INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA DE INTELIGÊNCIA ARTIFICIAL




Development of a Hybrid Recommendation System Focused on Serendipity for E-commerce

CAROLINE BASTOS
Outubro de 2024

P.PORTO

Development of a Hybrid Recommendation System Focused on Serendipity for E-commerce

Caroline Bastos

Aluno nº: 1220499

Dissertação para obtenção do Grau de

Mestre em Engenharia de Inteligência Artificial

Orientador: António Constantino Lopes Martins, Professor Adjunto do Instituto Superior de Engenharia do Porto do Instituto Politécnico do Porto

Júri:

Presidente:

Paulo Sérgio dos Santos Matos, Professor Adjunto do Instituto Superior de Engenharia do Porto do Instituto Politécnico do Porto

Arguente:

Ana Maria Neves Almeida Baptista Figueiredo, Professora Coordenadora do Instituto Superior de Engenharia do Porto do Instituto Politécnico do Porto

Orientador:

António Constantino Lopes Martins, Professor Adjunto do Instituto Superior de Engenharia do Porto do Instituto Politécnico do Porto

Porto, Outubro 2024

Resumo

No dinâmico mundo do comércio eletrônico, os Sistemas de Recomendação tornaram-se cruciais para orientar as escolhas dos usuários e melhorar as experiências de compra. Sistemas de recomendação aumentam os lucros do comércio eletrônico e elevam a satisfação do usuário. Também foi comprovada a influência positiva na satisfação do cliente de aspectos como diversidade, novidade e serendipidade na recomendação. Apesar disso, a maioria dos sistemas de recomendação confia principalmente em métricas de precisão. Este estudo propõe uma abordagem inovadora para aprimorar as recomendações serendipitosas, aproveitando técnicas de aprendizado de máquina e processamento de linguagem natural aplicadas às descrições dos itens.

Avaliar a serendipidade continua sendo um desafio devido à sua natureza subjetiva e, por isso, este estudo utiliza um conjunto de dados ground truth de serendipidade para desenvolver seus experimentos. Quatro modelos diferentes de embeddings—BERT, DistilRoBERTa, MPNet e BM25—foram empregados para gerar embeddings dos itens a partir das descrições textuais dos produtos. Uma análise inicial comparou a serendipidade calculada, derivada da similaridade de cosseno dos embeddings dos itens, com a serendipidade ground truth. Os resultados não revelaram correlação significativa entre a serendipidade calculada e os dados de ground truth, ressaltando a complexidade de medir com precisão a serendipidade usando métricas tradicionais.

Para endereçar esta limitação, um modelo de classificação de serendipidade foi desenvolvido, para prever a probabilidade de um item ser serendipitoso para um usuário com base na descrição do item. Na sequencia, três modelos diferentes de sistemas de recomendação foram construídos: o modelo XGBoost básico, que funciona como um sistema de recomendação baseado em conteúdo; o modelo XGBoost-Seren, que incorpora dados de serendipidade ground truth para priorizar itens serendipitosos; e o modelo XGBoost-Seren + Classificador, que substitui os dados de serendipidade ground truth pelas previsões do classificador de serendipidade. Cada modelo foi comparado com o SASRec, um sistema de recomendação sequencial, para avaliar tanto as métricas de acurácia quanto de serendipidade.

Os resultados demonstram que o modelo XGBoost-Seren + Classificador supera o SASRec tanto nas métricas de acurácia quanto de serendipidade. O estudo confirma que incorporar a serendipidade prevista nos algoritmos de recomendação aumenta a capacidade de sugerir itens inesperados sem comprometer a acurácia. As descobertas indicam que usar embeddings baseados em texto não apenas enriquece o processo de recomendação, mas também melhora a experiência geral do usuário ao integrar a serendipidade de forma mais eficaz.

Palavras-chave: Sistemas de Recomendação, Comércio Eletrônico, Serendipidade, Aprendizado de Máquina

Abstract

In the dynamic world of e-commerce, Recommender Systems have become crucial in guiding user choices and improving shopping experiences. Recommender Systems enhances e-commerce profits and increases user satisfaction. It has also been proved the positive influence in customer satisfaction of aspects like diversity, novelty and serendipity in recommendation. Despite it, most recommender systems rely mostly on accuracy metrics.

To address this issue, this study proposes a novel approach to enhance serendipitous recommendations by leveraging machine learning and natural language processing techniques applied to item descriptions.

Four different embedding models were employed to generate item embeddings from textual item descriptions. An initial analysis compared calculated serendipity, derived from the cosine similarity of item embeddings, with ground-truth serendipity. Followed by the development of a serendipity classification model. Three different recommendation models were constructed: the baseline XGBoost model, which functions as a content-based recommender system; the XGBoost-Seren model, which incorporates ground-truth serendipity data to prioritize serendipitous items; and the XGBoost-Seren + Classifier model, which replaces the ground-truth serendipity data with the predictions from the serendipity classifier. Each model was compared against SASRec, a state-of-the-art sequential recommender system, to evaluate both accuracy and serendipity metrics.

The results demonstrate that the XGBoost-Seren + Classifier model outperforms SASRec in both accuracy and serendipity metrics. The study confirms that incorporating predicted serendipity into recommendation algorithms enhances the ability to suggest unexpected items without compromising accuracy. The findings indicate that using text-based embeddings not only enriches the recommendation process but also improves the overall user experience by integrating serendipity more effectively.

Keywords: Recommender Systems, E-commerce, Serendipity, Machine Learning

Acknowledgments

Firstly, I would like to thank my parents. To my father for teaching me through his example the profound impact and importance of education. To my mother, for being my greatest supporter and always showing her pride in me.

I am deeply grateful to my fiancé for being the best life partner and for his immense patience during this process and in all other moments of our lives.

I am also incredibly thankful to so many friends who have always believed in me more than I believed in myself.

And finally, I wish to express my special gratitude to my advisor, Professor Constantino Martins, for all the guidance, availability, and patience throughout this journey.

Contents

List of Figures

List of Tables

Acronyms and Symbols

List of Acronyms

RC	Recommender Systems
NLP	Natural Language Processing
DSR	Design Science Research
ML	Machine Learning
MAE	Mean Absolute Error
RMSE	Root Mean Square Error
HR	HIT Radio
NDGC	Normalized Discounted Cumulative Gain

1 Introduction

This chapter presents the context of this study, justifying the relevance of the proposed recommender systems. It will present the research focus, question and the methodology that will be applied.

1.1 Context

The growth of e-commerce in recent years has been substantial, it is transforming the way people shop and hugely impacting the way businesses operate. In 2017 the retail e-commerce sales worldwide amounted to 2.3 trillion US dollars and the revenue of the top 3 online stores (amazon, apple and Walmart) amounted to almost 100 billion US dollars (Babenko et al., 2019). The COVID-19 pandemic also contributed to the expansion of this sector, in 2020 the ecommerce in Brazil had a growth rate of 68% (Rocha et al., 2021). This illustrates the potential of investing in an ecommerce business.

The diversity of available products in e-commerce is outstanding, offering a high number of options that can both overwhelm customers and consume their time. To solve this problem recommendation systems (RC) are applied. Recommendation systems are a way of predicting which product a user would like to buy, making the shopping experience pleasant for the user and impacting directly on a company's profit (Fayyaz et al., 2020; Kompan et al., 2022).

E-commerce recommendation systems have the potential to enhance customer satisfaction by improving perceived information quality and usability, with deep personalization directly contributing to increased customer satisfaction (Liu Qian & Ma Hui min, 2010). It is estimated that implementing a recommender system can increase 30-35% of the number of purchases in an ecommerce business (Kompan et al., 2022).

This study proposes the development of a hybrid recommendation system using machine learning and Natural Language processing, to provide diverse and effective recommendations for e-commerce users.

1.2 Objectives

In the current e-commerce landscape, customer satisfaction has become a critical competitive differentiator. Li et al. (2020) conducted a study on the evaluation of recommender systems in e-commerce, considering customer satisfaction. The study concluded that both recommendation system accuracy and diversity have a positive effect on customer satisfaction, providing insights for e-commerce companies to enhance customer satisfaction and promote sustainable development. Silveira et al. (2019) survey highlights that a good recommendation system is one that provides accurate and relevant recommendations to users, while also satisfying their desires for novelty, diversity, and serendipity.

Novelty refers to recommending items that are new or unfamiliar to the user, while serendipity involves suggesting unexpectedly valuable or pleasing items that the user was not looking for. Recommendation diversity ensures distinct types of items in the recommendations, satisfying the broad interests of a user (Kaminskas & Bridge, 2016).

The use of these factors by recommender systems in e-commerce helps alleviate user fatigue from repetitive and familiar recommendations. Additionally, it plays a significant role in enhancing customer loyalty (Alamdari et al., 2020).

Despite this evidence, most studies on e-commerce RCs focus on enhancing the accuracy of recommendations and other aspects such as security, response time, novelty, diversity, and serendipity are often overlooked in numerous publications (Alamdari et al., 2020).

When it comes to RC approaches, hybrid RSs can offer more accurate and efficient recommendations, especially for new or unrated items. The hybrid approach in a RC involves utilizing a variety of algorithms and combining different methods to gather and process user or item data. Therefore, using hybrid RC in e-commerce improves recommendation accuracy and customer satisfaction (Alamdari et al., 2020). The use of Machine learning techniques has also been successfully applied to generate more personalized recommendations by analysing individual behaviours and habits, as well as to address other challenges such as the cold start problem (Nagy et al., 2021).

Considering all these aspects, this study will focus on developing a recommendation system using machine learning and natural language processing, that approaches the metrics novelty, serendipity, and recommendation diversity. By doing so, the system aims to enrich the user experience, not just through relevance and accuracy, but also by introducing elements of surprise and personalization that resonate with the users' diverse and evolving preferences.

Therefore, the central research question of this study is:

How can a recommender system be designed to optimize user satisfaction in e-commerce by balancing accuracy with elements of novelty, serendipity, and diversity in its recommendations?

This question seeks to explore the intersection of advanced algorithmic approaches and user-centric design principles to create a recommender system that not only understands but also anticipates and delights the diverse user base of e-commerce platforms.

1.3 Contributions

This study aims to develop a recommender system for e-commerce using state of the art technologies. Applying and combining techniques such as text mining and leveraging items metadata, such as titles and descriptions, which are by this date, not commonly used in this type of system.

The recommender system developed will be able to give diverse, useful, and insightful recommendations. Providing the user with a novelty experience with the feeling of discovery, increasing users' satisfaction.

The main contributions of this study will be: (1) Explore ground-truth serendipity and compare it with calculated serendipity. This comparison is innovative and can bring light to the effectiveness of current metrics in capturing true serendipitous experiences. This understanding may lead to the development of more accurate serendipity measures and improve recommendation systems by aligning them more closely with actual user perceptions. (2) Development of a serendipity focused recommender system, based on ground-truth dataset. Literature lacks serendipity ground-truth dataset and studies that address it. This study establishes that serendipity could be potentially learned and used to make better recommendations. (3) Contribute to RC research. Using NLP and text mining in items title and description, for tracing items correlations with focus on enhancing serendipity recommendation, to the author's best knowledge, has not been explored yet. The results of this study can be a great reference value for future research in the field.

1.4 Research Methodology

This study will follow the Design Science Research (DSR) methodology. The DSR methodology is a research paradigm that focuses on the development and evaluation of artifacts to address real-world problems (Storey et al., n.d.). DSR is a highly suitable methodology for research in e-commerce due to its focus on creating innovative artifacts and methodologies to address practical problems (Au, 2001).

In the dynamic environment of e-commerce, where businesses are constantly seeking new ways to enhance efficiency, effectiveness, and strategic positioning, the need for innovative solutions is prominent. DSR provides a structured approach to developing and evaluating new artifacts, such as e-commerce systems and technologies, aligning with the evolving needs of businesses and consumers (Au, 2001).

The DRS methodology defines six steps to be followed (Peffers et al., 2007):

1. Identification of Problem and Motivation: This involves identifying and defining a specific problem. For this study, the problem is the lack of diversity and creativity in e-commerce recommendations. This issue becomes particularly evident when customers seek unique gift ideas or products beyond the scope of popular brands. For this purpose, the PRISMA methodology will be applied with the conduction of a systematic review on the subjective of exploring diversity, novelty and serendipity in e-commerce RCs.

2. Defining Solution Objectives: Once the problem is identified, the next step is to define the objectives for a solution. For this study, the objective was to create an e-commerce recommendation system that excels in offering innovative and less conventional product suggestions.

3. Design and Development: This phase involves the conceptualization and creation of the solution. In this case, the recommendation system was designed and developed using cutting-edge techniques to ensure it could generate novel and diverse recommendations.

4. Demonstration: After developing the artifact, it is essential to demonstrate its functionality and effectiveness in solving the identified problem. The system was tested using a real-world scenario dataset to validate its performance.

5. Evaluation: This step focuses on assessing the artifact's performance against the defined objectives. For the developed system, the proposed metrics were used to prove its effectiveness, and the system was compared to a baseline model.

6. Communication: The final step involves disseminating the findings and contributions of the study. This research was presented as a master's thesis in Artificial Intelligence, contributing to the fields of e-commerce and recommendation systems with new insights and evidence. Additionally, an article will be written and submitted for publication in relevant academic journals that focus on recommender systems areas, ensuring that the findings reach more professionals in the field.

Figure 1 displays a diagram of the methodology applied.

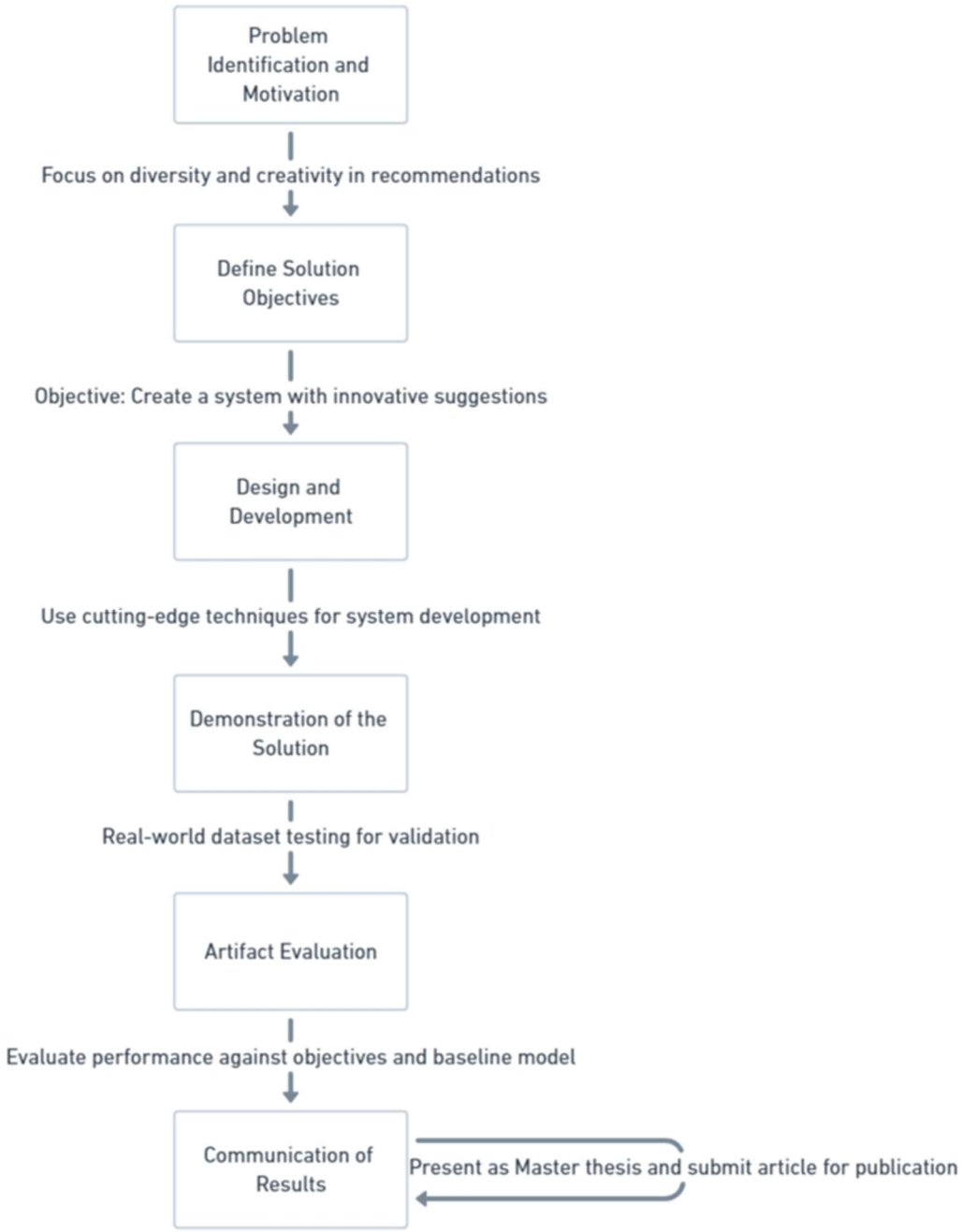

Figure 1 – Methodology diagram

1.5 Document Structure

This study is divided in 5 chapters with the following content:

The first chapter provides the context of this study and justifies the need for the proposed recommender system. It also establishes the contributions and research methodology used.

The second chapter presents the state of the art regarding recommender systems. It details various aspects of this technology followed by a systematic review with the purpose of answering the main research question of this study.

The third chapter presents the proposed model stating the methods and tools, evaluations and data protection, security and ethics concerns.

The fourth chapter describes the implementation, analysis and results discussions of the serendipity analysis comparison and the proposed model. The proposed model includes three different variations, and its results are displayed and compared in this chapter.

The last chapter presents the final conclusions that were drawn from the development and analysis of the results of the experiments carried out, it includes the limitations this study encountered, and future work ideas

2 State of the art

This chapter is divided into two subsections. The first section is a theoretical contextualization, it presents a summary of the state of the art in recommender systems, the main approaches and the techniques applied to its development, as well as the main problems and limitations, and how to evaluate them.

In the second section a systematic review is conducted, following the PRISMA methodology, to answer the proposed research questions that will serve to understand and guide in the novel aspects proposed for this study's recommender system.

2.1 Recommender Systems

Recommender system plays a vital role in a diverse types of domains like e-commerce, music and film streaming services, news media, and more. These systems are crucial for providing users with personalized recommendations, helping them navigate through the overwhelming amount of content available. Moreover, recommender systems increase user engagement and satisfaction, as well as boosting sales for content providers (Fayyaz et al., 2020).

The primary functions of recommender systems include collecting and representing user preferences, as well as predicting relevant items based on these preferences. These systems utilize different models, including content-based, collaborative filtering, and hybrid recommendation models, to achieve these functions (Nagy et al., 2021). Additionally, recommender systems face challenges in evaluating the accuracy and effectiveness of their predictions, highlighting the need for robust evaluation methods (Silveira et al., 2019b).

The evolution of recommender systems has seen the integration of advanced technologies such as deep learning and neural networks, which have enhanced the predictive capabilities and accuracy of these systems (Fu et al., 2024).

2.1.1 Recommender Systems Approaches

Recommender systems are typically categorized based on the type of information they use to suggest products or items to users. The literature primarily defines three main categories: content-based filtering, collaborative-filtering, hybrid filtering (Mansur et al., 2017). These three main approaches will be described in this section.

2.1.1.1 Content-based filtering

Content-based filtering (CBF) analyses the user's profile and recommends items based on the user's preferences and the attributes of the items themselves. This approach relies on the idea that if a user has liked certain items in the past, they are likely to enjoy similar items in the future (Sorde & Deshmukh, 2015).

This approach relies on data provided by the user, which can be either implicit or explicit. This user-related information, derived from the user's interactions, ratings, or responses to previous recommendations, forms the basis for future suggestions. As the user continues to interact with the system, providing more inputs and feedback, the recommender system progressively improves in its effectiveness and accuracy (Mansur et al., 2017).

CBF make recommendations analysing user profiles and extracting features from the content of items that the user has previously evaluated. To find similarities between documents and generate meaningful recommendations, CBF can employs different models. These include the Vector Space Model, such as Term Frequency-Inverse Document Frequency (TF/IDF), and probabilistic models like the Naïve Bayes Classifier, Decision Trees, or Neural Networks (Isinkaye et al., 2015).

This approach has known limitations such as the new user problem and Overspecialization. The new user problem refers to a new user in the system and there is not enough information about its preferences causing bad recommendations. Overspecialization happens because the system tends to recommend similar types of items, not recommending things the user has never seen before (Shah et al., 2016).

On the other hand, CBF does not rely on other users' data, ensuring privacy and security of user data. It can also overcome the cold-start problem if new items have sufficient features, recommending items that have not been rated by any user yet (Roy & Dutta, 2022).

2.1.1.2 Collaborative Filtering

Collaborative filtering (CF) identifies users with similar tastes and recommends items based on the preferences of similar users. This approach relies on the idea that users who have liked similar items in the past are likely to have similar preferences in the future (Sorde & Deshmukh, 2015).

CF focuses on users who have already interacted with items or products that are still unknown to the current user. collaborative filtering systems depend on gathering users' likes and dislikes regarding various products or items. These systems aggregate user opinions and then make

recommendations based on the degree of similarity in users' ratings. Users who share similar opinions on products contribute to the recommendation process (Mansur et al., 2017).

This approach does not need require any knowledge in item features, since it uses users' evaluations to make recommendations, on the other hand it will suffer from the cold-start problem. Also, since it relies on sharing users' data, privacy can be a concern (Roy & Dutta, 2022).

CF can be divided into two parts: Memory-based approach and Model-based approach.

Memory-based, also called neighbourhood-based methods can work in one of two ways:

User-user collaborative filtering: The algorithm calculates how similar users are by comparing their ratings on shared products or items. It then predicts a user's rating for a product or item based on the ratings given to it by users who are similar to the active user (Mansur et al., 2017).

Item-item collaborative filtering: the algorithm generates predictions based on the similarity of products or items, rather than calculating similarity between users a model of item similarities by collecting from the user-item matrix all items rated by an active user. The system then identifies items that are most similar to the target item, selecting the top 'k' items. For these selected items, it also determines the degree of similarity (Isinkaye et al., 2015).

Model-based recommendation systems create a smaller, specialized dataset known as a model. This model is developed by extracting specific information related to certain attributes from a larger database. Once created, the model is used for making recommendations instead of constantly accessing the large database, enhancing the system's speed and scalability. Model-based collaborative filtering (CF) algorithms employ techniques like Bayesian models, cluster-based CF, and regression-based methods (Nagarnaik & Thomas, 2015).

2.1.1.3 Hybrid Filtering

Hybrid filtering combine different recommendations techniques to leverage the best of each approach and to mitigate limitations and problems of individual techniques. Combining algorithms can provides more accurate and effective recommendations, increasing the RC performance. A hybrid algorithm can implement separately each algorithm and combine the results, or it can use content-based filtering in a collaborative method or use a collaborative filtering technique in a content-based method (Isinkaye et al., 2015; Roy & Dutta, 2022).

There are seven hybridization techniques: (Isinkaye et al., 2015)

Weighted hybridization: This technique combines the resulting score of each different techniques using a linear formula. In this way, it aggregates all the strengths in each approach. It uses the same weight for each recommendation at first, and then adjust it according to prediction confirmation or not.

Switching hybridization: Depending on the current requirements it switches between the different models using heuristics. It can avoid the specific methods problems, like cold-start

problem from CBF, switching to CF for example. These systems are sensitive to strengths and weaknesses of its recommenders; however, they can become complex due to the switching criteria.

Cascade hybridization: This technique uses an iterative process to refine a list of recommended items, where the recommendations of one technique is refined by another technique. It starts with one recommendation method that produces an initial, broad list of recommendations. This list is then refined by another recommendation technique, creating a more precise set of suggestions.

Mixed hybridization: Mixes the result of different recommendations techniques, and the combine output is given as recommendation. In this way, each item has more than one recommendation, presenting results coming from the different techniques. In this approach, the performance of individual technique does not necessarily dictate the overall effectiveness of the system.

Feature-combination: The features of one recommender system are fed into another recommendation technique. For example, the ratings of a collaborative filtering algorithm can be used as a feature in a content-based system.

Feature-augmentation: One recommendation model generates a rating or other information which is then used in another recommendation system to produce the final recommendation.

Meta-level: One recommendation technique generates an internal model, that is then used as an input for another recommendation technique. This approach can address problems like sparsity in collaborative filtering by utilizing the richer information content of the generated model.

2.1.2 Metrics for Recommender Systems Evaluation

Evaluating a recommendation system is a very important topic. Various metrics that measure and focus on different aspects of a RC have been proposed and used. The typically used ones will be explained in this section.

2.1.2.1 Accuracy

In the broader context of Artificial Intelligence, accuracy is a well-known metric, typically defined as the proportion of correct predictions to the total number of cases. In RS, this translates to the number of successful recommendations out of the total recommendations made (Hernández del Olmo & Gaudioso, 2008a).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

where TP and TN are the true positive and true negative results, and FP and FN are the false positive and false negative results.

Also, different metrics can be used to evaluate accuracy depending on the specific object and context, such as the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) for evaluating scoring predictions. MAE measures the average absolute deviation between the predicted rating and the actual user rating (Wu et al., 2012)(Hernández del Olmo & Gaudioso, 2008b).

$$MAE = \frac{\sum_{i=1}^{N} |P(u,i) - p(u,i)|}{N} \quad (2)$$

$$RMSE = \frac{\sqrt{\sum_{i=1}^{N} (P(u,i) - p(u,i))^2}}{N} \quad (3)$$

Where $p(u,i)$ represents the real rating of user u for item i, and $P(u,i)$ represents the predicted one. N is the number of ratings available. RMSE is biased to weight large errors more than small errors. A lower value of MAE and RMSE indicates high accuracy (Anitha et al., 2013).

Although accuracy is usually the primary metric for evaluating a RC performance, focusing exclusive in this aspect can cause the "filter bubble phenomenon", that is when the system is trapped in options that are too similar to the user's profile. To prevent that, latest studies emphasise the need to attend other metrics such as diversity, novelty and serendipity, that can, while also match users' preferences, improve user experience (L. Chen et al., 2019).

2.1.2.2 Precision

Precision is a measure of exactness, and it measures the fraction of relevant items retrieved out of all items retrieved (Anitha et al., 2013).

$$Precision = \frac{Number\ of\ sucessfull\ recommendation\ (TP)}{Number\ of\ recommendation\ (TP + FP)} \quad (4)$$

2.1.2.3 Recall

Recall is a measure of completeness, and it measures the fraction of relevant items retrieved out of all relevant items (Anitha et al., 2013).

$$Recall = \frac{Number\ of\ sucessfull\ recommendation\ (TP)}{Number\ of\ relevant\ items\ (TP + FN)} \quad (5)$$

2.1.2.4 F-measure

F-measure combines Precision and Recall into a single value for comparison purposes. It should give a more balanced view of the performance (Anitha et al., 2013).

$$F1 = \frac{2.Precision.Recall}{Precision + Recall} \quad (6)$$

2.1.2.5 Diversity

The concept of diversity in recommendation systems can be categorized into two types: individual diversity and aggregate diversity. Individual diversity refers to the variety within a single user's recommendation list, typically assessed by the average dissimilarity among the recommended items. In contrast, aggregate diversity looks at the diversity of recommendations across all users. Therefore, the coverage of recommendations across all users is a simple way to quantify it (Niu et al., 2019).

It's important to note that these two types of diversity are not directly correlated. For instance, a system might recommend diverse items to each user (high individual diversity) but only circulate a limited set of items among all users (low aggregate diversity) (Niu et al., 2019).

2.1.2.6 Novelty

Novelty is an attribute of recommendations that appears to be closely related to serendipity. While the concepts of novelty and serendipity might overlap, some authors emphasize the differences between the two. While a novel item is simply unknown to the user, it is not necessarily serendipitous. Serendipity requires an element of surprise, meaning that all serendipitous items are novel, but not all novel items are serendipitous (Kaminskas & Bridge, 2016).

On the other hand, it has become more common to define novelty in a way that is independent of a specific user. Generally, the novelty of an item is inversely related to its popularity (often measured by the number of ratings). Less popular items are considered more novel. According to this view, a novel item might not be serendipitous for a user, and a recommendation that is serendipitous does not necessarily have to be novel (Kaminskas & Bridge, 2016).

2.1.2.7 Serendipity

Serendipity can be defined as the finding of unexpected but valuable product-related information. In a RC serendipity occurs when users get recommended items that are both attractive and surprising to them. This means that for a recommendation to be considered serendipitous, it needs to have two aspects: unexpectedness and informational value (Grange et al., 2019).

Serendipity is crucial because it helps mitigate the limitations of traditional recommendation approaches that focus solely on relevance and popularity. By incorporating serendipity, Recommender Systems can offer diverse, novel, and unexpected recommendations, thereby enhancing user experience and satisfaction. This is particularly significant in addressing issues such as popularity bias and user acceptance of recommendations (Shrivastava et al., 2022).

Cultivating serendipity in online shopping environments can benefit both consumers and e-commerce platforms, because it provides a sense of excitement and discovery, making the shopping experience more enjoyable and memorable, and it could also lead to unexpected purchases, which can increase sales and revenue for e-commerce platforms (Grange et al., 2019).

2.1.3 Recommender Systems Challenges

Recommendation systems can face some challenges in their implementation and functionality. The main problems are listed below.

2.1.3.1 Cold Start

When a new user is created in the system with an empty profile, having not liked or rated any product or item, the system faces a challenge in determining the user's preferences. This scenario is known as the "cold start problem." In such cases, the recommender system lacks sufficient data to make accurate recommendations. To address this issue, one common approach is to create the user's information profile through a survey. This method helps in gathering initial data about the user's preferences, enabling the recommender system to start providing more personalized recommendations (Mansur et al., 2017).

Similarly, item-based systems encounter the cold start problem when new items are added to the system. These items suffer from a lack of historical user ratings or interactions, making it difficult for the system to assess their relevance for recommendation (Mansur et al., 2017).

Both cold start problems – for new users and new items – can be effectively mitigated using hybrid filtering techniques (Mansur et al., 2017). Hybrid systems combine the strengths of both content-based and collaborative filtering, allowing them to leverage alternative data sources and methods to provide recommendations even when direct user-item interaction data is sparse or non-existent.

2.1.3.2 Sparsity

When working with huge amount of data, which is the case for e-commerce, it is very likely that only a small number of items will be rated by the users, this is known as the sparsity problem. Because just a small proportion of items are rated, the matrix that represents user-item ratings becomes very sparse, that makes recommending items to the user a difficult task in collaborative filtering systems, since there may be not enough common items rated between two users to establish similarity (Bobadilla & Serradilla, 2009).

2.1.3.3 Scalability

The scalability problem occurs when the system needs to handle an increasing number of customers, products and interactions while maintaining accuracy and performance. In today’s internet and e-commerce environment, RCs needs to deal with huge amounts of data, making it a great challenge to recommend to a high number of users in online interactions (Pagare & A. Patil, 2013; Roy & Dutta, 2022). Model-based approaches often have higher scalability but poorer accuracy, whereas memory-based approaches have higher accuracy but lower scalability. (Pagare & A. Patil, 2013).

2.1.3.4 Privacy

Recommender systems, while enhancing user experience through personalized suggestions, can also bring significant privacy risks. A major concern is the potential exposure of personal user information. This exposure can result from unauthorized access to user data or the ability

to infer personal details from the recommendations made. It is crucial to find a balance between the need for sharing personal information to improve recommendation quality and the protection of user privacy. Techniques such as the value of information (VOI) are suggested for determining the optimal point at which to cease further data collection and to make informed decisions about discarding excessive user information (Lam et al., 2006). In addition, to ensure the privacy of the users, specialized algorithms and security programs can be applied (Mansur et al., 2017).

2.1.4 Recommender Systems Evaluation

In Recommender systems evaluation, there are three different evaluations protocols: offline evaluations, user studies, and online evaluations. These protocols can also be used together to obtain a more precise results about a RS performance (Zangerle & Bauer, 2023). The description for each of the evaluation types follows.

2.1.4.1 Offline Evaluation

This is most commonly used experiment type, and it consists of using pre-collet datasets that contains users' implicit or explicit feedback on items. Implicit feedback can be the items viewed or purchased by the user, and explicit feedback is usually the rating of items. This data in separated in training and tests, and the RS is evaluated in its capacity to predict users' behaviour in the test set data. In this way, there are no real users involved in the actual experiment. This method usually involves comparing the model performance to two or more RS algorithms (Zangerle & Bauer, 2023).

2.1.4.2 User Studies

User studies consist of getting a small group of people to test the RS by doing specific tasks. This helps to see how people use the system and what they think about it in real-time. The study can happen in a lab or in a real-world setting. The study tracks what users do, like how often they click on recommendations. It can also include asking the users questions to get more detailed feedback. It can be used to compare different systems or just look at how users interact with one system. (Zangerle & Bauer, 2023).

2.1.4.3 Online Evaluation

In online evaluations, the RS is deployed and used in a real-life setting. In this way, people using the system aren't given specific tasks but use it for their own real-world needs. This makes the evaluation very realistic because users interact with the system as they would normally. Online evaluations provide insights into how well the system performs for users who have actual needs for recommendations. User actions are tracked and recorded, and this data is used to figure out how accurate the recommendations are. Online evaluations are typically conducted as A/B testing to contrast the modified system with the original system (Zangerle & Bauer, 2023).

2.1.5 Recommender Systems in e-commerce

RSs enhance online shopping experiences by offering personalized product recommendations to users, the suggestions adapt and are unique to each user. This can enhance sales, as users who browse without intention of buying something, can eventually make a purchase if suggested a relevant recommendation. Different approaches to create recommendations can be adopted, such as recommending the top seller products, based on their overall popularity on the site, or creating recommendations derived from a customer's previous purchases (Fayyaz et al., 2020).

E-commerce RC can also increase profit by applying marketing strategies, such as cross-selling and upselling (Fadillah et al., 2021; Fayyaz et al., 2020). Cross-selling is when customers are offered extra products or services in addition to the main item they bought, it means selling different products along with the original product. Up-selling is when you encourage customers to buy more or a higher-priced version of what they're already buying, it is about enhancing the original product to a better version, often by offering a larger size or a premium variant (Fadillah et al., 2021).

Another approach that has been gaining popularity in the e-commerce recommenders is context-based recommendations. Considering additional factors such as user location, the type of gadget used, the time and even seasonal events or weather conditions can make the recommendations even more personalized and relevant. For example, a user browsing a clothe store can receive different recommendations during wintertime compared to what could be recommender in summertime. This type of approach has shown significant impact in customer satisfaction (Adomavicius et al., 2011).

In this section, a brief overview about the state-of-the-art regarding RS approaches, metrics, challenges and evaluations was conducted. This is intended to lay basic ground knowledge about RS, in order to axialite in the interpretations and insights regarding the systematic review results that follows.

2.2 Systematic Review

This systematic review aims to provides a comprehensive analysis of existing literature relevant to this study and to answer a research question. The review follows established guidelines for systematic reviews, including clearly defined search criteria, inclusion and exclusion criteria, and evaluation of the quality of selected studies, ensuring reliable basis for the concluded results.

2.2.1 Methodology

A systematic review is a mean of evaluating and interpreting all available research to a particular research question, topic area or phenomenon of interest (Kitchenham, 2004). This review will follow the PRISMA methodology with adaptations to gain understanding of the current state and provide insights on how to develop a novel recommendation system (Page et al., 2021).

Based on its guidelines, this study will take the following steps: Formulate the research questions that best fit to answer the central question of this study. Conduct a search for relevant studies, specifying the data sources and search terms employed. Screen and select based on the defined inclusion and exclusion criteria. Evaluate the quality of studies included using the established quality assessment criteria. Perform data extraction and analysis. Compile the evidence, interpret, and present the findings, answering the proposed research questions.

2.2.2 Research Questions

The main objective of this research is to develop a recommendation system for e-commerce that aims for customer satisfaction, and therefore focuses on the metrics novelty, serendipity, and recommendation diversity. Based on this goal, the following research questions were formulated:

RQ1: How can recommender systems in e-commerce apply serendipity, novelty and diversity in their algorithms?

RQ2: How can serendipity, novelty and diversity be evaluated in e-commerce recommender systems?

RQ3: What are the main limitations studies aiming for serendipity, novelty and diversity encounter?

2.2.3 Data Sources

Three electronic databases were chosen for the research of the relevant sources: The IEEE Xplore Digital Library[1], ScienceDirect[2], and ACM Digital[3]. IEEE Xplore, managed by the Institute of Electrical and Electronics Engineers, is renowned for its extensive resources in electrical engineering, computer science, and related fields. ScienceDirect, an Elsevier database, provides a broad spectrum of scientific and technical research in the physical sciences and life sciences among others, it has a comprehensive collection and user-friendly interface. The ACM Digital Library, from the Association for Computing Machinery, stands out for its focus on computing and information technology, hosting extensive amount of literature including pioneering work

[1] https://ieeexplore.ieee.org/Xplore/home.jsp
[2] https://www.sciencedirect.com/
[3] https://dl.acm.org/

in computer science. Collectively, these databases offer high-quality research materials, satisfying the standards to conduct this study.

2.2.4 Search Terms

The choice of the search query to be used in the database involved selecting the keywords related to the search domain, the object was to deeper the knowledge about recommender system in e-commerce and the metrics serendipity, novelty and diversity, so the following keywords were selected:

Recommendation system OR recommender system: these expressions are interchangeable, and both are present in the literature.

E-commerce OR electronic commerce: e-commerce is an abbreviation for electronic commerce and both forms are acceptable in literature.

Serendipity AND Novelty AND Diversity: the objective was to find sources that contained relevant data about those 3 metrics and their relationship.

All the keywords were combined in a query to perform the search. The IEEE database did not find any result for the combination of all the keywords together, so in this case, the search had to be done with three different queries, each one with one of the metrics. Table 1 displays the number of results found for each query in the databases.

Table 1 – Number of results found in the databases for each query

Query	IEEE	ACM	Science Direct
(“Recommendation System” OR “recommender system”) AND ("e-commerce" OR "Electronic commerce") AND "serendipity" AND "novelty" AND "diversity"	0	132	69
(“Recommendation System” OR “recommender system”) AND ("e-commerce" OR "Electronic commerce") AND "serendipity"	7		
(“Recommendation System” OR “recommender system”) AND ("e-commerce" OR "Electronic commerce") AND "novelty"	15		
(“Recommendation System” OR “recommender system”) AND ("e-commerce" OR "Electronic commerce") AND "diversity"	51		

2.2.5 Inclusion and Exclusion Criteria

The inclusion criteria to select a source to be part of the result was:

- The source describes how serendipity, novelty or diversity can be implemented in an e-commerce recommendation system.
- The source describes how serendipity, novelty or diversity can be implemented in recommendation system for other area, but it could be applicable or relevant for e-commerce.
- The source describes how serendipity, novelty or diversity can be measured in a recommendation system.

The following criteria excluded sources from the selection process:

- The focus of the source is not recommendation systems.
- The source does not approach serendipity, novelty or diversity as a relevant part of the study.
- The recommender system studied it not for e-commerce and cannot relate to the field in any way.
- The source was not published in the last 6 years.

2.2.6 Quality Assessment

After the paper passed the inclusion and exclusion criteria, its quality was assessed regarding these aspects:

- The relevance of the study for the e-commerce field.
- The clarity of the results. Each paper was evaluated whether it presents clear metrics, solid discoveries, and definitive conclusions.
- The number of citations the paper has.

2.2.7 Data Extraction and Synthesis

In total, 274 sources were found with the proposed research queries in the different databases. After the removal of the duplicated sources, 267 papers followed the abstract screening process. In that phase, the sources were assessed whether it met the inclusion or exclusion criteria. 215 papers were discarded, 36 were classified as potentially relevant and 16 as relevant. The potentially relevant sources went through an additional process of results and conclusions sections reading, resulting in 22 discarded and 14 classified as relevant. The resulting 30 papers classified as relevant went through the quality assessment process, resulting in 21 papers selected for the results of this review. Figure 2 displays a flow diagram of the selection process.

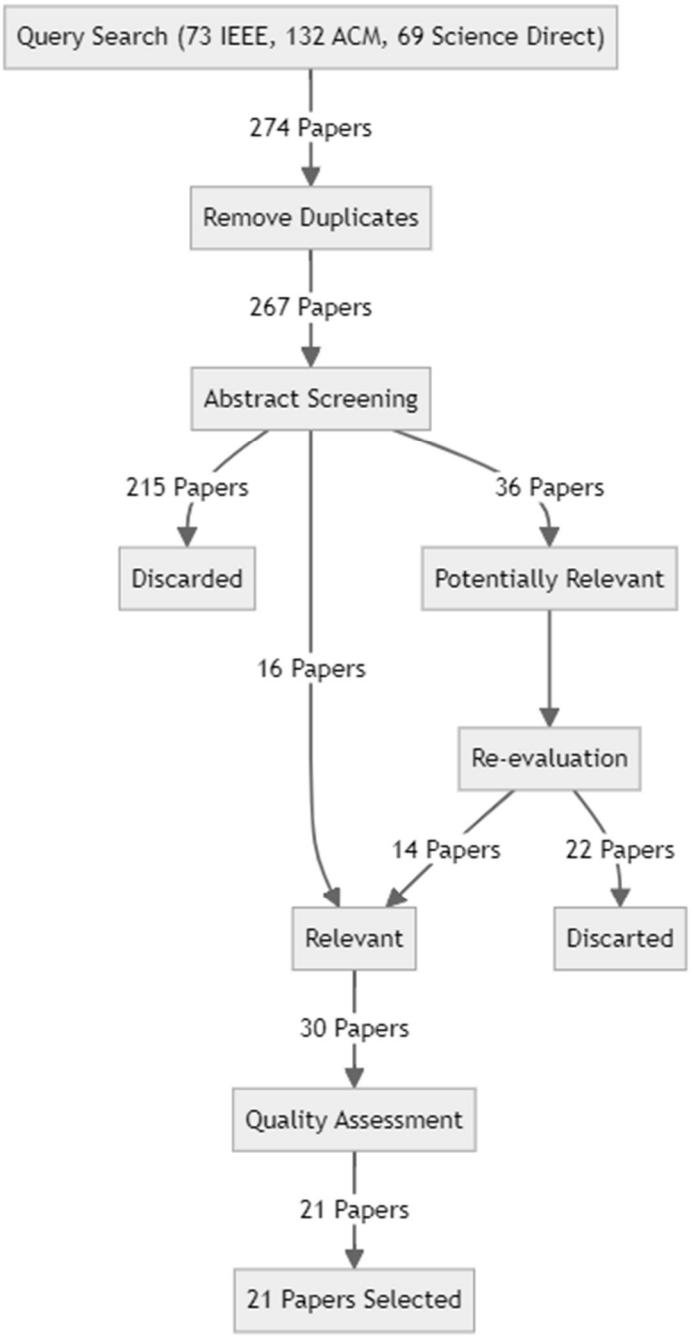

Figure 2 – Flow diagram of the paper selection process for the systematic review

2.2.8 Research Questions' Answers

After completing the data extraction and synthesis, the selected papers were used to answer the three research questions. Their content was thoroughly read and analyzed, resulting in the following conclusions.

RQ1: How do recommender systems in e-commerce apply serendipity, novelty and diversity in their algorithms?

Even though they are intrinsically related metrics, few studies focus on all three metrics at the same time (Herlocker et al., 2004). Although they are often mentioned together as an objective of improvement in the proposed RSs, the evaluation of the systems developed in the selected papers may have been done using either just one of the metrics or two combined or all three. Table 2 displays the selected papers, the datasets chosen for training and evaluation for each of them, and which of the three metrics was used for evaluation.

Focusing only on diversity improvement, many articles addressed the accuracy-diversity dilemma, that is how to improve diversity without decreasing accuracy. An et al. (2020) presented the Expert Tracking Approaches (ExTrA). Designed to identify and utilize 'fabricated experts' within RSs. Fabricated experts are users with an exceptional ability to assist others in discovering relevant and varied items. ExTrA methods distinguish these experts from regular users and prioritize them in the initial resource allocation phase of the Mass Diffusion (MD) model, a prominent bipartite network-based recommendation method. Carvalho et al. (2019) proposed a novel approach to address the accuracy-diversity dilemma using Particle Swarm Optimization (PSO). Their method involves a post-processing step that re-ranks items recommended by traditional RSs. (H. Yang et al., 2023) propose a novel method, LOAM, to improve long-tail session-based recommendation. LOAM consists of two main components: Niche Walk Augmentation (NWA) and Tail Session Mixup (TSM). NWA generates synthetic sessions considering the long-tail distribution, exposing the recommender model to various item transitions with global information. This improves item coverage in recommendations. TSM, on the other hand, interpolates sessions at the representation level, making the model more generalized and robust. It encourages the system to predict niche items with more diversity and relevance. (Zheng et al., 2021)proposed an innovative approach using Graph Convolutional Networks (GCN). Their method integrates diversification into the candidate generation stage, addressing the issue of information redundancy and improving user satisfaction. Their model employs strategies like rebalanced neighbor discovering, category-boosted negative sampling, and adversarial learning to ensure a variety of recommended items.

To improve aggregate diversity, Mansoury et al. (2020) introduced FairMatch, a graph-based algorithm that functions as a post-processing step after the generation of recommendations. The algorithm aims to improve aggregate diversity by iteratively identifying items that are rarely recommended yet are of high quality, and adding them to the users' final recommendation lists. This is achieved by solving the maximum flow problem on a recommendation bipartite graph.

Yu et al. (2019) aimed on improving individual diversity and aggregate diversity at the same time, developing an adaptive trust-aware recommendation model based on a user-item bipartite network. The model calculates the trust degree of a user towards a trustee and uses it to combine interests of the user and the trustee. It also quantifies the user's need for diversity calculating the dissimilarities among the items the user rated. Then it establishes an adaptive strategy to make trade-offs between accuracy and multiple dimensions of diversity. T

Shandhilya & Srivastava (2020) introduced a novel source of diversity in recommendation systems, focusing on within-item conceptual incongruity. The concept of 'unexpected' recommendations is typically measuring by the dissimilarity of items from what a primitive recommendation system would predict. In contrast, this study introduces a unique approach by quantifying conceptual incongruity within individual items using metadata. Each item is assigned an incongruity measurement that is independent of other items in the recommendation system's ranking. This innovative perspective allows for a more nuanced assessment of diversity, considering the inherent diversity within items themselves based on their conceptual incongruity.

Focusing on serendipity improvement, Shrivastava et al. (2022) proposed a method for improving the quality of top-n recommendation lists by integrating textual review-based rating metrics with user-item rating metrics. Their approach, the Two-Fold Algorithmic (TFA), calculates a serendipity score for each user-item pair, considering the uncertainties associated with item popularity and how closely items match user preferences and incorporates less common, long-tail items into the recommendations. Furthermore, the research introduces the Serendipity Objective Optimization-based Recommendation Framework (SOORF), an evolutionary optimization model designed to find a balance between different elements of serendipity, such as relevant, novel, and unexpectedly aligned with user preferences. Wang et al. (2019) proposed a termed innovator-based collaborative filtering (INVBCF) based on a user survey on the online shopping habits in China, which introduces the concept of "innovators." Innovators are users who can discover cold items without the help of a recommender system. By leveraging these innovators, the algorithm can balance serendipity and accuracy in recommendations.

(Hasan & Bunescu, 2023) introduced a content-based formulation that leverages Bayesian surprise to measure the serendipity of items after they are consumed and rated by users. This method was combined with a collaborative filtering component, which identifies similar users and recommends items that were serendipitous for them.

Ziarani & Ravanmehr (2021) developed a Deep neural network approach for a serendipity-oriented recommendations. Their method integrates a Convolutional Neural Network (CNN) with the Particle Swarm Optimization (PSO) algorithm. The CNN is employed to predict 'focus shift points', which are based on unexpectedness and relevance parameters, extracted from a ground-truth serendipity dataset, for each user. These points indicate the potential for serendipitous recommendations. The PSO algorithm then searches for recommendations close to these predicted focus shift points. To finalize the recommendation list, the Serendipitous Personalized Ranking (SPR) method is used to re-rank the candidate recommendations.

P. Li et al. (2020) proposed the Personalized Unexpected Recommender System (PURS), a deep learning model focused on incorporating unexpectedness into recommendations. The model employs a multi-cluster approach to model user interests in a latent space and utilizes a self-attention mechanism to capture personalized and session-based user perception of unexpectedness, taking into account users' varying preferences for novelty or familiarity and

their context-specific perceptions of unexpectedness, such as recommending familiar content in a series or diverse options after binge-watching. Additionally, an unexpected activation function is introduced to fine-tune the system's output, balancing between unexpectedness and business performance metrics like click-through rates.

(Xu et al., 2020) proposed the Neural Serendipity Recommendation (NSR) method, aimed at balancing accuracy and novelty. This method integrates Multi-Layer Perceptron (MLP) and Matrix Factorization (MF) to predict Serendipity, which the authors characterized by high user satisfaction and low initial interest. The NSR method operates in two phases: Serendipity Prediction and Personalized Recommendation. In the first phase, MLP captures the novelty aspect while MF ensures accuracy in recommendations. The second phase involves a candidate filtering method designed for personalized recommendation, which maximizes the knowledge acquired from predicted serendipity and alleviates the issue of data sparsity.

Aiming on improving diversity and novelty, Berbague et al. (2018) proposed the use of a weighted similarity measure in user based collaborative filtering. It considers the amount of unused information in a given user profile and exploit it to select a beneficial neighbourhood for each user. The adjustment to the similarity measure allows for the inclusion of unseen items in the counterpart candidate neighbour user. It also use an aggregated novelty measure to adjust users' pairwise similarity values, users who can positively affect the diversification task are made closer to the target user, while users who negatively affect diversification are excluded. (Niu et al., 2019) proposed a user-based two-step recommendation algorithm with popularity normalization. The method incorporates item popularity into both the similarity calculation and probability prediction processes to balance well-known and lesser-known items. The two-step approach involves first predicting items that users are likely to select and then estimating the ratings for these selected items, reflecting real user behavior. Additionally, two new evaluation metrics to measure diversity and novelty are proposed.

He et al. (2021) proposed a modified CF model named CF-S&WLT (CF-Strong & Weak ties for Long-Tail distribution) that integrates users' strong and weak ties and long-tail distribution items. Their model distinguishes user social ties with varying weights, maintaining accuracy through strong and weak social ties while improving diversity and novelty through incentive coefficients for long-tail items based on item ranking position and number of ratings.

Jain et al. (2020) proposed a genetic algorithm-based multi-objective optimization method that combines user-based and item-based collaborative filtering algorithms with a new probabilistic crossover operator and an extended similarity model based on the Bhattacharyya Coefficient. The method also considers both the order and frequency of parent genes in offspring generation to enhance the searchability of the algorithm.

Using novelty and serendipity as evaluation metrics, (Sun et al., 2020) introduced a new training framework based on Bayesian Graph Neural Networks (BGNNs) to address limitations in current graph neural network (GNN) based recommendation approaches. The proposed framework, named BGCF (Bayesian Graph Collaborative Filtering), incorporates a random graph generative model based on node-copying. This model produces sample graphs with sufficient diversity in

edges to promote better learning, addressing the uncertainty in observed user-item interaction records and bringing diversity into recommendation results.

Though the following models do not explicitly focus on increasing diversity, novelty or serendipity, they focus on improving RS performance and user experience in general, and in this way, they evaluate the proposed models using all three metrics.

Karthik & Ganapathy (2021) developed a fuzzy logic-based recommendation system that takes into account the current interest of the user, that is the end user of the search. The algorithm leverages sentiment analysis, demographic filtering, and fuzzy logic to generate nuanced recommendation lists tailored to the user's preferences and interests. The sentiment score, along with the average customer review score, number of reviews, and associated target user category, is used to calculate the overall product rating score for each end user group. The algorithm also considers demographic information such as age group and delivery location to filter products that are not known to the user, improving the recommendation list. Fuzzy logic is applied to generate highly recommended, recommended, and likely to be recommended lists based on the sentiment score and product rating score, providing varying levels of decisions for personalized recommendations. The algorithm uses NLP techniques such as tokenization, parsing, and parts of speech (POS) tagging to preprocess the customer reviews.

C. Yang et al. (2020) proposed the GANCF (Gated and Attentive Neural Collaborative Filtering) method, a neural network-based method that utilizes both item-level and list-level information for enhanced performance. The approach includes a representation learning network equipped with attention and gate mechanisms to concurrently learn user, item, and list embeddings. Additionally, they introduced an interaction network to learn the dynamics of user-item and user-list interactions. This network utilizes shared convolution layers for both types of interactions, contributing to improved overall performance.

M. Chen et al. (2021) proposed an approach to enhance user exploration in Reinforcement Learning (RL)-based recommender systems, focusing on the balance between exploration and exploitatio. Their methodology includes the implementation of Entropy Regularization, which encourages the system to explore content beyond the user's known interests by fostering a high-entropy output distribution, thereby preventing the system from prematurely converging to suboptimal deterministic policies. Additionally, they introduced an Intrinsic Reward mechanism, adding a novelty-based reward to the RL objective to incentivize the recommendation of novel items. Another significant aspect of their approach is Exploration via Diversification, which involves adjusting the softmax policy to boost the likelihood of recommending less popular items, thus enriching the diversity of the recommendations. Finally, they employed a Hybrid Exploration Strategy that synergistically combines these methods to effectively balance the aspects of exploration and exploitation.

Table 2 – Dataset and metrics used for evaluation of the selected studies

Study	Dataset	Diversity	Novelty	Serendipity
(He et al., 2021)	Hetrec2011-delicious-2k, Hetrec2011-lastfm-2k	x	x	
(Karthik & Ganapathy, 2021)	Amazon Review Data	x	x	x
(Jain et al., 2020)	MovieLens	x	x	
(Ziarani & Ravanmehr, 2021)	Serendipity Data 2018			x
(C. Yang et al., 2020)	Douban and Netease	x	x	x
(Shrivastava et al., 2022)	MovieLens, Serendipity Data 2018, IMDB movie review			x
(Yu et al., 2019)	Ciao, Yelp, Epinions	x	x	
(An et al., 2020)	MovieLens, Netflix, RYM	x		
(Berbague et al., 2018)	MovieLens	x	x	
(Niu et al., 2019)	MovieLens	x	x	
(Carvalho et al., 2019)	MovieLens	x		
(Wang et al., 2019)	Real-world e-commerce Alibaba Group dataset	x	x	x
(P. Li et al., 2020)	Yelp, MovieLens, Youku	x		x
(Xu et al., 2020)	Movielens, Yelp, BookCrossing			x
(Hasan & Bunescu, 2023)	Goodreads			x
(M. Chen et al., 2021)	Not mentioned	x	x	x
(Sun et al., 2020)	AmazonMovies, Amazon-Beauty, Amazon-CDs		x	x
(Shandhilya & Srivastava, 2020)	WikiPedia	x		
(Mansoury et al., 2020)	Epinions, MovieLens	x		
(H. Yang et al., 2023)	Nowplaying, Diginetica, Yoochoose,RetailRocket	x		
(Zheng et al., 2021)	Taobao, Beibei, Million Song Dataset	x		

RQ2: How can serendipity, novelty and diversity be evaluated in e-commerce recommender systems?

Innumerous approaches to measure those three metrics were found in the papers. Following, each metric is presented separately.

Diversity:

(He et al., 2021; C. Yang et al., 2020; Yu et al., 2019) uses cosine distance to measure the distance between items in a recommendation list. While the core concept of using cosine distance is similar, the specific formulations and contexts differ. Karthik & Ganapathy (2021) uses a similar approach where diversity is calculated as the average of the sum of similarities between pairs of items. Berbague et al. (2018) also using cosine similarity, defined intra-list diversity, computing the average distance between recommendations within the same list.

Aggregated diversity has also been approached with different definitions. Yu et al. (2019) calculate aggregated diversity as the total number of unique items in all users' top-N recommendation lists. An et al. (2020) focuses specifically on the number of distinct items in the recommendation lists of all users, with each list having a length of K (Diversity-in-top-K). Berbague et al. (2018)define it as the amount of promoted products among the recommendation sets.

An et al. (2020) also defined Intra-Diversity, the recommendation diversity for a single user using cosine similarity and Inter-Diversity, evaluates the difference between recommendation lists of each user pair using Hamming distance.

Some authors define diversity based on the characteristic of the items. Carvalho et al. (2019) used among others, the genre of movies. Jain et al. (2020) considered the distribution of items in different categories. M. Chen et al. (2021) considers diversity both on topic level and on content provider level.

Niu et al. (2019) proposed a new diversity metric, HitCOV, based on the coverage metric. The proposed metric measures the extent of distinct items recommended to users that align with their preferences. HitCOV enhances the traditional coverage metric by focusing not just on the variety of items suggested but also on their relevance to individual user interests.

(P. Li et al., 2020; Wang et al., 2019; H. Yang et al., 2023) measured diversity using the coverage metric, measured by the proportion of recommended items in the full item set, the higher coverage implies more items can be recommended to users. (H. Yang et al., 2023) also considered TailCoverage, showing the diversity of tail items.

(Mansoury et al., 2020; Zheng et al., 2021) used Coverage, Gini Index (The measure of fair distribution of recommended items) and Entropy (the uniformity of that distribution) for evaluating their diversity improvement proposed method.

Shandhilya & Srivastava (2020) proposed a novel diversity metric, incongruity. This metric considers the inherent diversity within items themselves based on their conceptual incongruity, that is quantified within individual items using metadata.

Novelty:

Popularity-based approaches are the most common one. The central idea is that an item's novelty is inversely proportional to its popularity. This popularity is often quantified in terms of the probability or frequency of the item being seen or chosen within the user base. The assumption is that items which are less frequently encountered or chosen in the system are more novel to users. Even following the same concept, the formulas used to calculate in each study are different. (Sun et al., 2020; C. Yang et al., 2020) novelty's is inversely proportional to the item's popularity. Karthik & Ganapathy (2021) calculated novelty depending on the probability of an item being observed. Similarly, He et al. (2021) defined the probability of an item being seen considering the frequency of choice of that item.

Jain et al. (2020) measured the proportion of unpopular (or new) items over the total number of items in the recommendation list, considering the total number of users in the recommendation system, the length of recommendation list and the number of users who rated each item.

M. Chen et al. (2021) novelty measurement focusses on the global popularity-based measurements, looking at how often each item is consumed by the entire user base.

Yu et al. (2019) defines novelty by the frequency of long-tailed items (items not frequently recommended or popular) in users' recommendation lists.

Wang et al. (2019) used the average popularity of recommended items metric to measure novelty. The lower average popularity implies higher novelty.

Niu et al. (2019) proposed a new metric for measuring novelty, named HitCIL, which is an extension of the coverage in long tail metric. This metric specifically evaluates the system's ability to recommend items that are not only less popular or less known, known as long tail items, but also relevant to the user's unique interests.

Serendipity:

Serendipity is commonly defined as a pleasant surprise, that is an item that combines the elements of surprise and relevant. However, different authors define surprise and relevance in different ways.

(Karthik & Ganapathy, 2021; Sun et al., 2020) calculates degree of surprise by considering the probability of a product that is recommended to the target end user and with how much probability that same product is recommended for any other customer.

C. Yang et al. (2020) and Shrivastava et al. (2022)measured serendipity considering the intersection between recommended items considered new or unexpected to the user, and

items that the user has consumed. This intersection is divided by the total number of recommendation items or lists. Although calculated in different ways, in essence both formulas quantify how many of the recommended items or lists are both unexpected and consumed by the user, relative to the total recommendations made.

M. Chen et al. (2021) considered the serendipity score of an item as 1 if the recommended item is not only relevant but also belongs to a different topic cluster than any other item in the user's interaction history, and 0 otherwise. Then used these scores to calculate the serendipity of the recommendation set.

Wang et al. (2019) evaluated serendipity using the average difference (AD) time and the average distance (AvgDistance) as evaluation metrics. The AD time metric measures the average temporal interval between when the recommender system suggests items and when users would have discovered these items on their own. AvgDistance is a novel proposed metric and measures the difference between the recommended items and the user's historical interests, considering the categories of the items.

P. Li et al. (2020) evaluated the concept of unexpectedness rather than serendipity, as the measures of item recommendations to users that are not included in their consideration sets and differs from the user's past behaviour or preferences.

Xu et al. (2020) proposed a novel Serendipity metric, S−p. This metric treats items with high Satisfaction but low Interest as Serendipity. The calculation involves the pair-wise distance between Interest and Satisfaction of each item. The idea is that if the Interest of an item is higher than Satisfaction, it harms the utility of Serendipity.

Ziarani & Ravanmehr (2021) used a ground-truth serendipity dataset, a dataset created using real users' feedback about the recommended items to evaluate the proposed model. They measured serendipity as the items that were considered both useful and unexpected by users' vote.

Hasan & Bunescu (2023) didn't evaluate the proposed model using a serendipity metric, instead they implemented serendipity by measuring the bayesian surprise of items in a real-world dataset and then evaluate the accuracy of the model against a manually annotated benchmark for serendipity.

RQ3: What are the main limitations studies aiming for serendipity, novelty and diversity encounter?

Common RS limitations were also found in the proposed models as data sparsity, bias (M. Chen et al., 2021), cold-start problem for new users and new items (Ziarani & Ravanmehr, 2021) and extensive requirement of computational and storage resources (Jain et al., 2020).

Although presenting overall good results, some algorithms did not achieve better performance than the compared algorithms in all measured metrics. C. Yang et al. (2020) outperformed in serendipity measurement, but in diversity and novelty it did not present improved results. Yu

et al. (2019) model focused on diversity but had better performance comparing to other algorithms in accuracy for cold-start users and long-tailed item. (An et al., 2020; He et al., 2021; Zheng et al., 2021) displayed acceptable accuracy loss in comparison with the start of the art algorithms. This demonstrates that novel approaches doesn't always reach their target completely, and highlights the complex and delicate process of maintaining accuracy while also improving other user satisfaction metrics.

Zheng et al. (2021) recognize that the trade-off between accuracy-diversity is generally more observed in offline evaluation, and that in online scenarios, diversity tends to increase user satisfaction. Mostly of the selected studies uses exclusively offline evaluation. Exceptions are (M. Chen et al., 2021; P. Li et al., 2020; Shandhilya & Srivastava, 2020) that conducted online experiments. And although offline, (Shrivastava et al., 2022; Ziarani & Ravanmehr, 2021) used a ground-truth serendipity dataset for evaluation, which enhances results' relations to real world scenarios.

Hasan & Bunescu (2023) tested their method by manually labeling surprise items. This methodology, while establishing reliable ground truth data, is very time-consuming and only covers a limited number of users.

Xu et al. (2020) pointed that one of the limitations of their proposed method, is that it relies solely on user-item ratings as input. This focus on user-item ratings means that their method might not be leveraging the full potential of available data. The authors suggest that future work could explore the use of multi-source, multi-view data, such as pictures, photos, texts, and audio, to enhance the serendipity recommendation for users. This observation can be applied to most of the encountered models.

2.3 Summary

The selected papers display a great number of different approaches to increase the focused user satisfaction metrics. Demonstrating how different techniques, such as pre-processing, post processing, re-raking, reinforcement learning, text mining and others, can be effective applied and display good results aiming on solving the same problem.

Exploring deep learning and text mining techniques has proven to be of great value. By leveraging textual reviews provided by users, for example, the model can extract valuable information even in scenarios where explicit ratings may be sparse (Shrivastava et al., 2022). This approach was only found in four of the selected studies though. And they focused on using textual reviews to sentiment analysis or score estimation. Leaving a gap to be explored, regarding items title and description. That could be used to create item-item and user-item relationships and be adapted to improve the systems with a chosen focus.

Even though a lot of the concepts are common, there is no consensus on how to measure any of the metrics. Because these concepts are intertwined and often dependent on each other, different authors have different definitions and different quantification approaches. That

makes it hard to compare the proposed model's performance with each other, making each result specifically related to its own study.

All the studies selected can be applied to e-commerce RS, but most of them do not use an e-commerce related database to train and evaluate the proposed models. In metrics like serendipity, that can be a problem, since its definition is often subject and the concept of pleasant surprise can be different in movies and e-commerce navigation, for example.

Serendipity has been the metric that was most connected to user satisfaction in the analyze results. Observations in live experiments concluded that serendipity-oriented models can improve long term user experience (M. Chen et al., 2021). In this way, the lack of common definition for serendipity calculation is a problem, and in many cases not even the core concept is similar between the studies. That emphasizes the need for more ground-truth serendipities datasets.

Based in these observations, this study will focus on improving the serendipity metric, as it inherently incorporates the concepts of diversity and novelty and is the most correlated to user satisfaction. NLP techniques play a key role by leveraging the semantic content of items, reflecting how users make choices in real-life scenarios. To simulate real-world interactions, the item's description will be used as its representation, focusing in recommending more serendipitous item and in stablishing serendipity based in its description, an innovative approach in the field. Additionally, as proven effective in the analyzed studies, the proposed model will incorporate Machine learning techniques in its development.

The study will also pursue the use of a serendipity ground-truth dataset, that captures directly the user definition of the serendipitous feeling in the e-commerce process. This approach seeks to address the existing challenges in defining and evaluating the serendipity metric, providing a more user-centred perspective on what constitutes a truly serendipitous recommendation.

3 Methods and Materials

This chapter outlines the methods and materials employed throughout the development of the experiments. It presents the dataset selected and the preprocessing steps applied to it. It also presents the experiments and the validation method. Additionally, it discusses ethical and security aspects that were taken into considerations in the data handling.

3.1 Introduction

L. Chen et al. (2019) performed a study where a survey involving over 3000 users in an industrial mobile e-commerce application, to measure users' perceptions of recommendations, in relation to serendipity novelty, diversity, user satisfaction, and purchase intention was conducted. The results proved the direct relationship that novelty and diversity have with serendipity and stated that serendipity has a significantly positive effect in user satisfaction and purchase intention. The study also compared four different algorithms approaches focused on popularity, relevance, novelty and serendipity. The serendipity-oriented algorithm performed better than the others in terms of novelty, relevance, timeliness, serendipity, user satisfaction, and purchase intention. This result stress the need to incorporate serendipity in e-commerce RSs, which focus relies in user satisfaction. The same conclusion can be taken from the state-of-the-art analysis.

Using semantic similarity to calculate an item incongruity, as proposed metric for diversity, has been proved in an online study to improve user experience, where participants were, on average, twice as likely to accept incongruity-centered recommendations than relevance-centered recommendations. Demonstrating that users who are casually browsing web media without a specific informational objective are likely to be intrigued and captivated by incongruity-centered recommendations, which offer serendipitous and creative content experiences (Shandhilya & Srivastava, 2020). That can be valuable insight for e-commerce RS, in situations where the user does not know exactly what he wants to buy, for example, when the user is searching for gift ideas. According to a user survey on online shopping habits in China, 38.1% of users love to browse the website in their spare time, and not only when they have clear demands (Wang et al., 2019).

Taking all this factor in consideration, this study proposes the development of a hybrid recommender system for e-commerce, that will focus on serendipity, and leverage NLP and text mining techniques to draw profiles of items themselves, and correlations between item and users, based on the titles and descriptions of items.

The idea is that similarly to Shandhilya & Srivastava (2020) work, individual items' metadata, in this case the title and description of an item, can be used to represent diversity and serendipity of a product itself, for example, a sandal for running is already a serendipitous item, since it is not obvious, as the usual choice for running is sneakers. The objective is to prove that based in an item title and description, the system can predict if for a specific user, it would be considered a serendipitous item or not. And use it to prioritize serendipitous items recommendations.

3.2 Method and tools

After state-of-the-art research and systematic review to investigate how to best apply diversity, novelty and serendipity to an e-commerce RS, this study proposes the development of a hybrid RS focused in serendipity, applying NLP and text mining techniques to items metadata, namely, title and description, in order to establish more serendipity-focused recommendations.

The development of RSs requires the model to be trained and evaluated using robust datasets. The proposed RS aims to improve the online shopping experience on e-commerce platforms. The SerenLens dataset (Fu et al., 2023) was chosen because, until this date, it is the only one that contains ground truth information about serendipity based on data from the Amazon Review Data dataset (McAuley et al., 2015). In this way, this dataset provides a realistic environment for training and evaluating the model, ensuring the reliability and relevance of the results.

The dataset will go through a preprocessing phase to ensure data quality and relevance. This includes cleaning the data and normalizing text data. After that, the proposed model will be trained and tested. The method includes comparing the performance of the model against one state-of-the-art recommender algorithms using strategic metrics. This comparison is important to validate the superiority of the developed RS.

3.3 Dataset

After further exploration for a serendipity ground truth dataset, the SerenLens dataset (Fu et al., 2023) was found. As it is the result of very recent research, the dataset did not appear in the previously conducted state-of-the-art research. The SerenLens dataset was developed by researchers to capture serendipity experiences in two different domains, Books and Movies, domains where personal taste and subjectivity play a significant role.

As Fu et al. (2023) did in their proposed model, this study will utilize the Books domain dataset. To create the dataset, they used Amazon Review Data (McAuley et al., 2015) as the foundation.

The goal was to classify the serendipity feeling using the reviews each user left for the items. In a ten months process, the researchers manually curated seven keywords related to serendipity, such as "stumble upon" and "find by chance," to identify relevant reviews. Then, they collected 41.340 user judgments through MTurk, to classify 10.000 selected reviews. Of these, 2.557 reviews were identified as serendipity experiences, linked to 2.346 users and 2.227 books. By incorporating the full review histories of these users, the dataset was expanded to include a total of 265.037 reviews and 113.876 books, providing a rich context for understanding serendipity. The SerenLens dataset is publicly accessible at SerenLens GitHub[4] (Fu et al., 2023).

Data format includes six fields, user_id, item_id, timestamp, review, rating and label (representing the item serendipity as 0 or 1). Table 3 shows data examples from the SerenLens dataset.

Table 3 – SerenLens dataset data examples

user_id	Item_id	timestamp	review	rating	label
a10e3f50diujee	0061148512	1196899200	i have always loved this book. plath's literature is exceptional, in my opinion. but the reason i bought this copy, even though i already own one, is because there are a few of plath's sketches included in this edition. it may seem insignificant, but for someone who likes plath, this should be quite a treat!	5	0
a9owlc66j3584	0151015392	1339718400	this is one of my all time favorite books. i stumbled upon the paperback in the high school library and was immediately hooked. this book is funny, dramatic, scary and even romantic. when they made the movie i was incredibly excited. this is probably the best adaptation i have ever seen. read this book, watch the movie and then repeat. I highly recommend it.	5	1

[4] https://github.com/zhefu2/SerenLens

As SerenLens is based in the Amazon Review Data (McAuley et al., 2015) , metadata about its items are available. A more up-to-date metadata is available in Amazon Reviews'23 (Hou et al., 2024), this metadata was chosen because it contains richer information about the items, such as more descriptive features. Since this research will consider the semantic content of the items, more accurate and richer information could be of great importance.

Metadata includes fourteen fields, such as *images*, *price*, *categories*, *description*, *details* and others, but only two fields, *title* and *features* will be used. Regarding the Book's dataset, the *features* field contain data about the books theme and content, which is in practice, the books' description.

The following process was applied to generate the final dataset used to train and test the proposed model:

- Initial Adjustments

The *timestamp* and *review* columns were removed, as they will not be considered in this study. The *label* column was renamed to *serendipity* to provide a clearer understanding within this context.

- Item Mapping

Each entry in the SerenLens dataset had its item_id mapped to the corresponding title and features extracted from the Amazon Review Book Metadata dataset.

- Data Cleaning

Items with empty features were removed.

- Text Preprocessing

Text preprocessing techniques were applied to the *title* and *features* fields. As preserving semantic content is important, only text cleaning was performed. The texts went through the following process: lists to strings conversion, HTML tags and link contents removal, and finally, removal of special characters and punctuation, preserving only the main ones that can affect text meaning (e.g., .,!?).

- Filtering

With clean texts, the dataset was filtered to include only items with features with length between 400 and 800 characters. That was done because text with less than 400 characters will probably not be so semantically rich. And text with more than 800 characters would slow down model performance, considering the processing limitations the machine used to develop this work had.

- Final Dataset Refinement

From the result of the filtered dataset was excluded users who did not have any serendipitous item associated with them.

Figure 3 shows a flow diagram of the preprocessing process.

Figure 3 – Flow diagram of the dataset preprocessing process

The process resulted in the final dataset including 11.977 records, 477 users, 8.023 items and 494 serendipitous items and the fields user_id, item_id, rating, title, features and serendipity. Table 4 shows data examples from the final dataset used in this study.

Table 4 – Final dataset data examples

user_id	item_id	rating	serendipity	title	features
a10f5lmypxqydf	380761319	5	0	the shadow and the star	from nationally acclaimed bestselling author laura kinsale comes a boldly original, breathlessly unforgettable tale of honour, adventure and undying love., the shadow is wealthy, powerful and majestically handsome, he is a man of dark secrets a master of the ancient martial arts of an exotic distant land. scarred by a childhood of shocking degradation, he has sworn to love chastely ... but burns with the fires of unfulfilled passion., the star is lovely, innocent and nearly destitute, and drawn to him by a fevered yearning she could never deny following her enigmatic shadow warrior into a dangerous world of desire and righteous retribution.
a10f5lmypxqydf	60090383	5	1	rachels holiday	the fast lane is much too slow for rachel walsh. and manhattan is the perfect place for a young irish female to overdo everything. but rachels love of a good time is about to land her in the emergency room. it will also cost her a job and the boyfriend she adores., when her loving family hustles her back home and checks her into irelands answer to the betty ford clinic, rachel is hopeful. perhaps it will be lovelyspa treatments, celebrities, that kind of thing. instead, she finds a lot of group therapy, which leads her, against her will, to some important selfknowledge. she will also find something that all women like herself fear a man who might actually be good for her.

3.4 Experimentation and Validation

Two different experiments were conducted. 1 - The items embeddings considering all the text fields were used to calculate serendipity using the often-used formula in Eq. 1, and the results were compared to the ground-truth serendipity. The purpose of this experiment was to validate and discuss the formula's credibility, given that serendipity is often considered a personal and subjective concept (Kotkov et al., 2016).

2 – Three model versions were developed, and the results were compared to each other and to the SASREC (Kang & McAuley, 2018) model, a next-item sequential recommendation method based on the Transformer's architecture. The initial intention was to compare the results to the SerenEnhance (Fu et al., 2023) model, the model developed by the creators of the SerenLens dataset. Due to machine processing limitations and unclarity of the model code available, it was not possible to run SerenEnhance in the final dataset to compare the results. Anyway, the metrics chosen to compare the results are the same used in the forementioned study, HIT Radio (HR) and Normalized Discounted Cumulative Gain (NDCG), and the proposed serendipity metrics HRseren and NDGCseren.

3.5 Data Protection, Security and Ethics

The ethical handling of data is extreme important, especially when dealing with datasets that may contain personal or identifiable information. The SerenLens dataset (Fu et al., 2023) does not contain user-specific data but includes review texts that could potentially contain identifiable information. The Amazon Review Data'23 (Hou et al., 2024) for user reviews contains fields such as rating, title, text, images, user id and timestamp. While usernames are not included, the dataset does contain user ids and review texts and images, which might inadvertently include personal information.

When utilizing these datasets, it is essential to consider the implications of the General Data Protection Regulation (GDPR) in the data handling practices. GDPR, a regulation enacted by the European Union, imposes strict guidelines on the collection, processing, and storage of personal data of individuals within the EU. To ensure compliance with GDPR, the study follows some key measures.

Even though user ids are included in the dataset, they're anonymized and don't directly reveal personal identities. The study ensures these ids can't be linked back to individual users, maintaining their anonymity. To further address privacy concerns, any information that could potentially identify someone— like review texts and images - is deliberately left out of the analysis. From the Amazon Review Data'23 (Hou et al., 2024), the study uses only the item metadata dataset, that doesn't include any personal information, only data about the items.

By focusing on non-identifiable data such as ratings and items' features, the study reduces privacy risks. By excluding review texts and images and working with anonymized user IDs that

can't be traced back to individuals, it ensures that no personal data is involved in the analysis. This approach minimizes the risk of accidentally processing personal information and aligns with the GDPR's principle of data minimization.

While direct user consent is not obtained due to the public nature of the data, user rights are respected by ensuring that data handling aligns with GDPR provisions. The study acknowledges the responsibility associated with using publicly available datasets and takes steps to maintain ethical standards.

Transparency is maintained throughout the research process. The methodology is clearly defined, and any limitations or potential ethical concerns are openly discussed. The data is used exclusively for research purposes, and results are reported honestly and accurately, adhering to principles of academic integrity.

In conclusion, the study is conducted with a deep commitment to ethical standards and legal compliance. Adhering to GDPR guidelines and ethical research practices, ensures the protection of individual privacy and maintains the integrity of the research. This approach to data handling reinforces the credibility of the research outcomes and demonstrates a responsible use of publicly available datasets for academic purposes.

3.6 Summary

In this chapter, the goal of the study was presented, to develop a hybrid recommender system for e-commerce, leveraging NLP and machine learning techniques to focus on serendipity recommendations. It will focus on the book domain and use the books description to improve the quality of the recommendations and predict serendipity.

The methods and tools carried out in the study were covered, along with detailed information about the datasets involved and the final dataset obtained. The choice and the process that led to the final dataset used is of great importance, as the quality of the data has a strong influence on the quality and reliability of the results.

It also explained how the experiments will be conducted and validated, comparing the results of the proposed model with another benchmark model and using the HR@k and NDGC@k metrics. Proper implementation and validation methods are essential to ensure the proposed model performance and relevance.

Additionally, the approach to data protection, security, and ethics was discussed. By following the necessary rules and guidelines, the proposed model aims to be accurate, serendipity focused and ethical. By sharing details about the methods, tools, datasets, and the ethical considerations taken, research transparency is ensured.

4 Implementation, Analysis and Results discussion

In this chapter, the implementation of the proposed experiments will be presented. The Serendipity comparison analysis will compare four different embeddings generation approaches with the ground-truth serendipity and discuss the results. The proposed serendipity focused models will be developed based on XGBoost and the results will be compared against the SASRec model.

4.1 Serendipity Comparison analysis

The goal is to compare calculated serendipity based on the item's characteristics and in the user history with the ground-truth serendipity. Serendipity is often defined as a subjective feeling (Kotkov et al., 2016), so this is a way of comparing and validating true information with assumption metrics.

The state-of-the-art research shows there are different ways to calculate serendipity across literature. This study will, for the purpose of this comparison, consider a popular definition, that is the combination of unexpectedness and relevance (Silveira et al., 2019a). Considering the dataset only contains relevant items, serendipity will be calculated as the cosine similarity between the item (i) and the history of consumption of the user ($U = \{u_{1,}u_{2,} \dots, u_{n,}\}$). A low cosine similarity represents more serendipity, which makes sense, as the serendipitous item should not be similar to the other compared items (Silveira et al., 2019a). To a straighter forward comparison to the ground-truth dataset, which represents serendipity as 0 or 1, Eq. 7 shows the serendipity formula utilized.

$$S = 1 - \frac{1}{n}\sum_{j=1}^{n} \cos(u_j, i) \tag{7}$$

Where $\cos(u_j, i)$ is the cosine similarity between the embedding of the idem u_j from the user's history and the embedding of the item i, and n is the number of items in the user's history.

4.1.1 Embeddings Generation Models

Different embeddings generation models will obviously generate different embeddings, and thus, the choice of the model can influence the serendipity and the recommendation system results. Because the description of each book is its representation, it is important to capture the semantic meaning of the content. To fulfill this objective and have solid trustworthy results, four different embedding generation models were selected, and the serendipity for each user-item interaction were calculated using each of them separately. For the first three models the Sentence Transformer Library was used. The following models were selected:

4.1.1.1 BERT

Bidirectional Encoder Representations from Transformers (BERT) is a groundbreaking language model presented by Devlin et al. (2018). BERT's innovation is its bidirectional training approach, which allows the model to consider the context from both left and right simultaneously. This bidirectional understanding enables BERT to capture deeper semantic and syntactic nuances in text compared to unidirectional models. Although currently deprecated in the sentence transformer package, this model was considered due to its still relevant presence in academic research.

4.1.1.2 DistilRoBERTa

DistilRoBERTa is a distilled version of the RoBERTa model, which is an optimized variant of BERT developed by (Liu et al., 2019). Distillation is a model compression technique that reduces the size of a large model while attempting to maintain its performance levels (Hinton et al., 2015). DistilRoBERTa achieves this by training on the outputs of the larger RoBERTa model, resulting in a model that is lighter and faster. In practical applications, DistilRoBERTa offers a balance between computational efficiency and performance. Because of its relation to the Bert model, and because it has one of the best classifications among the models provided by sentence transformers, this model was included in this study.

4.1.1.3 MPNET

Masked and Permuted Pre-training for Language Understanding (MPNET) is a language model that combines strengths of two pre-training objectives: masked language modeling, as used in BERT, and permuted language modeling. By integrating these approaches, MPNet addresses the limitations of each method, improving the model's ability to understand the context and dependencies within text. When generating embeddings, MPNet produces highly contextualized representations that capture both local and global semantic information (Song et al., 2020). This model was included in this study for having the best classification among the models provided by sentence transformers.

4.1.1.4 BM25

Okapi BM25, is a ranking function used in information retrieval systems, designed to estimate the relevance of documents to a given search query. BM25 achieves this by utilizing adjusted values of term frequency (TF) and inverse document frequency (IDF), along with a consideration for document length, to compute a relevance score between a document and a query. BM25 can be used to represent textual content effectively by capturing essential terms that contribute significantly to the meaning of the documents (Robertson & Walker, 1994).

Zhu et al. (2021) have demonstrated the efficacy of BM25 in recommendation systems. In their scholarly recommendation system, BM25 outperformed several advanced text representation techniques, including models like BERT. This suggests that traditional methods like BM25 remain competitive, and that is why it is included in this study.

4.1.2 Results

The results of the serendipity comparison analysis are illustrated in Figures 4 and 5. Figure 4 displays the ground-truth serendipitous interactions and the corresponding calculated serendipitous values of each model. Figure 5 displays the non-serendipitous interactions and the corresponding calculated serendipitous values of each model.

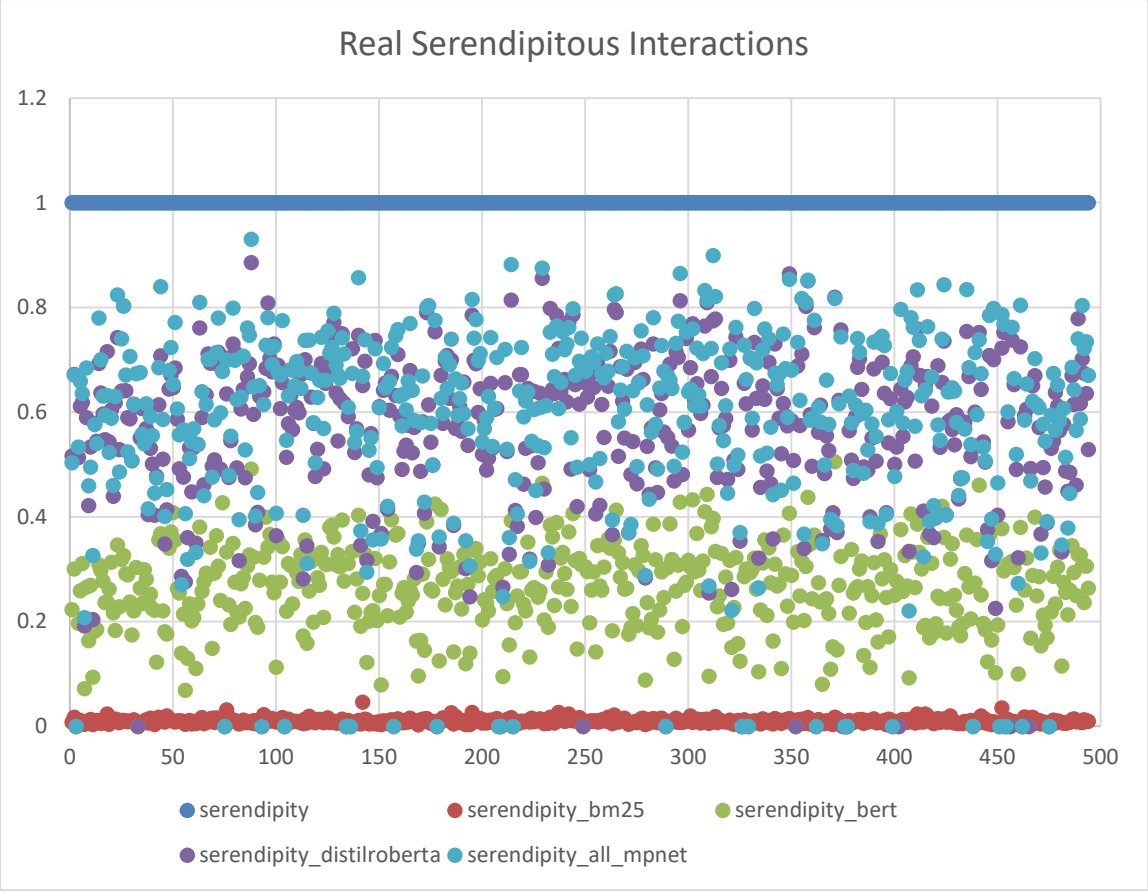

Figure 4 – Ground-truth vs calculated serendipity - serendipitous interactions

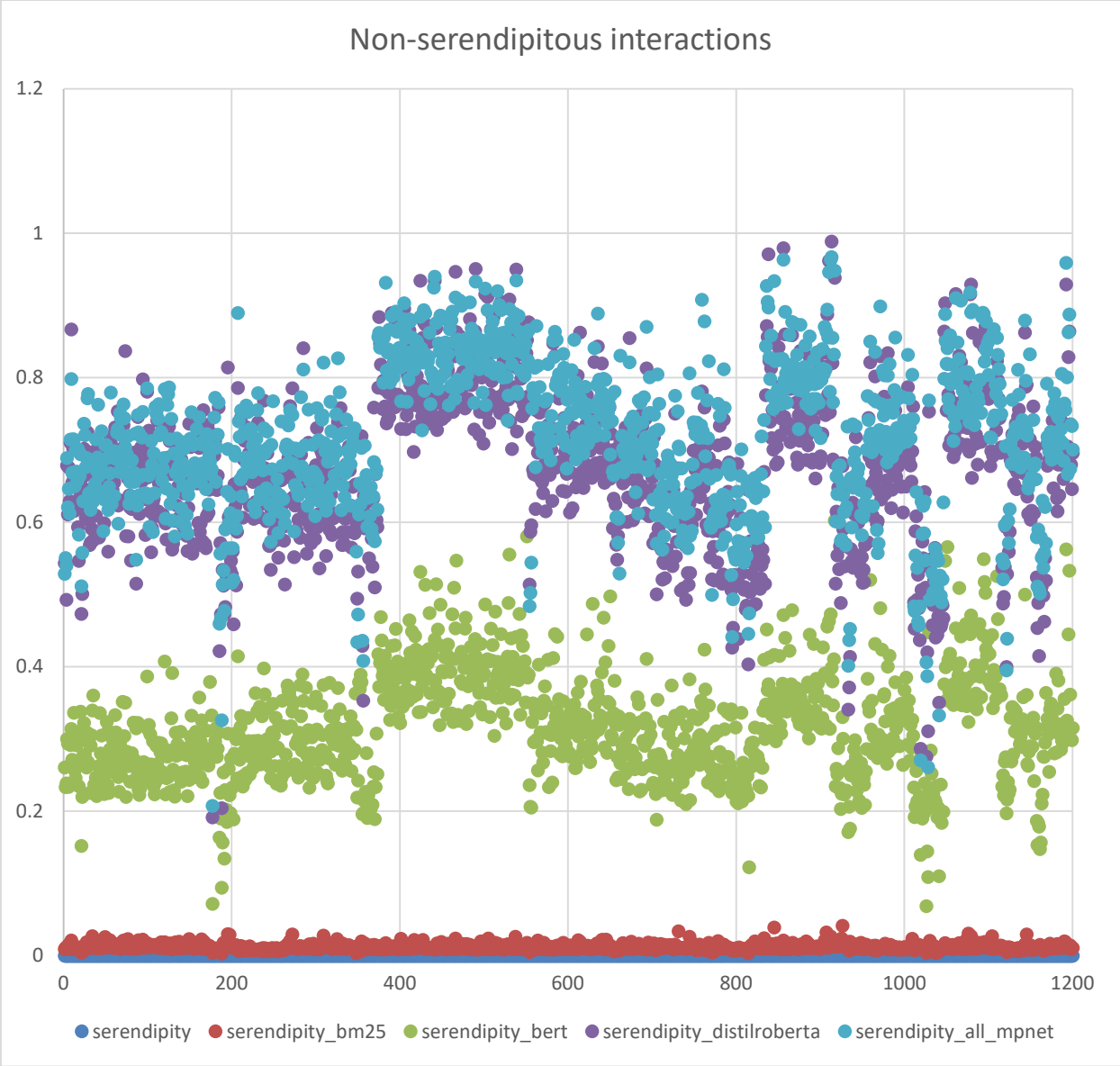

Figure 5 – Ground-truth vs calculated serendipity – non-serendipitous interactions

From the graphics it can be observed that the model which better captures embedding differences is MPNET. Its serendipity values are more widely dispersed in the graphic, indicating that the different contexts are being better separated. DistilRoberta shows similar results, consistently following a similar pattern, but with slightly lower values. Bert also follows a comparable pattern, but with even lower values, around half of DistilRoberta.

In contrast, BM25 shows almost no variation in the serendipity values. This could be explained because BM25 uses sparse, high-dimensional vectors based on term frequency. When calculating cosine similarity between these vectors, many items appear similar because they share common terms, leading to consistently high similarity scores. This makes it hard for BM25 to distinguish between different items. Additionally, it is worth mentioning that BM25 required

significantly more processing time because working with these large, sparse vectors is computationally intensive.

No clear correlation between real serendipity and calculated serendipity was observed in this experiment for any of the models. Both graphs—representing real serendipitous and non-serendipitous interactions—displayed random patterns. This outcome reinforces what has been expressed in previous research: serendipity is a personal and subjective experience, making it very difficult to quantify and calculate accurately.

Additionally, it should be highlighted that the similarity calculations could be performed using different formulas. And that utilizing different features of the items or different items representation forms might generate embeddings that better or worse capture the dissimilarities between them. But this experiment indicates that using direct similarity measures or other type of formulas alone, may not be sufficient for calculating serendipity. It may be necessary to involving more complex models to capture the nuanced nature of serendipity more effectively. Further validations and experiments could explore these possibilities.

4.2 Proposed model

Two different models were developed and then combined into a final model. The Serendipity classifier model is used to predict if a user-item combination is serendipitous or not. The Serendipity recommender system is used to predict a user-item combination rating, favoring serendipitous items. Finally, the serendipity classifier model is integrated into the Serendipity recommender system, predicting each user-item serendipity possibility and using its prediction as part of the input in the Serendipity RS. The eXtreme Gradient Boosting - XGboost algorithm was chosen to implement all models.

4.2.1 XGBoost

XGBoost is a powerful machine learning algorithm that is widely used for supervised learning tasks. It is an implementation of gradient boosted decision trees designed for speed and performance. It uses an ensemble approach, where it builds multiple decision trees in sequence, each correcting the errors of its predecessor, to make predictions. This boosting technique helps the model achieve high accuracy by optimizing a regularized objective function, which balances model complexity and training loss, thus preventing overfitting. XGBoost also supports sparse data handling, making it well-suited for datasets with missing values or high-dimensional sparse features. Due to these features, XGBoost has become a go-to tool for data scientists and has been widely adopted in numerous machine learning competitions and real-world applications for tasks such as classification, regression, and ranking. (T. Chen & Guestrin, 2016).

Bhavana et al. (2023) used XGBoost and a hybrid model combining XGBoost and Random Forest for developing a recommender system that leverage sentiment analysis, classifying sentiment on product reviews. Both models score more than 0.9 in the accuracy, precision, recall and F1 score metrics.

Shahbazi et al. (2020) developed a content-based recommender system based in the users click information and compared the XGBoost classifier with four other machine learning algorithms, namely Random Forest, SCM, KNN and Logistic regression. XGBoost had the best performance among them all resulting in 89.6% accuracy.

For the Serendipity classifier, the XGBoost classifier algorithm is applied. It receives as input user and item embeddings and aims to classify the combination into serendipitous or not.

The Recommender Systems uses the XGBoost Regression model. It receives as input the same user and item embedding and aims to predict the user-item rating. Three variations of this model were developed. 1- Simple content-based recommender system, do not consider serendipity as a differential, 2-The model uses the serendipity information from the dataset as part of the input and emphasizes the serendipitous combination. This model's purpose is only to serve as benchmark to the third model, once it uses real serendipity information in train and test data. 3-The third model is the second model modified to use the Serendipity classifier predictions as the source of the serendipity information, for train and test purposes, instead of the dataset. They will be identified as XGBoost, XGBoost-Seren and XGBoost-Seren + classifier.

4.2.1.1 Embeddings

To obtain the user and item embeddings, the technique of Truncated Singular Value Decomposition (Truncated SVD) was applied to the interaction matrix between users and items. This matrix factorization technique maps both users and items into a shared latent factor space where user-item interactions are modelled as inner products within this space. The latent factors aim to explain the observed ratings by representing both products and users through features automatically inferred from user feedback (Anitha et al., 2013).

The item embeddings originated from SVD were then combined with the item embeddings generated from the items title and description by the different text processing models stated in the previous section.

The dataset was divided, 20% test and 80% training. To avoid data leakage, the same training and test data were used in the serendipity classifier and all Serendipity RC models.

Finally, the next four figures illustrate the developed models.

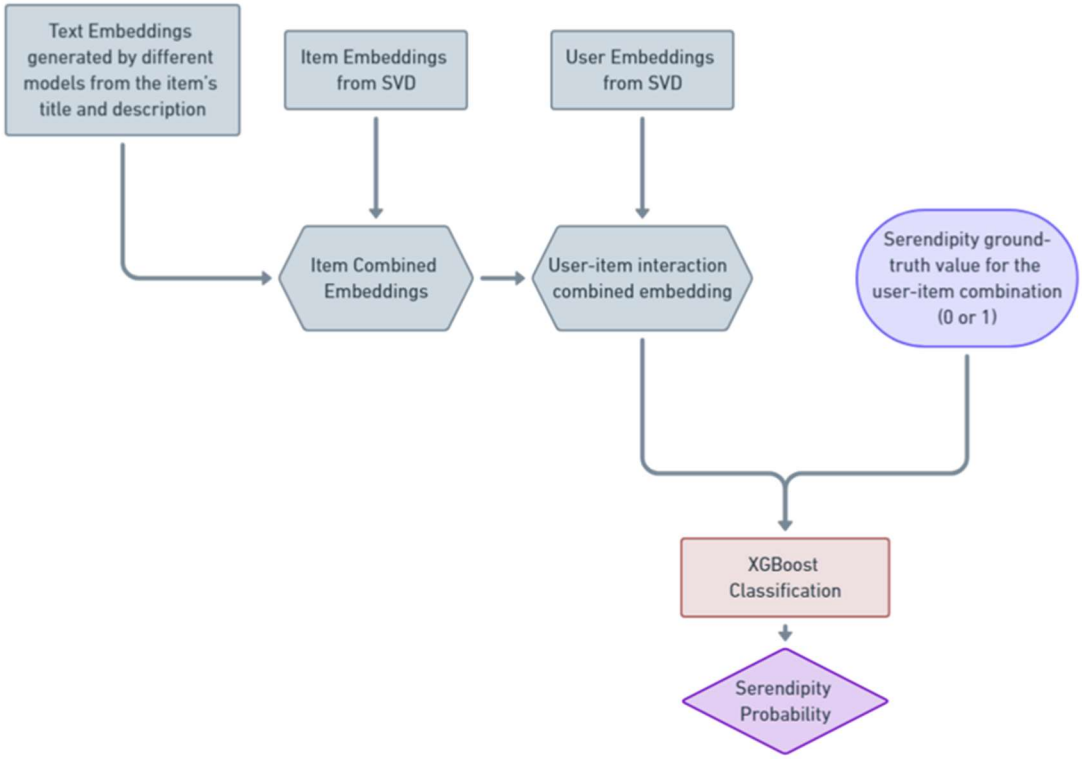

Figure 6 – Serendipity Classifier illustration

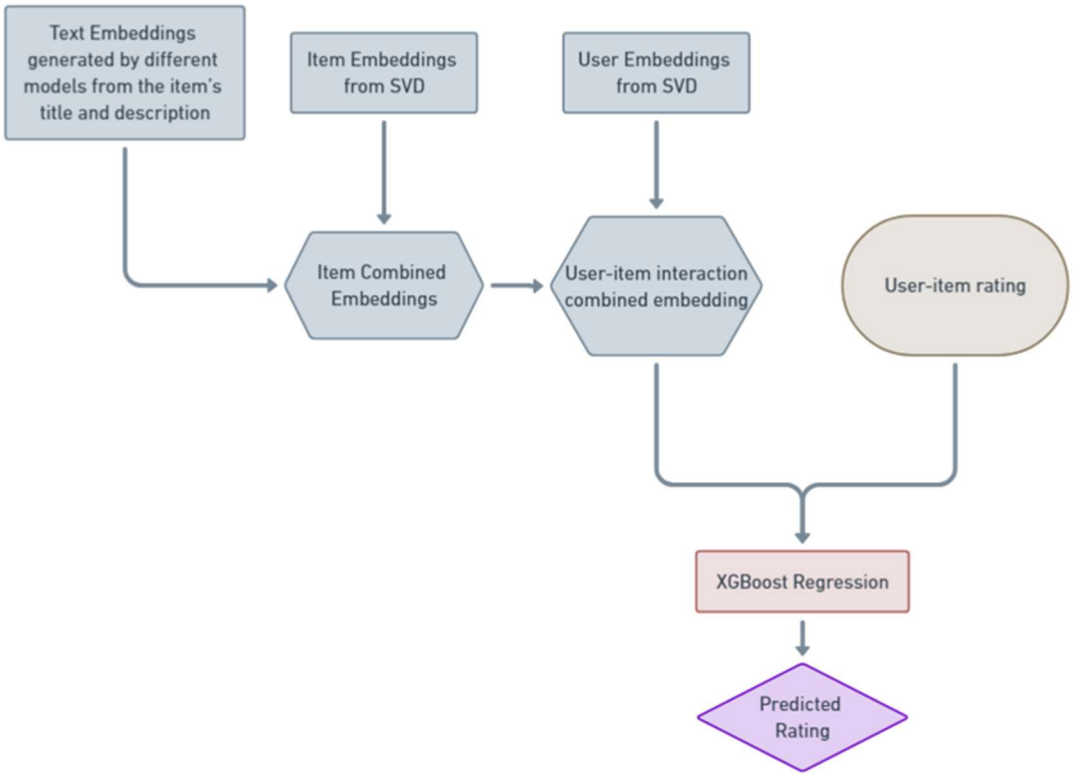

Figure 7 – XGBoost recommender model illustration

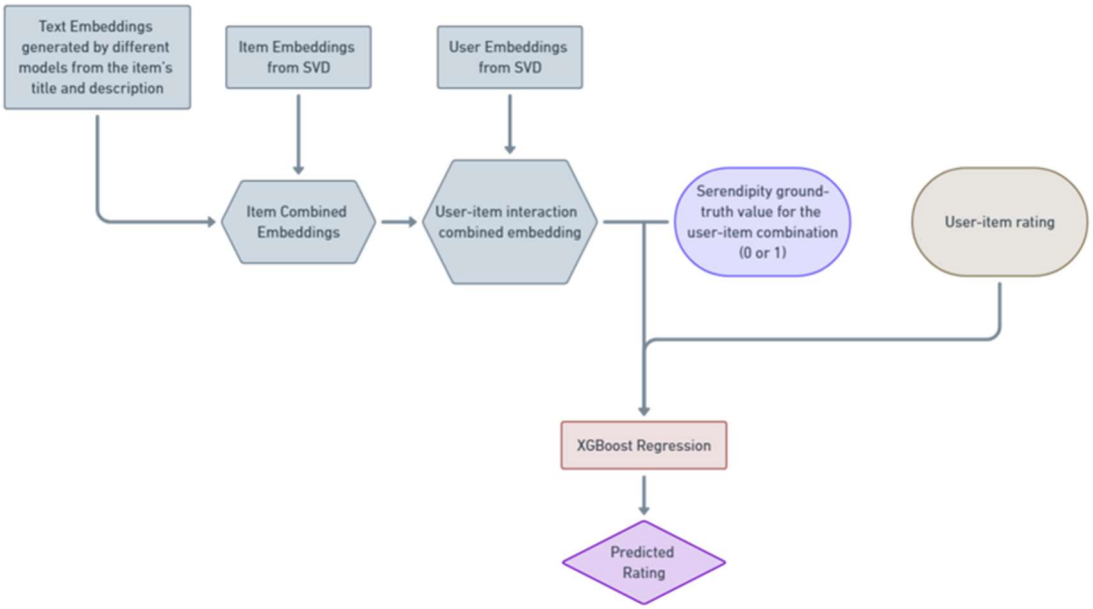

Figure 8 – XGBoost-seren recommender model illustration

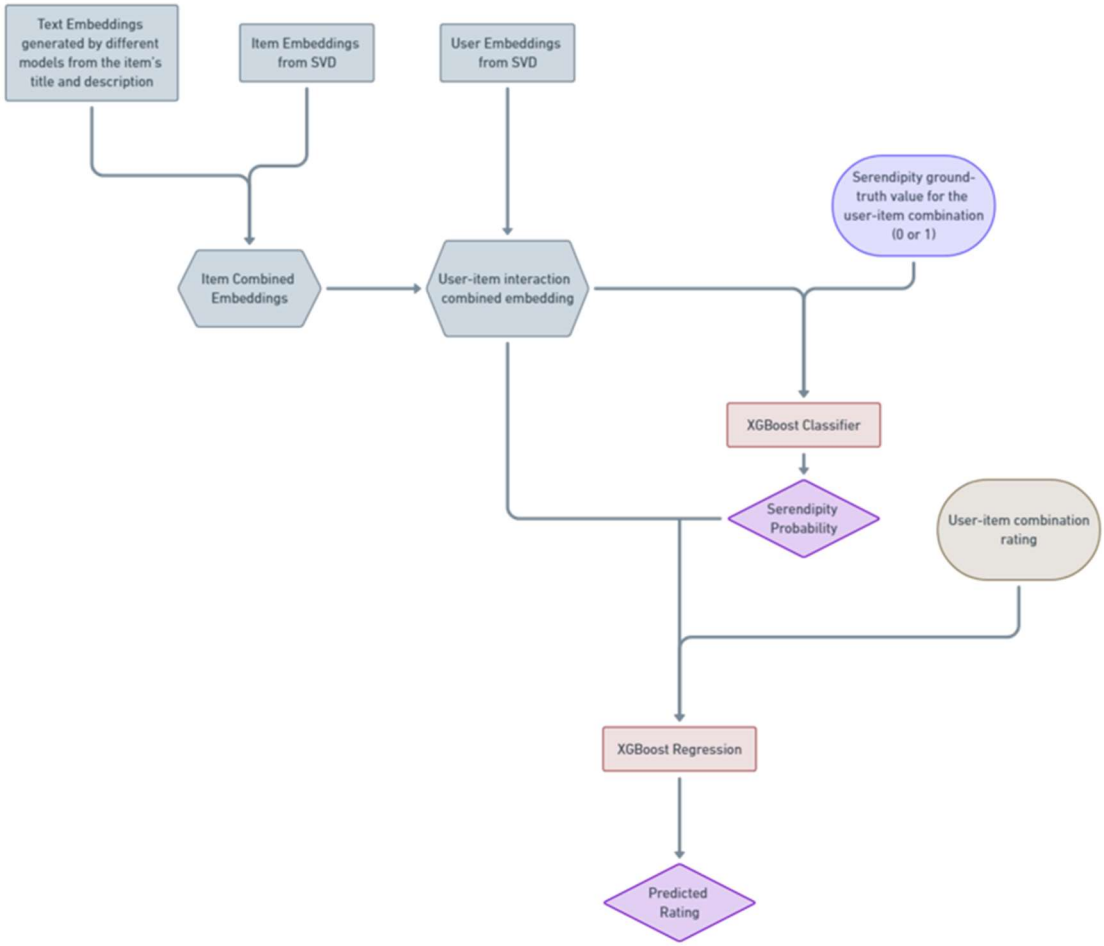

Figure 9 – XGBoost-seren + classifier recommender model illustration

4.2.2 Hyperparameters tuning

The optuna (Akiba et al., 2019) package was used to automate the hyperparameter optimization. A metric needs to be defined, so the package will look for the best hyperparameters to improve that metric.

For the serendipity classifier, F1-score was chosen as the metric to be improved. For the XGBoost models, mean square error (MSE) was chosen as the metric to be improved.

The parameters that were passed as intervals for optuna to find the optimized hyperparameters, can be seen in table 5.

Table 5 – Hyperparameters interval used in optuna

Hyperparameter	**Value interval**
Number of estimators	100 - 500
Max depth	3 - 10
Learning rate	0.01 - 0.3
subsample	0.6 - 1.0
Col sample by tree	0.6 - 1.0
alpha	0.1 - 10
lambda	0.1 - 10

4.2.3 Metrics

For the evaluation of the Serendipity classifier precision, recall and F1-score were considered. For evaluating and comparison between the RC models, MSE and MAE were used. To compare the models' performance to each other, and to SASRec and to evaluate serendipity, Hit Rate (HR) and Normalized Discounted Cumulative Gain (NDGC) at K was used. HR measures if the correct item is recommended among a list of not interacted items of k length, and NDGC measures the quality of this recommendation, considering the position of the correct item in the k-length recommendation list (Ma et al., 2024). See the formulas below.

$$HR@K = \frac{\sum_{i=1}^{n} hit_i}{N} \tag{8}$$

Where N is to total number of instances and hit_i is 1 if the relevant item appears within the top-k position for the i-th instance, and 0 otherwise.

$$DGC@K = \sum_{i=1}^{K} \frac{rel_i}{log_2(i+1)} \tag{9}$$

Where rel_i, is the relevance of the item at position i.

$$IDGC@K = \sum_{i=1}^{\min(K, number\ of\ relevant\ items)} \frac{1}{log_2(i+1)} \quad (10)$$

$$NDCG@K = \frac{DCG@K}{IDCG@K} \quad (11)$$

The SASRec model is meant to predict the next item interaction in a sequence, in this way HR and NDGC are only calculated for one single item in the test data of each user, which means that N in HR is the total number of users, and the number of relevant items in IDGC will always be one. Following the same approach, for the proposed models, it is considered a relevant item if its rating is above 3. All relevant items are separately classified along with a list of not interacted items, which means that N in this case, is the total number of relevant items in the test dataset.

Regarding serendipity, the two metrics were adapted to consider only serendipitous items. That means that this metric is only calculated when a serendipitous item is being predicted as the relevant item. In this case, for HR_{seren} and $NDGC_{seren}$, N will be the number of serendipitous items in the test data. The number of negative examples was set to 20 in all calculations.

4.2.4 Results

All four different models – Serendipity classifier, XGBoost, XGBoost-Seren and XGBoost-Seren + classifier were run for each of the four different text embeddings generation model – Bert, DistilRoBerta, MPPNET and BM25. The results are displayed in this section.

4.2.4.1 Serendipity Classifier

The serendipity classifier faces a very difficult problem to solve, that is to classify imbalanced classes. In the dataset utilized, class 1, that represents the serendipitous items is around 4% of the entries. What make this problem complex, is that in this case, it is not a matter of an imbalanced dataset, because that could be fixed with balancing techniques. It is really a matter of an imbalanced problem, as in the real-world, serendipitous item will be a real minority comparing to the vast number of recommendations. That makes very hard to achieve precision for the minority class.

The idea is to favour serendipitous items recommendations, so the main goal would be to correctly classify the items that are indeed serendipitous. Since recall measures the proportion of relevant items retrieved and considering that precision is very difficult to achieve in a disbalanced model, the recall metric was chosen to evaluate the model's performance.

Table 6 shows the results for the serendipity classifier for each embedding model. Although no significant different can be seen in the results, the DistilRoberta model presented the best results.

Table 6 – Serendipity classifier recall results

Serendipity Classifier	**Recall classe 0**	**Recall classe 1**
MPNET	0.58	0.64
BM25	0.58	0.58
Bert	0.61	0.61
DistilRoberta	**0.61**	**0.65**

Figure 10 and 11 shows the values obtained by the DistilRoberta serendipity classifier for the non-serendipitous interactions and real serendipity interactions.

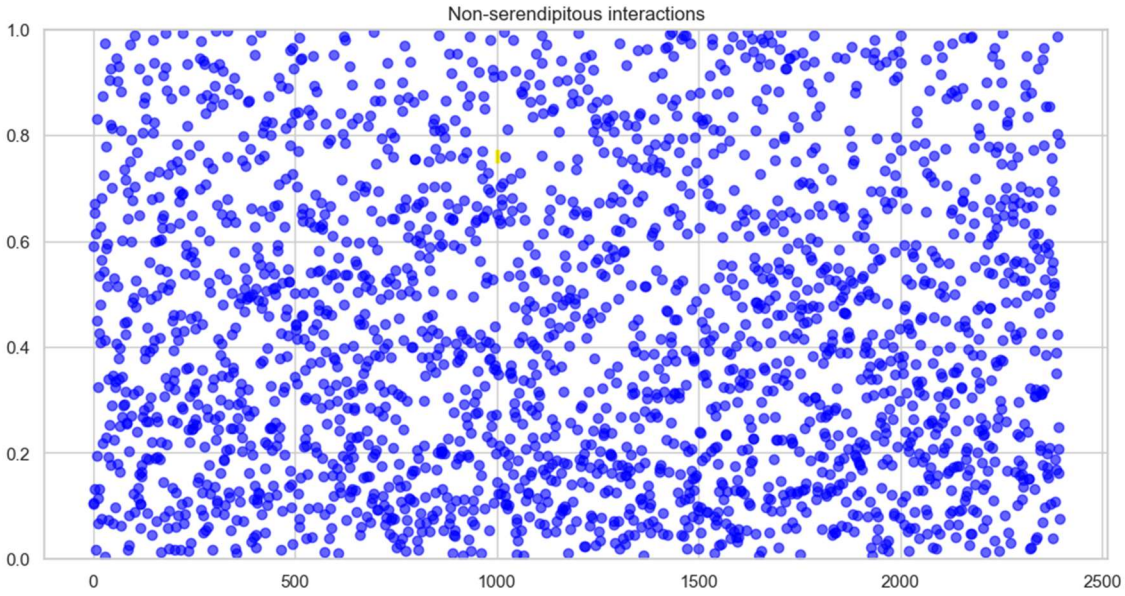

Figure 10 – DistilRoberta serendipity classifier non-serendipity interactions results

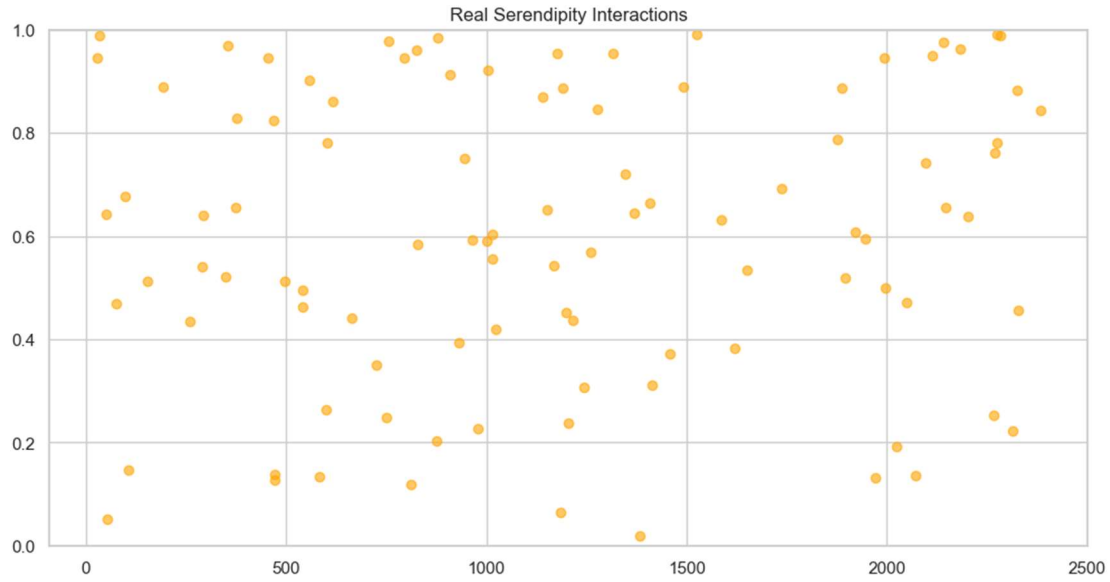

Figure 11 – DistilRoberta serendipity classifier real serendipity interactions results

Although a lot of the values are misclassified, there is a slightly bigger concentration of values under 0.5 for the non-serendipitous interactions and above 0.5 for the real serendipitous interactions. It can also be observed that, comparing to the calculated serendipity analysis, the results are sparser and more diverse, indicating that the classifier model captures more varied nuances and complexness for each item.

Table 7 displays the final hyperparameters selected by optuna to optimize the F1-score metric for the DistilRoBERTa Model in the serendipity classifier.

Table 7 – Final hyperparameter for the DistilRoberta serendipity classifier

Hyperparameter	**Serendipity Classifier DistilRoBERTa**
Number of estimators	434
Max depth	13
Learning rate	0.05
subsample	0.80
Col sample by tree	0.65

The results show that for the classification tasks alone, there were no significant differences between the different embeddings model. This indicates that, despite the differences in how these models generate embeddings, they all capture enough relevant information from the book descriptions to perform similarly in the classification task. It appears that the classification model is robust to the variations in embedding techniques, possibly because the task relies on high-level semantic features that are effectively captured by all models. Even though BM25

operates differently by focusing on term frequency, it still encodes sufficient information to distinguish between classes. This suggests that for this specific classification task, the choice of embedding model may be less critical, and traditional models like BM25 can perform comparably to more advanced neural embedding models.

The recall metric around 0.64 means that two thirds of the relevant items are being correctly classified. That is a very good indicator for the minority class, but this being an imbalanced problem, it also means that a third of the non-serendipitous items are being misclassified. Although the results can be improved, this classifier presents very reasonable results for an imbalanced problem. Considering the model is meant to be integrated into a recommender system model, and not to use the classification as a recommendation directly but to favor the items, this indicates that the model could be a valuable contribution to increase serendipitous recommendations.

4.2.4.2 XGBoost

The XGBoost model does not take into consideration the items serendipity at all, being only a content-based recommender system that uses items description as item representation. The serendipity metrics in this case are here for comparison purposes – with the total number of recommendations and with the other models developed.

Table 8 shows the accuracy metrics and table 9 shows the serendipity metrics results.

Table 8 – XBGoost accuracy metrics results

XGBoost	$HR@5$	$HR@10$	$NDGC@5$	$NDGC@10$
MPNET	0.21	0.34	0.125	0.167
BM25	0.21	0.34	0.130	0.173
Bert	0.22	0.35	0.127	0.169
DistilRoberta	**0.25**	**0.37**	**0.149**	**0.186**

Table 9 – XBGoost serendipity metrics results

XGBoost	$HR_{seren}@5$	$HR_{seren}@10$	$NDGC_{seren}@5$	$NDGC_{seren}@10$
MPNET	0.17	0.32	0.101	0.154
BM25	0.22	0.32	**0.159**	**0.194**
Bert	**0.26**	**0.36**	0.148	0.174
DistilRoberta	0.21	0.34	0.118	0.164

Once again, the results across the different embedding models are not significantly different but do show small variations. For the accuracy metrics, DistilRoberta outperformed the other models across all metrics, while MPNET presented the lowest results.

In the serendipity results, however, the BERT model achieved the best scores for the HR@k indices. And although BM25 had lower HR@k results compared to BERT, it showed better NDCG@k indices, this implies that the serendipitous items recommended by BM25 are ranked higher in terms of relevance and highlights that traditional models still hold value. The MPNET model again presented the lowest results.

Comparing the accuracy results with the serendipity results, we can see that the outcomes are quite similar. This indicates that, in general, the model neither favors nor disfavors serendipitous items. This suggests an opportunity for enhancement. Incorporating algorithms specifically designed to promote serendipity could improve the serendipity metrics and lead to a more engaging user experience. And that will the next step developed in the XGBoost-seren model.

4.2.4.3 XGBoost-seren

In the XGBoost-Seren model, the ground-truth serendipity of each user-item interaction was added to the XGBoost model as an item feature. Additionally, for serendipitous items, the item embeddings were multiplied by 2 to highlight and favor these item recommendations. This method amplifies the representation of serendipitous items in the feature space, increasing their influence during the training process. By enhancing the embeddings of serendipitous items, the model is more likely to prioritize them in its recommendations.

The results of this approach are presented in the following tables.

Table 10 – XBGoost-seren accuracy metrics results

XGBoost-Seren	$HR@5$	$HR@10$	$NDGC@5$	$NDGC@10$
MPNET	0.25	0.38	0.150	0.186
BM25	**0.28**	**0.39**	**0.176**	**0.210**
Bert	0.26	0.39	0.155	0.196
DistilRoberta	0.26	0.38	0.156	0.195

Table 11 – XBGoost-seren serendipity metrics results

XGBoost-Seren	HR_{seren}@5	HR_{seren}@10	$NDGC_{seren}$@5	$NDGC_{seren}$@10
MPNET	0.61	0.70	0.425	0.455
BM25	**0.97**	**0.98**	**0.806**	**0.806**
Bert	0.77	0.86	0.568	0.593
DistilRoberta	0.62	0.75	0.478	0.502

In the accuracy metrics, no significant differences are observed between the models. However, considering all metrics, BM25 demonstrated the best performance across all models, while MPNET had the lowest performance.

In the serendipity metrics, BM25 also consistently outperformed all other models, this time with significantly higher values. This can be explained because BM25 emphasizes term frequency and inverse document frequency, it captures the uniqueness of item descriptions by assigning higher weights to less common terms. When the embeddings of serendipitous items are multiplied, BM25's focus on distinctive terms becomes even more pronounced, leading to a greater differentiation of serendipitous items in the feature space. This results in the model more effectively identifying and recommending serendipitous items.

Comparing the serendipity results with the original XGBoost model, it can be concluded that introducing serendipity into the model and favouring items based on it was a successful experiment. All models outperformed their previous serendipity results when compared to the standard XGBoost model, with significant differences, and BM25 showing the most notable improvement.

Comparing the accuracy results with those of the original XGBoost model, the XGBoost-Seren model consistently achieved higher results. This can be explained by the great improvement in the serendipity metrics, which elevated the overall performance metrics.

These results demonstrate that, with reliable data on the serendipity of items, a model can favour the recommendation of these items, increasing user satisfaction while also enhancing overall accuracy metrics. This validates the thesis that incorporating serendipitous items into recommendation models brings substantial benefits. However, since in the real world there is no prior information about the serendipity of items, the next step is to incorporate the serendipity classifier into the model, replacing the previously used ground-truth serendipity data.

4.2.4.4 XGBoost-seren + Classifier

In the XGBoost-seren + Classifier model, all ground-truth serendipity data were replaced by the predict serendipity resulted from the serendipity classifier model. The serendipity classifier models predict the probability of an item being serendipitous, resulting into a number between 0 and 1. For each user-item interaction, the output number (between 0 and 1) of the serendipity

classifier was added as an item feature, and the item embedding were multiplied by 1 plus the result. This follows the same XGBoost-seren approach, but instead of having serendipity as a 0 or 1 value, in this case the value will be between 0 and 1. This decision was made to give the model more flexibility since the results of the serendipity classifier does not present great accuracy due to its imbalanced nature.

The results are presented in the following tables:

Table 12 – XBGoost-seren + classifier accuracy metrics results

XGBoost-seren + classifier	$HR@5$	$HR@10$	$NDGC@5$	$NDGC@10$
MPNET	0.23	0.35	0.138	0.177
BM25	**0.27**	**0.39**	**0.161**	**0.198**
Bert	0.23	0.36	0.135	0.176
DistilRoberta	0.25	0.38	0.150	0.189

Table 13 – XBGoost-seren + classifier serendipity metrics results

XGBoost-seren + classifier	HR_{seren} @5	$HR_{seren}@10$	$NDGC_{seren}@5$	$NDGC_{seren}@10$
MPNET	0.27	0.37	0.189	0.209
BM25	0.28	0.39	0.173	0.202
Bert	0.26	0.36	0.167	0.200
DistilRoberta	**0.38**	**0.43**	**0.221**	**0.241**

For the accuracy metrics, BM25 slightly outperformed the other models in all metrics. With MPNET and BERT sharing the lowest results.

For the serendipity metrics, DistilRoberta displayed the best results among all models.

Comparing the serendipity results with the XGBoost-Seren model, it's evident that the XGBoost-Seren + Classifier model has room for improvement. This was somewhat expected and can be easily justified by the moderate performance of the serendipity classifier. The XGBoost-Seren + Classifier serves as a benchmark model and provides an idea of what could be accomplished if serendipity were an easily identifiable concept. The key aspect of the analysis is to compare the XGBoost-Seren + Classifier results with the original XGBoost results.

The MPNET model showed consistently lower results compared to the other models in all the experiments conducted. That is, in a way, surprising, as this model was chosen because it had the best classification of the models provided by sentence transformers. However, when comparing the serendipity results with the XGBoost results, there was a significant increase in the serendipity metrics, even though this improvement wasn't strongly reflected in the accuracy

metrics, which remained similar. This proves that while MPNET does not stand out among the models, incorporating the serendipity classifier did enhance its ability to recommend serendipitous items.

The BM25 model showed increases in all metrics – both accuracy and serendipity- compared to the original XGBoost model. As previously observed, BM25's results stood out in the XGBoost-Seren model, demonstrating the model's capacity to highlight certain items. However, the values achieved by the XGBoost-Seren + Classifier, although higher than those of the original XGBoost, are significantly lower than the potential BM25 showed in the XGBoost-Seren model. This is a clear indicator that while BM25 has great potential to prioritize recommendations, it may not be as effective in identifying the serendipity of items when relying on the classifier's predictions.

The BERT model showed basically the same results in both the XGBoost-Seren + Classifier and the original XGBoost models. This indicates that BERT, in general, was less effective at classifying serendipity and leveraging it to enhance recommendations. There was no significant difference between incorporating the serendipity classifier or not. This could be due to BERT's embeddings not capturing the nuances necessary for the serendipity classifier to make impactful predictions. Additionally, the moderate performance of the serendipity classifier may not have been sufficient to improve BERT's recommendations.

Finally, the DistilRoberta model presented the best serendipity results among all models for the XGBoost-Seren + Classifier, also showing the greatest increase in serendipity metrics compared to the original XGBoost of all models. Despite this, the accuracy metrics remained similar when compared to the XGBoost results. This suggests that, overall, the DistilRoberta model was the most effective at classifying the serendipity of items and using this information to enhance recommendations. It successfully increased the recommendation of serendipitous items while maintaining the system's overall accuracy.

Finally, tables 14 and 15 present the comparison of the proposed models with the SASREC model. The results from the DistilRoberta embedding model are used for this comparison, as it demonstrated the best performance in the final experiments.

Table 14 – SASREC comparison accuracy metrics results

Model	$HR@5$	$HR@10$	$NDGC@5$	$NDGC@10$
SASREC	0.20	0.34	0.137	0.181
XGBoost	0.25	0.37	0.149	0.186
XGBoost-Seren	0.26	0.38	0.156	0.195
XGBoost-Seren + classifier	0.25	0.38	0.150	0.189

Table 15 – SASREC comparison serendipity metrics results

Model	HR_{seren}@5	HR_{seren}@10	$NDGC_{seren}$@5	$NDGC_{seren}$@10
SASREC	0.14	0.34	0.094	0.160
XGBoost	0.21	0.34	0.118	0.164
XGBoost-Seren	0.62	0.75	0.478	0.502
XGBoost-Seren + classifier	0.38	0.43	0.221	0.241

These tables clearly show that the final model proposed XGBoost-Seren + classifier outperform SASREC in both accuracy and serendipity metrics, indicating the study achieved better overall recommendation performance compared to the SASREC model.

When comparing the accuracy metrics, the proposed models showed improvements over SASRec. The XGBoost model increased HR@5 from 0.20 to 0.25 and NDCG@5 from 0.137 to 0.149 relative to SASRec. The XGBoost-Seren and XGBoost-Seren + Classifier models offered similar results with accuracy being very consistent across the models. This demonstrates the slight superiority of the XGBoost models over SASRec in terms of accuracy.

However, the most significant difference appears when comparing serendipity metrics. The XGBoost model already displayed superior results in HRseren@5 compared to SASRec. These results were further enhanced by the final XGBoost-Seren + Classifier model, increasing HRseren@5 from 0.14 in SASRec to 0.38. While the XGBoost-Seren model achieved even higher serendipity metrics, it's not a fair comparison to SASRec since it uses ground-truth serendipity data as a benchmark reference.

Overall, the proposed model demonstrated notable improvements over the traditional SASRec model. This validates the approach of incorporating a serendipity classifier into recommendation systems, suggesting that this method has practical potential for real-world applications.

4.2.4.5 Cold Start

In preparing the dataset, the train and test data were divided considering a proportional balance regarding serendipitous labels. Because of this, there were users and items in the test data, that were not part of the train data, leading to cold start scenarios where the model has no prior interaction information for these users or items.

To handle cold start users, a default user embedding was defined as the mean of the user embeddings generated by the SVD matrix factorization. For cold start items, their embeddings were created by combining a default item embedding—also the mean of the item embeddings from the SVD—and the item's description embedding.

Utilizing an independent item characteristic like the item description offers a significant advantage for cold items. Since descriptions are inherent to the items and do not rely on previous user interactions, they can be used to generate meaningful embeddings even when

the item is new to the system. This approach also benefits the serendipity classification, as it considers user-item interactions and leverages item descriptions to enhance the quality of recommendations for both cold users and items.

Table 16 and 17 presents the metrics for cold users' recommendations using the DistilRoberta model.

Table 16 – XBGoost-seren + classifier cold users accuracy metrics results

Model	$HR@5$	$HR@10$	$NDGC@5$	$NDGC@10$
XGBoost-Seren + classifier	0.125	0.22	0.071	0.099

Table 17 – XBGoost-seren + classifier cold users serendipity metrics results

Model	HR_{seren} @5	$HR_{seren}@10$	$NDGC_{seren}@5$	$NDGC_{seren}@10$
XGBoost-Seren + classifier	0.10	0.15	0.050	0.059

Although the obtained results for cold users are lower than the overall results, they are comparable to the performance of models like SASRec in general scenarios. This indicates that the proposed approach mitigates the cold start problem by leveraging item descriptions.

By incorporating item descriptions into the embedding generation process, the model can provide reasonable recommendations even when there is no historical interaction data for certain users or items. Using item descriptions helps address the cold start problem because they provide intrinsic information about the items that are always available.

This approach demonstrates that relying on item-specific data can compensate for the lack of interaction history. This is particularly beneficial in real-world applications where new users or items frequently enter the system.

5 Conclusions

In this chapter, an overview of the study developed, and the objectives achieved are presented. Following the study results, the limitations encountered, and future work suggestions are presented.

5.1 Summary and Objectives Achieved

This study proposed the development of a recommendation system for e-commerce using machine learning and NLP techniques, with a focus on improving serendipitous recommendations. The approach aimed to explore an out-of-the-box idea by leveraging item descriptions to prioritize serendipitous recommendations, addressing a less-explored theme in the literature.

The central research question of this study is *How can a recommender system be designed to optimize user satisfaction in e-commerce by balancing accuracy with elements of novelty, serendipity, and diversity in its recommendations?*

Initially, a contextual theoretical introduction about recommendations system was presented, highlighting the most common RC approaches, evaluation methods and metrics, and challenges of the field. A systematic review was carried out to answer the research question.

The result of the review showed numerous ways to incorporate serendipity, novelty, and diversity into recommender systems. However, it also highlighted a lack of consensus on how to measure these metrics, with the most significant challenge being how to balance them with the accuracy of the systems. Based on this analysis, the study focused specifically on serendipity, as it naturally involves aspects of novelty and diversity. Also, the decision was made to utilize a ground truth dataset due to the lack of consensus in defining these metrics. The objective was to develop a system that prioritizes serendipitous recommendations while maintaining good performance and utilizing item descriptions. This chapter was crucial in providing an updated

overview of the state of the art and directing the study towards a theme that is underexplored in the literature.

With the theme defined, the methods for developing the proposed model were presented. This included a detailed explanation of the dataset used, the implementation and validation methods, and considerations regarding data protection, security, and ethics. Clearly defining and communicating the methods and materials is of extreme importance to ensure transparency and understanding of the work presented.

Two main contributions were made during the development. 1 - An analysis comparing ground-truth serendipity with calculated serendipity. 2 - The development of three different recommendation system models, along with a serendipity classification model.

The serendipity comparison analysis aimed to understand whether calculating serendipity through traditional formulas aligns with the real serendipity captured by the ground truth dataset. Four different embedding generation models were utilized: BERT, DistilRoberta, MPNET, and BM25. The use of multiple embeddings was decided because the embeddings generations have great influence in the models results. So, the results differences between these models could be explored. The results indicated that no discernible pattern could be established in this comparison. None of the embeddings produced plausible results when compared with the ground truth, and no significant connections could be drawn. This reinforces the notion that serendipity is a subjective metric for each user and is difficult to calculate using traditional methods. Often efforts in balancing accuracy and serendipity can compromise the RC performance, and being it based on improving a metric that doesn't really represent real user feedback, the efficiency of the approach can be questioned. It underscores the need for more complex calculation methods, such as machine learning models.

Following this reasoning, a serendipity classification model was developed using the XGBoost classification algorithm. The model used embeddings of items and users as input features to predict whether an interaction would be serendipitous. Given that only 4% of the dataset entries were serendipitous items, the problem was highly imbalanced, presenting a challenge for accurate classification. The model achieved a recall of approximately 0.65, which means that although many items are still incorrectly classified, there is reasonable performance considering the nature of the problem. This demonstrates the potential of the developed model to correctly classify serendipity and contribute to improving the performance of a recommendation system focused on these items.

Subsequently, three proposed models were developed using XGBoost regression model: XGBoost, XGBoost-Seren, and XGBoost-Seren + Classifier. The XGBoost model served as a baseline content-based recommender system focused on item descriptions, without considering serendipity. The XGBoost-Seren model incorporated serendipity as a feature and prioritized serendipitous items using ground-truth serendipity data, serving as a benchmark. The XGBoost-Seren + Classifier model replaced the ground-truth serendipity data with the serendipity predicted by the previously developed classifier.

The results demonstrated that XGBoost-Seren, being a benchmark model with access to ground-truth serendipity, achieved excellent performance in the serendipity metrics, displaying very high HR@k results, with special mention to the BM25 model, that proved to have a great capacity at favoring specific items. The excellent serendipity results were also reflected in the accuracy results, enhancing them as well. This proves that serendipitous items can be prioritized without adversely affecting overall accuracy.

XGBoost-Seren + Classifier model showed improved serendipity results compared to the baseline XGBoost model, without compromising accuracy metrics. Although it underperformed compared to the XGBoost-Seren benchmark model, this was an expected result considering the serendipity classifier performance. This confirms that despite room for improvement in the serendipity classifier, the XGBoost-Seren + Classifier model was able to capture aspects of serendipity and utilize them to enhance recommendations. There were some variations in the results across the different embeddings generation models. The DistilRoberta model had the overall best performance, and the MPNET had the lowest performance.

All developed models had better performance than SASRec, both in accuracy and serendipity metrics. These findings indicate that the approach of using item descriptions and machine learning models to predict serendipity and prioritize serendipitous recommendations is effective.

It is important to note that it was not possible to fully apply the Design Science Research (DSR) methodology due to time constraints. There was no opportunity to iterate or revisit previous steps, nor was there sufficient time for interactive solution refinement.

In conclusion, the study successfully explored the use of item descriptions to prioritize serendipitous recommendations in e-commerce. By focusing on an innovative approach and addressing a less-explored area in the literature, the research contributes valuable insights into the development of recommendation systems that balance accuracy with serendipity, enhancing the overall user experience.

5.2 Limitations and Future work

Despite the promising results obtained in this study, several limitations need to be acknowledged. A significant limitation was the data imbalance regarding serendipitous interactions in the dataset. Only about 4% of the items were labelled as serendipitous, creating a highly imbalanced classification problem. This imbalance poses challenges for machine learning models, particularly in accurately predicting the minority class. Traditional techniques for handling imbalanced datasets, such as under-sampling or over-sampling, were not suitable in this context, in real-world scenarios, most items a user interacts with are not serendipitous, making the imbalance a natural characteristic of the problem rather than an artifact of the data collection process.

Furthermore, a critical limitation is the scarcity of ground-truth datasets regarding serendipity. Throughout the literature, only two such datasets were identified. This scarcity underscores the need for further research and development in this area. Although the concept of serendipity may naturally represent a small proportion of items, there is currently no definitive way to accurately determine what that proportion really is due to the highly limited sources.

Another limitation was related to computational resources. High-dimensional embeddings and large-scale datasets require substantial computational power and memory, which were limited in this study. The machine processing capabilities available constrained the complexity of models and the size of datasets that could be handled efficiently. This restriction prevented the exploration of more complex models and extensive datasets as well as the inclusion of additional features that could potentially improve the performance of the serendipity classifier and the recommendation system.

Future research can address these limitations and build upon the findings of this study in several ways. Regarding ground-truth datasets, employing different data collection techniques than those used in SerenLens (Fu et al., 2023) could result in more complete and balanced datasets. Additionally, exploring serendipity in other domains beyond book reading—such as fashion shopping, music recommendation, or other e-commerce sectors—could provide insights into how serendipity manifests differently across contexts and how models can be adapted accordingly.

Enhancing the serendipity classifier's accuracy is another critical opportunity for future work. Improving the classifier could involve experimenting with advanced machine learning algorithms, such as deep learning models, which may capture more complex patterns in the data. Incorporating additional features derived from user behaviour, item attributes, or contextual information could also enhance the model's predictive power. Utilizing other dataset that could potentially contain richer information could also improve the performance.

Additionally, there is an intention to disseminate the findings of this research by writing and publishing an article in relevant academic journals of the recommendation system field. Sharing the results with the research community contributes with the field development and encourage further exploration in this area.

In summary, the research confirms that serendipity is a valuable component in recommendation systems and that models can be designed to promote serendipitous discovery effectively. Incorporating serendipity not only enriches the user experience but can also maintain or even enhance the overall performance of the recommendation system. Addressing the identified limitations—such as data imbalance and the lack of ground-truth datasets—and pursuing the forementioned future work can lead to more robust and effective recommendation systems that better addresses users’ needs and preferences.

References

Adomavicius, G., Mobasher, B., Ricci, F., & Tuzhilin, A. (2011). Context-Aware Recommender Systems. *AI Magazine*, *32*(3), 67–80. https://doi.org/10.1609/aimag.v32i3.2364

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). *Optuna: A Next-generation Hyperparameter Optimization Framework*. http://arxiv.org/abs/1907.10902

Alamdari, P. M., Navimipour, N. J., Hosseinzadeh, M., Safaei, A. A., & Darwesh, A. (2020). A Systematic Study on the Recommender Systems in the E-Commerce. *IEEE Access*, *8*, 115694–115716. https://doi.org/10.1109/ACCESS.2020.3002803

An, Y. H., Dong, Q., Yuan, Q., & Wang, C. (2020). Improving Recommendation Diversity by Highlighting the ExTrA Fabricated Experts. *IEEE Access*, *8*, 64422–64433. https://doi.org/10.1109/ACCESS.2020.2984365

Anitha, L., Devi, M. K. K., & Devi, P. A. (2013). A Review on Recommender System. In *International Journal of Computer Applications* (Vol. 82).

Au, Y. A. (2001). Design Science I: The Role of Design Science in Electronic Commerce Research. *Communications of the Association for Information Systems*, *7*. https://doi.org/10.17705/1cais.00701

Babenko, V., Kulczyk, Z., Perevosova, I., Syniavska, O., & Davydova, O. (2019). Factors of the development of international e-commerce under the conditions of globalization. *SHS Web of Conferences*, *65*, 04016. https://doi.org/10.1051/shsconf/20196504016

Berbague, C. E., Karabadji, N. E. I., & Seridi, H. (2018). Recommendation diversification using a weighted similarity measure in user based collaborative filtering. *Proceedings of the 2018 13th International Symposium on Programming and Systems, ISPS 2018*, 1–6. https://doi.org/10.1109/ISPS.2018.8379011

Bhavana, B., Karthik, J., & Kumari, P. L. (2023). A Novel Approach for Product Recommendation using XGBOOST. *IDCIoT 2023 - International Conference on Intelligent Data Communication Technologies and Internet of Things, Proceedings*, 256–261. https://doi.org/10.1109/IDCIoT56793.2023.10053453

Bobadilla, J., & Serradilla, F. (2009). *The Effect of Sparsity on Collaborative Filtering Metrics*.

Carvalho, Di., Silva, N., Trotta, T., Pereira, A. C. M., Mourao, F., & Rocha, L. (2019). A Particle Swarm approach to mitigate the apparent diversity-accuracy dilemma in recommendation domains in recommendation domains. *2019 IEEE Congress on Evolutionary Computation, CEC 2019 - Proceedings*, 2183–2190. https://doi.org/10.1109/CEC.2019.8790039

Chen, L., Yang, Y., Wang, N., Yang, K., & Yuan, Q. (2019). How Serendipity Improves User Satisfaction with Recommendations? A Large-Scale User Evaluation. *The World Wide Web Conference*, 240–250. https://doi.org/10.1145/3308558.3313469

Chen, M., Wang, Y., Xu, C., Le, Y., Sharma, M., Richardson, L., Wu, S. L., & Chi, E. (2021). Values of user exploration in recommender systems. *RecSys 2021 - 15th ACM Conference on Recommender Systems*, 85–95. https://doi.org/10.1145/3460231.3474236

Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. https://doi.org/10.1145/2939672.2939785

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*.

Fadillah, A. R., Nurma Yulita, I., Pradana, A., & Suryani, M. (2021). Data Mining Implementation Using Frequent Pattern Growth on Transaction Data for Determining Cross-selling and Up-selling (Case Study: Cascara Coffee). *2021 International Conference on Artificial Intelligence and Big Data Analytics, ICAIBDA 2021*, 272–277. https://doi.org/10.1109/ICAIBDA53487.2021.9689752

Fayyaz, Z., Ebrahimian, M., Nawara, D., Ibrahim, A., & Kashef, R. (2020). Recommendation systems: Algorithms, challenges, metrics, and business opportunities. *Applied Sciences (Switzerland)*, *10*(21), 1–20. https://doi.org/10.3390/app10217748

Fu, Z., Niu, X., & Maher, M. Lou. (2024). Deep Learning Models for Serendipity Recommendations: A Survey and New Perspectives. *ACM Computing Surveys*, *56*(1), 1–26. https://doi.org/10.1145/3605145

Fu, Z., Niu, X., & Yu, L. (2023). Wisdom of Crowds and Fine-Grained Learning for Serendipity Recommendations. *SIGIR 2023 - Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 739–748. https://doi.org/10.1145/3539618.3591787

Grange, C., Benbasat, I., & Burton-Jones, A. (2019). With a little help from my friends: Cultivating serendipity in online shopping environments. *Information and Management*, *56*(2), 225–235. https://doi.org/10.1016/j.im.2018.06.001

Hasan, T., & Bunescu, R. (2023). Topic-Level Bayesian Surprise and Serendipity for Recommender Systems. *Proceedings of the 17th ACM Conference on Recommender Systems, RecSys 2023*, 933–939. https://doi.org/10.1145/3604915.3608851

He, W., Ai, D., & Wu, C. (2021). A recommender model based on strong and weak social Ties: A Long-tail distribution perspective. *Expert Systems with Applications*, *184*, 115483. https://doi.org/10.1016/j.eswa.2021.115483

Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, *22*(1), 5–53. https://doi.org/10.1145/963770.963772

Hernández del Olmo, F., & Gaudioso, E. (2008a). Evaluation of recommender systems: A new approach. *Expert Systems with Applications*, *35*(3), 790–804. https://doi.org/10.1016/j.eswa.2007.07.047

Hernández del Olmo, F., & Gaudioso, E. (2008b). Evaluation of recommender systems: A new approach. *Expert Systems with Applications*, *35*(3), 790–804. https://doi.org/10.1016/j.eswa.2007.07.047

Hinton, G., Vinyals, O., & Dean, J. (2015). *Distilling the Knowledge in a Neural Network*.

Isinkaye, F. O., Folajimi, Y. O., & Ojokoh, B. A. (2015). Recommendation systems: Principles, methods and evaluation. In *Egyptian Informatics Journal* (Vol. 16, Issue 3, pp. 261–273). Elsevier B.V. https://doi.org/10.1016/j.eij.2015.06.005

Jain, A., Singh, P. K., & Dhar, J. (2020). Multi-objective item evaluation for diverse as well as novel item recommendations. *Expert Systems with Applications*, *139*. https://doi.org/10.1016/j.eswa.2019.112857

Kaminskas, M., & Bridge, D. (2016). Diversity, serendipity, novelty, and coverage: A survey and empirical analysis of beyond-Accuracy objectives in recommender systems. In *ACM Transactions on Interactive Intelligent Systems* (Vol. 7, Issue 1). Association for Computing Machinery. https://doi.org/10.1145/2926720

Kang, W.-C., & McAuley, J. (2018). *Self-Attentive Sequential Recommendation*. http://arxiv.org/abs/1808.09781

Karthik, R. V., & Ganapathy, S. (2021). A fuzzy recommendation system for predicting the customers interests using sentiment analysis and ontology in e-commerce. *Applied Soft Computing*, *108*. https://doi.org/10.1016/j.asoc.2021.107396

Kitchenham, B. (2004). *Procedures for Performing Systematic Reviews*.

Kompan, M., Gaspar, P., MacIna, J., Cimerman, M., & Bielikova, M. (2022). Exploring Customer Price Preference and Product Profit Role in Recommender Systems. *IEEE Intelligent Systems*, *37*(1), 89–98. https://doi.org/10.1109/MIS.2021.3092768

Kotkov, D., Wang, S., & Veijalainen, J. (2016). A survey of serendipity in recommender systems. *Knowledge-Based Systems*, *111*, 180–192. https://doi.org/10.1016/j.knosys.2016.08.014

Lam, S. K., Frankowski, D., & Riedl, J. (2006). Do You Trust Your Recommendations? An Exploration of Security and Privacy Issues in Recommender Systems. *Emerging Trends in*

Information and Communication Security, *3995 LNCS*, 14–29. https://doi.org/10.1007/11766155_2

Li, P., Que, M., Jiang, Z., HU, Y., & Tuzhilin, A. (2020). PURS: Personalized Unexpected Recommender System for Improving User Satisfaction. *Fourteenth ACM Conference on Recommender Systems*, 279–288. https://doi.org/10.1145/3383313.3412238

Li, Q., Choi, I., & Kim, J. (2020). *Evaluation of Recommendation System for Sustainable E-Commerce: Accuracy, Diversity and Customer Satisfaction*. https://doi.org/10.20944/preprints202001.0015.v1

Liu Qian, & Ma Hui min. (2010). A study on the influence of recommendation models on customer satisfaction in B2C e-commerce. *2010 International Conference on Networking and Digital Society*, 452–455. https://doi.org/10.1109/ICNDS.2010.5479465

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*.

Ma, H., Gao, M., Wei, F., Wang, Z., Jiang, F., Zhao, Z., & Yang, Z. (2024). Stealthy attack on graph recommendation system. *Expert Systems with Applications*, *255*. https://doi.org/10.1016/j.eswa.2024.124476

Mansoury, M., Abdollahpouri, H., Pechenizkiy, M., Mobasher, B., & Burke, R. (2020). FairMatch: A Graph-based Approach for Improving Aggregate Diversity in Recommender Systems. *UMAP 2020 - Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, *20*, 154–162. https://doi.org/10.1145/3340631.3394860

Mansur, F., Patel, V., & Patel, M. (2017). A review on recommender systems. *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, 1–6. https://doi.org/10.1109/ICIIECS.2017.8276182

McAuley, J., Targett, C., Shi, Q., & Van Den Hengel, A. (2015). Image-based recommendations on styles and substitutes. *SIGIR 2015 - Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 43–52. https://doi.org/10.1145/2766462.2767755

Nagarnaik, P., & Thomas, A. (2015). Survey on recommendation system methods. *2nd International Conference on Electronics and Communication Systems, ICECS 2015*, 1603–1608. https://doi.org/10.1109/ECS.2015.7124857

Nagy, F., Haroun, A., Abdel-Kader, H., & Keshk, A. (2021). A Review for Recommender System Models and Deep Learning. *IJCI. International Journal of Computers and Information*, *8*(2), 170–176. https://doi.org/10.21608/ijci.2021.207864

Niu, K., Zhao, X., Li, F., Li, N., Peng, X., & Chen, W. (2019). UTSP: User-Based Two-Step Recommendation with Popularity Normalization towards Diversity and Novelty. *IEEE Access*, *7*, 145426–145434. https://doi.org/10.1109/ACCESS.2019.2939945

Pagare, R., & A. Patil, S. (2013). Study of Collaborative Filtering Recommendation Algorithm Scalability Issue. *International Journal of Computer Applications*, *67*(25), 10–15. https://doi.org/10.5120/11742-7305

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Systematic Reviews*, *10*(1), 89. https://doi.org/10.1186/s13643-021-01626-4

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Robertson, S. E., & Walker, S. (1994). Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. In *SIGIR '94* (pp. 232–241). Springer London. https://doi.org/10.1007/978-1-4471-2099-5_24

Rocha, A. B. da S., Meirim, M. O., & Nogueira, L. C. (2021). *Trends in the E-commerce and in the Traditional Retail Sectors During the Covid-19 Pandemic: an Evolutionary Game Approach*. http://arxiv.org/abs/2105.06833

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

Shah, L., Scholar, R., Gaudani, H., & Balani, P. (2016). Survey on Recommendation System. In *International Journal of Computer Applications* (Vol. 137, Issue 7).

Shahbazi, Z., Byun, Y., & Byun, Y.-C. (2020). Product Recommendation Based on Content-based Filtering Using XGBoost Classifier. *International Journal of Advanced Science and Technology*, *29*(04), 6979–6988. https://www.researchgate.net/publication/342864588

Shandhilya, T., & Srivastava, N. (2020). Using conceptual incongruity as a basis for making recommendations. *RecSys 2020 - 14th ACM Conference on Recommender Systems*, 557–561. https://doi.org/10.1145/3383313.3412231

Shrivastava, R., Sisodia, D. S., Nagwani, N. K., & BP, U. R. (2022). An optimized recommendation framework exploiting textual review based opinion mining for generating pleasantly surprising, novel yet relevant recommendations. *Pattern Recognition Letters*, *159*, 91–99. https://doi.org/10.1016/j.patrec.2022.05.003

Silveira, T., Zhang, M., Lin, X., Liu, Y., & Ma, S. (2019a). How good your recommender system is? A survey on evaluations in recommendation. *International Journal of Machine Learning and Cybernetics*, *10*(5), 813–831. https://doi.org/10.1007/s13042-017-0762-9

Silveira, T., Zhang, M., Lin, X., Liu, Y., & Ma, S. (2019b). How good your recommender system is? A survey on evaluations in recommendation. *International Journal of Machine Learning and Cybernetics*, *10*(5), 813–831. https://doi.org/10.1007/s13042-017-0762-9

Song, K., Tan, X., Qin, T., Lu, J., & Liu, T.-Y. (2020). *MPNet: Masked and Permuted Pre-training for Language Understanding*.

Sorde, R. K., & Deshmukh, S. N. (2015). Comparative Study on Approaches of Recommendation System. In *International Journal of Computer Applications* (Vol. 118, Issue 2).

Storey, V. C., Robinson, J. M., & Baskerville, R. L. (n.d.). *Computational Science: A Field of Inquiry for Design Science Research*. https://hdl.handle.net/10125/80043

Sun, J., Guo, W., Zhang, D., Zhang, Y., Regol, F., Hu, Y., Guo, H., Tang, R., Yuan, H., He, X., & Coates, M. (2020). A Framework for Recommending Accurate and Diverse Items Using Bayesian Graph Convolutional Neural Networks. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2030–2039. https://doi.org/10.1145/3394486.3403254

Wang, C. D., Deng, Z. H., Lai, J. H., & Yu, P. S. (2019). Serendipitous recommendation in e-commerce using innovator-based collaborative filtering. *IEEE Transactions on Cybernetics*, *49*(7), 2678–2692. https://doi.org/10.1109/TCYB.2018.2841924

Wu, W., He, L., & Yang, J. (2012). Evaluating recommender systems. *7th International Conference on Digital Information Management, ICDIM 2012*, 56–61. https://doi.org/10.1109/ICDIM.2012.6360092

Xu, Y., Yang, Y., Wang, E., Han, J., Zhuang, F., Yu, Z., & Xiong, H. (2020). Neural Serendipity Recommendation. *ACM Transactions on Knowledge Discovery from Data*, *14*(4), 1–25. https://doi.org/10.1145/3396607

Yang, C., Miao, L., Jiang, B., Li, D., & Cao, D. (2020). *Gated and attentive neural collaborative filtering for user generated list recommendation ☆*. *187*, 104839. https://doi.org/10.1016/j.knosys

Yang, H., Choi, Y. S., Kim, G., & Lee, J. H. (2023). LOAM: Improving Long-tail Session-based Recommendation via Niche Walk Augmentation and Tail Session Mixup. *SIGIR 2023 - Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 527–536. https://doi.org/10.1145/3539618.3591718

Yu, T., Guo, J., Li, W., Wang, H. J., & Fan, L. (2019). Recommendation with diversity: An adaptive trust-aware model. *Decision Support Systems*, *123*. https://doi.org/10.1016/j.dss.2019.113073

Zangerle, E., & Bauer, C. (2023). Evaluating Recommender Systems: Survey and Framework. *ACM Computing Surveys*, *55*(8), 1–38. https://doi.org/10.1145/3556536

Zheng, Y., Gao, C., Chen, L., Jin, D., & Li, Y. (2021). DGCN: Diversified recommendation with graph convolutional networks. *The Web Conference 2021 - Proceedings of the World Wide Web Conference, WWW 2021*, 401–412. https://doi.org/10.1145/3442381.3449835

Zhu, J., Patra, B., & Yaseen, A. (2021). Recommender system of scholarly papers using public datasets. *AMIA ... Annual Symposium Proceedings. AMIA Symposium*, *2021*, 672–679.

Ziarani, R. J., & Ravanmehr, R. (2021). Deep neural network approach for a serendipity-oriented recommendation system. *Expert Systems with Applications*, *185*. https://doi.org/10.1016/j.eswa.2021.115660