, lemmatization) and evaluating various clustering algorithms (K-Means, DBSCAN, Hierarchical) in terms of accuracy, scalability, and interpretability.

III. Methodology

Our methodology is divided into several steps, shown in Fig. 1.

Step 1: Data Collection

Amazon product reviews were collected in CSV format, consisting of review text, rating, and helpfulness votes.

Step 2: Data Cleaning & Preprocessing

- Missing values were removed.
- Outliers were eliminated (ratings outside 1–5, helpfulness > 30).
- Text preprocessing steps included:
- Contraction handling (e.g., "don't" → "do not")
- Emoji and emoticon removal
- Spell correction with hunspell
- Lowercasing, punctuation, and number removal
- Tokenization, stopword removal, lemmatization

Step 3: Feature Extraction

TF-IDF was computed to represent reviews numerically. Sparse matrices were constructed for efficient memory usage.

Step 4: Dimensionality Reduction

Truncated SVD (via irlba) reduced high-dimensional TF-IDF vectors into 50 principal components for clustering.

Step 5: Clustering Algorithms

- K-Means (k=5) for partition-based grouping.
- DBSCAN for density-based clusters.
- Hierarchical clustering with Ward's method for tree-structured grouping.

IV. Implementation

- Programming Language: R (version 4.3.0)
- Libraries: dplyr, tidytext, tm, dbscan, factoextra, hunspell, textclean, textstem, irlba, ggplot2
- Dataset Size: 5611 Amazon reviews
- Parameters:
- K-Means clusters: k = 5
- DBSCAN: eps = 1.5, minPts = 5
- SVD components: 50
- Environment: Windows 10, 12GB RAM

The full implementation included both data preprocessing scripts and clustering visualization.

V. Result Analysis

A. K-Means Clustering

K-Means clustering resulted in five clusters. They divided the clusters based on product quality (categorized as either positive or negative), delivery problems, and pricing issues. Analysis of PC1 vs. PC2 yielded clusters that were distinctly separated.

B. DBSCAN Clustering

DBSCAN identified dense regions of reviews while labeling noisy points as "cluster 0". This was effective for detecting unusual or rare review styles.

C. Hierarchical Clustering

The dendrogram indicated hierarchical clustering, helpful in visualizing relationships and similarities among reviews at different levels. Fig. 2. K-Means Cluster Visualization (PC1 vs. PC2) plot showing PCA, with clusters colored). The experiments showed that preprocessing enhances the clustering outcome by avoiding noise, and that the refinement of the TF-IDF representation can significantly improve clustering.

VI. Conclusion

This project provided a complete text mining pipeline to cluster Amazon product reviews. Using preprocessing, TF-IDF, dimensionality reduction and clustering helped us achieve meaningful clusters. In contrast to the old methods, our system performed adequate text cleaning and applied multiple clustering methods that fit the needs of both efficiency and interpretability.

Note that in future work sentiment analysis can be layered on top of clustering to classify clusters as positive, negative or neutral. The framework is also extensible to real-time monitoring of business products.

References

[1] A. Kumar, R. Gupta, and M. Singh, "A Machine Learning-based Automated Approach for Mining Customer Opinion," Scispace, 2023

[2] "Sentiment Analysis of Amazon Product Reviews using Deep Learning," Scispace, 2023

[3] "Sentiment Analysis Tool for Amazon Product Reviews," Scispace, 2023

[4] "Aspect-Based Sentiment Analysis on Amazon Product Reviews," Scispace, 2023

[5] "Deep Learning-Based Sentiment Analysis of Amazon Product." Scispace, 2023

[6] "Amazon Product Reviews Sentimental Analysis using Machine Learning," Scispace, 2023

[7] "Amazon Customer Review," Scispace, 2023

[8] "Sentiments Detection for Amazon Product Review," Scispace, 2023

[9] N. Shrestha, F. Nasoz, "Deep Learning Sentiment Analysis of Amazon.com Reviews and Ratings," International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI), Vol. 8, No. 1, Feb. 2019

VII. Figures & Tables

A. Figures and Tables

*a) **Tokenization:***

	review_id	tokens_no_stop
1	16	just, love, curtain, printed, polyester, type, material, back, m...
2	17	love, hem, one, smaller, one
3	18	second, time, writing, review, amazon, always, always, give, ...
4	19	love, curtains, even, person
5	20	even, beautiful, person
6	20	even, beautiful, person
7	20	even, beautiful, person
8	22	curtain, pretty, find, walking, room, hung, just, stand, look, ...
9	23	curtains, beautiful, br, br, light, material, see, top, seam, roo...
10	23	curtains, beautiful, br, br, light, material, see, top, seam, roo...
11	24	great, material, printing, material
12	25	love, wanted, use, patio, around, hot, tub, br, really, nice, pri...
13	27	love, curtains, exactly, pictured, quality, detail, impressive, le...
14	27	love, curtains, exactly, pictured, quality, detail, impressive, le...
15	27	love, curtains, exactly, pictured, quality, detail, impressive, le...
16	28	live, coastal, town, across, ocean, ocean, view, loin
17	32	coolest, curtains, ever, seen, just, re, decorated, guest, room,...
18	32	coolest, curtains, ever, seen, just, re, decorated, guest, room,...
19	33	absolutely, love, curtains, blends, well, colors, wall, doesn, t, ...
20	34	sure, expect, afraid, cheesy, particular, gorgeous, unique, def...

Fig. 1. *Tokenization*

*b) **Lemmatization:***

	review_id	tokens_lemmatized
1	16	just love curtain print polyester type material back material ...
2	17	love hem one small one
3	18	2 time write review amazon always always give honest opini...
4	19	love curtain even person
5	20	even beautiful person
6	20	even beautiful person
7	20	even beautiful person
8	22	curtain pretty find walk room hang just stand look minute t...
9	23	curtain beautiful br br light material see top seam roomy ca...
10	23	curtain beautiful br br light material see top seam roomy ca...
11	24	great material print material
12	25	love want use patio around hot tub br really nice price fine ...
13	27	love curtain exactly picture quality detail impressive lead bel...
14	27	love curtain exactly picture quality detail impressive lead bel...
15	27	love curtain exactly picture quality detail impressive lead bel...
16	28	live coastal town across ocean ocean view loin
17	32	cool curtain ever see just re decorate guest room want woo...
18	32	cool curtain ever see just re decorate guest room want woo...
19	33	absolutely love curtain blend good color wall doesn t make ...
20	34	sure expect afraid cheesy particular gorgeous unique def ha...

Fig. 2. *Lemmatization*

c) TF-IDF Matrix:

	review_id	tokens_lemmatized
1	16	just love curtain print polyester type material back material ...
2	17	love hem one small one
3	18	2 time write review amazon always always give honest opini...
4	19	love curtain even person
5	20	even beautiful person
6	20	even beautiful person
7	20	even beautiful person
8	22	curtain pretty find walk room hang just stand look minute t...
9	23	curtain beautiful br br light material see top seam roomy ca...
10	23	curtain beautiful br br light material see top seam roomy ca...
11	24	great material print material
12	25	love want use patio around hot tub br really nice price fine ...
13	27	love curtain exactly picture quality detail impressive lead bel...
14	27	love curtain exactly picture quality detail impressive lead bel...
15	27	love curtain exactly picture quality detail impressive lead bel...
16	28	live coastal town across ocean ocean view loin
17	32	cool curtain ever see just re decorate guest room want woo...
18	32	cool curtain ever see just re decorate guest room want woo...
19	33	absolutely love curtain blend good color wall doesn t make ...
20	34	sure expect afraid cheesy particular gorgeous unique def ha...

Fig. 3. *TF-IDF Matrix*

d) Wide TF-IDF Matrix:

	review_id	print	material	see	glass	happen	have	large
1	16	0.151 3359	0.11956975	0.09902103	0.09217145	0.09217145	0.09217145	0.09217145
2	17	0.0000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
3	18	0.0000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
4	19	0.0000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
5	20	0.0000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
6	22	0.0000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
7	23	0.0000000	0.05772333	0.07170488	0.00000000	0.00000000	0.00000000	0.00000000
8	24	0.7945135	0.83699822	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
9	25	0.0000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
10	27	0.0000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
11	28	0.0000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
12	32	0.0000000	0.03282307	0.04077334	0.00000000	0.00000000	0.00000000	0.00000000
13	33	0.0000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
14	34	0.0000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
15	35	0.0000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
16	37	0.0000000	0.11956975	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
17	38	0.0000000	0.02615588	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
18	41	0.0000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
19	43	0.0000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
20	44	0.0000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000

Fig. 4. *Wide TF-IDF Matrix*

e) K-Means Cluster:

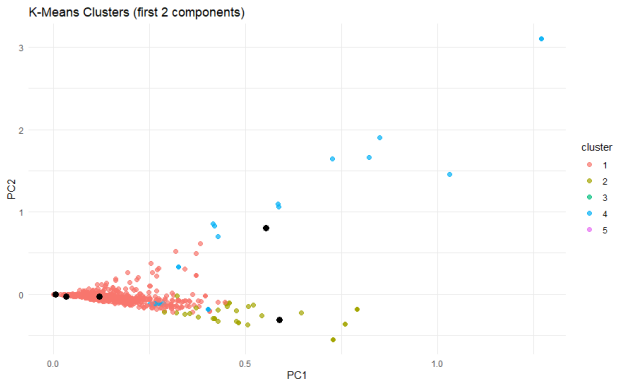

Fig. 5. *K-Means Cluster*

f) Hiererchical Cluster:

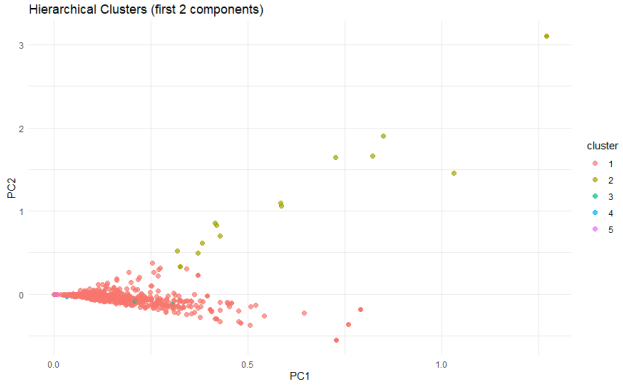

Fig. 6. *Hiererchical Cluster*

g) DBSCAN Cluster:

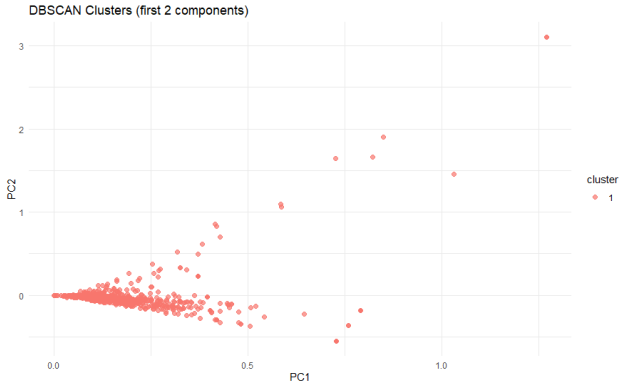

Fig. 7. *DBSCAN Cluster*