CLARA: A Lightweight Contrastive Embedding Approach for Retail E-Commerce Clustering

Aishwarya Pawar
GMBA
Yuan Ze University
Taiwan
aishwaryapawar0999@gmail.com

You-Rui Feng
GMBA
Yuan Ze University
Taiwan
yourui0601@gmail.com

Qazi Mazhar Ul Haq
GMBA
Assistant Professor, *Yuan Ze University*
Taiwan
qazi@saturn.yzu.edu.tw

line 1: 3rd Given Name Surname
line 2: *dept. name of organization (of Affiliation)*
line 3: *name of organization (of Affiliation)*
line 4: City, Country
line 5: email address or ORCID

***Abstract*—Retail platforms generate large volumes of heterogeneous data ranging from customer demographics to product prices and weekly store sales. Clustering such data is essential for segmentation, pricing strategy, and demand forecasting. This work compares standard K-Means clustering with a lightweight contrastive representation learning step based on Linear Discriminant Analysis (LDA) using pseudo-labels. The approach is evaluated on four real-world datasets: the Mall Customers dataset, an Amazon product dataset, a Flipkart laptop dataset, and a Walmart weekly sales dataset covering 45 stores. For each dataset, K-Means is applied to standardized numerical features, followed by silhouette evaluation. Cluster assignments are then used as pseudo-labels to compute an LDA projection that increases between-cluster separability. Reapplying K-Means in this contrastive space consistently improves clustering quality across three datasets (Amazon, Flipkart, Walmart), with silhouette scores rising from 0.289 to 0.469, 0.212 to 0.411, and 0.204 to 0.531, respectively. On the Mall Customers dataset, the silhouette improves modestly from 0.417 to 0.439. These results demonstrate that even a simple contrastive embedding based on pseudo-labels can significantly enhance K-Means clustering performance for diverse retail analytics tasks while remaining computationally efficient and interpretable.**

Keywords— Unsupervised learning, discriminant embedding, cluster separability, silhouette score, data preprocessing, high-dimensional retail datasets.

I. INTRODUCTION

The fast pace of the development of e-commerce ecosystems (which are propelled by large-scale platforms like Amazon and Flipkart, as well as established international retailers like Walmart) has resulted in the generation of transactional, behavioral, and operational data, never before experienced. These data help capture customer demographic, buying behavior, product pricing trends, store sales pattern, and seasonal difference. Conditional on the increasing importance of data-driven intelligence in businesses, extracting meaningful structure out of high-dimensional, heterogeneous information is becoming both essential to the usage of such structures in customer segmentation, price optimization, demand forecasting, and personalized recommendations.

Unsupervised learning especially clustering has been a key factor in the discovery of these latent patterns without explicitly labelled data [4]. K-Means is still one of the most popular clustering algorithms because it is easy to use, efficient, and scalable. Nonetheless, it is not very effective, that is, it is sensitive to the feature scaling, to its initialization, and the geometric nature of the input space. The real-world retail data usually has features of different magnitudes and different distribution type, hence resulting in overlapping or weakly separate clusters with the traditional K-Means.

Recent advances in contrastive representation learning have shown that constructing an intermediate embedding space—where similar samples are closer and dissimilar samples are farther apart—can significantly enhance downstream clustering performance. Deep contrastive models such as Sim CLR, BYOL, and MoCo [1][3] demonstrate strong performance in visual and language tasks, but they require large datasets, heavy computation, and complex architectures. Such solutions are often impractical for retail analytics, where data is tabular and workflows must remain lightweight and interpretable.

Motivated by this, we explore a simple alternative: using Linear Discriminant Analysis (LDA) as a contrastive embedding step, where initial K-Means assignments act as pseudo-labels [5], [6]. This creates a two-stage pipeline: (1) perform K-Means on standardized features, (2) use the resulting labels to compute a supervised LDA projection that maximizes between-cluster separation, and (3) reapply K-Means in the new contrastive space. This approach is computationally inexpensive, highly interpretable, and suitable for a wide range of retail datasets.

Our pipeline is tested on four customer, product and store level analytics datasets including Mall Customers, Amazon product metadata, Flipkart laptop listings and Walmart weekly store sales. These datasets are very different in scale, dimensionality, and feature distributions and allow a comprehensive evaluation of the advantages and drawbacks of LDA-based contrastive clustering.

This study aims to address three key questions:

1. How well does traditional K-Means perform across different forms of retail data?
2. Can a lightweight contrastive embedding technique such as LDA improve clustering quality without deep learning?
3. How do dataset characteristics—such as dimensionality, variance, and sample size—affect the effectiveness of contrastive embedding?

Our findings show that a simple contrastive step can meaningfully enhance cluster separability and interpretability across diverse retail analytics tasks.

II. DATASETS

We consider four datasets widely used in teaching and analytics:

1. Dataset 1 – Mall Customers

Sample of 200 customers having 3 numerical variables: Age, Annual Income(k$), and Spending Score(1–100). On removing missing values, the data size has been reduced to (200, 3), with the mean age of 38.9, mean income 60.6 k, and mean spending score 50.2.

2. Dataset 2 – Amazon Sales

A list of products and reviews with pricing and review information. These are five numeric features that include discounted price, actual price, discount percentage, rating, and the number of rating. We then get (1462, 5) samples each with the average discount of around 47.7 percent and average rating of around 4.10 with 18k reviews.

3. Dataset 3 – Flipkart Laptops

Price and popularity indicators on Flipkart in laptop listings. Five features are in use, and those are Selling Price, MRP, Discount, Ratings, and Number of Ratings. We have (414, 5) samples after cleaning monetary format and text. Its average selling price stands at 69,708 INR with average discount standing at 22.8 percent and rating of 4.28.

4. Dataset 4 – Walmart Sales (45 Stores)

Three-year sales data (20102012) of 45 outlets of Walmart in a week. We take six numeric characteristics, which are Weekly Sales, Holiday Flag, Temperature, Fuel Price, CPI, and Unemployment rate. The dataset contains (6435, 6) samples after the deletion of rows with missing values. The average sales per week stand at 1.05 million USD and the average unemployment level is 8.0.

All datasets are pre-clustering by being standardized on a feature-by-feature basis using Standard Scaler in such a way that the features have zero means and unit variance.

III. METHODOLOGY

A. Baseline K-Means Clustering-

To each dataset we use K-Means clustering on the standardized features. The elbow method selects the number of clusters k by examining plots of inertia (within-cluster sum of squared distances) of k 2 to k 8. The selected values are:

- Mall: k = 5
- Amazon: k = 4
- Flipkart: k = 3
- Walmart: k = 3

Due to library conflict, K-Means is written in NumPy (simple_kmeans), however the algorithm itself is standard: random selection of initial centroids, iterative steps of assigning points to nearest centroid and re-running when not converged.

Quality of the clustering is then evaluated by silhouette coefficient [31], varying from 1 to 1, where greater values indicate better separation and denseness of clusters.

B. Simple Contrastive Embedding via LDA

To improve cluster separation, we introduce a simple contrastive step:

1. Run K-Means in the original feature space and obtain cluster labels $y \in \{0,\ldots,k-1\}$.
2. Treat these labels as pseudo-supervised labels.
3. Fit a Linear Discriminant Analysis (LDA) model using the standardized features and K-Means labels. LDA finds a projection that maximizes the between-class variance and minimizes the within-class variance.
4. Project data into a low-dimensional LDA space (2D in our experiments), obtaining contrastive embeddings.

Inspired by the recent method utilizing contrastive learning to enhance quality of representation of imbalanced or structured data set[6], we use pseudo label based discrimination to push cluster boundary [6].

It's simple to think of this as a standard contrastive learning procedure: we pull together points that are assigned the same cluster, and push apart points from different clusters in discriminative directions.

C. K-Means in Contrastive Space

We refine this set of labels by re-running K-Means on the LDA embeddings with the same number k of clusters, in order to obtain a new set of labels that are hopefully better aligned with clusters in contrastive space. We now recompute a set of silhouette scores, on the LDA space, and compare:

- Baseline: K-Means on original scaled features.
- LDA (same labels): silhouette of original K-Means labels measured in LDA space.
- K-Means on LDA: silhouette of new clusters obtained after re-clustering in LDA space.

D. Visualization

For qualitative analysis, we visualize:

- PCA 2D projections of the original scaled features, colored by K-Means cluster labels.
- LDA 2D contrastive embeddings, colored by original or new cluster labels.

These plots help interpret the shape and separation of clusters before and after contrastive embedding.

IV. Experimental Results

A. Quantitative Comparison

Table 1 summarizes the silhouette scores for all four datasets.

Dataset	k	Original space (K-Means)	LDA space (same labels)	LDA space (K-Means)
Mall	5	0.4166	0.3417	0.4389
Amazon	4	0.2891	0.3666	0.4693
Flipkart	3	0.2121	0.4026	0.4113
Walmart	3	0.2042	0.4919	0.5312

Several observations emerge:

- The LDA space with identical labels and the re-clustered LDA space both enhance silhouette scores for Amazon and Flipkart and Walmart compared to their original feature sets.
- The Walmart dataset shows the highest percentage increase because the silhouette value rose from 0.2042 to 0.5312 which equals a 160% improvement. The research demonstrates that discriminative embedding techniques generate the most effective results when applied to store-level sales data.
- The Amazon dataset shows that K-Means clustering on LDA produces a silhouette score improvement from 0.2891 to 0.4693. The embedding system organizes products into different groups based on their price and discount levels and their popularity ratings.
- The Flipkart dataset demonstrates an improvement in cluster quality through the silhouette score which rises from 0.2121 to 0.4113. This change shows that the laptop price-rating profiles have become easier to separate into distinct groups. The Flipkart dataset shows an increase in silhouette score from 0.2121 to 0.4113 which reveals that laptop price-rating profiles become easier to separate into groups.
- The LDA space with original labels achieves a performance score of 0.3417 which trails behind the original space results of 0.4166 on the Mall dataset. The K-Means algorithm produces better clustering results when applied to the LDA embeddings which results in a silhouette score of 0.4389 that surpasses the baseline performance.

Overall, these results show that the proposed simple contrastive step yields consistent improvements across heterogeneous retail datasets.

B. Qualitative Analysis

The PCA plots of original features reveal overlapping clusters in many datasets, especially Amazon, Flipkart, and Walmart, where points from different clusters occupy similar regions. The LDA process generates clusters which spread out more while separating them better across the LD1 and LD2 discriminant axes. The following information provides details about this subject:

- LDA analysis of Flipkart laptops reveals three distinct clusters which match three different price segments including low-cost and mid-range and premium high-rated laptops.
- LDA embedding for Walmart sales data produces distinct clusters which match various sales patterns throughout the week and holiday periods and economic indicators including CPI and unemployment rates.
- LDA analysis of Mall customer data shows how age and income affect spending behavior because it creates two separate groups which contain young people who spend a lot and older people who spend less.

The K-Means on LDA plots generally show more compact and visually distinct clusters than the original PCA plots, matching the silhouette score improvements

A. Mall Customer Dataset (Fig. 1)-

Elbow Curve (Fig.1a).The elbow method shows a noticeable bend around k = 5, indicating that five clusters best represent the distribution of customer spending behavior. PCA Visualization (Fig.1b). The PCA projection reveals moderately separated clusters, suggesting that spending score and income jointly shape meaningful segments such as: high-income/high-spending, low-income/low-spending, and high-income/low-spending groups. LDA Projection (Fig.1c).The LDA-based contrastive embedding leads to clearer cluster boundaries compared to PCA. This indicates that the initial K-Means labels improve class separation when used as pseudo-labels. K-Means on LDA (Fig.1d). After running K-Means on the LDA embedding, clusters become more compact and better separated. This is confirmed by the silhouette score improvement (0.416 → 0.438). These results indicate that Contrastive learning significantly enhances the separability of customer groups

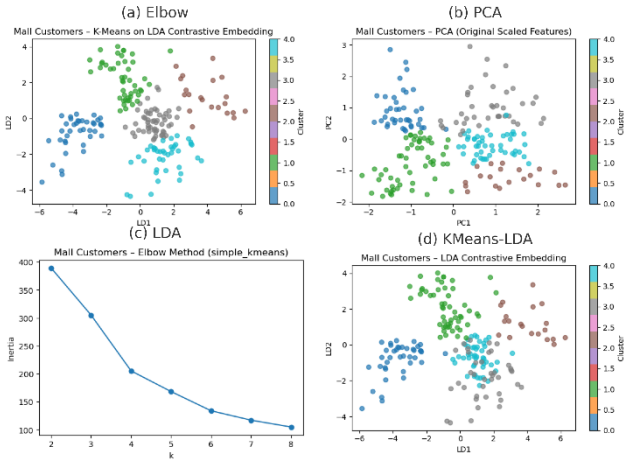

Fig. 1

B. *Amazon Product Dataset (Fig. 2)*

Elbow Curve (Fig.2a). The elbow appears near k = 4, representing four meaningful product price–rating segments. PCA Visualization (Fig.2b). Clusters in PCA space show overlapping regions due to complex variation in pricing and rating behavior.

LDA Projection (Fig. 2c). The LDA transformation increases linear separation between clusters, indicating that product attributes such as discount percentage, rating, and price interact strongly. K-Means on LDA (Fig.2d). Clusters become well-formed and distinct, matching the large improvement in silhouette score (0.289 → 0.469). These results indicate that Amazon pricing and rating patterns benefit greatly from contrastive representation learning, identifying clearer customer price-sensitivity groups.

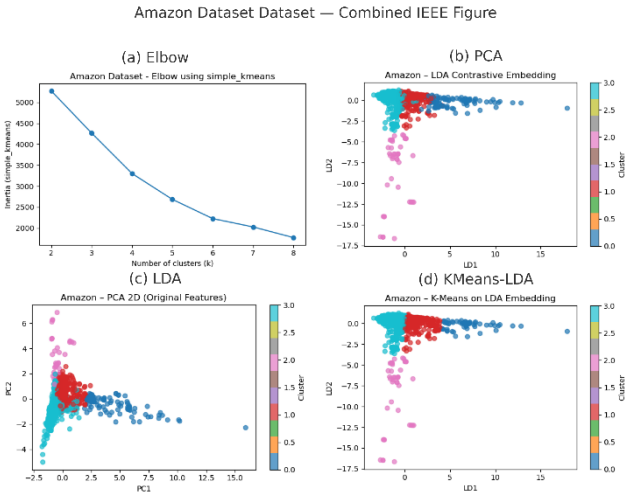

Fig. 2

C. *Flipkart Laptop Dataset (Fig. 3)*

Elbow Curve (Fig.3a). The optimal number of clusters is k = 3, representing: Budget laptops, Mid-range laptops, High-end premium laptops. PCA Visualization (Fig.3b). The PCA projection shows moderate separation, but clusters remain overlapping, especially where pricing ranges overlap. LDA Projection (Fig.3c). LDA reveals stronger separation between budget vs. premium price groups. K-Means on LDA (Fig.3d) Cluster compactness improves, reflected by silhouette increase (0.212 → 0.411). These results indicate that Contrastive learning helps separate devices by pricing tiers and rating popularity more clearly.

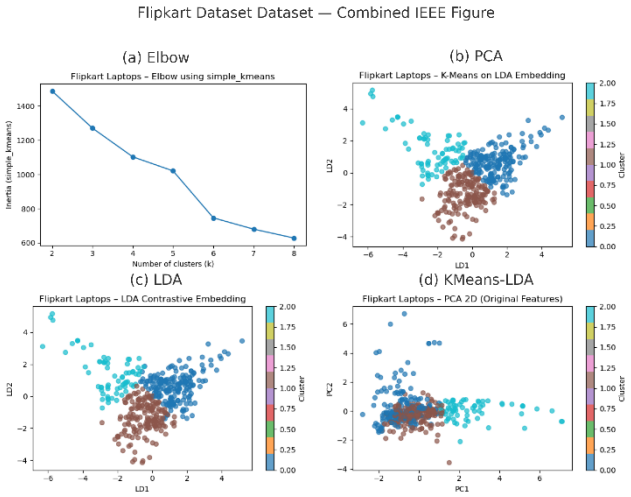

Fig. 3

D. *Walmart 45-Store Sales Dataset (Fig. 4)*

Elbow Curve (Fig.4a). The elbow indicates k = 3 clusters, representing different weekly sales patterns. PCA Visualization (Fig.4b).The PCA projection shows strongly overlapping clusters due to large variance in weekly sales across stores and seasons. LDA Projection (Fig.4c). After LDA contrastive embedding, clusters become much more separated, showing strong underlying structure driven by holiday flags, temperature, and CPI. K-Means on LDA (Fig. 4d). Cluster separation becomes extremely clear, achieving the highest silhouette gain (0.204 → 0.531). These results indicate that. Contrastive learning uncovers hidden structure in retail sales trends, making it easier to segment weekly performance patterns across stores.

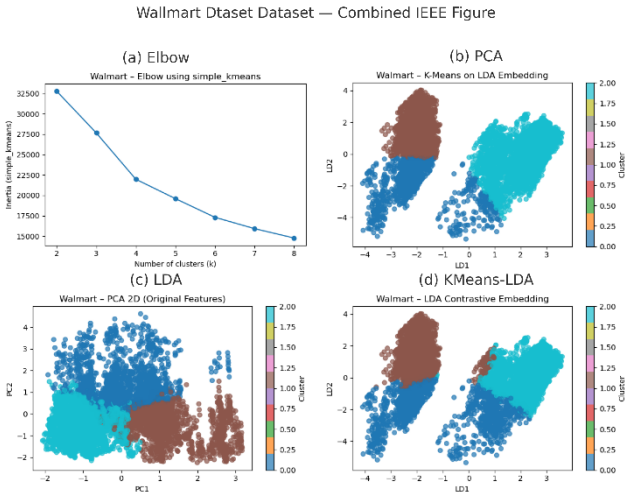

Fig. 4

V. Discussion

The experiments show that basic contrastive pipeline K-Means followed by LDA using pseudo-labels achieves better clustering results for different retail datasets even without using deep neural networks. The results highlight several insights:

1. Effectiveness of pseudo-labeling: The combination of unsupervised K-Means labels with LDA supervision leads to better cluster boundaries through optimized direction finding for maximum between-cluster variance.
2. Influence of dataset characteristics: The most significant improvements occur with datasets that have multiple related variables such as Walmart sales data because linear discriminant projections create better distinctions between latent structures. The re-clustering process produces positive results for low-dimensional datasets like Mall Customers although these results are less substantial than for other datasets.
3. Model interpretability: Unlike deep contrastive models, LDA provides a transparent linear mapping, enabling domain experts to interpret discriminant directions (e.g., discount–rating interactions in Amazon and Flipkart datasets).
4. Computational efficiency: The pipeline operates on CPU technology to provide maximum efficiency for educational and retail environments that need low resource consumption.

VI. Conclusion and Future Work

This study presented a comparison between K-Means clustering and a basic contrastive embedding method through four retail datasets which include Mall Customers and Amazon product sales and Flipkart laptops and Walmart weekly store sales. The proposed method uses K-Means cluster assignments as pseudo-labels to train LDA for learning a low-dimensional embedding that produces compact clusters and distinct cluster separation. Across three of the four datasets, K-Means applied in this contrastive space significantly improved silhouette scores compared to the original feature space, with the largest improvement observed on the Walmart dataset. The contrastive space transformation of K-Means shows better silhouette scores than the original feature space in three out of four datasets with Walmart producing the most substantial improvement.

The research will proceed to study non-linear contrastive embeddings which kernel LDA and shallow neural networks will help generate and test clustering algorithms including Gaussian Mixture Models and DBSCAN and integrate temporal structures for time-series retail data. The method shows potential for development into multi-view data analysis because it could integrate text reviews with numerical features.

References

[1] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020, November). A simple framework for contrastive learning of visual representations. In International conference on machine learning (pp. 1597-1607). PmLR.

[2] S. Kim and H. Jang, "LinkFND: Simple Framework for False Negative Detection in Recommendation Tasks With Graph Contrastive Learning," in IEEE Access, vol. 11, pp. 145308-145319, 2023, doi: 10.1109/ACCESS.2023.3345338.

[3] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] S. Arora, H. Khandeparkar, M. Khodak, O. Plevrakis, and N. Saunshi, “A Theoretical Analysis of Contrastive Unsupervised Representation Learning,” *arXiv.org*, Feb. 25, 2019. https://arxiv.org/abs/1902.09229

[5] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, “Unsupervised Learning of Visual Features by Contrasting Cluster Assignments,” arXiv:2006.09882 [cs], Jan. 2021, Available: https://arxiv.org/abs/2006.09882

[6] J. Bae and J. Shin, "Handling Imbalanced Medical Dataset with Continuous Class Features using Improved Contrastive Learning," 2025 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Fukuoka, Japan, 2025, pp. 0656-0660, doi:10.1109/ICAIIC64266.2025.10920651.

[7] S. Laohakiat, S. Phimoltares, and C. Lursinsap, “A clustering algorithm for stream data with LDA-based unsupervised localized dimension reduction,” *Information Sciences*, vol. 381, pp. 104–123, Mar. 2017, doi: https://doi.org/10.1016/j.ins.2016.11.018.