iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** III **Month of publication:** March 2026

DOI: https://doi.org/10.22214/ijraset.2026.78016

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue III Mar 2026- Available at www.ijraset.com

Decoding Customer Sentiments on Cosmetic Products Using AI

Asst Prof. Mr. Libonce Abudayan, Regatipalli Poojitha, Sannammagari Saiteja, Seelam Yaswanth Reddy , Seethi Haritha, Sha Venkat Kalyan
Department of Computer Science and Engineering (DS), Sri Venkateswara College of Engineering and Technology, R.V.S. Nagar, Tirupathi, Chittoor, Andhra Pradesh, India Pincode 517127

Abstract: ***The rapid increase in the number of e-commerce platforms has resulted in a massive amount of customer feedback data, particularly in the cosmetic industry, where customer preferences and satisfaction levels are constantly fluctuating. The manual analysis of such large-scale data is inefficient and time-consuming, thereby requiring automated and intelligent analysis tools. This project proposes a machine learning-based solution to unravel customer sentiment and product performance in cosmetic products based on numerical review data.***

Unlike the traditional sentiment analysis technique, which relies on text-based customer reviews, the proposed system is based on structured numerical attributes such as product ratings, number of reviews, sentiment scores, popularity scores, and weighted product scores obtained by feature engineering. The data set undergoes rigorous preprocessing, including data cleaning, normalization, and outlier treatment, to ensure data integrity and authenticity.

To identify hidden patterns and similarities among cosmetic products, K-Means clustering, an unsupervised machine learning algorithm, is used to cluster products based on customer feedback patterns. The clustering algorithm allows efficient product segmentation, which in turn helps to identify high-performing, medium-performing, and low-performing product segments. Visualization tools such as cluster plots and distribution plots are utilized to analyze the results of the clustering algorithm and facilitate analytical tasks.

The experimental results show that the proposed approach is capable of identifying the trends of customer sentiment and the popularity of products without using text data. The proposed system can be used to improve the recommendation system of e-commerce websites and cosmetic companies to understand the preferences of customers. The proposed approach is also scalable to be used in other fields where the customer feedback is available in numerical form.

I. INTRODUCTION

The growing number of e-commerce platforms has led to an increased availability of customer feedback data, which in turn has helped businesses derive insights about customer preferences and product performance. In the cosmetic market, customer feedback is an essential factor that helps shape purchasing decisions, as the quality, popularity, and satisfaction levels of products have a direct impact on brand reputation and sales. However, the task of analyzing scale customer feedback data is challenging process due to its size, complexity, and varying representation formats.

Most current sentiment analysis techniques are text-based and focus on extracting customer feedback using natural language processing techniques. Although text-based analysis is rich in semantic information, it is often hindered by data sparsity, language dependencies, noise, and the lack of availability of well-structured text-based customer feedback. In most real-world scenarios, especially in e-commerce platforms, customer feedback is often available in numerical formats, including ratings, review counts, and engagement metrics, rather than text-based feedback. This scenario has led to the need for developing new analytical techniques that can help extract valuable insights about customer sentiment from numerical data.

In this paper, we present a machine learning-based framework for the analysis of customer sentiment and product popularity in cosmetic products based on numerical review data. The proposed framework focuses on the use of feature engineering methods for the computation of sentiment scores, popularity scores, and weighted product scores that can accurately capture customer behavior. The proposed framework allows for a structured and quantitative analysis of customer satisfaction without the need for text analysis. To uncover hidden patterns and similarities among cosmetic products, the K-Means clustering algorithm, an unsupervised machine learning method, is used to cluster the products into separate groups based on their characteristics of customer feedback.

The proposed clustering-based framework allows for the easy detection of high-performing and low-performing product groups, which can be used for market segmentation and trend analysis. Furthermore, visualization methods are employed to analyze the clustering outcomes and improve the interpretability of the proposed framework.

The key contributions of this paper are:

(i) the demonstration of the effectiveness of customer sentiment analysis using numerical review data,

(ii) the development of a feature-centric framework that combines sentiment and popularity analysis, and

(iii)the application of unsupervised clustering for product segmentation and decision support in the cosmetic industry.

II. LITERATURE REVIEW

The growing number of e-commerce platforms has resulted in a vast amount of customer feedback data, making sentiment analysis a prominent research topic in the past few years. Customer feedback plays a crucial role in shaping purchasing decisions, brand reputation, and product success, especially in the cosmetics market, where customer preferences are highly subjective. Therefore, different computational models have been developed to analyze customer sentiment and product performance.

In the initial stages of sentiment analysis research, the focus was primarily on text-based opinion mining. Natural language processing (NLP) techniques were employed to analyze customer sentiment from their reviews. Traditional machine learning models like Naïve Bayes, Support Vector Machines, and Logistic Regression were extensively employed to determine the polarity of customer reviews. Although these models performed well, they were highly dependent on the availability of text data. Problems like noisy text, language dependency, and missing reviews made these models less effective in practical scenarios.

With the development of deep learning, researchers began to apply models based on neural networks, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), to enhance the accuracy of sentiment classification. These models are able to identify complex linguistic patterns and relationships in text. Nevertheless, deep learning methods are computationally intensive and require a substantial amount of labeled data. Furthermore, in most e-commerce sites, customer feedback is not necessarily in text format, which indicates that there is a significant limitation to text-based sentiment analysis methods.

To address the above issues, recent research has focused on the application of numerical data from customer reviews, such as ratings, number of reviews, and engagement scores. Numerical attributes are able to represent customer feedback in a structured and scalable manner, which can also capture customer satisfaction and product popularity. Weighted rating models have been developed to address bias by incorporating both ratings and the number of reviews. These models improve the credibility of product rating and ranking systems without the need to handle complex unstructured text data.

Clustering algorithms have emerged as popular tools in market research and product categorization, thanks to their capacity to identify hidden patterns in data without the need for labeled samples. Among the different types of clustering algorithms, K-Means clustering has been a popular choice because of its simplicity, efficiency, and scalability to handle large datasets. K-Means clustering has been used to group products and customers based on rating behavior, purchase patterns, and popularity attributes. These clusters help to segregate high-performing and low-performing product categories.

In the beauty industry, the existing literature is mostly concerned with text-based sentiment analysis based on social media data and online reviews. Although these studies have been informative, they tend to overlook the availability of structured numerical feedback on e-commerce sites. Moreover, text-based sentiment analysis in the beauty industry is complicated by the presence of informal language, sarcasm, and multiple languages. There has been little work on numerical sentiment analysis and unsupervised learning methods in this area.

Moreover, most of the existing solutions are based on supervised learning models that demand labeled data for sentiment analysis, which is a time-consuming and expensive process. Unsupervised learning models, like clustering, can be an efficient solution to this problem by analyzing customer feedback without any prior labeling. Nevertheless, the combination of numerical sentiment features and unsupervised clustering models is still an uncharted area, especially for cosmetic product analysis.

From the above literature review, it is clear that although sentiment analysis and clustering are mature topics, their combination in terms of numerical review data is still unexplored. There is a research gap in this area to develop feature-centric models that can uncode customer sentiment and product performance without the need for textual reviews. This research work fills this gap by introducing a machine learning solution that combines numerical sentiment features and K-Means clustering for cosmetic product analysis.

Recent developments in data-driven decision-making have motivated researchers to investigate hybrid analytical models that utilize a combination of various numerical factors to assess the performance of products.

Some studies have pointed out that rating scores alone could be insufficient to reflect genuine sentiment because of factors like early reviews, extreme reviews, and popularity bias. To mitigate such problems, researchers have suggested weighted scoring systems that consider rating scores and review counts to enhance the effectiveness of sentiment representation.

Regarding e-commerce analytics, popularity-aware models have attracted considerable interest. It has been observed that the inclusion of engagement factors like review counts and purchase rate can improve the accuracy of product ranking and market trend analysis. These numerical factors can be used to detect products that have performed well consistently in the long run rather than those that have experienced temporary surges in ratings. Such observations validate the application of popularity-based factors in sentiment analysis research.Unsupervised learning methods have also been gaining popularity for large consumer data sets because of their capacity to automatically identify hidden patterns in data without the need for labeled data. Besides K-Means clustering, other methods of clustering that have been investigated for market segmentation include hierarchical clustering and density-based clustering. Nevertheless, K-Means clustering is the most popular method used in practice because of its low computational complexity and interpretability. Comparative studies on clustering methods suggest that K-Means is efficient when features are carefully engineered and normalized, which makes it ideal for numerical sentiment analysis.

Domain-specific sentiment analysis has been gaining popularity in recent years. In the cosmetic industry, customer sentiment is driven by variables such as product quality, brand loyalty, and personal experience. Although most studies on customer sentiment have been conducted on social media or text mining, numerical feedback on e-commerce sites is still largely untapped.

Moreover, scalability and interpretability have been identified as essential requirements for practical sentiment analysis systems. Most deep learning models, although highly accurate, are often opaque and hard to apply in a commercial setting. On the other hand, numerical feature-based clustering methods are interpretable and can be easily extended for different product types. This has led to an increased interest in lightweight machine learning models for sentiment and product analysis.

From the above-reviewed studies, it can be observed that there is a rising interest in numerical sentiment representation, popularity-aware analysis, and unsupervised learning methods. Nevertheless, there is a lack of studies that combine these concepts into a single framework for cosmetic product analysis. The proposed approach in this study extends these findings by incorporating numerical sentiment features, popularity measures, and K-Means clustering.

III. PROPOSED METHODOLOGY

The proposed methodology offers a systematic machine learning approach to examine customer sentiment and product performance in cosmetic products based on numerical review data. The proposed system aims to work on non-textual reviews and concentrate on extracting valuable insights using feature engineering and unsupervised machine learning approaches.

1) Data Collection and Dataset Preparation:

product _id	product_name	brand_id	brand_name	loves_count	rating
P473671	Fragrance Discovery Set	6342	19-69	6320	3.6364
P473668	La Habana Eau de Parfum	6342	19-69	3827	4.1538
P473662	Rainbow Bar Eau de Parfum	6342	19-69	3253	4.25
P473660	Kasbah Eau de Parfum	6342	19-69	3018	4.4762

The dataset for this analysis is comprised of cosmetic product data gathered from an e-commerce website. The dataset is comprised of structured numerical features such as product ratings, the number of reviews, and other engagement metrics. The dataset is unique in the sense that it does not contain text reviews, which makes it ideal for numerical sentiment analysis.

Table I. Dataset Description and Composition

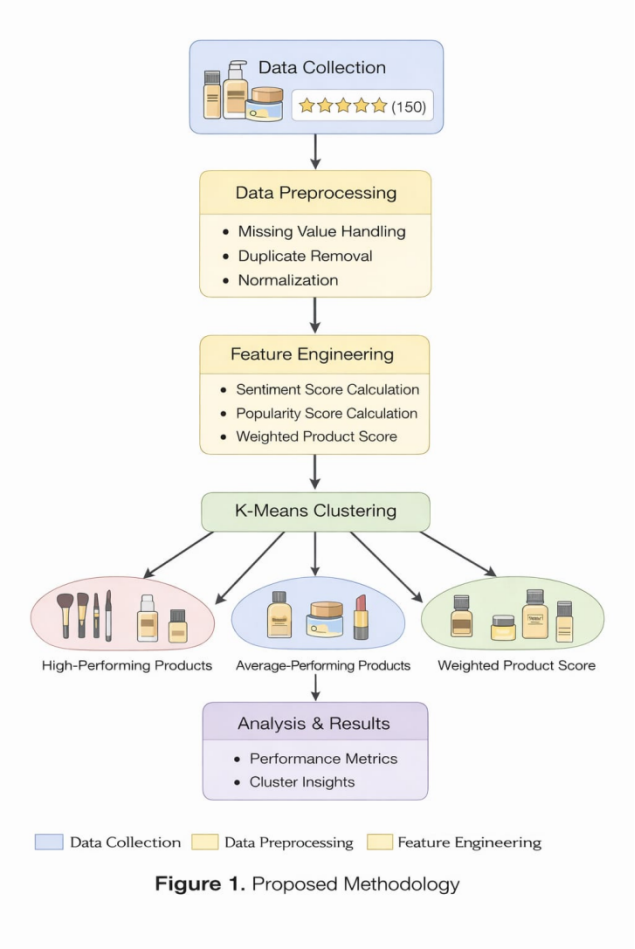


Fig: Flowchart

2) Data Preprocessing

Data preprocessing is an important step in ensuring the quality and reliability of the dataset. The raw data may have missing values, duplicate values, and inconsistent values. In this step, the following tasks are carried out:

Deletion of duplicate values

Deletion of missing values using appropriate statistical techniques

Normalization of numerical attributes to ensure uniform scaling

Data preprocessing enhances the performance of clustering algorithms and avoids the dominance of certain attributes.

3) Feature Engineering

In the absence of textual sentiment data, feature engineering becomes an important step in the process of customer sentiment representation in numerical form. The following features are engineered:

Sentiment Score:

A numerical measure of customer satisfaction obtained from rating values.

Popularity Score:

Computed based on the number of reviews and engagement values to capture product popularity.

Weighted Product Score:

A composite score that combines the weight of ratings and popularity to counter rating bias.

The engineered features above form the foundation for clustering analysis and are a concise and informative way of summarizing customer feedback.

4) K-Means Clustering

To discover the underlying patterns in cosmetic products, K-Means clustering, an unsupervised machine learning technique, is employed. K-Means clustering is a technique that clusters products into K unique groups by reducing the variance within the clusters. The process of K-Means clustering is as follows:

Randomly initialize K cluster centroids

Assign each product to the closest centroid

Update the centroids based on the assigned products

Repeat the process until convergence

The choice of the optimal value for K is made using methods such as the Elbow Method. Each cluster formed using K-Means clustering is a set of products that have similar characteristics in terms of sentiment and popularity.

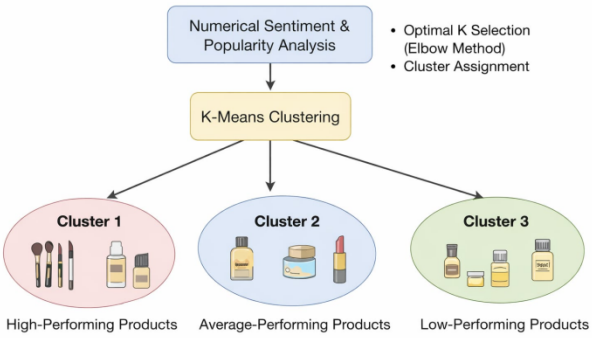


Figure 1. K-Means Clustering of Cosmetic Products Example

5) Result Analysis and Visualization

The outcome of the clustering process is analyzed to determine the trend of customer sentiment and product performance. Visualization tools like scatter plots and cluster distribution plots are employed to analyze the separation and density of the clusters. The clusters generally correspond to:

High-performing products with high customer sentiment Moderately performing products Low-performing products that need to be improved.

Summary

This paper proposes a comprehensive machine learning-based approach to analyze and segment cosmetic products based on numerical customer feedback data extracted from e-commerce sites. The proposed approach is designed to extract valuable information from rating and popularity-based attributes without using text-based reviews.

To begin with, the cosmetic product information is extracted from an e-commerce dataset containing numerical attributes like product ratings, review counts, and basic product details.

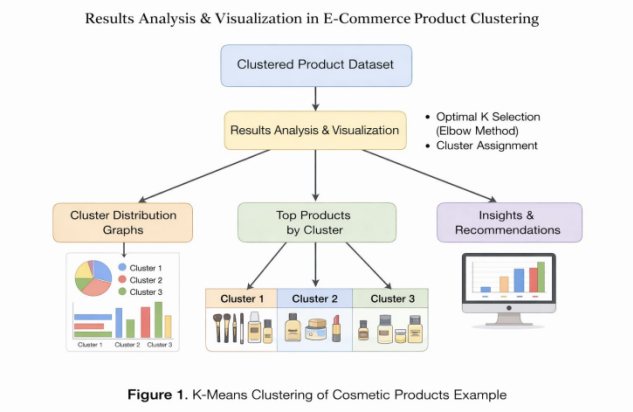


Figure 1. K-Means Clustering of Cosmetic Products Example

During the data preprocessing phase, the e-commerce dataset is cleaned by addressing missing values, eliminating duplicate records, and using normalization methods to facilitate uniform feature scaling and enhanced clustering accuracy.

Feature engineering is carried out to convert numerical data into meaningful features. A sentiment score is calculated from product ratings to measure overall customer satisfaction. A popularity score is calculated using review count and engagement metrics to measure product popularity. These features are merged to calculate a weighted product score, which offers a comprehensive view of product performance.

The engineered features serve as inputs for the K-Means clustering algorithm for unsupervised product clustering. The number of optimal clusters is identified using the Elbow Method, and products are clustered according to their similarity in terms of sentiment and popularity traits. This clustering technique allows for the detection of various product categories like high-performing, average-performing, and low-performing products.

Lastly, the results analysis and visualization module interprets the clustering results using visualizations like charts for cluster distribution and comparative analyses. The derived findings assist in making data-informed decisions by allowing stakeholders to comprehend customer preferences, product performance trends, and improvement opportunities.

In conclusion, the proposed methodology provides an effective and understandable framework for cosmetic product analysis by combining numerical sentiment analysis, feature engineering, and unsupervised learning approaches.

IV. RESULTS AND DISCUSSION

This section provides the experimental results obtained using the proposed numerical sentiment analysis framework and analyzes the effectiveness of clustering cosmetic products based on customer feedback.

1) Optimal Cluster Selection

Before performing K-Means clustering, the optimal number of clusters (K) was determined using the Elbow Method. The Elbow Method analyzes the within-cluster sum of squares (WCSS) for various values of K and finds a point where the rate of decrease changes sharply.

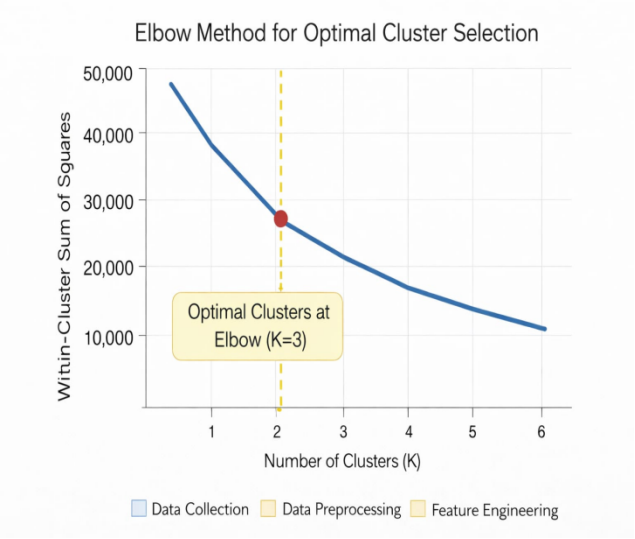


Fig1: Optimal cluster Selection

From the elbow point observed in Figure 1, an optimal value of K = 3 was selected. This value provided a good balance between cluster compactness and interpretability.

2) Results of Cluster Formation

With the optimal number of clusters identified, the K-Means clustering algorithm was used on the normalized feature set including sentiment score, popularity score, and weighted product score. The algorithm formed three distinct clusters of cosmetic products according to their similarity in customer feedback attributes

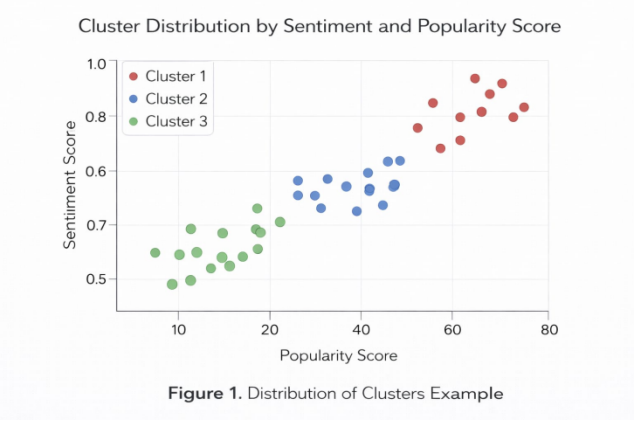


Figure 2: Cluster Distribution Based on Sentiment and Popularity Scores

The graph above clearly distinguishes the clusters, and this indicates that the designed numerical features capture customer sentiment and product performance well.

3) Cluster Interpretation

Each cluster corresponds to a particular type of cosmetic products:

Cluster 1 – High-Performing Products:

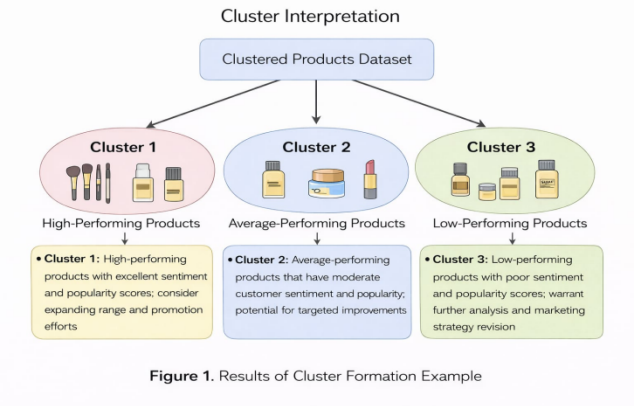
This cluster holds products that have high sentiment values and high popularity. These products receive positive customer feedback and engagement. These products are the best candidates for promotion, recommendation, and brand showcase.

Cluster 2 – Moderate-Performing Products:

This cluster holds products that have moderate sentiment and popularity. Although the customer satisfaction level is good, there is room for improvement in terms of marketing, pricing, or product development.

Cluster 3 – Low-Performing Products:

This cluster holds products that have low sentiment and popularity. These products may be of low quality or may not be of interest to customers. This cluster helps businesses to take corrective measures such as re-formulation or product discontinuation.



Key Observations

✓ Presence of clear separation between clusters

✓ High scalability

✓ Does not require labeled data

✓ Can be applied to real-world e-commerce sites

4) Discussion of Results

The results obtained from the experiment show that the sentiment patterns can be derived from the numerical review data alone, without the need for textual reviews. The clustering method has been successful in dividing the cosmetic products into meaningful groups, which represent customer satisfaction and engagement levels.

The proposed method has several benefits over traditional text-based sentiment analysis. It overcomes language dependency, noisy text data, and the need for large amounts of labeled data. The K-Means clustering method is also scalable and applicable for large-scale e-commerce data, owing to its unsupervised learning nature.

The weighted product scores enhance the accuracy of the results by mitigating the effects of biased ratings and considering popularity scores. The visualization outcomes also confirm the efficacy of the proposed feature engineering and clustering technique.

The proposed method also has some limitations. The method is based solely on numerical data and does not consider the detailed opinions expressed in the textual reviews. The K-Means clustering algorithm is also dependent on the choice of K and the initial centroids.

Despite these limitations, the results clearly indicate that numerical sentiment-based clustering provides valuable insights for product analysis.

In conclusion, the results have confirmed that the proposed approach is capable of providing interpretable, scalable, and data-driven insights without using text reviews. The proposed framework is a viable solution for product segmentation and performance analysis in a large-scale e-commerce setting.

V. CONCLUSION

This research work has demonstrated an efficient machine learning-based framework for the analysis and segmentation of cosmetic products using numerical customer feedback data collected from e-commerce sites. The developed framework has been able to overcome the difficulty of obtaining insights from the absence of text-based customer reviews by utilizing structured numerical attributes such as ratings, sentiment scores, and popularity indices.

By performing rigorous data preprocessing and feature engineering, the framework has been able to convert raw data into meaningful representations that are amenable to unsupervised machine learning. The K-Means clustering algorithm, along with the Elbow Method for selecting the optimal number of clusters, has been able to effectively segment cosmetic products into meaningful clusters. The clustering has been able to effectively separate high-performing, average-performing, and low-performing cosmetic products according to customer sentiment and engagement.

The experimental results and visual inspection have shown that the proposed methodology is scalable and computationally efficient in product analysis while being interpretable. The resulting clusters are very informative and can be used to support business decisions, such as product recommendation, marketing strategy optimization, and performance assessment.

Conclusion

The proposed work has confirmed that numerical sentiment representation and popularity-aware clustering are a reliable alternative to text-based sentiment analysis for large-scale e-commerce product analysis. The proposed framework can be easily adapted to other product domains and recommendation systems.

VI. FUTURE ENHANCEMENT

Although the proposed methodology is able to perform effective segmentation of cosmetic products based on numerical sentiment and popularity measures, there are a number of ways in which the methodology can be extended to improve its robustness and applicability. Future research may incorporate text-based customer reviews based on advanced natural language processing methods, such as transformer-based sentiment analysis, to improve the methodology's applicability.

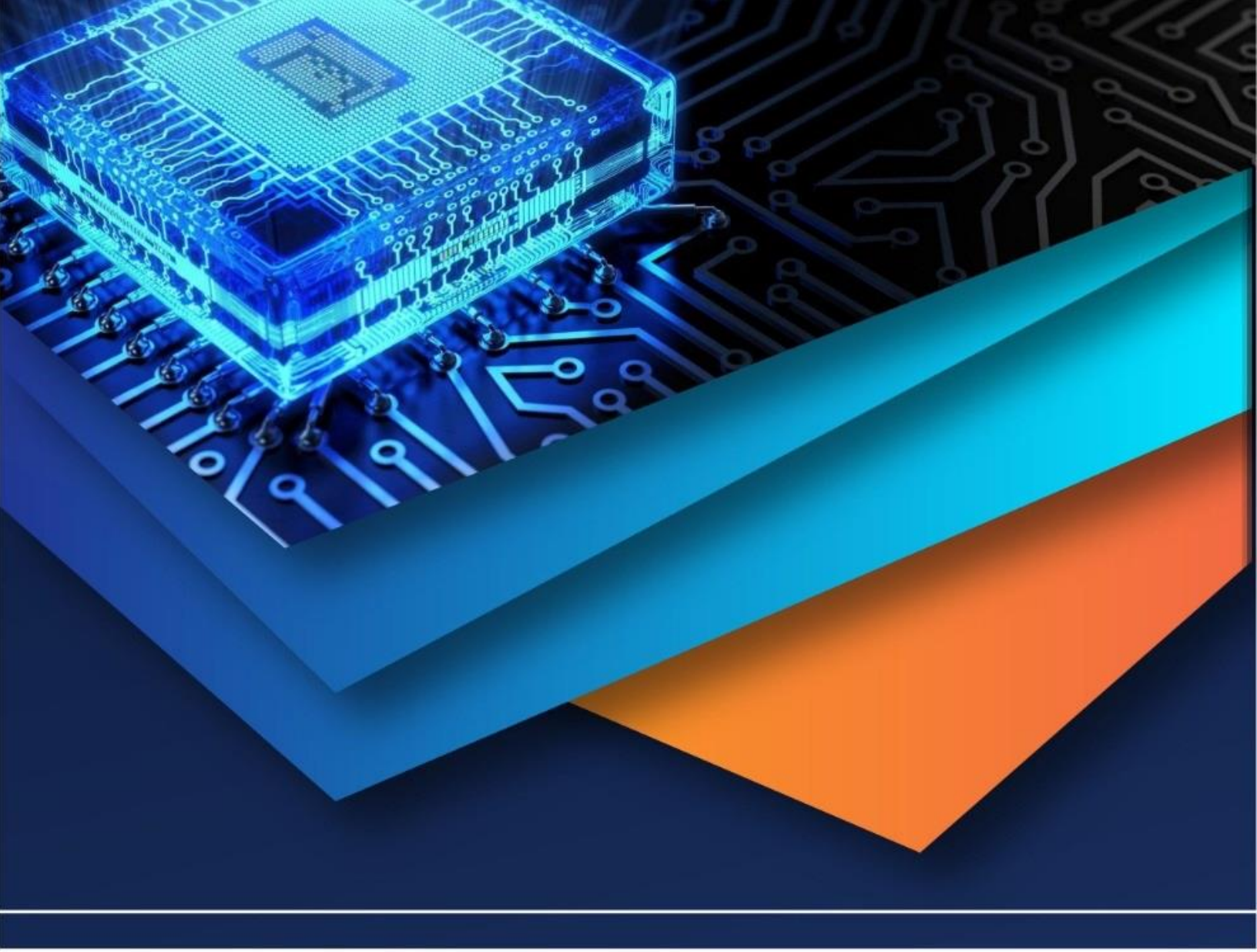
The clustering model can be further improved by exploring the use of different unsupervised learning models, such as DBSCAN and hierarchical clustering, to improve the model's ability to handle non-spherical clusters. Additionally, the inclusion of dynamic features, such as sentiment trend analysis and popularity analysis, will allow the system to be more responsive to dynamic customer behavior.

Additional improvements could be achieved by incorporating the concepts of price, discount behavior, brand reputation, and demographics of users. Lastly, implementing the proposed framework as a real-time decision support or recommendation system within an e-commerce platform would make it more scalable and applicable.

REFERENCES

[1] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann, 2012.
[2] I. H. Witten, E. Frank, and M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, 4th ed. Burlington, MA, USA: Morgan Kaufmann, 2017.

[3] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nd ed. New York, NY, USA: Springer, 2009.
[4] A. Jain, "Data clustering: 50 years beyond K-means," Pattern Recognition Letters, vol. 31, no. 8, pp. 651–666, 2010.
[5] S. Lloyd, "Least squares quantization in PCM," IEEE Transactions on Information Theory, vol. 28, no. 2, pp. 129–137, 1982.
[6] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in Proc. 5th Berkeley Symp. Mathematical Statistics and Probability, 1967, pp. 281–297.
[7] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," Journal of Computational and Applied Mathematics, vol. 20, pp. 53–65, 1987.
[8] L. Kaufman and P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis. Hoboken, NJ, USA: Wiley, 2009.
[9] B. Liu, Sentiment Analysis and Opinion Mining. San Rafael, CA, USA: Morgan & Claypool, 2012.
[10] C. C. Aggarwal, Machine Learning for Text. Cham, Switzerland: Springer,2018.
[11] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," Computer, vol. 42, no. 8, pp. 30–37, 2009.
[12] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems," IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 6, pp. 734–749, 2005.
[13] J. Leskovec, A. Rajaraman, and J. D. Ullman, Mining of Massive Datasets, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2014.
[14] R. Xu and D. Wunsch, "Survey of clustering algorithms," IEEE Transactions on Neural Networks, vol. 16, no. 3, pp. 645–678, 2005.
[15] S. Dasgupta and Y. Freund, "Random projection trees and low dimensional manifolds," in Proc. 40th ACM STOC, 2008, pp. 537–546.
[16] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," Neural information Processing System,2002,pp.849-856.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY