Course Title: BEMM466 Business Project

Module Convenor: Dr Stuart So

University *of* Exeter

Title: Dissertation - A Hybrid Product Recommendation Model Based on Sentiment Analysis and Product Clustering

Date of Submission: 29.08.2024

Author: Mukundh Srikanth (ms1452@exeter.ac.uk)

Student ID: 730038980

Acknowledgements

Research is a journey that one undertakes with a thought and an interest in a topic, driven by the desire to learn more about it. Along the path, one confronts tremendous confusion, frustration, and roadblocks at first, but eventually, the fruit of hard work and consistency is sweet. Even though the idea originates in the researcher's mind, do not for once think that producing a work of quality is possible without the help, support, and guidance of others.

First and foremost, I would like to thank *you*, the reader, for taking the time to read this work. A great deal of time and effort was spent performing this research, and it would go in vain if not for you taking the time out of your busy schedule to read, understand, and contemplate this work. As a researcher, my aim is to ensure that this work is engaging throughout and worthy of your attention.

I would like to thank Dr. Stuart So, my professor, who has guided me every step of the way. He has been instrumental in helping me structure my thoughts and devise actionable steps that could be taken to make incremental progress throughout this journey. His key insights and experience have guided me in producing quality work and have helped me avoid common pitfalls and traps that a researcher may be prone to.

Next, I would like to thank Miss G.V. Kameshwari. She has been my inspiration and my rock throughout this journey, someone with whom I could always bounce my ideas off and have productive discussions. She, herself being an accomplished individual in the scientific community, has shown me the tools and techniques to conduct research in a structured way, and without her contributions and support, none of this would have been possible.

A special mention to all my well-wishers, just to name a few, Mr. Gokul K. Nair, Mr. Roshan Shekhar, and Mr. Karthik Narayanan, who have provided me with the necessary motivation and an outlet for much-needed breaks when the journey became daunting and during times of pressure.

I would especially like to thank my family, Mr. and Mrs. Srikanth Natarajan and Miss Lavanya Srikanth, whose love, support, and hard work afforded me the opportunity to pursue a master’s degree. It is due to the foundation they have given me that I am able to utilize my faculties in this productive manner and produce this work of research. I will always be indebted to them.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

RS – Recommender Systems

CFRBCS - Collaborative Filtering Recommendation algorithm Based on Clustering and Sentiment analysis

CFRBS - Collaborative Filtering Recommendation algorithm Based Sentiment analysis

e-commerce – Electronic Commerce

UBCF - User-Based Collaborative Filtering

IBCF - Item-Based Collaborative Filtering

CIBCF - Context-similarity Item-Based Collaborative Filtering recommender model

VADER—Valence Aware Dictionary and Sentiment Reasoner

SVM – Support Vector Machine

CNN – Convolutional Neural Network

RNN - Recurrent Neural Network

MMN – Min-Max Normalisation

WCSS – Within Cluster Sum of Squares

PCA – Principal Component Analysis

Glossary

Term	**Definition**
Recommendation engine / Recommender systems / Recommendation systems / Recommendation models (Used interchangeably in the work)	*A platform, engine, or algorithm that is a subclass of information filtering system and provides suggestions for items that are most pertinent to a particular user.*
Sentiment Analysis	*The process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral.*
Product Clustering	*Product clustering groups similar products based on shared attributes like price, popularity, or features to identify patterns and insights.*
Cold Start	*The cold start problem in recommender systems occurs when there is insufficient data on new users or items, making it difficult to generate accurate recommendations.*
Sephora	*Sephora is a global beauty retailer known for its extensive selection of cosmetics, skincare, and personal care products.*
Similarity Matrix	*In item-based collaborative filtering, a similarity matrix is a grid that measures and represents the similarity between items based on user ratings or interactions.*
Hyperparameter Tuning	*Hyperparameter tuning is the process of systematically adjusting the settings/inputs of a machine learning algorithm to optimize its performance and improve predictive accuracy.*

Executive Summary – CFRBCS model: Advancing E-commerce Recommendations

Choices are abundant in today's digital marketplaces, making it feel almost like finding a needle in a haystack to discover that perfect product. It has become commonplace to experience confusion and frustration while scrolling for a specific item or service, as the suggestions received tend to be banal and impersonal. Recent trends indicate that 72% of consumers are more likely to purchase products from companies that offer them personalized recommendations or offers. On its part, this has led to growth in the recommendation engine market, forecasted to reach 15.13 billion USD by 2026 (Skovhøj, 2022). Companies are always on the lookout for data-driven ways to drive the personalization of their recommender systems, and new and innovative work is called for in this context.

The focus of this study is on increasing the significance and reliability of recommendation systems in e-commerce (Electronic Commerce). Specifically, the need is addressed through the introduction of the CFRBCS model: Collaborative Filtering Recommendation Algorithm Based on Clustering and Sentiment Analysis. In contrast to traditional systems that only make use of user ratings, the CFRBCS model explores ways to integrate multiple data dimensions to form a cohesive recommendation engine capable of making personalized suggestions.

Development of the CFRBCS model follows a hybrid, multistage methodology that begins with data collection and preprocessing. In this work, the research is based on the Sephora dataset containing more than 8,000 beauty products and nearly a million user reviews (Inky, 2023). The data is rigorously preprocesses for consistency and quality, after which the product data is segmented into meaningful clusters using the K-means algorithm, grouping similar products together based on their pricing and popularity. This is a crucial step in the process and serves the dual purpose of making the CFRBCS model more accurate in its predictions and providing valuable insights into the characteristics of the product inventory. Another facet of this work is applying sentiment analysis to gauge the polarity of customer feedback and integrating this information into the recommendation generation process.

The novelty of this research lies in the way the CFRBCS model calculates the comprehensive similarity between two products. Along with historical user ratings, the overall sentiment score and cluster information of the products are taken into consideration while assessing their similarity. This similarity score between products is a vital factor that recommender systems depend on when generating product recommendations.

The reliability of the CFRBCS model has been tested through rigorous statistical vetting, comparing it with the performance of other existing models. This comparative analysis included traditional recommendation systems that rely solely on user ratings, as well as other advanced methods such as the CFRBS (Collaborative Filtering Recommendation algorithm Based Sentiment analysis) model developed by Jian Zhen Yu et al. (2018). The evaluation metrics used encompassed several key indicators, including Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Mean Squared Logarithmic Error (MSLE). Results from these extensive tests show that the CFRBCS model has better overall accuracy and drastically reduced RMSE to an error rate of about 0.003, whereas for the traditional model and CFRBS model, it was around 0.007. Again, performance in MAE of the CFRBCS was above the others, having a value of about 0.00006, indicating that it reduced the average error by 40%, making its predictions more accurate. Higher accuracy in predictions implies that the model can better estimate a user's potential review score for a product, thereby

providing a stronger foundation for delivering personalized product recommendations that the user is highly likely to purchase.

The substantial practical implications and key findings of the research are discussed with regard to various stakeholders in the e-commerce sector. For instance, marketing professionals in organizations can utilize the product clusters discovered to devise targeted marketing campaigns. Product managers and analysts can explore ways to integrate the CFRBCS algorithm into their existing systems to generate personalized recommendations that reflect both product characteristics and user sentiment.

The research also covers ethical considerations in the case of emerging technologies like AI, referencing guidelines that put forward responsible AI practices (European Commission, 2021; Gov.UK, 2024). It identifies the key requirement of ethical standards in the handling of data and AI at the time of deployment, given the consequences at both business and end-user levels. This helps ensure transparency, accountability and integrity during the research process and ensures that the user privacy remains protected.

This research also aims to address the gap in existing literature, where there is a lack of studies combining clustering techniques with sentiment analysis to enhance recommendation engines. Serving as a guide, this work discusses future research avenues, such as incorporating additional data dimensions or exploring alternative sentiment analysis techniques to further improve the model's effectiveness, and hopes to provoke interest, thought, and investment in the domain of recommendation systems. The next step of the work includes building a web application in which real-world end users can play with anybody interested in the Sephora product catalogue to get generated product recommendations from the CFRBCS algorithm. Next, the author will introduce data from other companies and then train the algorithm so that it generalizes its application in this work.

Figure 1 presents the infographic of the executive summary, providing stakeholders with a bird’s eye view of the research. This visual summary encapsulates the aim and objectives, importance, methodology, and results, enabling a quick yet comprehensive understanding of the study’s essential insights.

Executive Summary

CFRBCS Model: Advancing E-commerce Recommendations

Combining Clustering and Sentiment Analysis for Personalized Product Suggestions

Aim and Objectives

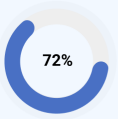
To increase the significance and reliability of recommendation systems in e-commerce by introducing the **CFRBCS** model.

Obj 1 : Analyse user sentiment to reveal product preferences.

Obj 2 : Cluster products by price and popularity for tailored recommendations.

Obj 3 : Validate the performance of the newly proposed hybrid model.

Why this research is important?



72% of consumers prefer personalized recommendations while shopping online (Skovhøj, 2022).

75% of business executives believe personalisation is crucial for digital experiences (Artug, 2022).

Research gap : This research fills an existing methodology gap by integrating sentiment analysis and clustering, which haven't been combined before, especially using product attributes.

Methodology Overview

The hybrid model is developed using a multistage methodology.

Dataset : Sephora dataset (Inky, 2023).

Stage 1 : Product Cluster Analysis

Stage 2 : Comprehensive Similarity Score Calculations

Stage 3 : Recommendation score prediction

Results, Significance and Impact

Identified 5 distinct product clusters, each defined by specific factors such as price and popularity, providing valuable insights into customer preferences.

The CFRBCS model outperforms traditional models in predicting user ratings, excelling across all evaluation metrics like RMSE, MAE, MAPE, and MSLE.

Greater prediction accuracy enhances understanding of customer preferences, leading to more personalized and effective product recommendations.

Product managers and analysts can integrate the CFRBCS algorithm into existing systems to deliver tailored recommendations based on comprehensive data analysis.

Future work includes exploring advanced clustering algorithms like DB-SCAN for more refined clusters and leveraging deep learning models for enhanced sentiment analysis.

References :

- Artug, E. (2022, December 20). 57 Personalization Statistics & Facts for 2022 You Shouldn't Ignore. Ninetailed.io. https://ninetailed.io/blog/personalization-statistics/
- Inky, N. (2023). Sephora Products and Skincare Reviews. Www.kaggle.com. https://www.kaggle.com/datasets/nadyinky/sephora-products-and-skincare-reviews
- Skovhøj, F. Z. (2022, January 12). The Power of Product Recommendations: 30 Must-Know Statistics for 2022. Clerk.io. https://www.clerk.io/blog/product-recommendations-statistics#:~:text=49%25%20of%20consumers%20said%20they

Figure 1 - Executive summary of the research. Adapted from: Author's own work.

Chapter 1 – Introduction

This dissertation explores the development and improvement of RS (Recommender Systems) within an e-commerce setting, looking toward improving predictive accuracy and enhancing product recommendation by integrating product segmentation and sentiment analysis.

At the dawn of the Internet era, companies realized that early adoption of selling their products online would be a great way to attract the broadest possible base of customers. This led to online shopping becoming the norm in our modern world, resulting in the rapid development of e-commerce platforms showcased by the fact that online retail sales have increased from 1,336 billion USD in 2014 to nearly 6,330 billion USD in 2024 (Chevalier, 2024) (Figure 2).

Consumers have also responded to this move by migrating their purchasing to the Internet where clothing, household electronics, and cosmetics have witnessed a high increase in sales within online commerce (Abu-AlSondos et al., 2023).

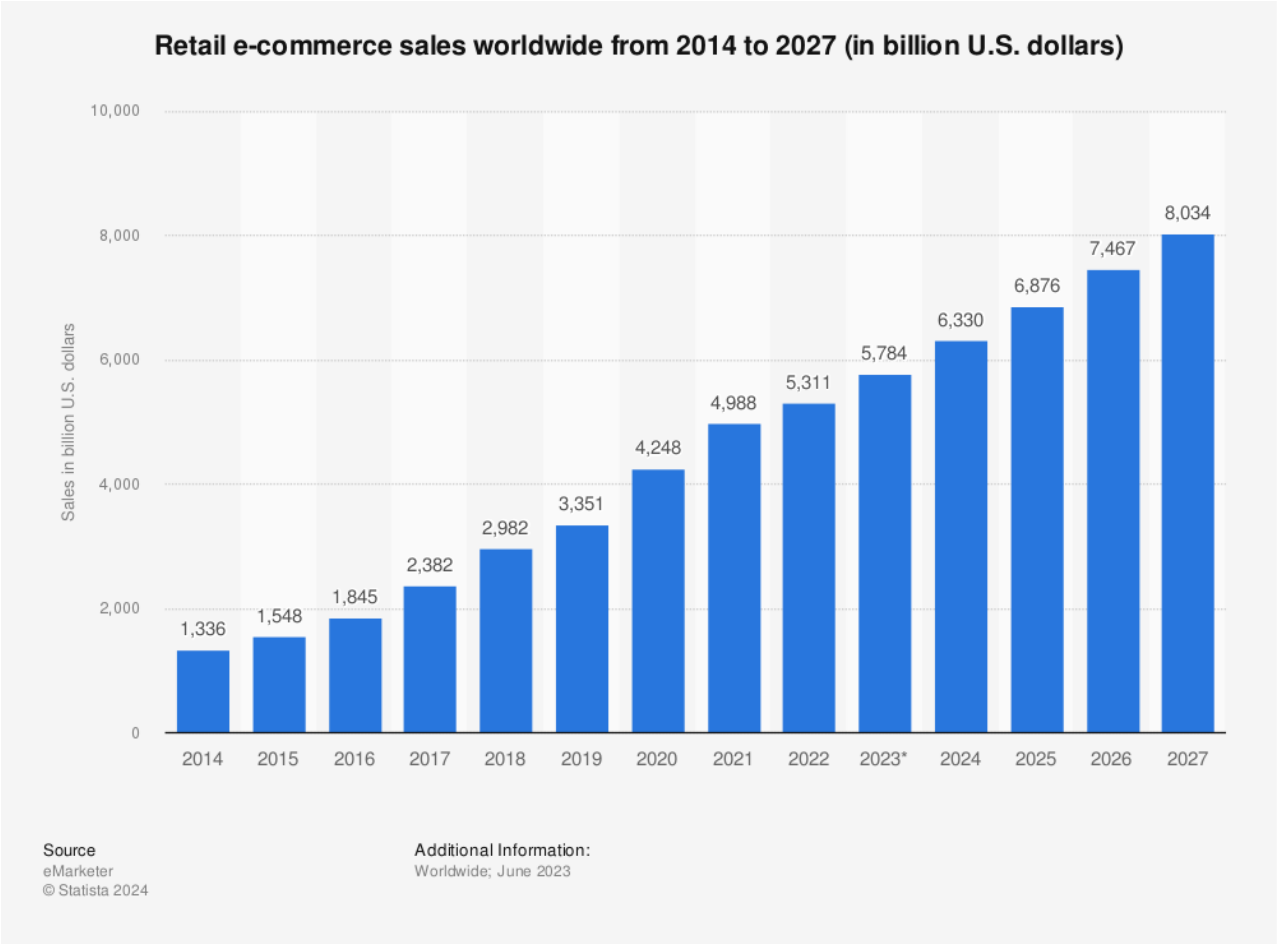


Figure 2 - Global trend in retail e-commerce sales from 2014 to 2027. Adapted from: Chevalier, S. (2024). Global retail e-commerce sales 2014-2027 [Review of Global retail e-commerce sales 2014-2027]. https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/

Modern shoppers, in recent times, enjoy a host of advantages while shopping online. Which ranges from experiencing convivence, to easy price comparison, no crowds, along with better value for money (Wang et al., 2019). However, with the recent explosion in the number of products available online, it has become increasingly challenging to choose the most appropriate one (Maheshwari, 2023).

A survey conducted in 2024 by a premium American press release company, Businesswire, shows that nearly 46% of global consumers start their search on e-commerce marketplaces for a product they desire (businesswire, 2024). However, their searching experience is found to be extremely overloading with irrelevant and inaccurate listings of products. Thus, increasing customer dissatisfaction and choice fatigue (Wang et al., 2023).

Realizing this fact, modern businesses are integrating sophisticated marketing techniques with data analyses to help match their offerings with the tastes and needs of their audience, helping them stand out in the marketplace (Micol Policarpo et al., 2021). In this regard, RS are probably the most ubiquitous technology that e-commerce players use while acquiring valuable insights (Klimashevskaia et al., 2023). These systems include sophisticated algorithms designed to analyse the behaviour and preferences of the users, with the aim of individualizing product recommendations to the customer (Bellini et al., 2022).

1.1 Purpose of the Study

Aim: This research work is aimed at improving the accuracy and relevance of recommendation systems in e-commerce.

Objectives: The objectives or this research project are stated below:

Objective 1: To perform sentiment analysis on user comments to uncover hidden sentiments that influence product preferences.

Objective 2: To conduct product clustering based on price and popularity to segment products effectively for personalized recommendations and to aid in recommendations presented to new users.

Objective 3: To integrate the insights from sentiment analysis and product clustering with traditional recommender systems to enhance predictive accuracy and recommendation relevance.

1.2 Research Questions (RQs)

The following RQs have guided this research toward better understanding: How does one understand product segmentation, sentiment analysis, and recommendation models, and how can these be integrated to improve predictive accuracy of hybrid systems?

RQ1: Identification of Key Product Segments

What are the key product segments that can be identified through data analysis, and how do these segments influence product preferences?

This question focuses on how products can be grouped into distinct segments, mainly based on data-driven criteria such as popularity and price attributes. The identification of the segments is quite important in tailoring the marketing strategies or product recommendations (Tuboalabo et al., 2024). Businesses can optimize their offerings to best meet the varied needs of their customers by understanding how different segments influence product preferences.

RQ2: Sentiment Analysis of Customer Reviews

In what ways does sentiment analysis of customer reviews contribute to understanding product preferences and enhancing recommendation models?

Sentiment analysis is a useful tool to gauge the feelings and attitude customers have towards a product (Lakshay Bharadwaj, 2023). By answering this question, it is possible to identify patterns and trends that reflect customer satisfaction or dissatisfaction. This information can greatly enrich recommendation models with emotional and subjective dimensions of customer feedback.

RQ3: Comparative Analysis of Recommendation Models

How does the hybrid recommendation model, that combines product segmentation with sentiment analysis perform in comparison with traditional recommendation models on predictive accuracy?

The third research question helps validate the predictive proficiency of the hybrid recommendation model. This research is an attempt to augment the predictive power of traditional models using a hybrid approach. A hybrid model, therefore, by design, may be expected to come up with more accurate and hence personalized recommendations, thereby improving overall customer experience by combining the merits of product segmentation and sentiment analysis.

1.3 Significance of the Study

Many studies have been conducted on the integration of RS with techniques from sentiment analysis and/or clustering (Dang et al., 2021) (Bellini et al., 2022) (Bhaskaran & Marappan, 2021). However, most of the researchers focus either on clustering based on user profiles and analysis of their comments (Xu et al., 2024) (Shankar et al., 2023), or on how clustering can be used to boost systems' performance once sentiment analysis has been performed (Karn et al., 2022) (Lakshay Bharadwaj, 2023). There is still a lack of research regarding the combination of insights drawn from product cluster statistics with sentiment analysis (Yarasu Madhavi Latha & B. Srinivasa Rao, 2023). The purpose of this study is to fill this gap and hence contribute towards the theoretical body of literature on hybrid models of recommendation by addressing some of the limitations of traditional RS techniques.

Few of the results obtained in this work have important implications for a variety of stakeholders, but in particular for marketing professionals and product managers in the online retail industry. Accurate predictions from this model would show the strategic way of resource allocation that e-commerce managers can adopt, enabling them to stock items with high ratings and maintain quick order-to-delivery cycles. These suggestions incoporated into operational changes can help deal with demand patterns such that overstocking and under-stocking can be avoided, again leading to reduced costs, better quality of service as well as greater customer loyalty for companies with diverse products.

1.4 Scope of the Study

One of the key tenets of this research, is to quantify the impact of various product segments and understand how they influence recommendations. To achieve this, the study utilizes and is focused on Attribute-Based clustering for products (Palmatier & Sridhar, 2017), and of the various methods present in sentiment analysis, this research is limited to implementing Lexicon-Based Approaches (Mitra, 2020).

The hybrid model is developed against the Sephora dataset (Inky, 2023), which contains very detailed information about products and comprehensive customer reviews. To gain insights on trending products and their affordability, the scope is limited to product popularity and price features.

1.5 Organisation of the Study

The rest of this study is organized as follows. Chapter 2 presents a literature review in this research area. Chapter 3 describes the methodology for the proposed CFRBCS model (Collaborative Filtering Recommendation Algorithm Based on Clustering and Sentiment analysis). Chapter 4 – Results and Analysis and Chapter 5 – Discussion outlines the results and discussion, and Chapter 6 – Conclusion presents the final conclusions.

Chapter 2 – Literature Review

2.1 Evolution and Importance of Recommender Systems

RS are information filtering tools that provide relevant and personalized information to user. The very essence of a recommender system is to decrease, as much as possible, a user's searching effort and time in finding meaningful information over the internet (Panda & Ray, 2022).

E-commerce giants like Amazon, have utilized the strengths of RS in recommending to their consumer pool, and have netted an increase of almost 35% in volume from online-only sales (Xu & Sang, 2022). Netflix, the popular online streaming service for movies and television series, uses a multi-layered hybrid RS in guiding their users' selections, accounting for about 80% of the content consumed (Gomez-Uribe & Hunt, 2016). The major usage of RS by these companies is to learn patterns of user interaction with their system and gain valuable insights into product and user similarity. Data from these RS is effectively utilized in making relevant suggestions, cross-selling or upselling similar products, and inventory management, which enhance the chances of conversion, and customer satisfaction. (Prijic, 2023).

2.2 Types of Recommender Systems

Broadly, there are three major categorizations of RS: collaborative filtering, content-based filtering, and hybrid approaches (Roy & Dutta, 2022). Figure 3 shows the detailed cataloguing of RS.

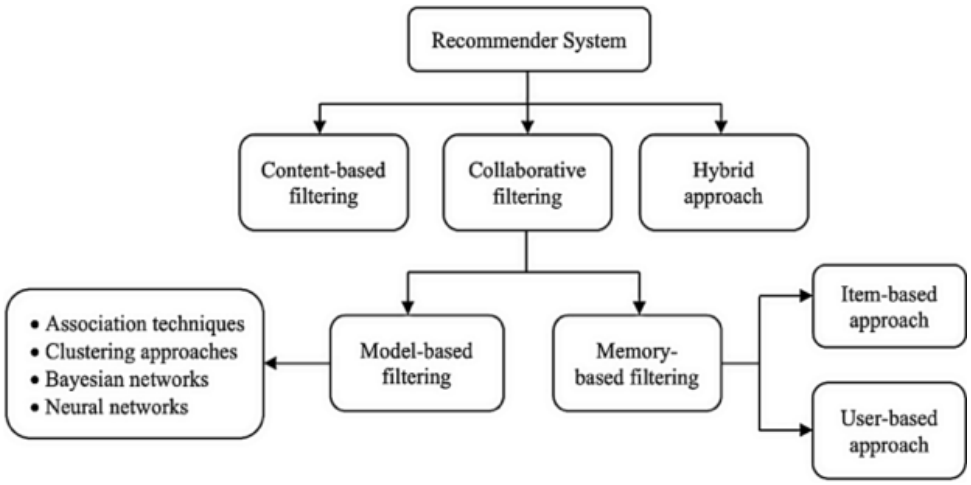


Figure 3 - Types of Recommender Systems. Adapted from: (Roy & Dutta, 2022)

Conceptually, content-based RSs work on the principle of utilizing user and item profiles to generate recommendations (Khanal et al., 2019). Item profiles are built based on product descriptions and its features. For example, in cosmetic products, brand, ingredients, and product type (e.g., moisturizer, lipstick) can used to group items into different item profiles (Ko et al., 2022).

When an item is rated positively by a user, all the other items present in that item profile are aggregated together to build the user profile, as show in Figure 4. As the user profile is a combination of highly favoured item profiles, the products belonging this user profile can be used as recommendations for the user (Roy & Dutta, 2022).

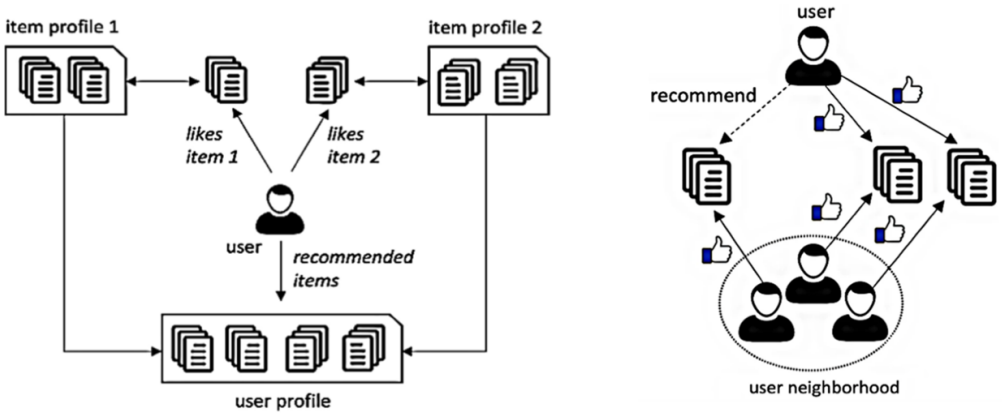


Figure 4 - Working of Content Based RS and Collaborative Filtering RS respectively. Adapted from: (Roy & Dutta, 2022)

The content-based recommendations are generated specific to each user profile, and formulating these recommendations is independent of any data from other users, as they have no effect on the individual recommendations (Fayyaz et al., 2020). Thus, implementing this RS can ensure that a particular user's information is secure and private (Roy & Dutta, 2022). Another positive aspect of this algorithm is that it can dynamically adjust itself to changing user preferences over time (Jannach et al., 2021). However, the major shortcoming of this approach is that for generating highly accurate recommendations, in-depth knowledge of item features is required, and this data may not be available all the time (Roy & Dutta, 2022).

Figure 4 also demonstrates the working of collaborative filtering RS where recommendation is generated based on measures of similarity between the users. The algorithm computes a set of users—termed "neighbourhood"—who are most similar in preference to a target user. Recommendations for the target user are derived from the preferences of the neighbourhood, giving more importance to items highly rated by its members (Zheng & Wang, 2021). A study by Wu (2022), shows that collaborative filtering RS are widely preferred in e-commerce applications and tend to outperform content-based RS in most scenarios.

Collaborative filtering RS can further be classified into model-based approaches and memory-based approaches (Roy & Dutta, 2022). Model-based approaches use machine learning techniques to infer user preferences from the learned patterns in the data (Tran et al., 2020). These models get trained on the history of user-item interactions to come up with recommendations. One major advantage associated with model-based approaches is their capability to scale on large datasets effectively and provide accurate recommendations; however, they require huge computational resources and time for the model training process (Roy & Dutta, 2022). Memory-based approaches are also called heuristic-based methods, as they use the complete user-item interaction datasets to provide recommendations (Shambour, 2021). They directly compute the similarities between users, normally at runtime, to generate recommendations. When compared to model-based approaches, memory-based approaches are preferred for their simplicity and ease of implementation (Roy & Dutta, 2022).

Lastly memory-based approaches are sub-divided into two subcategories: user-based and item-based (Roy & Dutta, 2022). User-based collaborative filtering User-Based Collaborative Filtering (UBCF) finds users that are similar to form neighbourhoods and recommends items that these neighbours have liked. In contrast, item-based collaborative filtering (IBCF) finds items like the ones with which the target user has interacted and recommends these similar items (Ajaegbu, 2021), as shown in Figure 5.

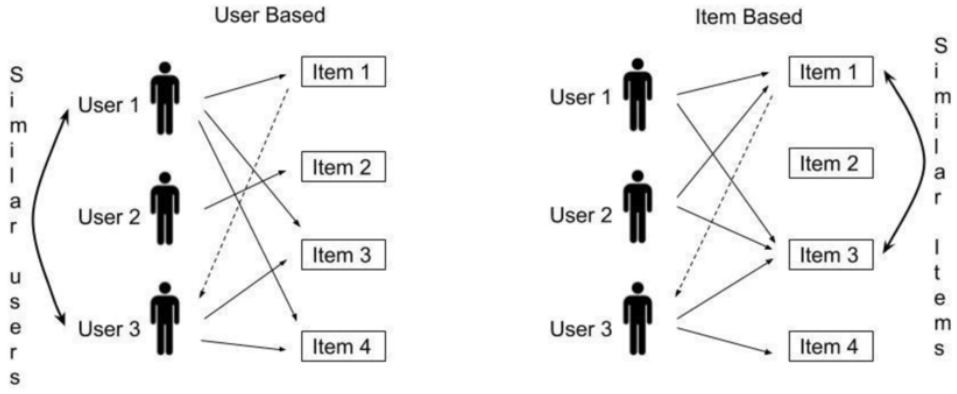


Figure 5 - User-based and Item-based collaborative filtering. Adapted from:(Upadhyaya, 2023)

2.3 Comparison of User-based and Item-based collaborative filtering

The main difference between UBCF and IBCF, is the basis on which similarity matrices are calculated, used for making recommendations. The foundation for both these approaches is a two-dimensional *user-item matrix* with rows as the users, columns that represent products and entries are the ratings provides by these users for corresponding items as shows in Figure 6 (A. Almohsen & Al-Jobori, 2015).

		Items					
		1	*2*	...	*i*	...	*m*
Users	*1*	5	3		1	2	
	2		2				4
	:			5			
	u	3	4		2	1	
	:					4	
	n			3	2		

Figure 6 - User-Item matrix. Adapted from: (A. Almohsen & Al-Jobori, 2015)

For UBCF, a *user-user similarity matrix* is established to study the similarity between users, whereas IBCF considers relationships between items, generating recommendations from an *item-item similarity matrix* (Singh et al., 2022). Authors Yan et al. (2019) implement a *Weighted Slopeone-IBCF Algorithm* to improve recommendation accuracy and during the development prioritise IBCF over UBCF as it is more stable and scalable. Their justification behind this choice is that similarities between items often remain more consistent over time, whereas user preferences can be more volatile.

Another major problem with UBCF is that there is a cold start issue wherein new users who have very little interaction data cannot be matched to similar users, hence making the recommendation less reliable (Roy & Dutta, 2022). IBCF does not suffer as much from the cold start problem since new items can quickly be compared to existing items based on their attributes, even if they have few interactions initially and for this reason Huynh et al. (2020) have used IBCF to develop the context-similarity item-based collaborative filtering recommender model (CIBCF) that outperforms traditional UBCF methods.

Authors Sineglazov and Yuriy Oliinuk (2021), explore the upper hand IBCF enjoys over UBCF in the context of e-commerce. Since e-commerce platforms frequently introduce new products, the ability of IBCF to quickly integrate and

recommend new items based on their similarity to existing items is crucial. Secondly, IBCF is generally faster in generating recommendations because it operates on a smaller subset of item similarities rather than computing user similarities across potentially millions of users.

2.4 Types of Similarity Calculations

To come up with reliable recommendations, a RS has to calculate similarity correctly. Two of the most used techniques are Pearson Correlation (Su & Khoshgoftaar, 2009), (Liu et al., 2014) and Cosine Similarity (Adomavicius & Tuzhilin, 2005); each has different strengths in different scenarios. Pearson correlation measures the degree of linear relationship between two variables, returning a value between -1 and 1. In RS, this technique calculates how similar two users or items are concerning their ratings. The Pearson correlation coefficient $r_{A,B}$ for two vectors, $A\ and\ B$, representing user ratings, is given by:

$$r_{A,B} = \frac{\sum(A_i - \bar{A})(B_i - \bar{B})}{\sqrt{\sum(A_i - \bar{A})^2 \sum(B_i - \bar{B})^2}} \tag{1}$$

Where $\bar{A}\ and\ \bar{B}$ are the mean rating of users $A\ and\ B$, respectively. This method captures the direction and strength of the linear relationship between the ratings of two users or items; hence, it can be applied in situations where different scales are used by users while rating items (Nicholas & Francis, 2019).

On the other hand, cosine similarity is a measure of the cosine of the angle between two vectors, considering their orientation and not their magnitude (Fkih, 2021). It is primarily useful for the comparison of various items or users in a multi-dimensional space, where every dimension corresponds to a different item or user (Fkih, 2021). Cosine similarity $sim_{A,B}$ between vectors $A\ and\ B$ is calculated as:

$$sim_{A,B} = \frac{\sum(A_i \cdot B_i)}{\sqrt{\sum {A_i}^2 \sum {B_i}^2}} \tag{2}$$

Cosine similarity ranges from 0 to 1, where 1 indicates identical orientation and 0 orthogonality, thus no similarity. This method comes in handy, especially when the data is sparse (which is quite a common situation in recommender systems), since it handles vectors with a lot of zero entries very efficiently (Zhang et al., 2019). Computational simplicity makes fast calculations possible with cosine similarity, which is very important in real-time recommendation generation (Bobadilla et al., 2013). All these advantages of cosine similarity calculations make it a preferred choice in many e-commerce applications (Md Nadeem Noori et al., 2024).

2.4 Sentiment Analysis in the context of Recommender Systems

Sentiment Analysis is the act of determining whether a sentiment in a certain text conveyed is either positive, negative, or even neutral (Chiranjeevi & Rajaram, 2023). Bhavitha et al. (2017), capture the insights into the preferences of customers at the sentence, document, feature levels using sentiment analysis. As per Wankhade et al. (2022), there are primarily three main techniques for sentiment analysis: lexicon-based techniques, machine learning-based techniques, and hybrid approaches.

Lexicon-based methods, which are categorized into dictionary-based and corpus-based, are based on pre-defined word lists with respect to sentiment (Adomavicius & Tuzhilin, 2005). In particular, the method of VADER—Valence Aware

Dictionary and Sentiment Reasoner—has been outstanding in catching the intensity of sentiments in review text (Hutto & Gilbert, 2014). Machine learning-based techniques, which include traditional algorithms like Support Vector Machines (SVM) and Naive Bayes as well as advanced models such as Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN), classify text based on sentiment (Bing Liu, 2012). Hybrid approaches combine the strengths of both lexicon-based and machine learning-based techniques to improve accuracy (Adomavicius & Tuzhilin, 2005).

Sentiment analysis has been applied to a wide range of domains in order to improve the performance of recommender systems (Elahi et al., 2023). In e-commerce, for example, it has been applied in the optimization of product recommendation—conducting analysis on customer reviews to enhance satisfaction and conversion rates (Karabila et al., 2023). Preethi et al. (2017) used recursive neural networks in order to achieve better restaurant and movie recommendation results on a cloud-based system by analysing review sentiments. Wang et al. (2018) proposed a hybrid recommender system integrating sentiment analysis into its utility list of recommended results for its optimization. Similarly, Panda et al. (2020) combined collaborative filtering, content-based filtering, and sentiment analysis over movie tweets to make the system efficient. Aida Osman and Azman Mohd Noah (2018) integrated the sentiment-based analysis into their work on IMDb and Movie Lens datasets, and the improvements have been registered ever since. Although most RS system rely on explicit ratings provided by users, incorporating sentiment analysis of text reviews helps a great deal in boosting recommendation accuracy and reliability as an implicit source of feedback (Devlin et al., 2019).

2.5 Clustering in the context of Recommender Systems

Clustering techniques can definitely become an important solution to address some of the challenges faced by recommender systems. Grouping similar items together into clusters can enhance accuracy and relevance in recommendation (Roy & Dutta, 2022). One such algorithm used for clustering is K-means, having wide application, especially proving to be very effective in partitioning a data set into different clusters based on feature similarity (Cui et al., 2020). In recommender systems, K-means can be used for clustering products or items on features such as popularity and price. This, in turn, helps reduce issues related to scalability and sparsity, hence improving the recommendation quality (Abbasi-Moud et al., 2021).

K-means runs efficiently with a time complexity that is linear with respect to the number of items and clusters; hence, it is useful for large data sets, as shown by Ikotun et al. (2022). Every product is assigned by the algorithm to a cluster whose centre is the mean of the features, and the system can recommend products within the same cluster to users with similar preferences, as shown by Anitha and Patil (2019). For instance, Dara et al. (2019). demonstrated that integrating K-means clustering with collaborative filtering is an approach to enhance the recommendation accuracy since grouping similar items and users reduces computational complexity, resulting in more relevant recommendations. Another approach was taken by Zhang et al. (2020), who applied clustering in the grouping of items based on user interactions and item attribute data, hence yielding more personalized and contextually relevant recommendations.

Also, clustering items by popularity and price can address the cold-start problem and improve recommendation precision. For new or popular products, clustering will ensure that they are recommended together with other similar items from the same cluster, thus raising their visibility and chances of being chosen by users (Roy & Dutta, 2022). Moreover, clustering takes care of sparsity since it gives information that is meaningful even when data from the interaction between the user and the items are less (Hajer Nabli et al., 2023). In general, K-means clustering can be used in RS to enhance reorganisation of products to be according to the user's preference and patterns of behaviour (Roy & Dutta, 2022).

Chapter 3 – Methodology

3.1 Research Desing

Figure 7 presents the proposed research design that adopts a mixed mythology approach for developing a comprehensive framework to build the CFRBCS model. It incorporates different stages of data processing and analytical techniques to come up with relevant and accurate recommendations for products. The structure of the research design can be demarcated into three primary stages that succeed the crucial phase of data preprocessing.

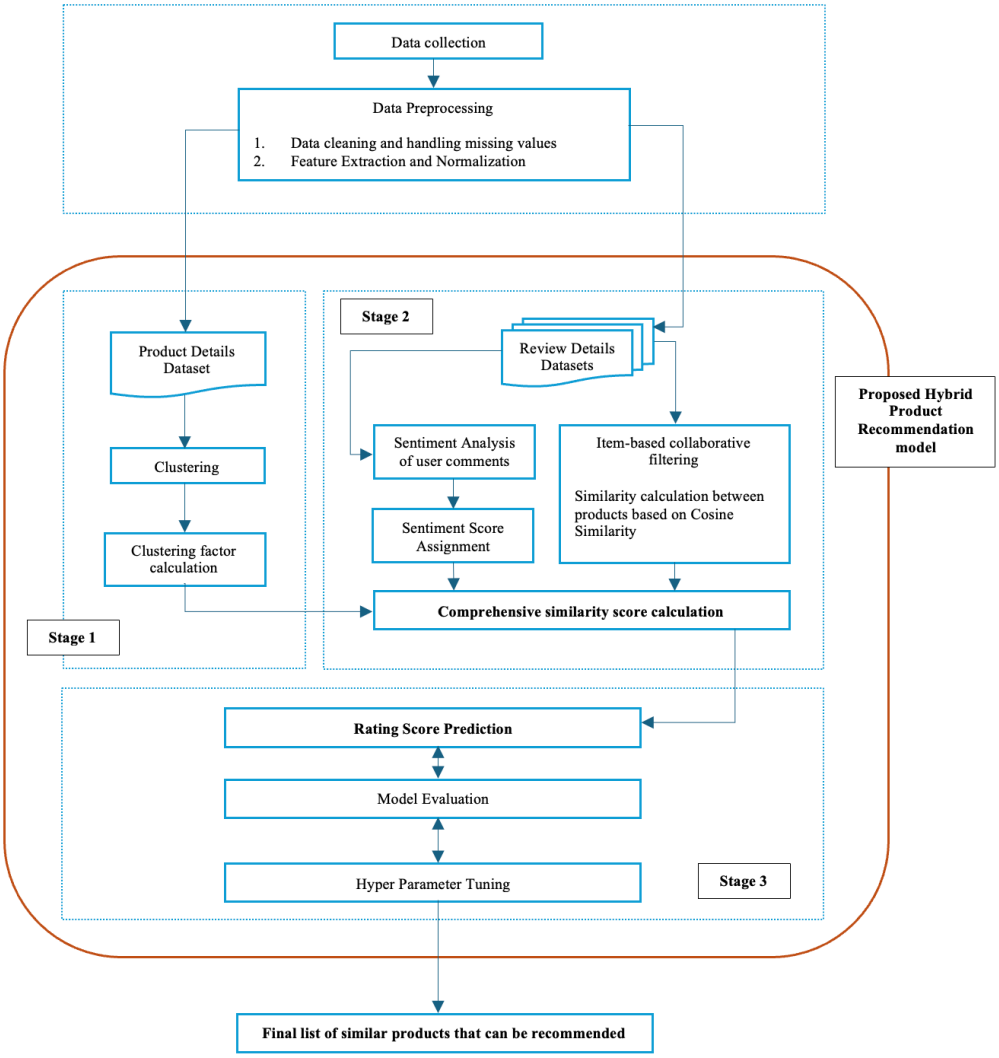


Figure 7- CFRBCS Model Methodology. Adapted from: Author's own work.

3.1.1 Data Collection

Data collection is the first process in the implementation. This step involves selecting a dataset that is appropriate to be used for further analysis. For the hybrid recommendation system, the Sephora dataset from Kaggle (Inky, 2023) is used. This dataset was collected in March 2023 using a Python scraper and contains all information on more than 8,000 beauty

products listed on Sephora's online store. The data includes information on product and brand names, price ranges, ingredients, ratings, and several other features.

Moreover, it comprises around one million reviews by users against over 2,000 skincare products, which include all the details regarding the appearances of the users along with corresponding ratings. Table 1 summarizes all the datasets.

Table 1 - Description of datasets. Adapted from: Author's own work.

Data set name	Description	Number of columns	Number of rows
Product_info.csv	Contains the Product data content.	27	8,494
reviews_0-50.csv	Contains the Reviews data content from March 2023.	19	6,02,130
reviews_250-500.csv	Contains the Reviews data content from March 2023.	19	2,06,725
reviews_500-750.csv	Contains the Reviews data content from March 2023.	19	1,16,262
reviews_750-1250.csv	Contains the Reviews data content from March 2023.	19	1,19,317
reviews_1250-end.csv	Contains the Reviews data content from March 2023.	19	49,977

3.1.2 Data Preprocessing

Before the main stages of the recommendation model, an important preliminary phase in the research design is *Data Preprocessing*. These are the two major activities involved in this preliminary step: *Cleaning of data and handling missing values*, *Extraction of features and normalization*.

Data cleaning is the process of checking for inconsistencies or gaps in the collected information and ensuring that the dataset is complete and consistent (Wang et al., 2020). The features relevant for price and popularity of the products are retained for clustering, while other features, such as review comments of the users and their ratings, are used in sentiment analysis and similarity calculations. In sum, this preprocessing phase is quite important to enhance the quality and consistency of the input data for improved overall performance and reliability of the proposed recommendation model.

3.1.3 Stage 1: Product Cluster Analysis

The first stage of the research design involves processing the Product Details Dataset, focusing on two main tasks. First, a clustering algorithm is applied to group similar products based on their pricing and popularity attributes. This helps organize the product catalogue into clear, meaningful segments. The algorithm selected for this purpose is the K-means algorithm.

K-means has many advantages compared to other clustering algorithms such as hierarchical clustering and DBSCAN. For instance, it is computationally efficient (Ikotun et al., 2022), making it perfect for Sephora product data. Linear in time complexity, k-means can iterate over the thousands of products with quite minimal problems in performance (Naeem & Aishan Wumaier, 2018). Another advantage associated with K-means is its simplicity, which makes it easy to implement and, more importantly, easy to understand (Ikotun et al., 2022). The method differs from hierarchical clustering, which is computationally demanding and can be hard to visualize for large numbers of data points (Aastha Gupta et al., 2021). Hierarchical clustering arranges clusters as a multilevel hierarchy; this may be very awkward to handle for large data sets (Shetty & Singh, 2021). Although powerful in the sense that it is able to find clusters of any shape and is robust against

noise, DBSCAN needs careful adjustment of parameters like epsilon, ε, and the minimum number of points, which are not easy to adjust properly without domain knowledge (Braune et al., 2014). K-means, on the other hand, requires only the specification of k, the number of clusters, which can be determined in many ways by well-established statistical methods (Ikotun et al., 2022).

The Elbow method can be used to estimate the optimal number of clusters (Sammouda & El-Zaart, 2021). It plots the value of distortion score on the y-axis, and the number of clusters on the x-axis. In essence, it is a measure of the squared distances between each point and the centroid of its assigned cluster; thus, it shows the compactness of the clusters. The "elbow" point, where the marginal gain decreases sharply is indicative of the optimal number of clusters; this method is visual and intuitive to select k, making sure that the chosen number of clusters really corresponds to the underlying structure of the data.

The Silhouette method is also applied to ensure the validity of the chosen number for the clusters. The Silhouette value measures the similarity of an object to its own cluster compared with other clusters and, hence, is a measure for cluster cohesion and separation (Shutaywi & Kachouie, 2021). A high average Silhouette score implies the data points are well-clustered; a clear separation is made for each group. Such well-structured clustering becomes a strong basis for the following stages of the recommendation model, ensuring the ability to group similar products together to make better recommendations.

Next, for each of the clusters identified a corresponding cluster factor C_a is calculated as follows.

$$C_a = \frac{Average\ popularity\ of\ products\ in\ the\ cluster}{Average\ price\ of\ products\ in\ the\ cluster} \tag{3}$$

Since the features used for clustering are measured on varied scales, features with larger scales can sometimes dominate, leading to bias in the clusters. To counter this bias and calculate the normalized cluster factor, C_{a_norm} , Min-Max Normalization (MMN) is performed. This method is particularly useful for preserving the relationships among the original input data, unlike normalization methods based on the mean and standard deviation, which may vary over time (Singh & Singh, 2019).

$$C_{a_{norm}} = \frac{C_a - minvalue}{maxvalue - minvalue} \tag{4}$$

3.1.4 Stage 2: Comprehensive Similarity Score Calculations

In the second stage, the research design dives into the analysis of the information in the datasets that make up the review details. Sentiment analysis of user comments is the first step, where the polarity of each comment is determined by a lexicon-based approach, after which the sentiment score is normalized. The VADER method is used for this exercise since it is a rule-based model that is conventionally efficient in terms of computational effort, thus able to efficiently handle large e-commerce reviews (Tripathy & Rath, 2017). The sentiment scores from these reviews are meaningful and useful only when comparisons can be done. The *MinMaxScaler* library is, therefore, used to transform the data into a fixed scale range of [0, 1]. This is one of the normalization techniques that will ensure that the sentiment scores obtained are consistent and comparable if scales involved in the original data are different (Raju et al., 2020). This process not only identifies the general sentiment regarding user reviews but also gives a standardized way to compare the sentiments.

The main outcome of the second stage, and the novelty of this research, lies in the comprehensive similarity score calculation. Traditionally, IBCF RS calculate the similarity between products solely based on user ratings. As discussed earlier, this approach often fails to capture the inherent relationship between product features and the sentiments related to the products. To address this, the similarity between products is also calculated based on the cluster to which the product belongs, and the sentiment score it has received. The similarity calculations for these factors, along with the traditional rating similarity, are based on cosine similarity, as demonstrated by the following equations.

Traditional similarity based on rating:

$$sim(a,b) = \frac{r_a \cdot r_b}{||r_a|| \cdot ||r_b||} \tag{5}$$

Where,

$$sim(a,b) = \text{Similarity between items } a \text{ and } b,$$
$$r_a = Rating\ for\ prduct\ a,$$
$$r_b = Rating\ for\ product\ b,$$
$$r_a \cdot r_b = Dot\ product\ of\ vectors\ r_a\ and\ r_b,$$
$$||r_a|| = The\ Euclidean\ norm\ (magnitude)\ of\ vector\ r_a,$$
$$||r_b|| = The\ Euclidean\ norm\ (magnitude)\ of\ vector\ r_b,$$

Similarity based on sentiment score:

$$sim'^{(a,b)} = \frac{s_a \cdot s_b}{||s_a|| \cdot ||s_b||} \tag{6}$$

Where,

$$sim'(a,b) = \text{Similarity between items } a \text{ and } b,$$
$$s_a = Sentiment\ score\ for\ prduct\ a,$$
$$s_b = Sentiment\ score\ for\ product\ b,$$
$$s_a \cdot s_b = Dot\ product\ of\ vectors\ s_a\ and\ s_b,$$
$$||s_a|| = The\ Euclidean\ norm\ (magnitude)\ of\ vector\ s_a,$$
$$||s_b|| = The\ Euclidean\ norm\ (magnitude)\ of\ vector\ s_b,$$

Similarity based on cluster factor:

$$sim''^{(a,b)} = \frac{c_a \cdot c_b}{||c_a|| \cdot ||c_b||} \tag{7}$$

Where,

$$sim''(a,b) = \text{Similarity between items } a \text{ and } b,$$
$$c_a = Cluster\ factor\ for\ prduct\ a,$$
$$c_b = Cluster\ factor\ for\ product\ b,$$
$$c_a \cdot c_b = Dot\ product\ of\ vectors\ c_a\ and\ c_b,$$
$$||c_a|| = The\ Euclidean\ norm\ (magnitude)\ of\ vector\ c_a,$$
$$||c_b|| = The\ Euclidean\ norm\ (magnitude)\ of\ vector\ c_b,$$

The comprehensive similarity score, $SIM(a, b)$ is calculated combining the equations 5, 6, 7 along with their respective weighting factors as shown below.

$$\boldsymbol{SIM(a,b) = \alpha * sim(a,b) + \beta * sim'(a,b) + \gamma * sim''(a,b)} \quad (8)$$

3.1.5 Stage 3: Rating score prediction

The last phase of the research design focuses on using the CFRBCS model for predicting a rating score. In the IBCF framework, an estimated rating for any item a by user x, can be obtained considering the mean ratings for all the other times the user has rated weighted by their similarity.

For example, assuming a user rated four items, say A, B, C, and D, and the model is to predict the estimated rating for a new item E, it would do this based on ratings given to each of these items and their relationship to item E in the similarity matrix. This ensures that the prediction uses both the user's past rating behaviour and the relationships between items to make an informed prediction. Equation 9 shows the mathematical expression for the computation of this estimated rating.

$$r_{xa} = \frac{\sum_{b \in N(a;x)} SIM(a,b) * r_{xb}}{\sum_{b \in N(a;x)} SIM(a,b)} \quad (9)$$

Where,

$$r_{xb} = Rating\ for\ item\ b\ by\ user\ x,$$

$$r_{xa} = Estimated\ rating\ for\ item\ a\ by\ user\ x,$$

$$SIM(a,b) = Comprehensive\ similarity\ between\ item\ a\ and\ b,$$

$$N(a;x) = Set\ of\ items\ rated\ by\ user\ x\ and\ similar\ to\ item\ a,$$

The output of stage 2 is the comprehensive similarity score, which is utilised in stage 3 to optimise the predicted ratings. These scores indicate the likelihood of a product being a suitable recommendation for a given user, particularly in the context of e-commerce product recommendations.

3.1.6 Model training and testing

This section focuses on how exactly data was prepared in order to train and test the model. In this respect, the following dataset were employed: 'reviews_0-50.csv', 'reviews_250-500.csv', 'reviews_500-750.csv' and 'reviews_750-1250.csv'. After concatenation of all the above files, one comprehensive dataset of nearly 1,044,434 rows was obtained. This large dataset serves as the basis for training and testing the model. To account for system performance and reducing compilation time the dataset has been further restrained to 521,496 rows.

The train_test_split function was utilised from the *sklearn.model_selection* package to enable an effective training process. This function made an 80-20 split of data possible, where 80% of the data is for training and 20% kept for testing. This split is very important in ensuring that the CFRBCS model is well-trained but at the same time has enough data to validate its performance.

In the second stage, the comprehensive similarity matrix with the given training data was constructed. As seen previously, this matrix is very important in understanding the different relationships amongst the items and is used in predicting future

ratings. In the third stage, predictions were made using the test data, helping in assessing the accuracy and reliability of the CFRBCS product recommendation model.

3.1.7 Model Evaluation

The performance of the developed model needs to be tested to ensure that it is accurate and reliable. In this section, the developed CFRBCS product recommendation model is validated against different statistical parameters such as RMSE, MAE, MAPE, and MSLE. All of these give useful insights into the different aspects related to the predictive abilities of the model.

3.1.7.1 Root Mean Square Error (RMSE):

One majorly used accuracy metric is the RMSE, which is the root mean square error. It's a squared root of the average of the squared differences between predicted and actual values. The RMSE is given by the following expression:

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i})^2} \tag{10}$$

Here, $\mathrm{y_i}$ represents the actual values, $\hat{\mathrm{y}}_\mathrm{i}$ represents the predicted values, and n is the number of observations. RMSE is particularly useful for understanding the magnitude of prediction errors, with lower values indicating better model performance (Chai & Draxler, 2014).

3.1.7.2 Mean Absolute Error (MAE):

Mean Absolute Error (MAE) is another important metric that measures the average absolute difference between the predicted and actual values. It is given by the formula:

$$\mathrm{MAE} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}|\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i}| \tag{11}$$

Where, $\mathrm{y_i}$ represents the actual values, $\hat{\mathrm{y}}_\mathrm{i}$ represents the predicted values, and n is the number of observations. For the average forecast error, MAE gives an interpretable meaning. The smaller the values, the better the model accuracy. It is less sensitive to outliers than RMSE (Willmott & Matsuura, 2005).

3.1.7.3 Mean Absolute Percentage Error (MAPE):

Mean Absolute Percentage Error (MAPE) expresses the accuracy of predictions as a percentage. It is calculated as follows:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100 \tag{12}$$

Where, y_i represents the actual values, $\hat{y}_i$ represents the predicted values, and n is the number of observations. One of the major advantages of MAPE is that it helps in understanding relative prediction errors; due to it being a relative measure,

performance comparison across different scales can be performed. However, it creates some problems when actual values are close to zero (Kim & Kim, 2016).

3.1.7.4 Mean Squared Logarithmic Error (MSLE):

The Mean Squared Logarithmic Error is the average of the squares of the differences between log values of the predicted and actual values. The formula is:

$$MSLE = \frac{1}{n}\sum_{i=1}^{n}(\log(1+y_i) - \log(1+\hat{y}_i))^2 \quad (13)$$

Here, y_i represents the actual values, $\hat{y}_i$ represents the predicted values, and n is the number of observations. MSLE is particularly useful where the model's performance needs to be evaluated on a relative scale since MSLE penalizes under-estimations more than overestimations (Hodson et al., 2021).

3.1.7.5 Comparative Analysis with other models:

The performance of the CFRBCS model was compared with two models: the traditional Item-Based Collaborative Filtering (IBCF) model and the CFRBS (Collaborative Filtering Recommendation Algorithm Based on Sentiment analysis) model, proposed by Jian Zhen Yu et al. (2018). The performance was tested and benchmarked against the IBCF and CFRBS models, through the statistical parameters mentioned in the preceding section.

3.1.8 Hyperparameter tuning

This section describes how the hyperparameter tuning was done in an effort to obtain better model accuracy and performance. Essentially, hyperparameter tuning is that part of a machine learning process whereby parameters pertaining to the training process itself need adjusting. In the case of the hybrid model, this involved the optimization of weights (α, β, γ) for individual similarity factors, as mentioned in equation 8. The weights were tried over different combinations, with their sum equating to 1. The idea was to get an optimal combination that would minimize the RMSE by trying different values for α, β and γ.

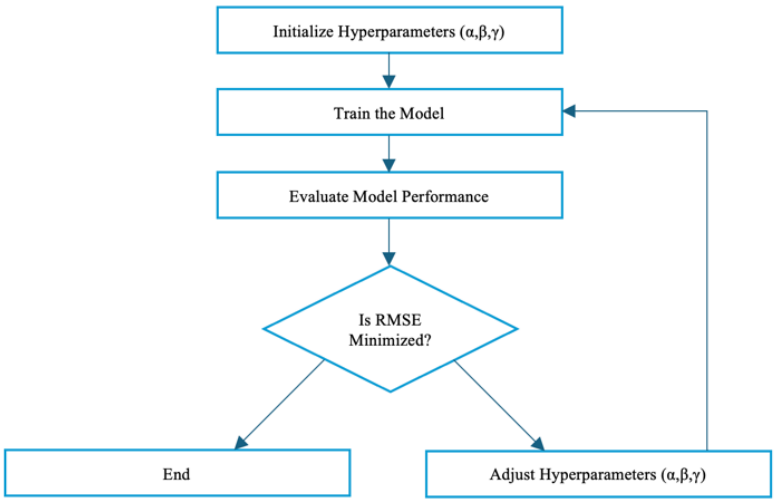


Figure 8 - Systematic approach towards hyperparameter tunning. Adapted from: Author's own work.

This iterative process of tuning and evaluation is illustrated in Figure 8. It therefore means that the process was repeated several times to ensure continuous adjusting based on the results obtained for best performance. The cycles reiterated model performance at every turn and called for necessary adjustments on the hyperparameters, stating a dynamic and responsive approach taken in this study.

3.1.9 Final Output

This final result of the research design is the output of a "Final list of similar products that can be recommended." This output represents the practical application of the CFRBCS model in which actionable product recommendations are given based on comprehensive analysis conducted through these different stages.

3.2 Pseudocode of CFRBCS algorithm

Table 2 and Table 3 provide an overview of the steps involved in the CFRBCS algorithm. One of the goals of this algorithm is to predict the rating a user will give to a target product based on similarities with other rated products. Starting with processing the input, the next step retrieves the similarities between the target product and S similar products, along with their respective ratings. These ratings and the similarity information are used to make an informed estimate of the predicted rating. Another goal is to generate a list of product recommendations specific to a target author based on a target product they have previously purchased.

Table 2 - The pseudocode of CFRBCS algorithm to make predictions. Adapted from: Author's own work.

Algorithm: The CFRBCS algorithm
// Input: Target author, Target product, Item user matrix, Product similarity matrix, S (number of similar products)
// Output: predicted rating: R
1. Processing the input data.
2. Retrieving similarities between S similar products and target product.
3. Retrieving relevant ratings of similar products.
4. Computing the predicted rating using a weighted average of ratings and the similarity factor.
5. Returning the final prediction.

Table 3 - The pseudocode of CFRBCS algorithm to generate recommendations. Adapted from: Author's own work.

Algorithm: The CFRBCS algorithm
// Input: Target author ID, Item-user matrix, Cosine similarity matrix, S (number of similar products), N (number of products to recommend)
// Output: List of top N recommended products and their predicted ratings
1. Retrieve all product IDs from the from the item-user matrix.
2. Compute the predicted rating for the products using the function to predict ratings.
3. Select the top N products based on their predicted ratings;

3.3 Use of AI in Methodology

The methodology was developed and put into practice, with AI tools to make the process smooth and efficient. More precisely, ChatGPT, Claude AI, and Google Gemini LLM have been leveraged for assistance in different coding tasks. These AI models have assisted in generating code snippets, debugging errors, and *optimizing the hybrid algorithm*. ChatGPT especially helped a lot during the early development stages with the *generation of code templates* and insightful suggestions. It is through the use of Claude AI that effective solutions to problems were identified, reducing debugging time. Google Gemini LLM was helpful in *fine-tuning the code* and making sure it followed the best coding practices. This way, using the AI models together really proves how AI can assist in reducing development time and making the methodology both robust and efficient.

3.4 Ethical considerations of the study

3.4.1 Data Handling

This research is based on the Sephora dataset available in Kaggle (Inky, 2023) under Attribution 4.0 International License (CC BY 4.0), which gives rights to share and adapt under proper attribution. The quality checking of the data was performed to ensure that product information and customer reviews were scraped accurately off the official website of Sephora (Sephora, 2023). Figure 9 and Figure 10 illustrate source of the official product details and customer review information available on the website.

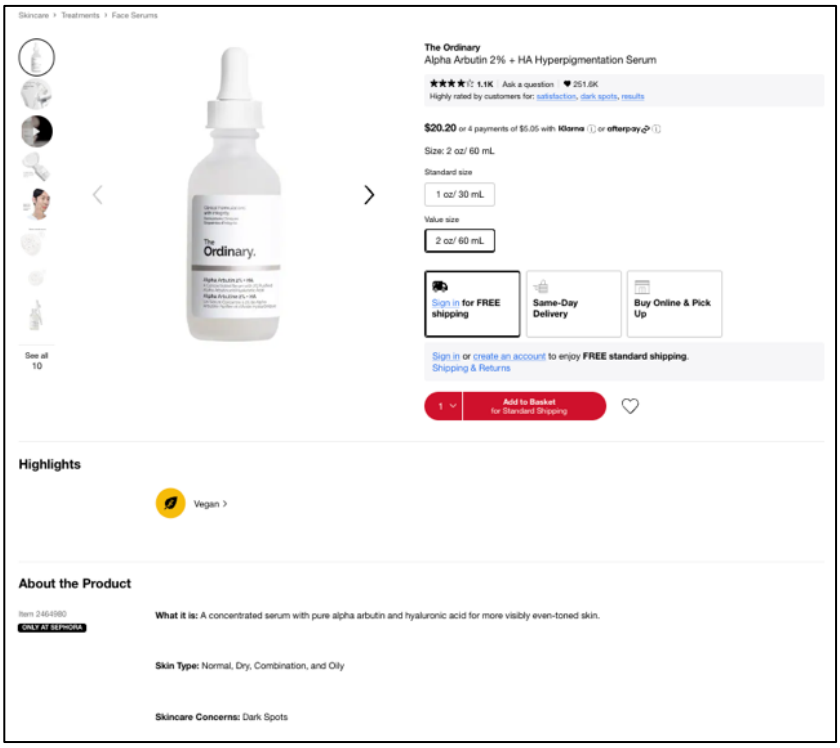


Figure 9 - Official product information on Sephora's website. Adapted from: Author's own work . Sourced from: https://www.sephora.com/product/the-ordinary-deciem-alpha-arbutin-2-ha-P427412?skuId=2464980&icid2=products%20grid:p427412:product

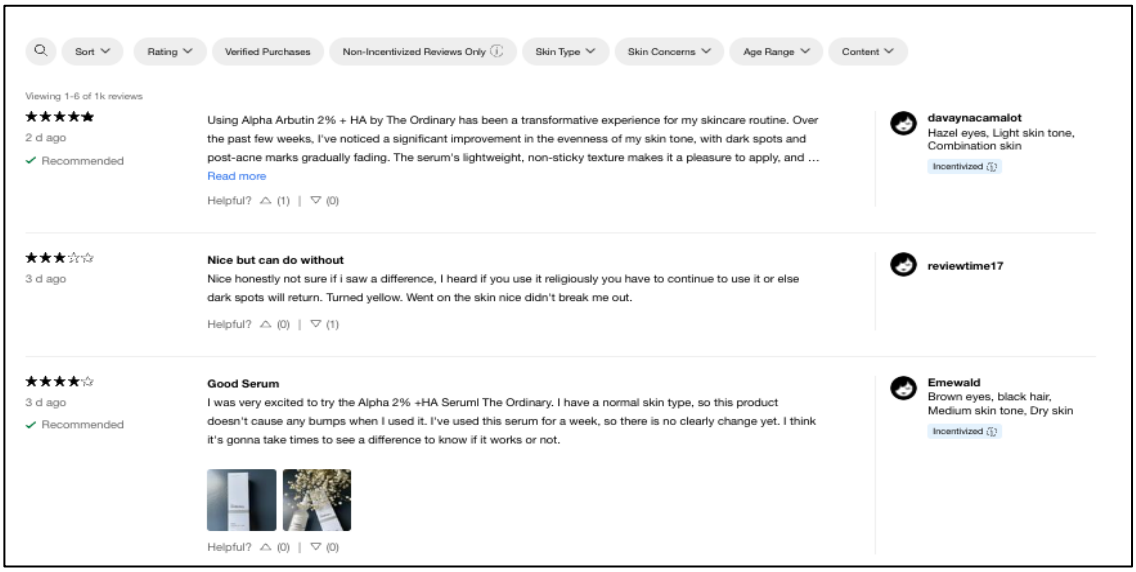


Figure 10 - Official customer reviews on Sephora's website. Adapted from: Author's own work. Sourced from: https://www.sephora.com/product/the-ordinary-deciem-alpha-arbutin-2-ha-P427412?skuId=2464980&icid2=products%20grid:p427412:product

In an effort to maintain high ethical standards, and manage the perceived ethical issues, the research focuses on data anonymization and transparent data ethics. The Anonymity of users is ensured by using *Author Ids* to represent product comments. It is difficult to publicly identify the user based on the systems internal Author Id, as show in Figure 11.

Figure 11 - Numeric Author id used to represent the users in the dataset. Adapted from: (Inky, 2023)

To ensure all user's data has been ethically collected, an in-depth investigation of Sephora privacy policy was conducted, and it is observed that, consent from users is obtained through clear privacy disclosure guidelines and opt-out mechanisms are also provided (Privacy Policy, 2021). The dataset is retained only during the course of the study and deleted once the study has been completed, following ethical best practice regarding the retention and disposal of data.

3.4.2 Ethics of AI Techniques

There are also ethical considerations with the K-means clustering and Lexicon-Based Sentiment Analysis approach. Though K-means clustering could effectively segment the products, it must be treated carefully not to promote biases in the data. As underlined by the EU Artificial Intelligence Act, ethical AI practices have to be based on responsible use of algorithms that can foresee and prevent possible unwanted effects and guarantee fairness (European Commission, 2021).

The Lexicon-Based Sentiment Analysis method used here depends on predefined word lists; hence, it has a limited capacity in capturing the full range of sentiment expressed in user reviews. It may also limit the system's accuracy in giving an interpretation of sentiments. Proposals for AI regulation put forward by the UK government (Gov.UK, 2024) recommend

that transparency and accountability be observed in AI applications, thus creating the need to further understand and consider the limitations of the Sentiment Analysis method.

3.4.3 Overall Study Ethics

During the course of the research, high ethical standards are maintained by bringing into light issues of anonymization, data quality, and transparency. Since the study is based on secondary data, ethical concerns regarding the collection of new data from human subjects are reduced. The study follows the data protection regulations and ethical guidelines, especially the Artificial Intelligence Act of the EU (European Commission, 2021), with special emphasis on data protection, transparency, and accountability. The AI regulatory proposals for the UK (Gov.UK, 2024) further underline the need for responsible AI practices, with stringent oversight to guard against misuse. By aligning with these guidelines, the study ensures that ethical considerations are integral to its research methodology and execution.

This study also adheres to the GDPR of the European Union and the UK's Data Protection Act 2016, both of which emphasize the need for the lawful, fair, and transparent processing of personal data, which is central to handling the data in this research. The general principles of data minimization, purpose limitation, and the right to be forgotten under Article 17 of the GDPR are considered. The study ensures that any use of the Sephora dataset complies with the legal requirements for data protection and privacy (European Union, 2016).

Similarly, the UK Data Protection Act, as influenced by the Brexit provisions in the GDPR, imposes stringent conditions for data use, including the necessity of explicit consent and the implementation of security measures to protect personal information. This research observes these principles by anonymizing user data, limiting the retention period to the duration of the research, and arranging for secure disposal upon completion (Gov.Uk, 2018).

Chapter 4 – Results and Analysis

4.1 Presentation of Findings and Interpretation

4.1.1 Exploratory Data Analysis

Figure 12 and Figure 13 present the distribution of data in numerical and categorical variables of the Product_info dataset, respectively. These plots are were instrumental in understanding the general trends of data in the dataset. For example, the rating of most the products is between 3 to 5, and the price of most of the products are in the range of 0 to 500 USD. The analysis also indicated that the perfume products were in the majority, followed by moisturizers and face serums.

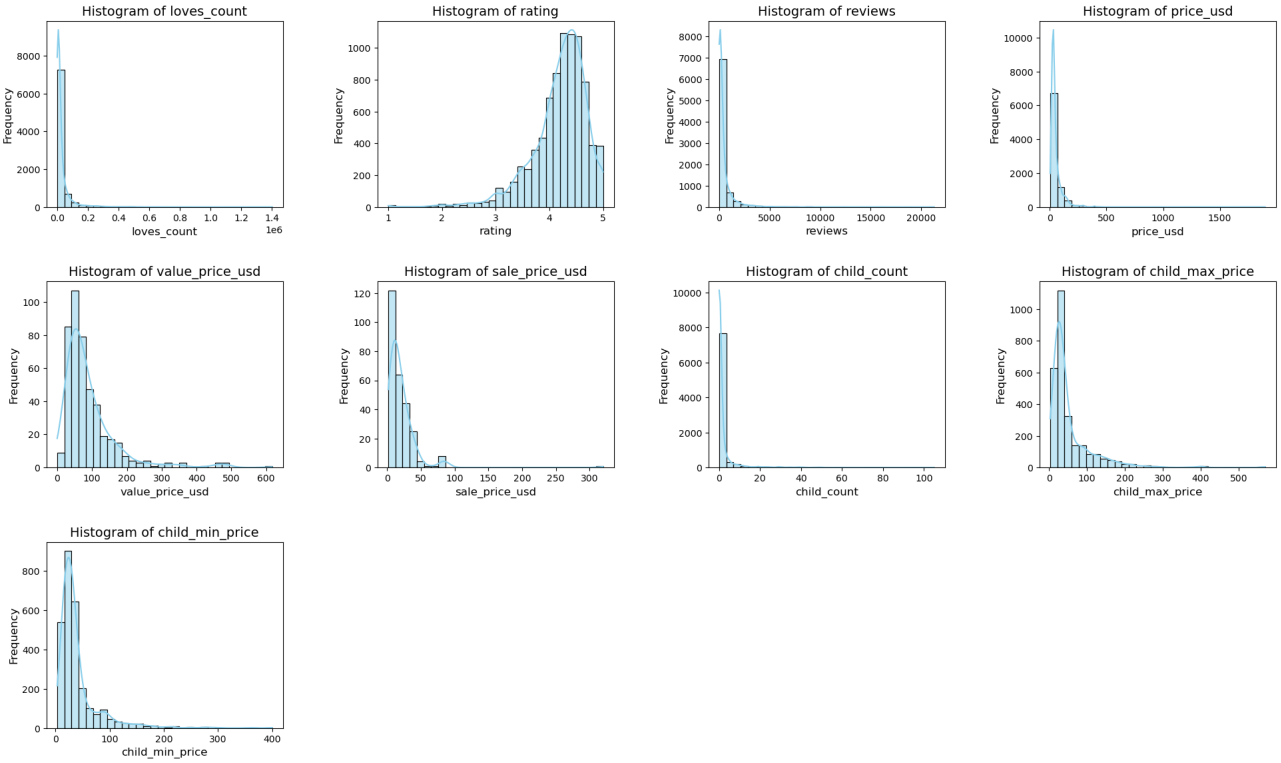


Figure 12 - Distribution of numerical data in Product_info dataset. Adapted from: Author's own work.

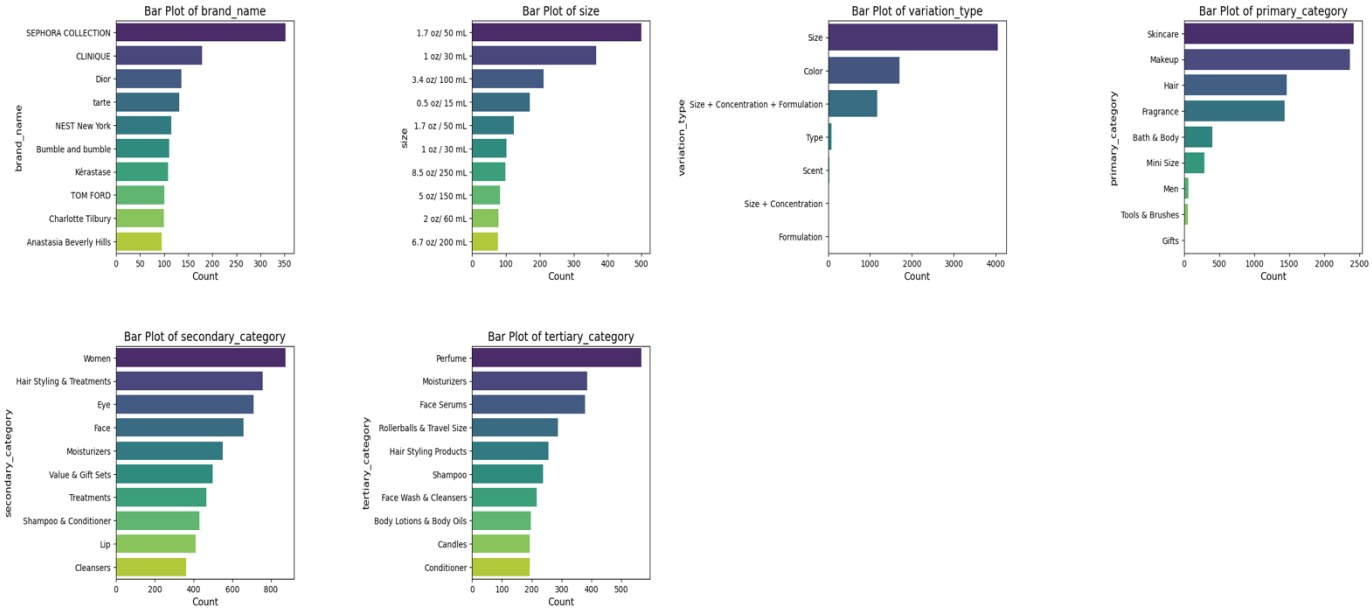


Figure 13 - Distribution of categorical data in Product_info dataset. Adapted from: Author's own work.

Figure 14 and Figure 15 clearly depict the null value distribution across columns for the *Product_info* dataset and the combined *review* dataset, respectively. These null values need to be identified so that the integrity of the data can be maintained. Those rows where 'reviews' or 'review_text' columns are missing were deleted to maintain the quality of the data. An inconsistency had also been identified in the length of some *author_ids*, which should be either 9 or 10 characters. Figure 16 highlights this issue, showing the *author_ids* with invalid lengths, which were subsequently removed from the dataset to preserve consistency.

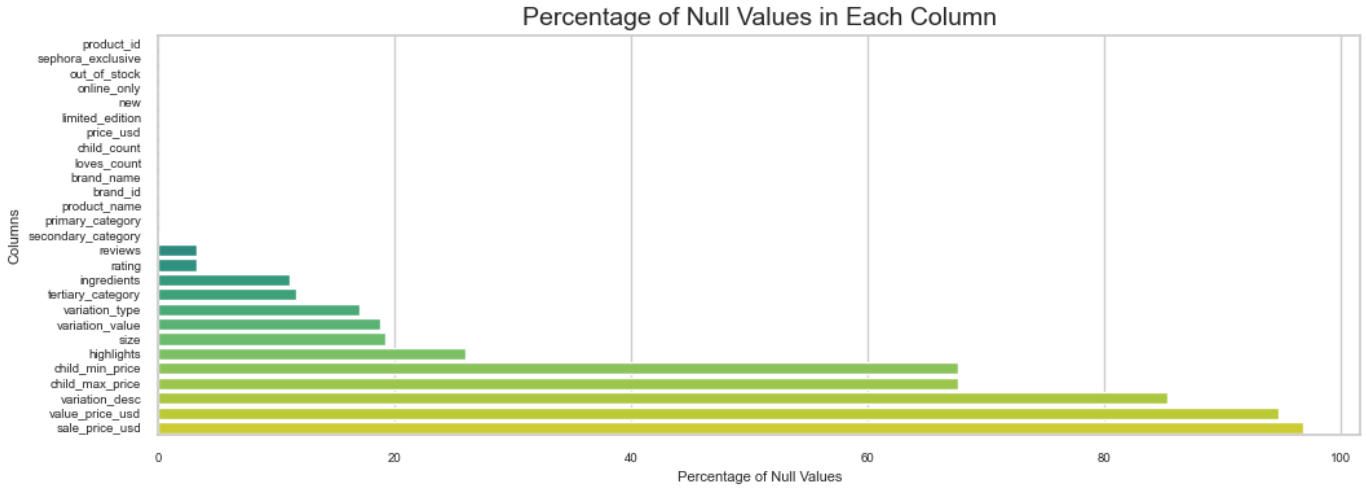


Figure 14 - Percentage of null values in Product_info dataset. Adapted from: Author's own work.

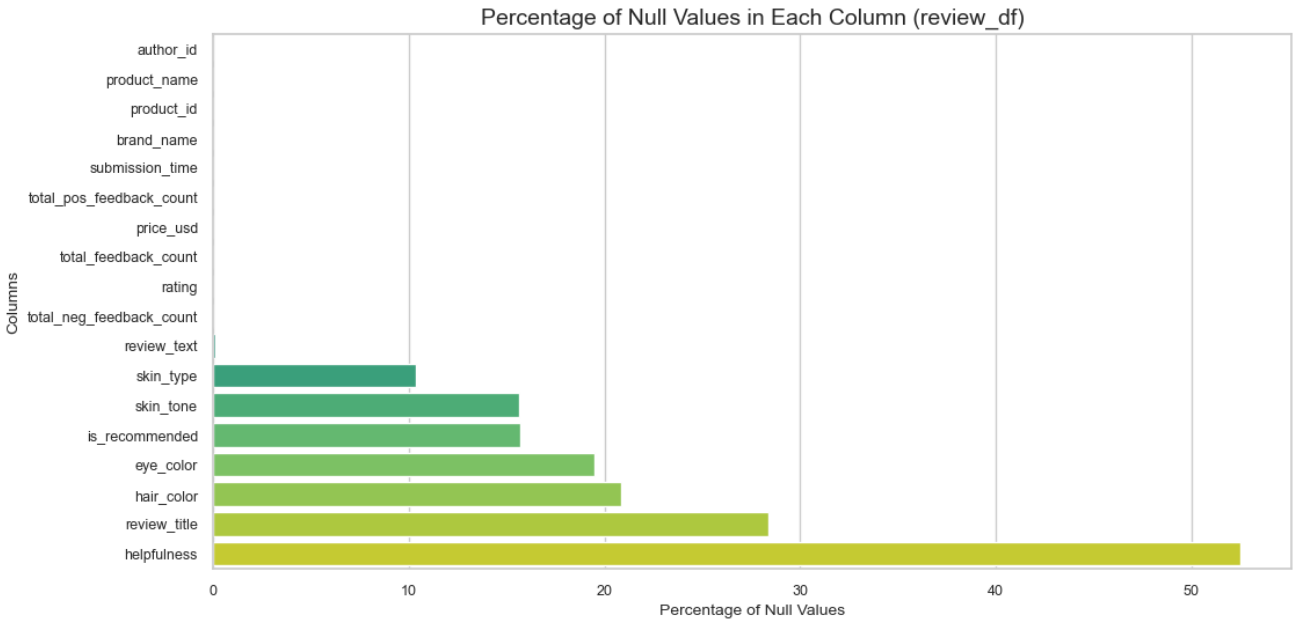


Figure 15 - Percentage of null values in the combined review data-frame. Adapted from: Author's own work.

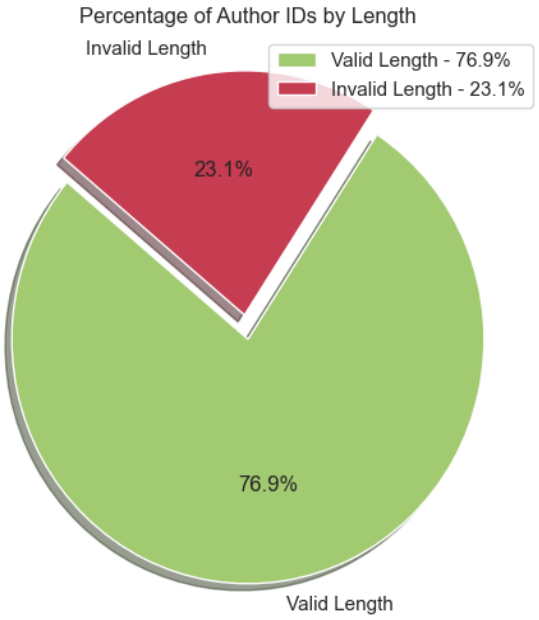


Figure 16 - Percentage of Author_ids by length. Adapted from: Author's own work.

4.1.2 Product Clustering Analysis

This section presents the results on product clustering, the first stage of the methodology. As mentioned in Section 3.1.3, K-means algorithm was employed to cluster products into groups using their pricing and popularity features. Figure 17 plots the within-cluster sum of squares (WCSS) also known as distortion score against every value of K. At K=5, an elbow point can be visibly noted in the graph, after which further decreases in WCSS values are gradual. As such, this method returns 5 as the optimum number of clusters. The graph also plots fit time at each value of K for additional information.

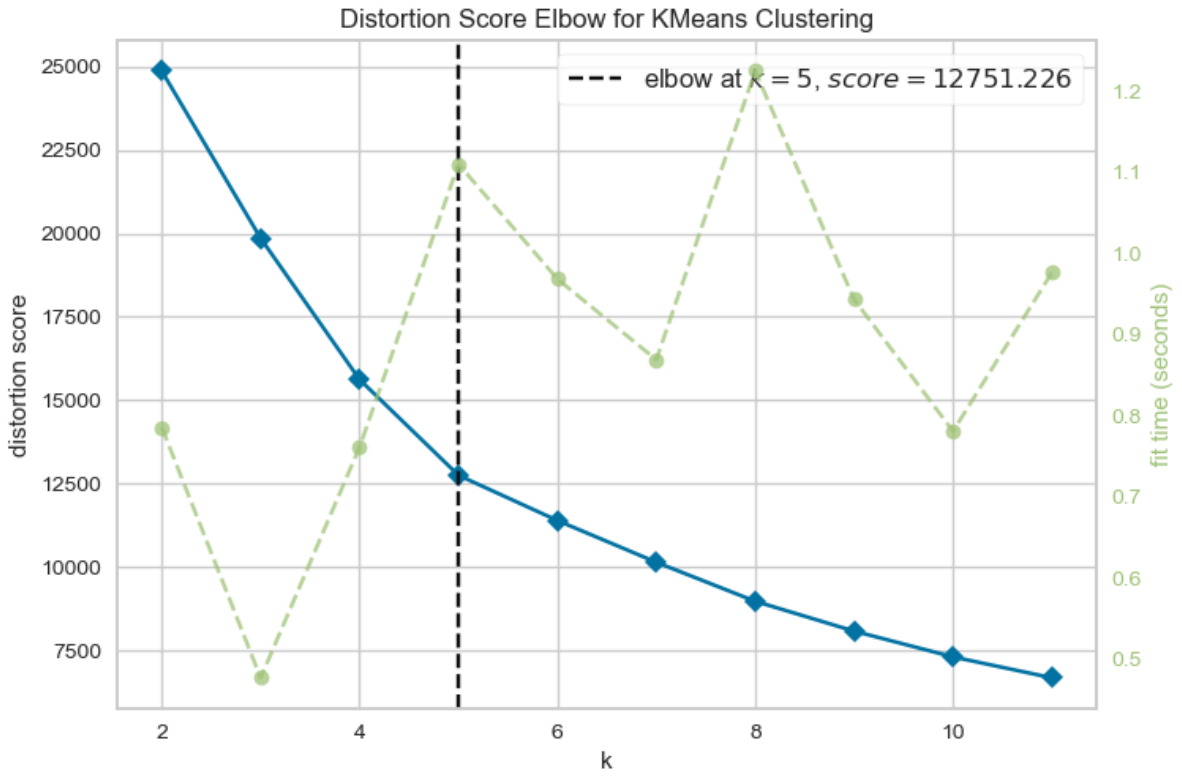


Figure 17 - Distortion score Elbow method for K-means clustering. Adapted from: Author's own work.

To further validate the number of optimal clusters, a silhouette analysis was performed. According to this analysis, K=5 indeed is the correct optimum number of clusters for this section of data. As shown in Figure 18, excluding K=2 since it gives only two clusters, K=5 corresponds to a silhouette score of 0.4 that is higher compared to other K values. Therefore, it can be concluded that K=5 corresponds to the optimal number of clusters.

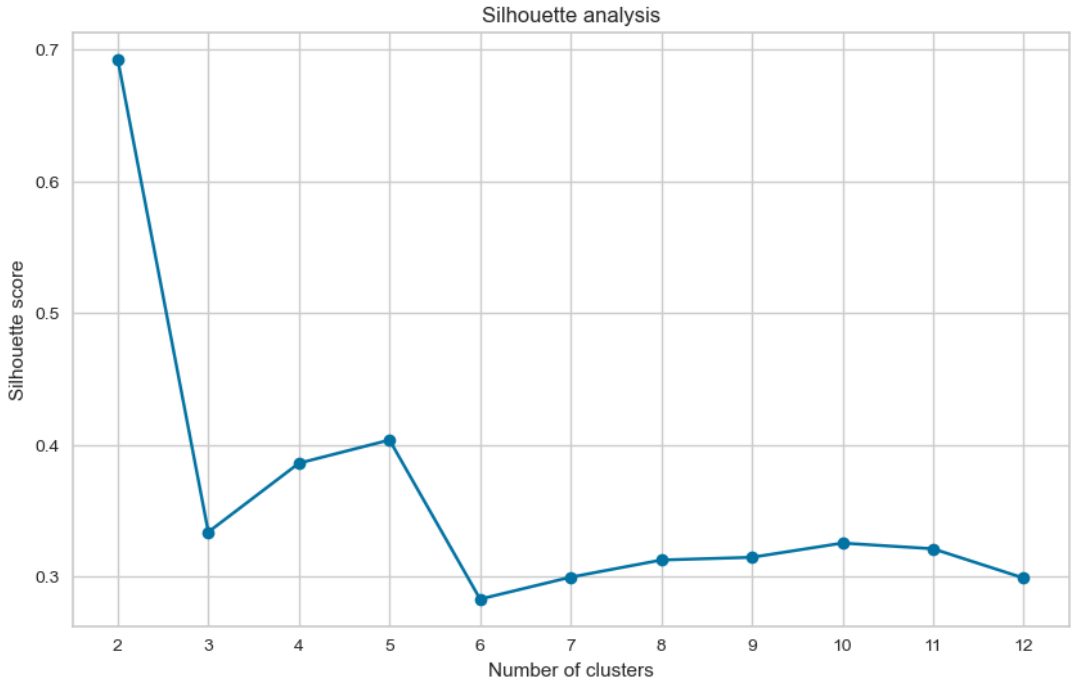


Figure 18 - Silhouette Analysis showing high score at k=5. Adapted from: Author's own work.

The products were clustered based on the following features: 'price_usd', 'rating', 'reviews', 'loves_count'. However, to visualize these clusters, it has to be in lower dimensions. Hence Principal Component Analysis (PCA), was conducted using the *PCA* library from *sklearn's decomposition* module. PCA is an empirically strong technique that reduces the dimensionality of data while keeping its variance for efficient visualization of clusters in lower space. It is particularly useful for reduction in high-dimensional data since it manages to bring out key patterns while losing a minimum amount of information (Jolliffe & Cadima, 2016). Figure 19 illustrates the visualisation of the product clusters after applying PCA, providing a clear and insightful view of the underlying structure in the data.

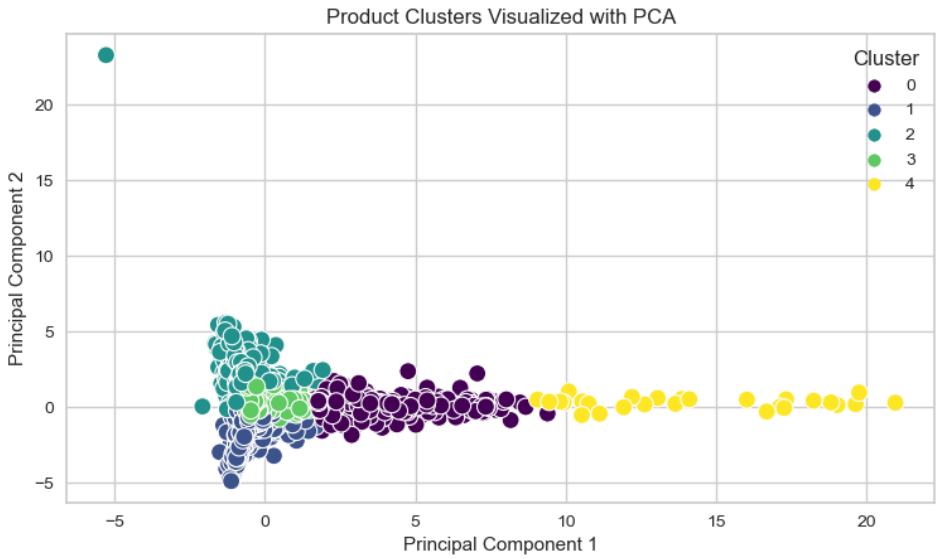


Figure 19 - Visualisation of product clusters using PCA. Adapted from: Author's own work.

The results of the cluster analysis are shown in Figure 20. The meaning of these different clusters can be understood quantitatively by looking at the averages of different features in these clusters. The average popularity of products is determined by considering a combination of 'rating,' 'reviews,' and 'loves_count,' the cluster factor of each cluster is determined by Equation 3, and it is normalized according to Equation 4 as described in Section 3.1.3. The analysis allows drawing the following conclusions with regards to the relationship of these five clusters with product popularity and pricing characteristic.

Cluster	0	1	2	3	4
Price_usd	-0.260365	-0.300372	2.373757	-0.20869	-0.419355
rating	0.147406	-1.479049	0.188671	0.420767	0.272149
reviews	2.141421	-0.274228	-0.172954	-0.122136	11.167902
loves_count	2.554085	-0.217962	-0.256337	-0.151425	9.09262
Avg Popularity	1.614304	-0.657079667	-0.080206667	0.049068667	6.844223667
Cluster Description	Very Popularand Mid-Rang	Moderately Popular and Affordable	Least popular		Highly Popular and Affordable
			Expensive	Mid-Range	
Cluster factor	6.200157471	2.18755299	0.033788912	0.235127062	16.32083477
Cluster factor Normalized	0.379	0.128	0	0.013	1

Figure 20 – Quantitative results of cluster analysis. Adapted from: Author's own work.

Cluster 0 can be labelled as "Very Popular and Mid-Range." It has an average popularity score of 1.614304, quite high as compared to other clusters in this analysis. The cluster has a high rating and several reviews, meaning that the products in this cluster are very popular with customers. More interestingly, although very popular, such items show mid-range pricing, possibly indicating where quality meets affordability.

In contrast, Cluster 1 would be the "Moderately Popular and Affordable" products. The items belonging to this cluster have lower ratings and fewer reviews compared to those in Cluster 0 but are able to stay competitive because of their much more accessible pricing. This could turn out to be an attractive market for budget-oriented consumers who will give up a bit of popularity in return for cost savings.

Now, Cluster 2 is an interesting case of being the "Least Popular and Expensive" group. Herein, having the highest average price, this group contains products with lower ratings, fewer reviews and lowest love counts. This may point to a price-perceived value mismatch or indicate this as a combination of niche products aimed at a smaller fraction of the market.

Products in Cluster 3 correspond to a very unique position: the "Least Popular Mid-Range" with a normalized cluster factor of 0.013. These products cost more compared to most clusters, except Cluster 2, while they suffer lower popularity compared to other clusters, notably Cluster 0 and Cluster 1. Products of this sort could be of moderate quality or have some brand reputation backing them up to a certain extent, thereby attracting consumers who would pay more for perceived benefits.

Finally, Cluster 4 is the "Highly Popular and Affordable" category. The rating and love count of a product in this category are high, while the price is low and highly competitive. What is likely be found in this cluster are the best value propositions, providing a compelling mix of quality, popularity, and affordability.

The computation of the cluster factors gives more context and thus a quick glance at how the clusters are performing. Cluster 4 has a normalized cluster factor of 1, meaning that generally it performs best amongst all the features analysed, while products within Cluster 0 present a relatively weak alternative with a cluster factor of 0.379. Such analysis could turn out to be very useful in context of pricing strategies, product positioning, and understanding consumer preferences within a market.

4.1.3 CFRBCS Results and Analysis

4.1.3.1 Hyperparameter Tunning

Section 3.1.8 of the methodology emphasizes the importance of hyperparameter tuning in improving the accuracy of the hybrid model. This is relevant to find the optimal weights for each individual similarity factor, as expressed in Equation 8. To this end, the RMSE was computed for a number of sets of α, β and γ, as graphically shown in Figure 22. Note that the sum of the weights is constrained to equal one.

The analysis showed that the combination that returned the lowest RMSE was 0.00275 for $\alpha = 0.1$, $\beta = 0.1$, and $\gamma = 0.8$. However, to have a more realistic balance in the distribution of weights, the combination chosen was $\alpha = 0.3$, $\beta = 0.1$, and $\gamma = 0.6$, which corresponded to an RMSE of 0.00310, still very close to the minimum value and is also very accurate. This balance keeps the model both accurate and practical. A 3D scatter plot of α, β and γ about the RMSE was generated to improve visualizing this relationship. In the plot (Figure 22), the darker the shade, the more accurate and lower the corresponding RMSE value is.

The other important aspect of optimization is finding the right value of S—the number of similar products used to predict the rating of a target product in the CFRBCS algorithm. For the determination of the optimum value, MAE was computed for different values of S, for the chosen set of α, β, and γ. As shown in Figure 21, MAE can be considered one of the critical measures of prediction accuracy, the smaller the value, the more accurate the prediction. The results of the analysis showed

that CFRBCS had the highest accuracy at an S setting of 2, which corresponded to a value of MAE that was the lowest and equal to 0.00008.

Furthermore, the behaviour of the MAE as the number of similar products (denoted as S) increases was closely examined. The results indicated that the MAE graph tends to stabilize at approximately 0.00017 as S increases beyond a certain point. This stabilization suggests that after reaching an optimal number of similar products, further increases in S have little to no effect on prediction accuracy. This finding emphasizes the importance of carefully selecting the optimal value of S rather than simply increasing it, as a strategy for refining the model and ensuring its reliability in producing accurate ratings.

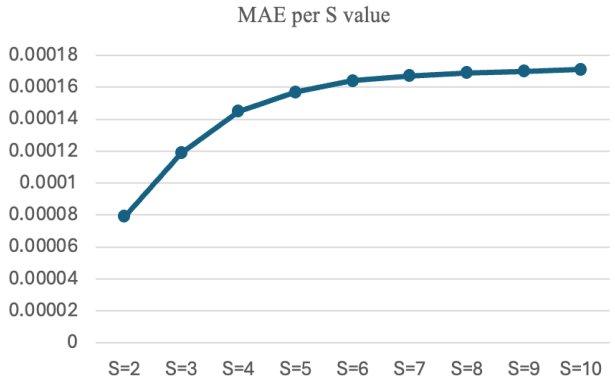
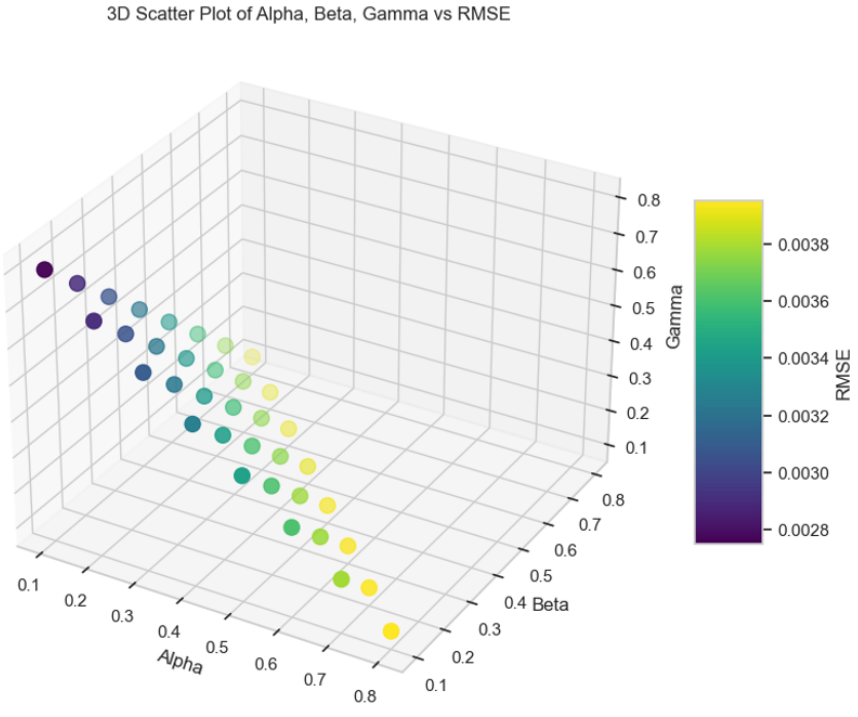


Figure 21 – Graph between MAE vs S values. Adapted from: Author's own work



α	β	γ	RMSE
0.8	0.1	0.1	0.003951
0.7	0.1	0.2	0.003784
0.7	0.2	0.1	0.003943
0.6	0.1	0.3	0.003611
0.6	0.2	0.2	0.003776
0.6	0.3	0.1	0.003937
0.5	0.1	0.4	0.003441
0.5	0.2	0.3	0.003606
0.5	0.3	0.2	0.003775
0.5	0.4	0.1	0.00393
0.4	0.1	0.5	0.00326
0.4	0.2	0.4	0.00343
0.4	0.3	0.3	0.00360
0.4	0.4	0.2	0.00374
0.4	0.5	0.1	0.00390
0.3	**0.1**	**0.6**	**0.00310**
0.3	0.2	0.5	0.00323
0.3	0.3	0.4	0.00340
0.3	0.4	0.3	0.00357
0.3	0.5	0.2	0.00374
0.3	0.6	0.1	0.00391
0.2	0.1	0.7	0.00290
0.2	0.2	0.6	0.00306
0.2	0.3	0.5	0.00323
0.2	0.4	0.4	0.00339
0.2	0.5	0.3	0.00357
0.2	0.6	0.2	0.00374
0.2	0.7	0.1	0.00390
0.1	0.1	0.8	0.00275
0.1	0.2	0.7	0.00290
0.1	0.3	0.6	0.00306
0.1	0.4	0.5	0.00322
0.1	0.5	0.4	0.00339
0.1	0.6	0.3	0.00356
0.1	0.7	0.2	0.00374
0.1	0.8	0.1	0.00390

Figure 22 - 3D plot of Alpha, Beta and Gamma vs RMSE along with tabulation of RMSE values for different weight combinations. Adapted from: Author's own work.

4.1.3.2 CFRBCS Model Evaluation

Predictions of the CFRBCS model were verified on a test dataset of 104,300 records. This model was evaluated using four key criteria: RMSE, MAE, MAPE, and MSLE, as described in section 3.1.7. The results, as shown in Figure 23, make very interesting reading and comparison among the Traditional, CFRBS, and CFRBCS models, bringing out clearly the advantages of the latter.

Figure 24 indicates that, among all the evaluation metrics, the CFRBCS model outperforms the others. On the RMSE, it reduces the error to about 0.003, way below the 0.007 attained using both Traditional and CFRBS models. This drop in error thus reflects the efficiency of CFRBCS model in minimizing large deviations, leading to more accurate predictions overall. This trend also extends into the MAE, where the CFRBCS model retains an approximate value of 0.000074 against the higher 0.0001 for the other methods. This reduction in average error of nearly 40% demonstrates that precision for CFRBCS remains steady and returns more reliable results from across the board.

Comparative plots of MAPE and MSLE further underline the supremacy of CFRBCS. More precisely, the MAPE for the CFRBCS model is close to 0.000023, much better than that obtained for traditional and CFRBS methods, which remains close to 0.000034. Hence, this reduction in percentage error points toward the reliability of the model in applications for which relative accuracy is paramount. Probably the most striking comparison comes from the MSLE, where CFRBCS manages to reduce the error to just 0.556e-6 against the rest, which stay close to 6.63e-6. A large fraction of this reduction can be accounted for by the fact that CFRBCS is robust against data with varying scales; in particular, it includes exponential growth and strongly skewed distributions.

The quantitative evidence is clear: CFRBCS model does not just close the gap in performance with traditional methods but does so with a level of precision that establishes it as an effective tool in predictive modelling that's both accurate and reliable. Its consistent outperformance across all metrics leaves no doubt that CFRBCS is the most effective model among those tested, solidifying its status as the top choice for predicting customer ratings for the given test data.

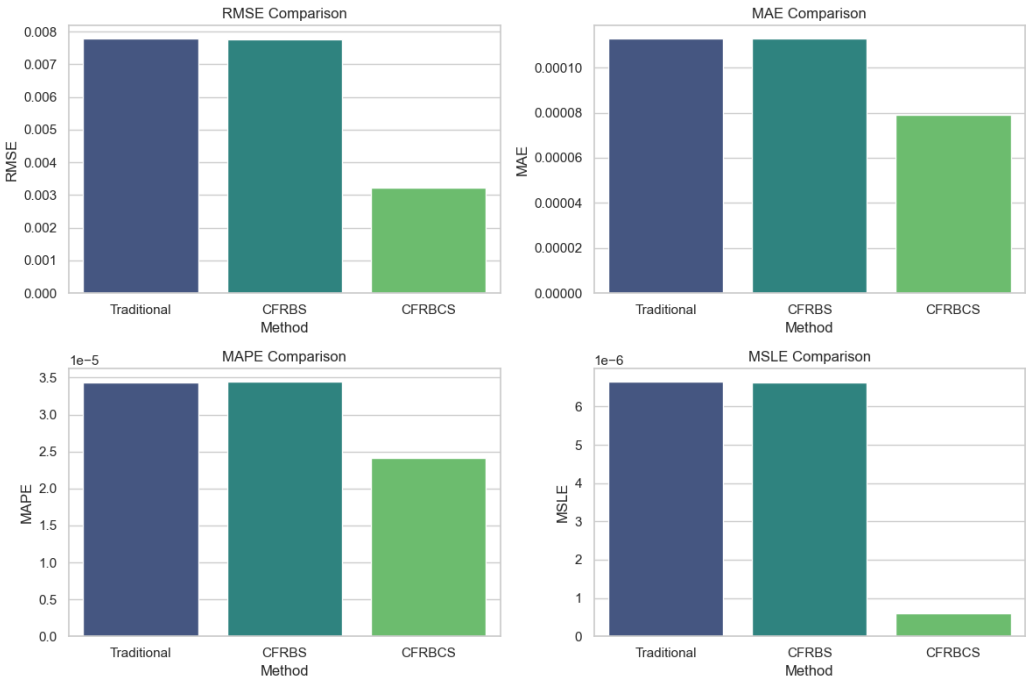


Figure 23 - Model Evaluation Comparison. Adapted from: Author's own work.

Method	RMSE	MAE	MAPE	MSLE
Traditional	0.007792	0.000113	0.000034	6.65E-06
CFRBS	0.007765	0.000113	0.000034	6.63E-06
CFRBCS	0.003097	0.000074	0.000023	5.56E-07

Figure 24 - Tabular comparison of the three methods across various evaluation metrics. Adapted from: Author's own work.

4.1.3.3 CFRBCS Model Recommendations

Figure 25 shows how the CFRBCS model is used to obtain a recommendation list of the top N products against a target user. Based on Section 3.2, the CFRBCS model predicts ratings that a target user would give to a certain product. From these predicted ratings, the top N products having the highest scores are picked as recommendations. For example, for target user '1696370280,' the following are recommended as the top 10 products: 'P446938', 'P461933', 'P4015', 'P427410', 'P439055', 'P426829', 'P462699', 'P297516', and 'P464239.' Recommendations are provided incorporating the purchase history of the user, the general sentiment score of every individual product, and their cluster factors.

```
target_user_id = '1696370280'

# Call the function to get the top N recommended products
top_recommendations = recommend_top_n_products(
    target_author_id=target_author_id,
    item_user_matrix=filtered_test_item_user_rating_matrix,
    cosine_similarity_matrix=combined_similarity_with_cluster,
    K=2,
    N=10
)

# Display the recommendations
print("Top Recommended Products:")
print(top_recommendations)
```

[444] ✓ 0.4s

```
Top Recommended Products:
    Product_ID  Predicted_Rating
518    P446938               5.0
279    P422257               5.0
715    P461933               5.0
165      P4015               5.0
316    P427410               5.0
412    P439055               5.0
308    P426829               5.0
725    P462699               5.0
70     P297516               5.0
740    P464239               5.0
```

Figure 25 - CFRBCS model product recommendations. Adapted from: Author's own work.

Chapter 5 – Discussion

5.1 Discussion of Results in Relation to Research Questions

The results of this research provide very insightful answers to the core research questions (Section 1.2) and help gain a better understanding of the advantages of this hybrid approach towards RS.

5.1.1 RQ1: Identification of Key Product Segments

The first research question sought to identify key product segments and establish how these segments drive product preferences. Through the application of the K-means clustering algorithm, five distinct product segments were identified, each with unique characteristics based on price and popularity features.

The results are very interesting and highlight that the identified segments are unique, helping towards understanding consumer preferences, as evidenced by varying cluster factor levels among the clusters. For instance, products belonging to Cluster 4 (Cluster factor = 1) fall under the 'Highly Popular and Affordable' category, as they are not only highly rated but also more budget-friendly for consumers, making them a great choice for recommendations. On the other hand, this segment contrasts sharply with Cluster 2 (Cluster factor = 0), 'Least Popular and Expensive,' which indicates a potential price-perceived value mismatch, and products in this cluster would not make great recommendations. Segmentation of this kind enhances the categorization of products, and the CFRBCS leverages this information to make more accurate suggestions and predictions. This approach can help reduce the cold start problem faced by RS systems (Guan et al., 2024). For instance, with the help of segmented data, an effective recommendation (any item from 'Highly Popular and Affordable' cluster) can be made to new users who have the least amount of historical data, thus easily and effectively improving the overall recommendation efficiency of the system.

5.1.2 RQ2: Consumer Reviews Sentiment Analysis

The second research question guided the study in finding out how the sentiment analysis of customer reviews could help in understanding the product preferences and enhancing RS. Sentiment analysis helps retrieve very important dimensions of emotions and subjectivity in customer feedback, which are not captured by metrics, such as ratings alone. This research has enhanced the prediction capabilities of the hybrid recommendation model by incorporating sentiment analysis into the model for more and accurate predictions. As shown in the performance analysis for the CFRBCS model, this model performed better than the traditional models in all measures: RMSE, MAE, MAPE, and MSLE. Such lower error rates obtained by incorporating sentiment analysis prove that understanding the emotional tone of customer reviews, be it positive, negative, or even neutral, adds great value during the recommendation process.

An important caveat to note here is that while sentiment analysis did elevate the efficacy of the CFRBCS model, according to hyperparameter tuning results, heavy weight should not be placed on this factor since it would significantly reduce accuracy.

This also intuitively makes a lot of sense as customer sentiment is very personal. While one consumer may like a product, another may not, and hence this factor alone cannot be used as a way to provide recommendations.

5.1.3 RQ3: Comparative Analysis of Recommendation Models

The third research question was aimed at determining the predictive accuracy of the proposed Hybrid model, which integrates product segmentation and sentiment analysis, against baseline models of recommendation and the CFRBS model by Jian Zhen Yu et al. (2018).

Results were compelling in that the CFRBCS model, had an upper hand as compared to traditional models relying on ratings only. The outcome was that the CFRBCS model drastically reduced the RMSE and considerably lowered the average error for MAE, which is indeed reflective of the efficiency of the CFRBCS model over others in predicting customer ratings.

Secondly, it's clear that one of the major strengths of the model it's robustness in error reduction particularity with regards to MSLE, showing the model can handle data with varying scales and distributions. The comparative analysis mainly goes to prove the hypothesis: using a combination of product segmentation with sentiment analysis will yield a more accurate and reliable recommendation system.

5.2 Practical Implications and Significance of the Findings

These results have important implications for stakeholders, especially marketing professionals and product managers in the online retail industry. The hybrid recommendation model can prove beneficial to marketing professionals due to the insights generated on product clusters. Using these insights, targeted marketing campaigns can be created to appeal specifically to budget-conscious customers. For instance, products identified in the "Highly Popular and Affordable" cluster could be recommended to cost-sensitive customers in a more target-oriented and effective way.

The managers in e-commerce companies can use these results to refine their inventory control. Since the hybrid model can help with accurate predictions, it can be used as a blueprint for resource allocation and strategic planning. This would mean that businesses can focus their stocking effort into those goods likely to receive higher ratings and thus ensure a faster turnover from an order to delivery. Such productivity directly results in increased satisfaction and retention levels of customers, which is very important for long-term success.

Additionally, the insights that can be used to greater extent especially in companies with a great number of products on offer. If companies have information on which products are likely to be in high demand, they would avoid overstocking and understocking, saving a lot in the process. Such operational efficiency brings extra value to the bottom line and enhances the quality of service offered, further entrenching the company in customer loyalty.

5.3 Limitations of the Study

The following points are few limitations of the study. First, this study limits sentiment analysis to only Lexicon-Based Approaches and does not incorporate other techniques, such deep learning models, using which more accurate sentiment interpretations from customer reviews can be obtained.

Another limitation is that the model was trained on Sephora website data only, which may affect generalizability. The model was very accurate with Sephora data, but its accuracy could be reduced when running data from other sources or companies. Additional finetuning would have to be conducted in order for the model to fit other datasets, especially very different patterns in customer behaviour or characteristics of the products. Therefore, though the results are very robust in this particular context, further research with adjustments will be needed to generalize this model over more e-commerce websites for its implementation effectively.

Chapter 6 – Conclusion

6.1 Summary of the Study

This research was performed to find out how to improve recommendation system in terms of accuracy and relevance for the domain of e-commerce. Detailed literature presented in Chapter 2 – Literature Review offered an in-depth insight into the present state of recommender systems and provided the shortcomings and/or possible areas for improvement. While the original approach was to use sentiment analysis solely to enhance RS, a gap was identified: product cluster analysis together with sentiment analysis is relatively unexplored in current studies. Such an observation led to the development of the CFRBCS algorithm to merge the strengths of these two techniques for the betterment of recommendation systems.

Chapter 3 – Methodology elaborated on the methodology followed to develop the CFRBCS algorithm, at different stages. In clustering techniques, it was found that the K-means algorithm is robust and efficient to segment the product data into distinct clusters. Finally, even though the scope was restricted to Lexicon-Based Approaches for sentiment analysis, the VADER method was used to extract sentiments from the user comments and include this information in making recommendations. From the results and discussions of Chapter 4 – Results and Analysis and Chapter 5 – Discussion, it becomes clear that the calculation of cluster factors for every segment was an important factor to enhance the accuracy and reliability of the CFRBCS algorithm. The CFRBCS model was developed with an iterative process for hyperparameter tuning, thus distributing the weights optimally and proved that though important and useful, the sentiment analysis factor is not worth overemphasizing at the expense of other factors. Section 5.1 Discussion of Results in Relation to Research Questions shows how the results help answer the research questions and Section 5.2 Practical Implications and Significance of the Findings points out a few of the major implications for marketing professionals and product managers should they decide to use the hybrid recommendation model, since it has very great promise to facilitate targeted marketing, better inventory management, and overall operational efficiency improvement. The result of these improvements will then increase customer satisfaction and customer loyalty. Finally, Section 3.4 Ethical considerations of the study also examines broader ethical risks and consequences associated with the use of AI tools in recommendation systems.

6.2 Key Takeaways, Recommendations and Future Scope

A key takeaway from this research focuses on how integration of product clustering with sentiment analysis can power recommendation systems. It is indicated in the study how e-commerce platforms can leverage such insights to help in the betterment of their operations. Given the modern age competitive landscape where firms are battling for market share, data analytics can help businesses garner a better understanding of the preferences of their customers and outperform other competitors.

The research is also indicative of the fact that clustering products, in the inventory, based on selected features proves beneficial for more than one reason. This segmentation step is curial towards increasing the robustness of the CFRBCS algorithm; and in the process sheds light of various product cluster statistics that help drive marketing insights.

While this study is primarily intended to showcase the efficiency and practicality of the CFRBCS algorithm to marketing professionals and product managers, there are a few recommendations that researchers in the field of analytics can implement as a future scope. The latest research trends indicate that Deep Learning-Based Methods are becoming

increasingly more valuable (Kaur & Sharma, 2024). Methods like CNN (Krishna Kumar Mohbey et al., 2023), RNN (Imad Zyout & Mo'ath Zyout, 2024), and transformer models like BERT (Wu et al., 2024) can be used to perform sentiment analysis instead of the VADER model used in this research, to obtain a more sentiment-aware version of the CFRBCS algorithm. Similarly, the scope of this project is limited to K-means based clustering, and future research can look to integrate and compare other forms of clustering such as DB-SCAN. Exploring different permutations and combinations of various AI techniques for clustering and sentiment analysis, while conducting a comparative analysis, is another avenue that can be explored in the future.

To fully benefit from the advantages of this research, the author's future scope is to build a web application that real-world end users (anyone interested in the Sephora product catalogue) can interact with to get product recommendations generated by the CFRBCS algorithm. In the upcoming iteration of this work, the author plans to integrate data from other companies and train the CFRBCS algorithm to make its application more universal.

Researchers, marketing analysts, and other stakeholders, if interested, can draw inspiration from the workings of the CFRBCS algorithm and implement a similar approach with modifications tailored to their specific scenarios. To allow for transparency and accountability, the source code for this project is presented as two Jupyter Notebook files (Code Files and Readme File) and can be found at (https://github.com/Mukundh141-exeter/BEMM466.git).

As a final word, this research study was conducted with the aim of improving the accuracy and relevance of recommendation systems in e-commerce. The research justifies that combining product clustering and sentiment analysis with item-based collaborative filtering is an innovative way to achieve this aim, forming the basis for the developed CFRBCS algorithm. The practical implications and significance of the CFRBCS algorithm are explored, along with a discussion on ethics, which is especially important in the current context of AI development and implementation. This study hopes to facilitate and provoke future research and development in the field of recommender systems and showcase the potential of combining clustering and sentiment analysis techniques.

References

Almohsen, K., & Al-Jobori, H. (2015). Recommender Systems in Light of Big Data. *International Journal of Electrical and Computer Engineering (IJECE)*, *5*(6), 1553. https://doi.org/10.11591/ijece.v5i6.pp1553-1563

Aastha Gupta, Himanshu Sharma, & Anas Akhtar. (2021). A COMPARATIVE ANALYSIS OF K-MEANS AND HIERARCHICAL CLUSTERING. *EPRA International Journal of Multidisciplinary Research (IJMR)*, *7*(8), 412–418. https://doi.org/10.36713/epra8308

Abbasi-Moud, Z., Vahdat-Nejad, H., & Sadri, J. (2021). Tourism recommendation system based on semantic clustering and sentiment analysis. *Expert Systems with Applications*, *167*, 114324. https://doi.org/10.1016/j.eswa.2020.114324

Abu-AlSondos, I. A., Alkhwaldi, A. F., Salhab, H. A., Shehadeh, M., & Ali, B. J. A. (2023). Customer attitudes towards online shopping: A systematic review of the influencing factors. *International Journal of Data and Network Science*, *7*(1), 513–524. https://doi.org/10.5267/j.ijdns.2022.12.013

Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, *17*(6), 734–749. https://doi.org/10.1109/tkde.2005.99

Afoudi, Y., Lazaar, M., & Al Achhab, M. (2021). Hybrid recommendation system combined content-based filtering and collaborative prediction using artificial neural network. *Simulation Modelling Practice and Theory*, *113*(102375), 102375. https://doi.org/10.1016/j.simpat.2021.102375

Aida Osman, N., & Azman Mohd Noah, S. (2018, March 1). Sentiment-Based Model for Recommender Systems. *IEEE Xplore*. https://doi.org/10.1109/INFRKM.2018.8464694

Ajaegbu, C. (2021). An optimized item-based collaborative filtering algorithm. *Journal of Ambient Intelligence and Humanized Computing*, *12*, 10629–10636. https://doi.org/10.1007/s12652-020-02876-1

Anitha, P., & Patil, M. M. (2019). RFM model for customer purchase behavior using K-Means algorithm. *Journal of King Saud University - Computer and Information Sciences*, *34*(5). https://doi.org/10.1016/j.jksuci.2019.12.011

Bellini, P., Palesi, L. A. I., Nesi, P., & Pantaleo, G. (2022). Multi Clustering Recommendation System for Fashion Retail. *Multimedia Tools and Applications*, *82*. https://doi.org/10.1007/s11042-021-11837-5

Bhaskaran, S., & Marappan, R. (2021). Design and analysis of an efficient machine learning based hybrid recommendation system with enhanced density-based spatial clustering for digital e-learning applications. *Complex & Intelligent Systems*, *9*(8). https://doi.org/10.1007/s40747-021-00509-4

Bhavitha, B. K., Rodrigues, A. P., & Chiplunkar, N. N. (2017). Comparative study of machine learning techniques in sentimental analysis. *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)*. https://doi.org/10.1109/icicct.2017.7975191

Bing Liu. (2012). *Sentiment analysis and opinion mining*. Morgan And Claypool.

Bobadilla, J., Ortega, F., Hernando, A., & Gutiérrez, A. (2013). Recommender systems survey. *Knowledge-Based Systems*, *46*, 109–132. https://doi.org/10.1016/j.knosys.2013.03.012

Braune, C., Besecke, S., & Kruse, R. (2014). Density Based Clustering: Alternatives to DBSCAN. *Partitional Clustering Algorithms*, 193–213. https://doi.org/10.1007/978-3-319-09259-1_6

businesswire. (2024, July 8). Decision Fatigue is Ruining Online Shopping, But Existing Tech Can Help. Silicon UK. https://www.silicon.co.uk/press-release/decision-fatigue-is-ruining-online-shopping-but-existing-tech-can-help

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

Chevalier, S. (2024, February 6). Global Retail e-commerce Market Size 2014-2027. Statista. https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/

Chiranjeevi, P., & Rajaram, A. (2023). A lightweight deep learning model based recommender system by sentiment analysis. *Journal of Intelligent & Fuzzy Systems*, *44*(6), 1–14. https://doi.org/10.3233/jifs-223871

Cui, Z., Xu, X., Xue, F., Cai, X., Cao, Y., Zhang, W., & Chen, J. (2020). Personalized Recommendation System Based on Collaborative Filtering for IoT Scenarios. *IEEE Transactions on Services Computing*, *13*(4), 685–695. https://doi.org/10.1109/tsc.2020.2964552

Dang, C. N., Moreno-García, M. N., & Prieta, F. D. la. (2021). An Approach to Integrating Sentiment Analysis into Recommender Systems. *Sensors*, *21*(16), 5666. https://doi.org/10.3390/s21165666

Dara, S., Chowdary, C. R., & Kumar, C. (2019). A survey on group recommender systems. *Journal of Intelligent Information Systems*, *54*, 271–295. https://doi.org/10.1007/s10844-018-0542-3

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, May 24). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv.org (Cornell University)*. https://doi.org/10.48550/arXiv.1810.04805

Elahi, M., Khosh Kholgh, D., Kiarostami, M. S., Oussalah, M., & Saghari, S. (2023). Hybrid recommendation by incorporating the sentiment of product reviews. *Information Sciences*, *625*, 738–756. https://doi.org/10.1016/j.ins.2023.01.051

EUROPEAN COMMISSION. (2021, February 10). *The Act*. The Artificial Intelligence Act. https://artificialintelligenceact.eu/the-act/

European Union. (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance)*. Europa.eu. https://eur-lex.europa.eu/eli/reg/2016/679/oj

Fayyaz, Z., Ebrahimian, M., Nawara, D., Ibrahim, A., & Kashef, R. (2020). Recommendation Systems: Algorithms, Challenges, Metrics, and Business Opportunities. *Applied Sciences*, *10*(21), 7748. https://doi.org/10.3390/app10217748

Fkih, F. (2021). Similarity measures for Collaborative Filtering-based Recommender Systems: Review and experimental comparison. *Journal of King Saud University - Computer and Information Sciences*, *34*(9). https://doi.org/10.1016/j.jksuci.2021.09.014

Gomez-Uribe, C. A., & Hunt, N. (2016). The Netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems*, *6*(4), 1–19. https://doi.org/10.1145/2843948

Gov.UK. (2024, February 6). *A pro-innovation Approach to AI regulation: Government Response*. GOV.UK. https://www.gov.uk/government/consultations/ai-regulation-a-pro-innovation-approach-policy-proposals/outcome/a-pro-innovation-approach-to-ai-regulation-government-response

Gov.Uk. (2018). *Data Protection Act 2018*. Legislation.gov.uk. https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted

Guan, J., Chen, B., & Yu, S. (2024). A hybrid similarity model for mitigating the cold-start problem of collaborative filtering in sparse data. *Expert Systems with Applications*, *249*, 123700–123700. https://doi.org/10.1016/j.eswa.2024.123700

Hajer Nabli, Raoudha Ben Djemaa, & Amor, B. (2023). Improved clustering-based hybrid recommendation system to offer personalized cloud services. *Cluster Computing*, *27*, 2845–2874. https://doi.org/10.1007/s10586-023-04119-2

Hodson, T. O., Over, T. M., & Foks, S. S. (2021). Mean Squared Error, Deconstructed. *Journal of Advances in Modeling Earth Systems*, *13*(12). https://doi.org/10.1029/2021ms002681

Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), 216–225. https://doi.org/10.1609/icwsm.v8i1.14550

Huynh, H. X., Phan, N. Q., Pham, N. M., Pham, V.-H., Hoang Son, L., Abdel-Basset, M., & Ismail, M. (2020). Context-Similarity Collaborative Filtering Recommendation. *IEEE Access*, *8*, 33342–33351. https://doi.org/10.1109/access.2020.2973755

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2022). K-means Clustering Algorithms: a Comprehensive Review, Variants Analysis, and Advances in the Era of Big Data. *Information Sciences*, *622*(622). https://doi.org/10.1016/j.ins.2022.11.139

Imad Zyout, & Mo'ath Zyout. (2024). Sentiment analysis of student feedback using attention-based RNN and transformer embedding. *IAES International Journal of Artificial Intelligence*, *13*(2), 2173–2173. https://doi.org/10.11591/ijai.v13.i2.pp2173-2184

Inky, N. (2023). *Sephora Products and Skincare Reviews*. Www.kaggle.com. https://www.kaggle.com/datasets/nadyinky/sephora-products-and-skincare-reviews

Jannach, D., Manzoor, A., Cai, W., & Chen, L. (2021). A Survey on Conversational Recommender Systems. *ACM Computing Surveys*, *54*(5), 1–36. https://doi.org/10.1145/3453154

Jian Zhen Yu, An, Y., Xu, T., Gao, J., Zhao, M., & Yu, M. (2018). Product Recommendation Method Based on Sentiment Analysis. *Web Information Systems and Applications*, 488–495. https://doi.org/10.1007/978-3-030-02934-0_45

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Karabila, I., Darraz, N., El-Ansari, A., Alami, N., & El Mallahi, M. (2023). Enhancing Collaborative Filtering-Based Recommender System Using Sentiment Analysis. *Future Internet*, *15*(7), 235. https://doi.org/10.3390/fi15070235

Karn, A. L., Karna, R. K., Kondamudi, B. R., Bagale, G., Pustokhin, D. A., Pustokhina, I. V., & Sengan, S. (2022). Customer centric hybrid recommendation system for E-Commerce applications by integrating hybrid sentiment analysis. *Electronic Commerce Research*, *23*(10). https://doi.org/10.1007/s10660-022-09630-z

Kaur, G., & Sharma, A. (2024). Automatic customer review summarization using deep learning-based hybrid sentiment analysis. *International Journal of Power Electronics and Drive Systems/International Journal of Electrical and Computer Engineering*, *14*(2), 2110–2110. https://doi.org/10.11591/ijece.v14i2.pp2110-2125

Khanal, S. S., Prasad, P. W. C., Alsadoon, A., & Maag, A. (2019). A systematic review: machine learning based recommendation systems for e-learning. *Education and Information Technologies*, *25*, 2635–2664. https://doi.org/10.1007/s10639-019-10063-9

Kim, S., & Kim, H. (2016). A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting*, *32*(3), 669–679. https://doi.org/10.1016/j.ijforecast.2015.12.003

Anastasiia Klimashevskaia, Dietmar Jannach, Elahi, M., & Trattner, C. (2024). A survey on popularity bias in recommender systems. *User Modeling and User-Adapted Interaction*. https://doi.org/10.1007/s11257-024-09406-0

Ko, H., Lee, S., Park, Y., & Choi, A. (2022). A Survey of Recommendation Systems: Recommendation Models, Techniques, and Application Fields. *Electronics*, *11*(1), 141. https://doi.org/10.3390/electronics11010141

Krishna Kumar Mohbey, Meena, G., Kumar, S., & K. Lokesh. (2023). A CNN-LSTM-Based Hybrid Deep Learning Approach for Sentiment Analysis on Monkeypox Tweets. *New Generation Computing*, *42*, 89–107. https://doi.org/10.1007/s00354-023-00227-0

Lakshay Bharadwaj. (2023). Sentiment Analysis in Online Product Reviews: Mining Customer Opinions for Sentiment Classification. *International Journal for Multidisciplinary Research*, *5*(5). https://doi.org/10.36948/ijfmr.2023.v05i05.6090

Liu, H., Hu, Z., Mian, A., Tian, H., & Zhu, X. (2014). A new user similarity model to improve the accuracy of collaborative filtering. *Knowledge-Based Systems*, *56*, 156–166. https://doi.org/10.1016/j.knosys.2013.11.006

Maheshwari, R. (2023, June 14). *Online Shopping And Its Advantages*. Forbes Advisor. https://www.forbes.com/advisor/in/credit-card/advantages-of-online-shopping/

Md Nadeem Noori, Ahamed, J., & Ahmed, M. (2024). Matrix Factorization and Cosine Similarity Based Recommendation System For Cold Start Problem in E-Commerce Industries. *International Journal of Computing and Digital System/International Journal of Computing and Digital Systems*, *15*(1), 775–787. https://doi.org/10.12785/ijcds/150156

Micol Policarpo, L., da Silveira, D. E., da Rosa Righi, R., Antunes Stoffel, R., da Costa, C. A., Victória Barbosa, J. L., Scorsatto, R., & Arcot, T. (2021). Machine learning through the lens of e-commerce initiatives: An up-to-date systematic literature review. *Computer Science Review*, *41*, 100414. https://doi.org/10.1016/j.cosrev.2021.100414

Mitra, A. (2020). Sentiment Analysis Using Machine Learning Approaches (Lexicon based on movie review dataset). *Journal of Ubiquitous Computing and Communication Technologies*, *2*(3), 145–152. https://doi.org/10.36548/jucct.2020.3.004

Naeem, S., & Aishan Wumaier. (2018). Study and Implementing K-mean Clustering Algorithm on English Text and Techniques to Find the Optimal Value of K. *International Journal of Computer Applications*, *182*(31), 7–14. https://doi.org/10.5120/ijca2018918234

Nicholas, J. S., & Francis, F. S. (2019). A Comprehensive Survey of Neighborhood-Based Recommendation Methods used in E-Learning Recommender Systems. *International Journal of Computer Sciences and Engineering*, *7*(1), 443–450. https://doi.org/10.26438/ijcse/v7i1.443450

Palmatier, R. W., & Sridhar, S. (2017). *Marketing Strategy : Based on First Principles and Data Analytics* (2nd ed.). Oxford Macmillan Education Palgrave.

Panda, D. K., & Ray, S. (2022). Approaches and algorithms to mitigate cold start problems in recommender systems: a systematic literature review. *Journal of Intelligent Information Systems*, *59*(4). https://doi.org/10.1007/s10844-022-00698-5

Panda, S. K., Bhoi, S. K., & Singh, M. (2020). A collaborative filtering recommendation algorithm based on normalization approach. *Journal of Ambient Intelligence and Humanized Computing*, *11*(11), 4643–4665. https://doi.org/10.1007/s12652-020-01711-x

Preethi, G., Krishna, P. V., Obaidat, M. S., Saritha, V., & Yenduri, S. (2017, July 1). Application of Deep Learning to Sentiment Analysis for recommender system on cloud. *IEEE Xplore*. https://doi.org/10.1109/CITS.2017.8035341

Prijic, M. (2023, March 21). AI Based eCommerce Recommendation System Use Cases. *IT Convergence*. https://www.itconvergence.com/blog/how-ai-based-recommendation-systems-are-transforming-ecommerce/

Privacy Policy. (2021). *SEPHORA*. https://www.sephora.com/beauty/privacy-policy

Raju, V. N. G., Lakshmi, K. P., Jain, V. M., Kalidindi, A., & Padma, V. (2020, August 1). Study the Influence of Normalization/Transformation process on the Accuracy of Supervised Classification. *IEEE Xplore*. https://doi.org/10.1109/ICSSIT48917.2020.9214160

Roy, D., & Dutta, M. (2022). A systematic review and research perspective on recommender systems. *Journal of Big Data*, *9*(1). https://doi.org/10.1186/s40537-022-00592-5

Sammouda, R., & El-Zaart, A. (2021). An Optimized Approach for Prostate Image Segmentation Using K-Means Clustering Algorithm with Elbow Method. *Computational Intelligence and Neuroscience*, *2021*, 1–13. https://doi.org/10.1155/2021/4553832

Sephora. (2023). *Cosmetics, Beauty Products, Fragrances & Tools | Sephora*. Sephora. https://www.sephora.com/

Shambour, Q. (2021). A deep learning based algorithm for multi-criteria recommender systems. *Knowledge-Based Systems*, *211*, 106545. https://doi.org/10.1016/j.knosys.2020.106545

Shankar, A., Perumal, P., Subramanian, M., Naresh Ramu, Deepa Natesan, Kulkarni, V. R., & Stephan, T. (2023). An intelligent recommendation system in e-commerce using ensemble learning. *Multimedia Tools and Applications*, *83*(11). https://doi.org/10.1007/s11042-023-17415-1

Shetty, P., & Singh, S. (2021). Hierarchical Clustering: A Survey. *International Journal of Applied Research*, *7*(4), 178–181. https://doi.org/10.22271/allresearch.2021.v7.i4c.8484

Shutaywi, M., & Kachouie, N. N. (2021). Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering. *Entropy*, *23*(6), 759. https://doi.org/10.3390/e23060759

Sineglazov, V., & Yuriy Oliinuk. (2021). Algorithms for the Formation of Recommendations in the Information System. *Elektronìka Ta Sistemi Upravlìnnâ*, *2*(68), 26–30. https://doi.org/10.18372/1990-5548.68.16088

Singh, D., & Singh, B. (2019). Investigating the impact of data normalization on classification performance. *Applied Soft Computing*, *97*, 105524. https://doi.org/10.1016/j.asoc.2019.105524

Singh, P. K., Sinha, S., & Choudhury, P. (2022). An improved item-based collaborative filtering using a modified Bhattacharyya coefficient and user–user similarity as weight. *Knowledge and Information Systems*, *64*(3), 665–701. https://doi.org/10.1007/s10115-021-01651-8

Skovhøj, F. Z. (2022, January 12). *The Power of Product Recommendations: 30 Must-Know Statistics for 2022*. Clerk.io. https://www.clerk.io/blog/product-recommendations-statistics#:~:text=49%25%20of%20consumers%20said%20they

Tran, T. N. T., Felfernig, A., Trattner, C., & Holzinger, A. (2020). Recommender systems in the healthcare domain: state-of-the-art and research issues. *Journal of Intelligent Information Systems*, *57*, 171–201. https://doi.org/10.1007/s10844-020-00633-6

Tripathy, A., & Rath, S. K. (2017). Classification of Sentiment of Reviews using Supervised Machine Learning Techniques. *International Journal of Rough Sets and Data Analysis*, *4*(1), 56–74. https://doi.org/10.4018/ijrsda.2017010104

Tuboalabo, A., Buinwi, J. A., Buinwi, U., Okatta, C. G., & Johnson, E. (2024). Leveraging business analytics for competitive advantage: Predictive models and data-driven decision making. *International Journal of Management & Entrepreneurship Research*, *6*(6), 1997–2014. https://doi.org/10.51594/ijmer.v6i6.1239

Upadhyaya, P. (2023). User-based and Item-based Collaborative Filtering. In *researchgate*. https://www.researchgate.net/figure/User-based-and-Item-based-Collaborative-Filtering_fig2_366902172

Wang, D., Liang, Y., Xu, D., Feng, X., & Guan, R. (2018). A content-based recommender system for computer science publications. *Knowledge-Based Systems*, *157*(10), 1–9. https://doi.org/10.1016/j.knosys.2018.05.001

Wang, J., Wang, X., Yang, Y., Zhang, H., & Fang, B. (2020). A Review of Data Cleaning Methods for Web Information System. *Computers, Materials & Continua*, *62*(3), 1053–1075. https://doi.org/10.32604/cmc.2020.08675

Wang, Y., Anderson, J., Joo, S.-J., & Huscroft, J. R. (2019). The leniency of return policy and consumers' repurchase intention in online retailing. *Industrial Management & Data Systems*, *120*(1), 21–39. https://doi.org/10.1108/imds-01-2019-0016

Wang, Y., Mo, D. Y., & Ho, G. T. S. (2023). How Choice Fatigue Affects Consumer Decision Making in Online Shopping. *IEEE Xplore*. https://doi.org/10.1109/ieem58616.2023.10406866

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, *55*(55). https://doi.org/10.1007/s10462-022-10144-1

Willmott, C., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, *30*(1), 79–82. https://doi.org/10.3354/cr030079

Wu, X. (2022). Comparison Between Collaborative Filtering and Content-Based Filtering. *Highlights in Science, Engineering and Technology*, *16*, 480–489. https://doi.org/10.54097/hset.v16i.2627

Wu, Y., Jin, Z., Shi, C., Liang, P., & Zhan, T. (2024, March 12). Research on the Application of Deep Learning-based BERT Model in Sentiment Analysis. *ArXiv.org (Cornell University)*. https://doi.org/10.48550/arXiv.2403.08217

Xu, K., Zhou, H., Zheng, H., Zhu, M., & Xin, Q. (2024). Intelligent Classification and Personalized Recommendation of E-commerce Products Based on Machine Learning. *ArXiv (Cornell University)*, *1*(3). https://doi.org/10.48550/arxiv.2403.19345

Xu, L., & Sang, X. (2022). E-Commerce Online Shopping Platform Recommendation Model Based on Integrated Personalized Recommendation. *Scientific Programming*, *2022*, 1–9. https://doi.org/10.1155/2022/4823828

Yan, W., Wang, D., Liu, J., Ma, L., & Li, Z. (2019). Multi-Channel and Fusion Encoding Strategy Based Auto Encoder Model for Video Recommendation. *IEEE Access*, *7*, 86004–86017. https://doi.org/10.1109/access.2019.2925653

Yarasu Madhavi Latha, & B. Srinivasa Rao. (2023). Product recommendation using enhanced convolutional neural network for e-commerce platform. *Cluster Computing*, *27*(4). https://doi.org/10.1007/s10586-023-04053-3

Zhang, Kudo, Murai, & Ren. (2019). Enhancing Recommendation Accuracy of Item-Based Collaborative Filtering via Item-Variance Weighting. *Applied Sciences*, *9*(9), 1928. https://doi.org/10.3390/app9091928

Zhang, Q., Lu, J., & Jin, Y. (2020). Artificial intelligence in recommender systems. *Complex & Intelligent Systems*, *7*(1). https://doi.org/10.1007/s40747-020-00212-w

Zheng, Y., & Wang, D. (Xuejun). (2021). A Survey of Recommender Systems with Multi-Objective Optimization. *Neurocomputing*. https://doi.org/10.1016/j.neucom.2021.11.041

Appendix 1A

Code Files and Readme File

This section presents the Jupyter notebook files used to conduct this research. There are two files (**EDA_Stage1.ipynb** and **Stage2_Stage3.ipynb**) – that need to be executed in a sequential order. Please refer to the Readme file for information on any dependencies in running the code.

Source : https://github.com/Mukundh141-exeter/BEMM466.git

Appendix 1B

Research Proposal

Enhancing E-commerce Product Recommendations: Integrating Product Segmentation and Sentiment Analysis to develop a hybrid recommendation system

Mukundh Srikanth
MSc Business Analytics
University of Exeter, Exeter, UK

ms1452@exeter.ac.uk

Abstract— *In the rapidly evolving e-commerce landscape, ensuring optimal product recommendations is essential for enhancing customer satisfaction and driving sales growth. However, traditional approaches frequently fail to fully grasp the intricate and ever-changing nuances of customer preferences. This study explores integrating product segmentation and sentiment analysis to overcome these inherent limitations. Utilizing clustering algorithms and sentiment analysis techniques on product data, a hybrid recommendation model is proposed. This model can accurately predict user rating scores for products by considering their product profiles, similarities with other products, and the users' overall sentiments toward the products, thereby enhancing personalized recommendations. The proposed methodology is expected to yield significant improvements in recommendation accuracy, thereby optimizing the overall shopping experience and boosting e-commerce performance. Furthermore, by leveraging these insights, e-commerce platforms can not only meet but exceed customer expectations, fostering long-term loyalty and sustainable growth in an increasingly competitive market.*

Introduction & Background to the problem

In the rapidly evolving e-commerce landscape, providing personalized and relevant product recommendations is crucial for enhancing customer satisfaction and driving sales. Despite advancements, many recommendation systems fail to accurately predict customer preferences, resulting in missed opportunities and suboptimal user experiences. Traditional systems primarily rely on collaborative filtering and content-based filtering techniques, which often fall short in capturing the dynamic and multifaceted nature of customer sentiments (Bobadilla et al., 2013). These methods only consider the relations between users and products. Research has been conducted to integrate sentiment analysis with different product recommendation systems; however, a gap still exists in integrating these methods along with product segmentation (Huang, Jin, & Liu, 2020). This study explores how integrating product segmentation and sentiment analysis can significantly enhance the accuracy and relevance of product recommendation systems. Specifically, it investigates combining insights from product segmentation with sentiment analysis of customer comments to develop a more robust product recommendation system. The development of this hybrid recommendation model aims to predict the likelihood of a user purchasing a product. The study's objective is to demonstrate that a recommendation system enriched with segmentation and sentiment analysis will lead to enhanced product recommendations and increased sales. By identifying specific product segments based on popularity and price sensitivity, and integrating this information with sentiment analysis, e-commerce platforms can tailor their recommendations more precisely to individual customers and their preferences.

This proposal will detail the methodology for implementing the enhanced recommendation system, including data collection, segmentation steps, sentiment analysis methods, and the integration of these components into a recommendation engine. Furthermore, it will discuss expected outcomes such as improved recommendation accuracy. This research seeks to contribute to the ongoing evolution of e-commerce, offering valuable insights and practical solutions for businesses aiming to optimize their recommendation systems.

Significance of the Problem

In e-commerce, recommendation systems (RS) rely heavily on collaborative filtering algorithms. These systems provide product recommendations based on the similarity of users' purchase behaviour, often overlooking nuanced customer preferences that can be found in the sentiment analysis of their reviews. Recent advancements, such as those explored by Jian Zhen Yu et al. (2018), have begun integrating sentiment analysis into recommendation systems. However, a significant gap remains in integrating segmentation alongside sentiment analysis within these systems (Yu et al., 2018). This research aims to address these gaps by proposing a framework that integrates segmentation and sentiment analysis into existing recommendation systems.

Research Questions

RQ1: How can product segmentation and sentiment analysis be integrated to improve the accuracy of product recommendations in e-commerce systems?

RQ1.1: What are the key product segments that can be identified through data analysis, and how do these segments influence product preferences?

RQ1.2: In what ways does sentiment analysis of customer reviews contribute to understanding product preferences and enhancing recommendation models?

RQ1.3: How does a hybrid recommendation model, which incorporates product segmentation and sentiment analysis, compare to traditional recommendation models in terms of predictive accuracy?

Aim and Objectives

To enhance the accuracy and relevance of e-commerce product recommendation systems by integrating product segmentation and sentiment analysis.

Obj1: Segment products using clustering algorithms and analyse their popularity and price characteristics to gain deeper understanding.

Obj2: Apply sentiment analysis to product reviews to understand user sentiments in order to integrate these insights with recommendation systems.

Obj3: Develop a hybrid recommendation model combining **Item-based collaborative filtering, product segmentation, and sentiment analysis**.

Data and Datasets Source

The following datasets will be used, to build and test the recommendation system.

Table 4 - List of datasets

Data set name	Description	Number of columns	Number of rows
Product_info.csv	Contains the Product data content.	27	8,494
reviews_0-50.csv	Contains the Reviews data content from March 2023.	19	6,02,130
reviews_250-500.csv	Contains the Reviews data content from March 2023.	19	2,06,725
reviews_500-750.csv	Contains the Reviews data content from March 2023.	19	1,16,262
reviews_750-1250.csv	Contains the Reviews data content from March 2023.	19	1,19,317
reviews_1250-end.csv	Contains the Reviews data content from March 2023.	19	49,977

Product_info.csv contains the following fields related to the products as shown in Table 2. The features highlighted in green will be used to obtain product popularity and price details.

Table 5 - Product Information dataset features

Feature	Description
product_id	The unique identifier for the product from the site
product_name	The full name of the product
brand_id	The unique identifier for the product brand from the site
brand_name	The full name of the product brand
loves_count	The number of people who have marked this product as a favourite
rating	The average rating of the product based on user reviews
reviews	The number of user reviews for the product
size	The size of the product, which may be in oz, ml, g, packs, or other units depending on the product type
variation_type	The type of variation parameter for the product (e.g. Size, Color)
variation_value	The specific value of the variation parameter for the product (e.g. 100 mL, Golden Sand)
variation_desc	A description of the variation parameter for the product (e.g. tone for fairest skin)
ingredients	A list of ingredients included in the product, for example: [‘Product variation 1:’, ‘Water, Glycerin’, ‘Product variation 2:’, ‘Talc, Mica’] or if no variations [‘Water, Glycerin’]
price_usd	The price of the product in US dollars
value_price_usd	The potential cost savings of the product, presented on the site next to the regular price
sale_price_usd	The sale price of the product in US dollars
limited_edition	Indicates whether the product is a limited edition or not (1-true, 0-false)
new	Indicates whether the product is new or not (1-true, 0-false)
online_only	Indicates whether the product is only sold online or not (1-true, 0-false)
out_of_stock	Indicates whether the product is currently out of stock or not (1 if true, 0 if false)
sephora_exclusive	Indicates whether the product is exclusive to Sephora or not (1 if true, 0 if false)

highlights	A list of tags or features that highlight the product's attributes (e.g. [‘Vegan’, ‘Matte Finish’])
primary_category	First category in the breadcrumb section
secondary_category	Second category in the breadcrumb section
tertiary_category	Third category in the breadcrumb section
child_count	The number of variations of the product available
child_max_price	The highest price among the variations of the product
child_min_price	The lowest price among the variations of the product

reviews_0-50.csv etc, contain the following fields related to reviews as shown in Table 3. The features highlighted in blue will be used to obtain sentiment scores and will be used in collaborative filtering algorithm.

Table 6 - Review Information dataset features

Feature	**Description**
author_id	The unique identifier for the author of the review on the website
rating	The rating given by the author for the product on a scale of 1 to 5
is_recommended	Indicates if the author recommends the product or not (1-true, 0-false)
helpfulness	The ratio of all ratings to positive ratings for the review: helpfulness = total_pos_feedback_count / total_feedback_count
total_feedback_count	Total number of feedback (positive and negative ratings) left by users for the review
total_neg_feedback_count	The number of users who gave a negative rating for the review
total_pos_feedback_count	The number of users who gave a positive rating for the review
submission_time	Date the review was posted on the website in the 'yyyy-mm-dd' format
review_text	The main text of the review written by the author
review_title	The title of the review written by the author
skin_tone	Author's skin tone (e.g. fair, tan, etc.)
eye_color	Author's eye color (e.g. brown, green, etc.)
skin_type	Author's skin type (e.g. combination, oily, etc.)
hair_color	Author's hair color (e.g. brown, auburn, etc.)
product_id	The unique identifier for the product on the website

Verification of Data Quality

The dataset, sourced from Kaggle, was collected using a Python scraper in March 2023 from Sephora's official website.

Link of dataset : (https://www.kaggle.com/datasets/nadyinky/sephora-products-and-skincare-reviews?select=product_info.csv)

Official website link : (https://www.sephora.com/product/the-ordinary-deciem-alpha-arbutin-2-ha-P427412?skuId=2464980&icid2=products%20grid:p427412:product)

The webpage contains product-related data and review information, as shown in Figures 1 and 2 respectively. Therefore, the data is valid and obtained from the official source directly. The dataset, is sourced from Kaggle and governed by the **Attribution 4.0 International (CC BY 4.0) license**, therefore it permits sharing and adaptation with appropriate credit.

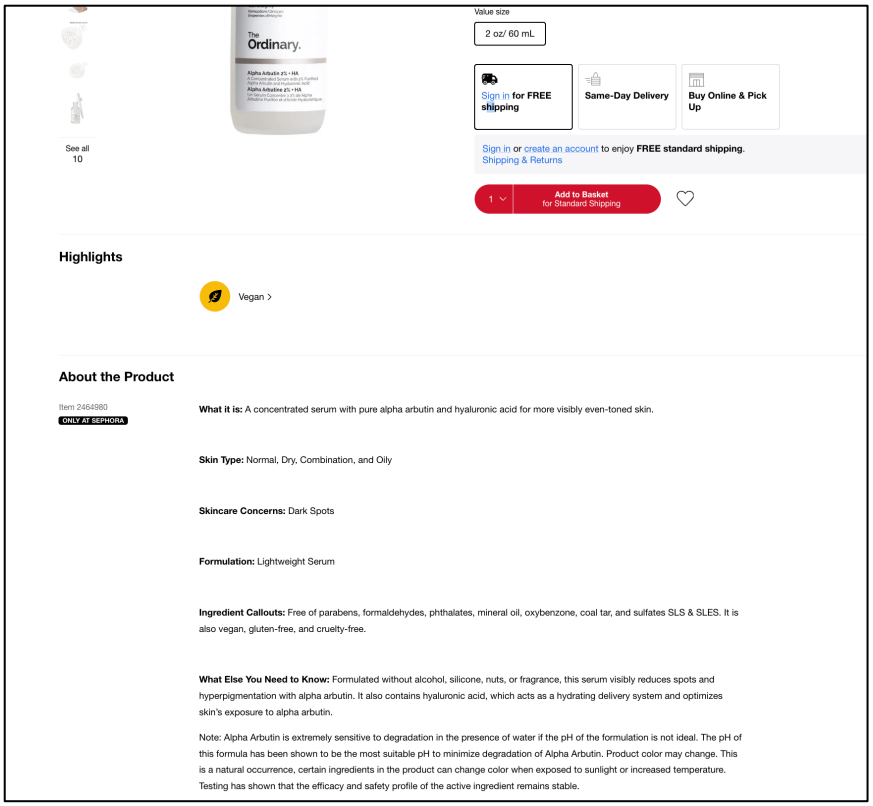


Figure 26 - Product Information on official website

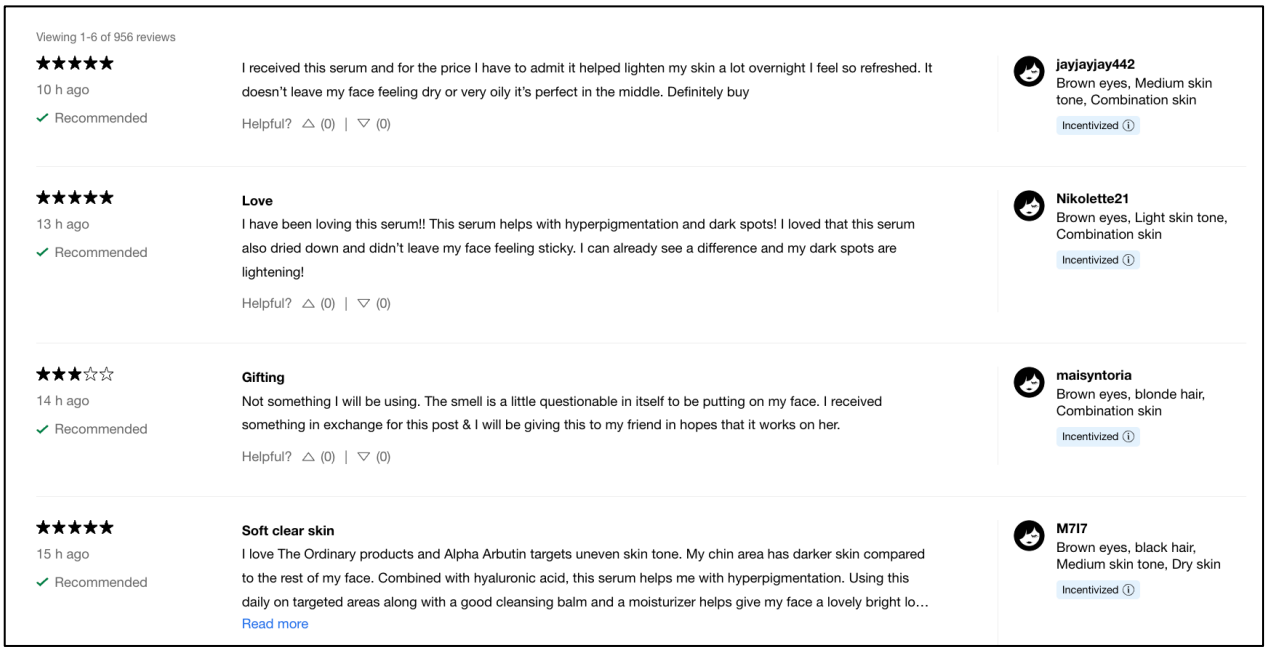


Figure 27 - Review information on official website

Methodology

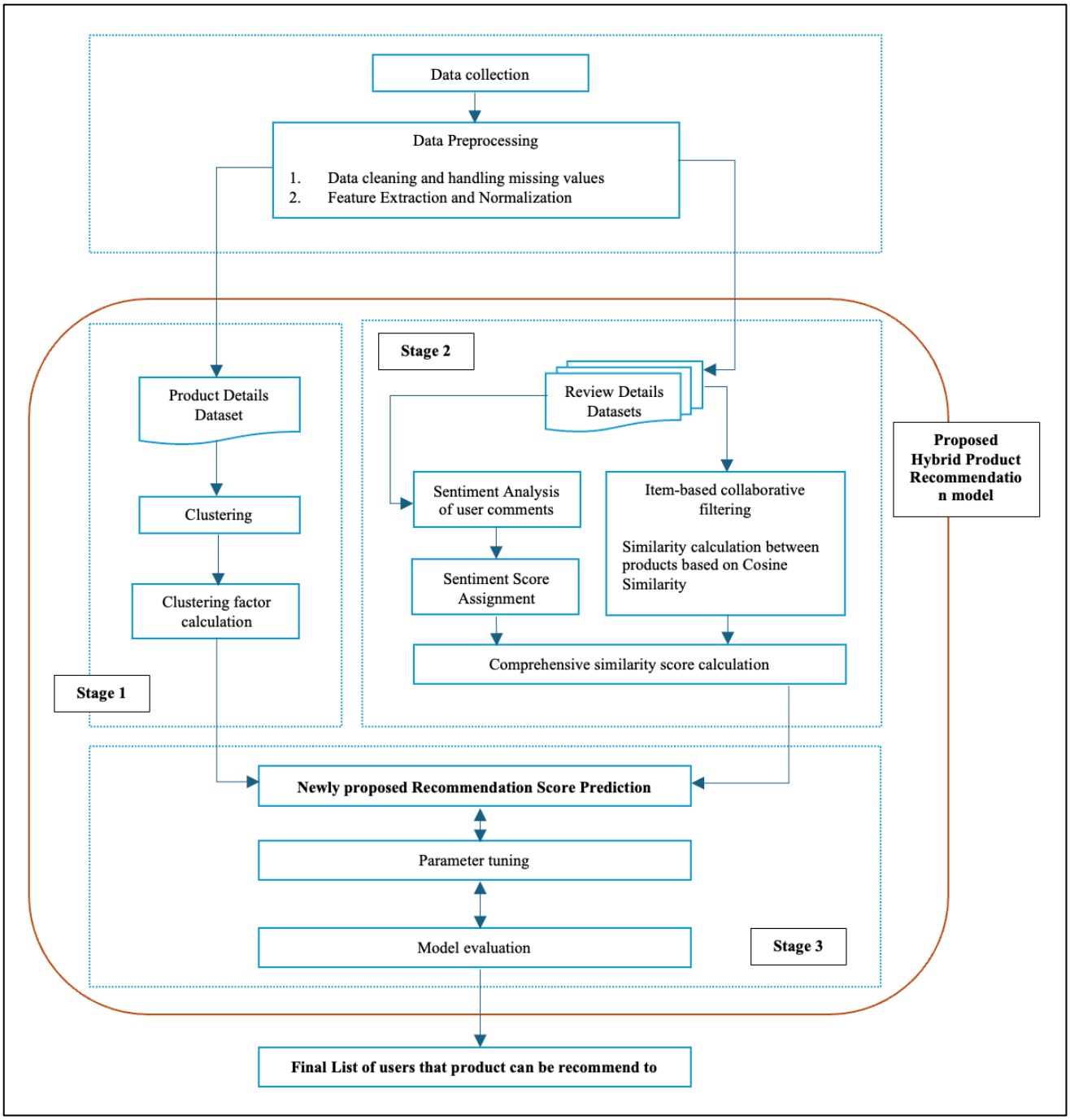
The methodology is to implement a mixed method analysis combining, **cluster** and **sentiment analysis** to improve product recommendation based on **item-based collaborative filtering**. The proposed methodology is depicted in the below flowchart.

The data is collected and pre-processed.

Stage 1 - Product related data will be used for clustering the products and cluster factor calculation.

Stage 2 - Review data will be used for sentiment analysis, and Item-based collaborative filtering.

Stage 3 - Combines the output from stage 1 and 2 to provide improved user-product recommendation.



Data cleaning and handing missing values

Exploratory data analysis has been performed to understand the dimensions, attributes of the datasets.

Only the target features as discussed in following section will be retained.

Entries containing null in the target feature columns will be dropped.

As there are nearly 1 million customer reviews on over 2,000 products, only a limit range of data will be used for building and testing the model as per the experiment results.

Feature extraction and normalization

The following features will be used to extract information related to product popularity and price for clustering purposes. The next set of features extracted from the review data will be used for sentiment analysis and item-based collaborative filtering. Normalization of the features will be carried out to ensure they are comparable.

Feature selection for Product Dataset (Product_info.csv): Popularity: loves_count, rating, reviews ; Price: price_usd

Feature selection for Review Datasets (reviews_0-50.csv etc) : Sentiment Analysis : review_text ; Item-based collaborative filtering : author_id, product_id, rating

Stage 1 – Product cluster analysis and cluster factor calculations

Based on the popularity and price features, various product clusters will be identified, and each product will be assigned a cluster factor score C based on its cluster membership. This cluster factor assigned to each product is crucial for Stage 3, where it will be used to predict ratings for products by users.

Cluster factor calculations :

For each identified cluster, the cluster factor will be calculated as proposed.

$$C_a = \frac{Average\ popularity\ of\ products\ in\ the\ cluster}{Average\ price\ of\ products\ in\ the\ cluster} \quad \text{Equation (1)}$$

For example, if a product (a) belongs to the (Popular and Cheap cluster) it will receive a higher value, indicating products in this cluster offer better perceived value for their price compared to products in other clusters say (Less Popular and Expensive).

Stage 2 – Sentiment analysis and Item-based Collaborative filtering(Using cosine similarity)

Inspired by the work of Jian Zhen Yu et al. (2018), the comprehensive similarity between products will be calculated using equations (2) and (3).

Traditional similarity based on rating:

$$sim(a,b) = \frac{r_a \cdot r_b}{||r_a|| \cdot ||r_b||} \quad (2)$$

Were,

$sim(a,b) = Similarity\ between\ items\ a\ and\ b,$

$r_a = Rating\ for\ prduct\ a,$

$r_b = Rating\ for\ product\ b,$

$r_a \cdot r_b = Dot\ product\ of\ vectors\ r_a\ and\ r_b,$

$||r_a|| = The\ Euclidean\ norm\ (magnitude)\ of\ vector\ r_a,$

$||r_b|| = The\ Euclidean\ norm\ (magnitude)\ of\ vector\ r_b,$

Similarity based on sentiment score:

$$sim'^{(a,b)} = \frac{s_a \cdot s_b}{||s_a|| \cdot ||s_b||} \quad (3)$$

Were,

$sim'(a,b) = Similarity\ between\ items\ a\ and\ b,$

$s_a = Sentiment\ score\ for\ prduct\ a,$

$s_b = Sentiment\ score\ for\ product\ b,$

$s_a \cdot s_b = Dot\ product\ of\ vectors\ s_a\ and\ s_b,$

$||s_a|| = The\ Euclidean\ norm\ (magnitude)\ of\ vector\ s_a,$

$||s_b|| = The\ Euclidean\ norm\ (magnitude)\ of\ vector\ s_b,$

Similarity based on cluster factor:

$$sim''^{(a,b)} = \frac{c_a \cdot c_b}{||c_a|| \cdot ||c_b||} \quad (4)$$

Were,

$sim''(a,b) = Similarity\ between\ items\ a\ and\ b,$

$c_a = Cluster\ factor\ for\ prduct\ a,$

$c_b = Cluster\ factor\ for\ product\ b,$

$c_a \cdot c_b = Dot\ product\ of\ vectors\ c_a\ and\ c_b,$

$||c_a|| = The\ Euclidean\ norm\ (magnitude)\ of\ vector\ c_a,$

$||c_b|| = The\ Euclidean\ norm\ (magnitude)\ of\ vector\ c_b,$

The comprehensive similarity score, $SIM(a,b)$ is calculated combining the equations 5, 6, 7 along with their respective weighting factors as shown below.

$$\boldsymbol{SIM(a,b) = \alpha * sim(a,b) + \beta * sim'(a,b) + \gamma * sim''(a,b)} \quad (5)$$

Stage 3 – Enhanced Recommendation Score Calculation

Based on equations (1), (2), (3), (4) and (5), the predicted rating for product a by user i is calculated as shown below. The accuracy of the model will be tested based on these predictions. **Equation 6 embodies the novelty of this research by combining clustering, sentiment analysis, and item-based collaborative filtering for product recommendations.**

$$r_{xa} = \frac{\sum_{b \in N(a;x)} SIM(a,b) * r_{xb}}{\sum_{b \in N(a;x)} SIM(a,b)} \quad (6)$$

Were,

$r_{xb} = Rating\ for\ item\ b\ by\ user\ x,$

$r_{xa} = Estimated\ rating\ for\ item\ a\ by\ user\ x,$

$SIM(a,b) = Comprehensive\ similarity\ between\ item\ a\ and\ b,$

$N(a;x) = Set\ of\ items\ rated\ by\ user\ x\ and\ similar\ to\ item\ a,$

Ethical Considerations

This research utilizes an existing dataset sourced from Kaggle, governed by the Attribution 4.0 International (CC BY 4.0) license, permitting sharing and adaptation with

appropriate credit. The dataset, collected from Sephora's official website, contains publicly available product and review information without any personal identifiers. Sephora's privacy policy ensures user consent for data collection, focusing on transparency and user rights.

Perceived Ethical Issues and Management:

Data Anonymization: The dataset does not include personal identifiers, ensuring anonymity. Numeric author IDs prevent the extraction of personal details.

Privacy and Consent: The data, collected from a public domain, adheres to Sephora's privacy policies, ensuring user consent and data protection.

Data Protection: Any data used for analysis will be strictly utilized only for the duration of the experiment and will be deleted upon completion. This ensures no long-term storage or misuse of data.

Use of Secondary Data: As the research relies on secondary data, it ensures that the original data collection adhered to ethical standards, and no new data collection from human subjects is required. Please refer to section "Verification of Data Quality" for more details.

Avoiding Identifiability: Ensuring no individuals are identifiable through data combinations, complying with ethical guidelines to protect personal data.

By adhering to these practices, this research maintains high ethical standards, ensuring privacy, consent, and data protection throughout the study.

Expected Outcomes

Improved Recommendation Accuracy & Higher Sales Conversion Rates

Outcome: Enhanced prediction of customer preferences resulting in more accurate product recommendations.

Objective Link: Obj3 - Develop a hybrid recommendation model combining Item-based collaborative filtering, product segmentation, and sentiment analysis.

Impact: By integrating the hybrid recommendation system companies will better understand and predict what products are likely to purchase, increasing overall recommendation precision.

Better Product Segmentation Insights

Outcome: Identification of distinct products segments and based on their popularity and price.

Objective Link: Obj1: Segment products using clustering algorithms and analyse their popularity and price characteristics to gain deeper understanding.

Impact: Understanding specific products segments that allows for targeted marketing strategies and better resource allocation.

Enhanced Sentiment Understanding

Outcome: Deep insights into customer sentiments derived from reviews, contributing to more nuanced product recommendations.

Objective Link: Obj2: Apply sentiment analysis to product reviews to understand user sentiments in order to integrate these insights with item-based collaborative filtering.

Impact: Incorporating sentiment analysis helps in capturing the emotional and subjective aspects of customer feedback, enhancing the relevance of recommendations.

Practical Real-World Impact

Improved Customer Experience and Satisfaction

Impact: By enhancing recommendation accuracy through advanced techniques like hybrid recommendation models integrating item-based collaborative filtering, product segmentation, and sentiment analysis, businesses can offer more personalized and relevant product suggestions to their customers.

Practical Impact: This can lead to improved customer satisfaction as customers are more likely to find products that meet their preferences and needs, thereby enhancing their overall shopping experience.

Increased Sales and Revenue

Impact: With more accurate product recommendations, businesses can increase sales conversion rates as customers are guided towards products they are more likely to purchase.

Practical Impact: Higher sales conversion rates directly translate to increased revenue and profitability for e-commerce platforms and retailers, demonstrating tangible financial benefits from implementing advanced recommendation systems.

Optimized Marketing Strategies

Impact: Insights gained from better product segmentation and understanding customer sentiments can inform targeted marketing strategies.

Practical Impact: Businesses can allocate marketing resources more efficiently by focusing on specific customer segments identified through clustering and tailoring promotional efforts based on sentiment analysis of customer reviews. This leads to more effective campaigns that resonate with the target audience.

Enhanced Product Development and Inventory Management

Impact: Understanding product popularity and price characteristics through segmentation allows businesses to optimize product development and inventory management strategies.

Practical Impact: By identifying and focusing on popular product segments, businesses can prioritize product development efforts and manage inventory levels more effectively, reducing costs associated with overstocking or understocking.

Technology Used

Python, Scikit-learn, Pandas, NumPy, and other graphic libraries.

Bibliography

Bobadilla, J., Ortega, F., Hernando, A., & Gutiérrez, A. (2013). Recommender systems survey. Knowledge-Based Systems, 46, 109-132. https://doi.org/10.1016/j.knosys.2013.03.012

Chen, L., Chen, G., & Wang, F. (2019). Recommender systems based on user reviews: The state of the art. User Modeling and User-Adapted Interaction, 29(2), 167-209. https://doi.org/10.1007/s11257-019-09231-7

Jian Zhen Yu, An, Y., Xu, T., Gao, J., Zhao, M., & Yu, M. (2018). Product Recommendation Method Based on Sentiment Analysis. Web Information Systems and Applications, 11242, Page 480. https://doi.org/10.1007/978-3-030-02934-0_45

Yu, J. Z., Guo, Y., & Hua, X. S. (2018). A review of the role of artificial intelligence in e-commerce recommendation systems. International Journal of Artificial Intelligence & Applications, 9(3), 65-80.

Huang, M., Jin, H., & Liu, Y. (2020). Sentiment analysis meets recommendation systems: A survey and new perspectives. ACM Transactions on Management Informantion