A Comparative Study of Clustering Techniques on an Online Marketplace Dataset

Dr N.Prabhakaran

Associate Professor, Department of Computer Science & Engineering, Anurag Engineering college, kodada, suryapeta, Telangana.

Submitted: 19-04-2025 *Accepted: 16-05-2025* *Published: 23-05-2025*

Abstract

Online marketplaces increasingly rely on large volumes of textual metadata to describe products, services, and digital assets. Inconsistent or manual categorization of such content often leads to reduced discoverability and suboptimal user experience. This paper presents a comparative analysis of prominent clustering techniques applied to an online marketplace dataset consisting of text-rich item descriptions. Two vectorization strategies—Term Frequency–Inverse Document Frequency (TF-IDF) and sentence-level transformer embeddings—are evaluated in combination with K-Means, HDBSCAN, and Agglomerative Clustering algorithms. Cluster quality is assessed using quantitative metrics, including Silhouette Score and Davies–Bouldin Index, supported by qualitative semantic interpretation and visualization. Experimental results demonstrate that sentence embeddings combined with density-based clustering yield more coherent and semantically meaningful groupings, highlighting their suitability for automated marketplace organization.

Keywords: Clustering techniques, Online marketplaces, Sentence embeddings, HDBSCAN, Text analytics

I. Introduction

Online marketplaces such as application stores, browser extension repositories, and creative tool platforms host thousands of items described primarily through textual metadata. As these platforms scale, manual categorization becomes impractical and prone to inconsistency. Poorly structured categories hinder search effectiveness, reduce user satisfaction, and negatively impact content visibility.

Unsupervised machine learning techniques, particularly clustering, offer a promising approach to automatically organizing such content. By grouping items based on textual similarity, clustering can reveal latent structures within datasets and support scalable content management. This study investigates how different combinations of text representation methods and clustering algorithms perform when applied to an online marketplace dataset. The objective is to identify approaches that balance computational efficiency with semantic interpretability.

II. Related Work

Clustering has been extensively studied in the contexts of document classification, customer segmentation, and product analysis. Prior research has explored traditional algorithms such as K-Means and hierarchical clustering for retail and e-commerce data. Recent advances in natural language processing have further enabled the use of contextual embeddings for representing textual data.

Studies focusing on marketplace analytics commonly rely on keyword-based representations, which capture surface-level similarity but often fail to reflect semantic relationships. Density-based approaches such as DBSCAN and HDBSCAN have shown promise in handling noisy, real-world datasets with irregular cluster structures. However, comparative evaluations of

clustering techniques on text-heavy marketplace metadata remain limited, particularly with respect to modern transformer-based embeddings.

III. Dataset Description

The dataset used in this study comprises approximately 300 marketplace items. Each item is described using multiple textual attributes, including title, short summary, detailed description, and a set of keywords. Additionally, a binary attribute indicates the presence or absence of a specific feature.

All textual fields were concatenated into a single document per item to create a unified representation. Standard preprocessing steps were applied, including lowercasing, punctuation removal, and stop-word elimination. This preprocessing aimed to reduce noise while preserving informative content relevant for clustering tasks.

Table 1: Dataset Attributes

Attribute	Description
Title	Name of the marketplace item
Summary	Brief overview of functionality
Description	Detailed textual description
Keywords	Developer-provided tags
Feature Flag	Binary indicator of a specific feature

IV. Methodology

The proposed methodology consists of three main stages: text vectorization, clustering, and evaluation.

A. Text Vectorization

Two distinct vectorization strategies were employed to transform textual data into numerical representations.

1) TF-IDF Representation

TF-IDF captures the importance of words relative to their frequency across the dataset. The Scikit-learn TfidfVectorizer was used with a maximum vocabulary size of 5,000 features to maintain a balance between expressiveness and computational efficiency.

2) Sentence Embeddings

To capture semantic context, sentence embeddings were generated using a pre-trained transformer model (all-MiniLM-L6-v2). This approach encodes entire text sequences into dense vectors that reflect contextual meaning beyond individual word frequencies.

Table 2: Comparison of Text Representation Techniques

B. Clustering Algorithms

Method	Dimensionality	Semantic Awareness	Sparsity
TF-IDF	High	Low	Sparse
Sentence Embeddings	Fixed (dense)	High	Dense

Three clustering algorithms were applied independently to both TF-IDF and embedding-based representations.

1)K-Means

K-Means partitions data into a predefined number of clusters by minimizing intra-cluster variance. In this study, the number of clusters was fixed at 10 to ensure consistency across experiments.

2)HDBSCAN

HDBSCAN is a density-based clustering algorithm capable of identifying clusters of varying shapes and densities without requiring prior specification of cluster count. It also labels low-density points as noise, making it robust to outliers.

3)AgglomerativeClustering

Agglomerative clustering follows a bottom-up hierarchical approach, merging the most similar clusters iteratively. Average linkage was used to balance sensitivity to cluster shape and size.

Table 3: Characteristics of Evaluated Clustering Algorithms

Algorithm	Cluster Shape	Noise Handling	Requires k
K-Means	Spherical	No	Yes
HDBSCAN	Arbitrary	Yes	No
Agglomerative	Hierarchical	Limited	Optional

C. Evaluation Metrics

Cluster quality was evaluated using both quantitative and qualitative methods.

- **Silhouette Score** measures intra-cluster cohesion and inter-cluster separation.
- **Davies–Bouldin Index** evaluates average similarity between clusters, where lower values indicate better separation.

Additionally, UMAP was employed to project high-dimensional vectors into two dimensions for visual inspection of cluster structure.

V. Results and Discussion

A. Quantitative Evaluation

Clustering approaches using sentence embeddings consistently outperformed TF-IDF-based methods. HDBSCAN applied to embedding vectors achieved the highest Silhouette Score (0.5826), indicating strong cohesion and separation. K-Means produced comparatively low scores across representations, suggesting limited suitability for semantically complex text data.

Table 4: Clustering Performance Metrics

Representation	Algorithm	Silhouette Score	Davies–Bouldin
TF-IDF	K-Means	Low	High
TF-IDF	Agglomerative	Moderate	Moderate
Embeddings	K-Means	Low	High
Embeddings	HDBSCAN	High	Low

B. Qualitative Analysis

Embedding-based clusters demonstrated clear semantic coherence, naturally grouping items into intuitive categories such as accessibility tools, media enhancement utilities, and analytics extensions. TF-IDF clusters, while keyword-consistent, often lacked meaningful semantic alignment.

C. Discussion

The findings highlight the importance of aligning feature representation with clustering strategy. Sentence embeddings provide a semantically rich space that benefits density-based clustering. HDBSCAN, in particular, adapts well to the irregular structure and noise inherent in real-world marketplace data, making it a strong candidate for automated content organization.

VI. Conclusion

This study presents a systematic comparison of clustering techniques for organizing online marketplace items using textual metadata. Results indicate that embedding-based representations combined with density-based clustering significantly outperform traditional methods in both quantitative and qualitative evaluations. Future work will explore multilingual datasets, larger-scale experiments, and integration with recommendation systems to further enhance marketplace intelligence.

References

[1] Ulfa, A. H., Maulidina, S. N., & Chandra, H. (2021). Product clustering analysis on the marketplace using K-means approach. *Asian Journal of Science and Engineering*, 1(2), 73–78.

[2] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of EMNLP*, 3982–3992.

[3] Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. *PAKDD*, 160–172.

[4] Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2), 224–227.

[5] Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65.

[6] Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1), 21–27.

[7] Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering: An overview. *WIREs Data Mining and Knowledge Discovery*, 2(1), 86–97.