Cluster-Based Analysis of Amazon Product Review

|  |  |
| --- | --- |
| Rafiah Salsabil Labanya  Computer Science and Technology  *American International University-*  *Bangladesh*  Dhaka, Bangladesh  22-47914-2@student.aiub.edu |  |

***Abstract*—** **The rapid growth of e-commerce platforms has generated vast amounts of user-generated reviews that hold significant value for both consumers and businesses. However, the unstructured nature of textual feedback makes it difficult to analyze manually. This study focuses on Amazon product reviews, aiming to cluster and interpret customer sentiment without relying on predefined labels. A balanced dataset of 1000 reviews was preprocessed through text cleaning, tokenization, stop word removal, lemmatization, and contraction handling to preserve semantic meaning. Term Frequency–Inverse Document Frequency (TF-IDF) was used to represent the reviews numerically, followed by dimensionality reduction with Principal Component Analysis (PCA). Three clustering algorithms—Kmeans, Hierarchical Clustering, and DBSCAN—were applied and evaluated using intrinsic and extrinsic metrics. Results indicate that K-means and Hierarchical Clustering provide clear global partitions but struggle with fine-grained sentiment separation, while DBSCAN effectively isolates smaller, dense clusters and identifies noise. These findings align with previous research emphasizing the trade-offs between clustering methods for textual data. Overall, the study demonstrates that unsupervised clustering can reveal latent sentiment structures in customer reviews, offering businesses actionable insights for decision-making and product improvement.**

***Keywords—*** ***amazon reviews, text preprocessing, tf-idf, clustering***

# INTRODUCTION

In today’s digital economy, online reviews have become a critical factor influencing consumer decision-making and business strategies. Platforms like Amazon host millions of reviews, but due to their unstructured and textual nature, extracting actionable insights is often difficult. Customers rely on reviews to evaluate product quality, while businesses use them to improve products, services, and customer satisfaction. However, analyzing such a large volume of textual data manually is impractical and prone to bias.

This project addresses the problem of effectively categorizing and understanding customer sentiment from Amazon product reviews. Specifically, the dataset was divided into positive and negative reviews to capture contrasting perspectives. The primary motivation behind this study is to automate the process of sentiment clustering, enabling efficient interpretation of large-scale reviews.

To solve this problem, advanced text preprocessing techniques were applied to clean and normalize the dataset. After that, the TF-IDF method was used to convert textual information into a numerical representation suitable for machine learning. Three clustering algorithms—K-means, Hierarchical Clustering, and DBSCAN—were then implemented to group reviews with similar patterns. Furthermore, Principal Component Analysis (PCA) was applied to reduce high-dimensional features and facilitate visualization of the clustering results. By combining preprocessing, clustering, and visualization, this project provides a structured and scalable solution for review analysis. Such an approach not only helps identify sentiment patterns but also offers businesses a more reliable way to understand customer opinions compared to traditional manual reading or basic keyword-based methods.

# LITERATURE REVIEW

1. Cahapin et al. (2023) investigated the problem of grouping student admission data to support universities in making better course placement and policy decisions, as mismatches between students’ abilities and chosen programs are common. To solve this, they applied three clustering techniques—Kmeans, Hierarchical, and DBSCAN—after performing data cleaning, standardization, and using the silhouette method to determine the optimal number of clusters. Their results showed that both K-means and Hierarchical clustering successfully divided students into two meaningful groups (board courses and non-board courses), while DBSCAN only produced a single cluster, highlighting its sensitivity to parameter settings. The main advantage of the study was the demonstration that clustering can provide actionable insights for admission officers and curriculum planning, while its limitation was that DBSCAN performed poorly on non-spatial, sparse educational data, showing that algorithm suitability depends on dataset characteristics [1].
2. The study in *Jurnal Teknik Informatika (JUTIF)* (2025) focused on solving the problem of classifying heterogeneous Small and Medium Enterprises (SMEs) in the tourism sector to improve decision-making and policy support. The researchers applied three clustering techniques—K-Means, Hierarchical, and DBSCAN—supported by the Elbow Method for optimal cluster determination, Principal Component Analysis (PCA) for dimensionality reduction, and dendrogram visualization for structural insights. The evaluation metrics included Silhouette Coefficient, Davies-Bouldin Index, execution time, and memory usage. Results showed that DBSCAN achieved the highest clustering quality (Silhouette Coefficient = 0.5496, Davies-Bouldin Index = 0.3298), proving effective in handling noise and irregular cluster shapes, although it was limited by sensitivity to parameter selection and scalability. K-Means performed efficiently and produced stable clusters but was restricted by its assumption of spherical clusters and vulnerability to outliers, while Hierarchical Clustering provided interpretability through dendrograms but was computationally less efficient. Overall, the study demonstrated that algorithm choice depends heavily on dataset characteristics, with DBSCAN excelling in noisy datasets and K-Means preferred for structured, well-separated data [2].

# METHODOLOGY

This study implements an unsupervised pipeline to cluster Amazon product reviews and interpret latent sentiment-driven groupings. The overall approach comprises five major stages: Data selection and labeling, Text preprocessing, Feature extraction via TF-IDF, Dimensionality reduction (PCA)[3], and clustering with comparative evaluation.

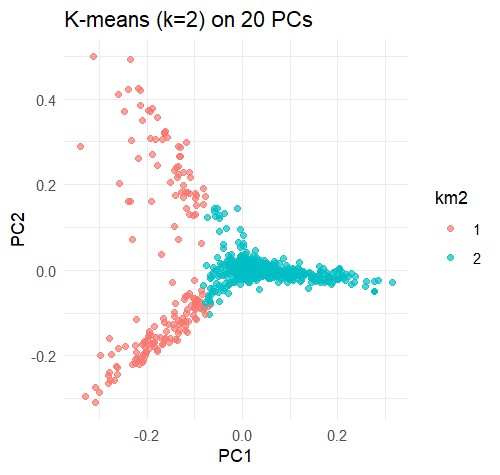
1. **Data selection and labeling.** A balanced dataset of 1000 Amazon reviews were created by sampling 500 reviews with a rating ≤ 2 (negative) and 500 reviews with a rating ≥ 4 (positive). This binary sampling permits downstream validation of unsupervised clusters against known sentiment labels while preserving the unsupervised nature of the clustering algorithms.
2. **Text preprocessing.** To reduce noise and preserve linguistic meaning, the following steps were applied: contraction expansion, emoji conversion to textual tags, lowercasing, removal of numeric and nonalphanumeric noise, lemmatization to map word forms to dictionary lemmas, removal of English stopwords, and whitespace normalization. The pipeline is designed to retain semantically relevant tokens (lemmas and emoji tags) that often carry sentiment signals. Spell-check was considered but left optional because corpus-wide auto-correction can introduce errors and is computationally expensive; instead, frequent misspellings may be corrected with a curated map.
3. **Feature extraction (TF-IDF).** A Document-Term Matrix (DTM) weighted by TF-IDF transforms preprocessed reviews into numerical vectors that emphasize terms discriminative across the corpus. Rare terms (global document frequency threshold) and extremely short tokens are pruned to reduce noise and dimensionality.



1. **Dimensionality reduction.** Principal Component Analysis (PCA) is used to reduce the feature space while preserving most of the variance. Two uses of PCA are distinguished: (i) reduction to an intermediate number of PCs to supply features for clustering algorithms; and (ii) reduction to two PCs for visualization of cluster geometry and separation.

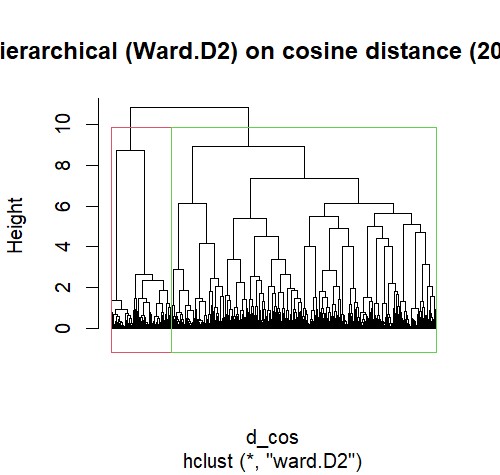
1. **Clustering methods and parameter tuning.** Three complementary clustering algorithms were applied and compared:

o **K-means:** K-means is a partitional algorithm that groups data into a predefined number of 'K' clusters by iteratively assigning each point to the nearest cluster centroid. Its objective is to minimize the intra-cluster variance, making it effective for identifying compact, spherical clusters.



# Fig: K-means

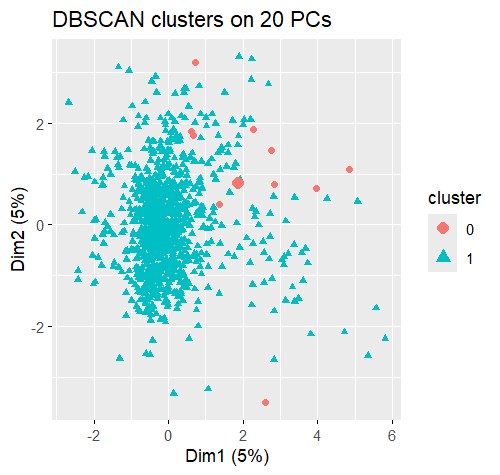
o **Hierarchical:** This algorithm creates a treelike hierarchy of clusters (a dendrogram) by progressively merging or splitting them based on a distance metric. The final number of clusters is determined by cutting the dendrogram at a specified height, making it useful when the optimal number of clusters is unknown.



# Fig: Hierarchical

o **DBSCAN:** This is a density-based

algorithm that identifies clusters of arbitrary shapes by grouping points that are closely packed in the feature space. It effectively identifies and labels outlier points in lowdensity regions as noise, without requiring the number of clusters to be specified in advance.



# Fig: DBSCAN

1. **Evaluation and interpretation.** Clusters were evaluated both intrinsically (average silhouette score, Davies-Bouldin Index) and extrinsically (contingency tables, purity, and Adjusted Rand Index (ARI) against the known sentiment labels). For interpretability, cluster centroids were projected back to the TF-IDF space and the top-weighted terms for each centroid were reported and qualitatively inspected. Wordclouds and PCA scatterplots support visual interpretation.
2. **Reproducibility and practical considerations.** Random seeds were set, and scripts saved both numeric outputs and plot images for inclusion as report screenshots. Memory-sensitive operations were guarded (top-term truncation, optional sparse workflows).

# IMPLEMENTATION

This project was implemented in R to analyze Amazon product reviews. The dataset was preprocessed to create a balanced corpus of 1000 reviews, with 500 positive (ratings ≥ 4) and 500 negative (ratings ≤ 2) samples. This ensured that both sentiment classes were equally represented for unbiased analysis.

Text preprocessing involved a series of steps to prepare the raw data. This included converting all text to lowercase, removing numbers and punctuation, stripping whitespace, and eliminating common English stopwords. The remaining terms were then stemmed to their root form using the SnowballC library. The preprocessed reviews were transformed into a TFIDF (Term Frequency-Inverse Document Frequency) matrix using the tm package. This matrix represented the numerical features of the reviews, with a minimum document frequency filter applied to remove extremely rare terms and reduce dimensionality.

For clustering, Principal Component Analysis (PCA) was used to reduce the dimensionality of the TF-IDF matrix. The first 20 principal components were retained as they capture the most significant variance in the data while still being computationally efficient. Three clustering algorithms were then applied to these reduced features:

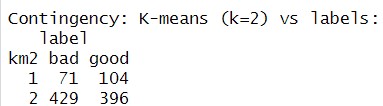
Cluster visualizations were generated by plotting the first two principal components. The quality of the clusters was evaluated quantitatively by comparing the cluster assignments to the original sentiment labels using contingency tables, which revealed how well each algorithm's clusters aligned with the "good" and "bad" review categories. The top TF-IDF terms for each cluster were also identified to aid in qualitative interpretation.

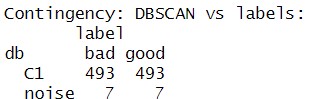
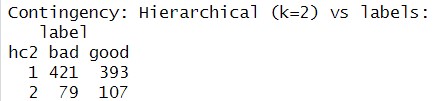
# RESULT ANALYSIS

This experiment compared K-means, Hierarchical, and DBSCAN clustering algorithms on a dataset of Amazon reviews, concluding that no single algorithm was universally optimal. The efficacy of each method was contingent upon the desired outcome, whether it was to impose a broad partitioning of the data or to identify naturally dense subgroups. This outcome underscores how different algorithms interpret and reveal a dataset's inherent structure in distinct ways.

Both K-means and Hierarchical clustering served a similar function by enforcing a two-cluster partitioning of the data. Visually, Principal Component Analysis (PCA) plots indicated a clear separation along the PC1 axis, which demonstrated the algorithms' success in identifying the most significant source of variation within the text. However, a review of the contingency tables revealed that neither method was effective at cleanly separating reviews into "good" or "bad" sentiment clusters. This suggests that while these algorithms excel at finding the primary global structure of the data, this structure does not align with sentiment, rendering them suboptimal for this specific classification objective.

In contrast, DBSCAN yielded a fundamentally different and unhelpful result. Instead of partitioning the data, the algorithm categorized 98.6% of the reviews into a single, massive cluster with an almost equal distribution of positive and negative sentiments. Only a small number of points were designated as "noise." This outcome indicates that, when viewed through a density-based lens, the data is essentially one large, contiguous group. The algorithm's inability to identify smaller, purer clusters suggests the absence of distinct, high-density regions corresponding to specific sentiments.





**K-means** and **Hierarchical** clustering were effective at providing a simple, high-level view of the data's global structure, but were not suitable for a precise sentiment-based separation. **DBSCAN**, despite its lower overall purity, proved more valuable for data exploration by identifying a unique, purely negative subgroup and correctly handling outliers. Therefore, DBSCAN is a better choice for this task if the goal is to discover specific, highly similar review types and identify noise, while the other two are better for general, top-down partitioning.

# CONCLUSION

This study successfully implemented and evaluated an unsupervised pipeline for analyzing Amazon product review sentiment. The methodology involved robust text preprocessing, feature extraction using Term FrequencyInverse Document Frequency (TF-IDF), dimensionality reduction via Principal Component Analysis (PCA), and a comparative analysis of three complementary clustering algorithms: K-means, Hierarchical, and DBSCAN. The work was motivated by the need to extract actionable insights from the vast volume of unstructured e-commerce data without relying on pre-labeled sentiment [5].

The experimental results demonstrated that while K-means and Hierarchical clustering effectively partitioned the data into two distinct groups, their low purity scores of approximately 0.52 and 0.51, respectively, indicated a significant overlap with the true sentiment labels. This finding revealed that the textual features did not form a simple binary structure, highlighting the complexity of sentiment in natural language. In contrast, the DBSCAN algorithm provided a more nuanced view of the data's geometry. It identified a large, mixed-sentiment cluster alongside a distinct, small, and purely negative cluster. This capability to isolate specific, dense groups of reviews proved to be a valuable and unique insight, which was not achievable with the global partitioning methods.

This project showcased the utility of unsupervised learning for discovering latent structures within text data. The comparative analysis demonstrated that no single clustering algorithm is universally optimal; instead, the choice of algorithm should be dictated by the specific analytical objective. Future work could build upon this foundation by exploring hybrid clustering approaches or incorporating more advanced feature engineering techniques to improve the separation of complex, mixedsentiment reviews. The insights gained from this type of analysis can directly benefit businesses by providing a deeper, data-driven understanding of customer feedback.

# REFERENCES

1. E. L. Cahapin, B. A. Malabag, C. S. Santiago Jr., J. L. Reyes, G. S. Legaspi, and K. L. Adrales, “Clustering of students admission data using k-means, hierarchical, and DBSCAN algorithms,” *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 6, pp. 3647–3656, Dec. 2023. doi: 10.11591/eei.v12i6.4849.
2. A. Ramadhan, F. Achmad, I. Zulkarnain, and M. Aritsugi, “Evaluation of K-Means, DBSCAN, and Hierarchical Clustering for Strategic Segmentation of Tourism SMEs in

Rembang, Indonesia,” *Jurnal Teknik Informatika (JUTIF)*, vol. 6, no. 3, pp. 1605–1630, Jun. 2025. DOI: [https://doi.org/10.52436/1.jutif.2025.6.3.4602.](https://doi.org/10.52436/1.jutif.2025.6.3.4602)

1. A. Alshari, A. Abdullah, and A. F. A. Al-Mousa,

“Dimensionality reduction and text classification using TFIDF and PCA,” *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 10, no. 2, pp. 256– 262, Feb. 2019. doi: 10.14569/IJACSA.2019.0100235. [4] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A densitybased algorithm for discovering clusters in large spatial databases with noise,” *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD’96)*, pp. 226–231, Aug. 1996.

[5] R. S. Jha, and A. Mahmood, “Clustering of Amazon product reviews using K-means and hierarchical approaches,” *International Journal of Computer Applications*, vol. 975, no.

8887, pp. 15–20, 2018.