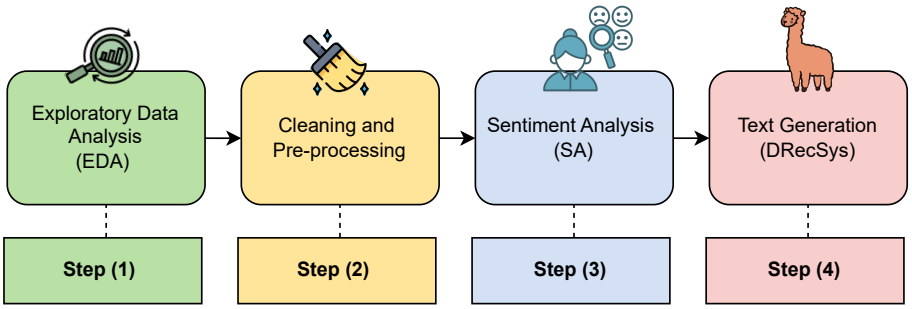


Drug Recommendation System Based on Symptoms and User Sentiment Analysis [DRecSys-SUSA]

ANA SOFIA SIMÕES PINTO

(Licenciada em Engenharia Eletrotécnica)

Dissertação para obtenção do Grau de Mestre em

Engenharia Informática e de Computadores

Orientadores: Doutora Matilde Pós-de-Mina Pato
Doutor Nuno Datia

Júri:

Presidente: Doutor José Manuel de Campos Lages Garcia Simão
Vogais: Doutora Vânia Patrícia Padrão Mendonça
Doutora Matilde Pós-de-Mina Pato

Fevereiro de 2025

Drug Recommendation System Based on Symptoms and User Sentiment Analysis [DRecSys-SUSA]

ANA SOFIA SIMÕES PINTO

(Grau de Licenciada em Engenharia Eletrotécnica)

Dissertação para obtenção do Grau de Mestre em

Engenharia Informática e de Computadores

Orientadores: Doutora Matilde Pós-de-Mina Pato, ISEL
Doutor Nuno Datia, ISEL

Júri:

Presidente:	Doutor José Manuel de Campos Lages Garcia Simão, ISEL
Vogais:	Doutora Vânia Patrícia Padrão Mendonça, FCUL
	Doutora Matilde Pós-de-Mina Pato, ISEL

Fevereiro de 2025

Acknowledgements

First, my deepest gratitude goes to my advisors, Prof. Matilde Pato and Prof. Nuno Datia, for their guidance, patience, and support throughout this journey. Their insights have been instrumental in both my academic achievements and my personal growth. I am also deeply grateful to Instituto Superior de Engenharia de Lisboa (ISEL) for providing an encouraging environment for my studies.

To my colleagues, cherished friends, and family at ISEL, you have filled my days with both happiness and productivity since the moment I arrived. The challenges we faced together were made less daunting by your support and friendship.

On a more personal note, I am deeply grateful to my family—my incredible parents, whose love and sacrifices have always inspired me. To my grandfather, Caetano, who taught me that hard work often replaces the need for luck, your wisdom has always guided me throughout my academic journey. And to my grandmothers, who made me feel lucky every day. Lastly, to my amazing sister, your constant belief in me holds a special place in my heart.

I owe immense thanks to all my extraordinary friends, whose kindness, support, and laughter have made every tough day better. To my wonderful and kind friend João Póvoas, your support and candid insights have been invaluable. And to Mónica Queimado, for being nothing short of absolutely amazing.

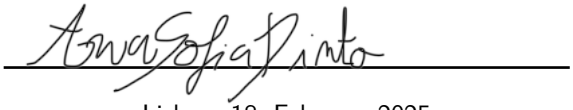
To everyone—family, friends, advisors, professors, and colleagues—who has been part of this journey, you have my deepest appreciation. Your support has shaped this achievement and for that, I am forever thankful.

From the bottom of my heart, thank you.

Statement of integrity

I declare that this dissertation is the result of my personal and independent research. Its content is original, and all sources listed in the bibliographic references were consulted and are duly mentioned in the text. I further declare that all scientific and technical references relevant to the development of the work are duly cited and included in the bibliographic references.

The author



Lisbon, 18, February 2025

Drug Recommendation System Based on Symptoms and User Sentiment Analysis [DRecSys-SUSA]

This document was created using the (pdf)LaTeX processor, based in the "iselthesis" template [84], developed at the DEETC of ISEL-IPL.

Abstract

The rapid growth of user-generated content on multiple online platforms has opened opportunities for improving decision-making across various domains, including healthcare. This dissertation focuses on the development of our Drug Recommendation System based on user-generated content (DRecSys-SUSA), designed to assist healthcare professionals and patients by providing personalized drug recommendations and supporting informed decision-making.

Our research leverages the UCI ML Drug Review dataset as the foundation for developing an advanced recommendation system. Our solution utilizes a combination of modern AI techniques, including Exploratory Data Analysis (EDA), data pre-processing, sentiment analysis (SA), and text generation using a fine-tuned Large Language Model (LLM).

We design and propose a recommendation system framework, within which we implement multiple variants of DRecSys-SUSA using different combinations of AI techniques. Each variant generates medically relevant suggestions to user-specific inputs such as age, symptoms, and current medications. Through an iterative process of implementation and evaluation using an LLM-as-judge methodology with AI-generated real-world scenarios, we identify which AI techniques are most beneficial for providing clinically appropriate and user-friendly drug recommendations.

The resulting insights contribute to the advancement of AI-driven healthcare tools by establishing effective approaches for leveraging user-generated content in medical recommendation systems.

Keywords: Natural Language Processing, Drug Recommendation System, Sentiment Analysis, Lexicon-based techniques, Data pre-processing, Transfer Learning, Large Language Models.

Resumo

O rápido crescimento de conteúdo gerado por utilizadores em múltiplas plataformas online abriu oportunidades para melhorar a tomada de decisões em vários domínios, incluindo no domínio da saúde. Esta dissertação centra-se no desenvolvimento do nosso Sistema de Recomendação de Medicamentos baseado em conteúdo gerado por utilizadores (DRecSys-SUSA), concebido para auxiliar profissionais de saúde e pacientes, fornecendo recomendações personalizadas de medicamentos e apoiando a tomada de decisões informada.

A nossa investigação utiliza o conjunto de dados UCI ML Drug Review como base para desenvolver um sistema de recomendação avançado. A nossa solução utiliza uma combinação de técnicas modernas de Inteligência Artificial (IA), incluindo Análise Exploratória de Dados, pré-processamento de dados, análise de sentimentos e geração de texto utilizando um Modelo de Linguagem de Grande Escala especializado.

Concebemos e propomos uma estrutura de sistema de recomendação, na qual implementamos múltiplas variantes do DRecSys-SUSA utilizando diferentes combinações de técnicas de IA. Cada variante gera sugestões medicamente relevantes para inputs específicos do utilizador, como idade, sintomas e medicação atual. Através de um processo iterativo de implementação e avaliação utilizando uma metodologia *LLM-as-judge* com cenários reais gerados por IA, identificamos quais as técnicas de IA mais benéficas para fornecer recomendações de medicamentos clinicamente apropriadas e de fácil utilização.

Os resultados obtidos contribuem para o avanço das ferramentas de saúde baseadas em IA, estabelecendo abordagens eficazes para aproveitar o conteúdo gerado por utilizadores em sistemas de recomendação médica.

Palavras-chave: Processamento de Linguagem Natural, Sistema de Recomendação de Medicamentos, Análise de Sentimentos, Técnicas baseadas em Léxico, Pré-processamento de Dados, Aprendizagem por Transferência, Modelos de Linguagem de Grande Escala.

Contents

List of Figures

List of Tables

Acronyms

ADRs	Adverse Drug Reactions 22
AI	Artificial Intelligence 1, 4, 5, 26, 27, 35, 60, 68, 76
BERT	Bidirectional Encoder Representations from Transformers 8, 12, 14, 15, 24, 78
CBF	Content-based Filtering 7, 17, 18, 19
CF	Collaborative Filtering 7, 9, 17, 19, 20, 21
CNN	Convolutional Neural Network 12, 22
DL	Deep Learning 10, 12, 13, 14, 15, 17, 20, 22
DRecSys	Drug Recommendation System xv, 1, 2, 3, 4, 5, 10, 17, 20, 22, 23, 24, 27, 29, 30, 31, 33, 34, 35, 37, 38, 42, 48, 49, 50, 52, 53, 54, 57, 59, 61, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 75, 76, 77, 78, 91, 92, 93
DRecSys-SUSA	Drug Recommendation System Based on Symptoms and User Sentiment Analysis 1, 2, 4, 5, 20, 27
EDA	Exploratory Data Analysis 2, 3, 4, 5, 7, 8, 17, 24, 27, 29, 31, 37, 46, 72, 75, 76
EMR	Electronic Medical Records 20
GAN	Generative Adversarial Networks 14
GPT	Generative Pre-trained Transformer 8, 12, 14, 15, 24, 26, 35, 60
HR	Hit Rate 25, 67, 68, 76
HRS	Health Recommendation Systems 4, 7, 18, 20, 21, 22
Llama	Large Language Model Meta AI 3, 4, 5, 8, 12, 14, 15, 16, 24, 25, 26, 34, 35, 48, 49, 51, 53, 66, 75, 76, 77
LLM	Large Language Model 1, 2, 3, 4, 5, 7, 8, 14, 15, 16, 17, 24, 25, 26, 27, 32, 33, 34, 44, 48, 49, 50, 53, 58, 62, 64, 68, 73, 75
LoRA	Low-Rank Adaptation 16, 48, 51
LSTM	Long Short-Term Memory 13, 22

ML Machine Learning 7, 8, 10, 11, 12, 14, 18, 23
MMLU Massive Multitask Language Understanding 35
MRR Mean Reciprocal Rank 25, 26, 67, 68, 76

NDCG Normalized Discounted Cumulative Gain 67, 68, 76
NER Named Entity Recognition 77
NLP Natural Language Processing 3, 8, 10, 12, 13, 14, 15, 24, 32, 44

PEFT Parameter-Efficient Fine-Tuning 48
PHR Personal Health Records 18
PIL Patient Information Leaflet 77

QLoRA Quantized Low-Rank Adaptation 16, 48, 51

RNN Recurrent Neural Networks 12, 13
RS Recommendation System 2, 3, 4, 5, 7, 8, 9, 10, 15, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 29, 34, 38, 40, 48, 75

SA Sentiment Analysis 2, 3, 4, 5, 7, 8, 10, 11, 12, 13, 15, 17, 22, 23, 24, 27, 29, 30, 31, 32, 33, 37, 44, 45, 46, 48, 49, 50, 59, 66, 67, 75, 76, 77
SKR Semantic Key Retrieval xv, xvii, 34, 35, 48, 49, 50, 53, 54, 58, 59, 61, 62, 63, 64, 65, 66, 67, 68, 69, 73, 75, 76, 91, 92
SVD Singular Value Decomposition 18
SVM Support Vector Machines 12

TRL Transformers Reinforcement Learning 48

VADER Valence Aware Dictionary and sEntiment Reasoner 3, 12, 22, 32, 33, 37, 44, 46, 48, 49, 50, 51, 60, 61, 65, 66, 67, 69, 70, 71, 72, 73, 75, 76, 77, 78
VAE Variational Autoencoders 14

1 Introduction

This chapter provides a detailed explanation of the research domain and its significance, highlighting the relevance of this dissertation at the intersection of medicine and technology. It outlines the motivation for this dissertation, offers an overview of the proposed solution, describes its main contributions, and concludes with a summary of the document's structure.

In today's healthcare landscape, technological advancements have opened new possibilities for enhancing clinical decision-making, improving patient outcomes, and advancing healthcare management. With the growth of online platforms and the increasing availability of large datasets—such as user reviews on the effectiveness of certain drugs—there remains significant under-explored potential to utilize this information to benefit both healthcare professionals and patients, [80, 104, 126].

This dissertation focuses on three main objectives:

The first objective is to develop a structured framework for Drug Recommendation System (DRecSys) that provides a systematic approach for integrating user-generated content with modern Artificial Intelligence (AI) techniques. This framework serves as a foundation for developing medical recommendation systems that can effectively process and utilize patient feedback data.

The second objective is to implement Drug Recommendation System Based on Symptoms and User Sentiment Analysis (DRecSys-SUSA), a personalized DRecSys that demonstrates the practical application of our framework. The system processes user-specific inputs such as age, sex, current medications, and symptoms to provide relevant drug suggestions, while exploring the potential of fine-tuned Large Language Model (LLM)s in healthcare applications.

The third objective is to systematically evaluate different variants of DRecSys-SUSA to identify the most effective approaches for drug recommendation systems. Through an iterative process of implementation and evaluation using an LLM-as-judge methodology with AI-generated real-world scenarios, we explore and compare different configurations and combinations of AI techniques. This evaluation aims to provide insights into which approaches are most beneficial for delivering clinically appropriate and user-friendly drug recommendations.

1.1 Motivation

The need for RSs focused on the healthcare sector have become especially pertinent in light of fairly recent global events, as the COVID-19 pandemic placed enormous pressure on healthcare systems, making it difficult for many individuals to access timely and personalized medical advice [44]. As a result, some patients were forced to make important healthcare decisions without sufficient information or guidance, underscoring the need for automated systems that can assist in selecting appropriate treatments. Additionally, new users are often more inclined to trust the experiences and recommendations of peers who have shared similar health conditions and outcomes. This dynamic presents an ideal use case for a drug recommendation system based on real-world data from user reviews.

An important aspect of this dissertation lies in its effort to use user-generated content, drug reviews, to provide valuable insights into drug efficacy, side effects, and overall performance. Over the last decade, user reviews have significantly increased with the growth of easily accessible online platforms, though unstructured, these reviews hold valuable insights that can be leveraged for various healthcare applications. The emergence of these platforms has created vast datasets across various domains, including healthcare, providing a rich source of data for the development of RS, [55, 80, 104]. In the healthcare domain, where patients and healthcare professionals must navigate an often overwhelming amount of treatment options, a system capable of analyzing relevant user feedback into actionable recommendations can become essential. Using user-generated content presents an opportunity to bridge the gap between real-world user feedback and the sophisticated algorithms that contribute for healthcare decision-making. This dissertation aims to contribute to the development of innovative healthcare technologies by prioritizing the integration of diverse techniques and ensuring their practical value and real-world applicability in the context of RS.

The increasing demand for personalized and context-aware, solutions highlights the potential of integrating RS, Sentiment Analysis (SA), LLM, and Exploratory Data Analysis (EDA). A context-aware system, in this case, refers to a RS that adapts to a user's medical needs by analyzing personal factors like medication history and current symptoms, while fine-tuning the model on real-world user experiences and medication reviews to enhance recommendation relevance. Each of these components has demonstrated significant value individually; however, their combined use offers unique opportunities to address more complex challenges, particularly in healthcare, where accurate and reliable recommendations can directly impact patient outcomes, [51, 74, 80, 104, 126]. In this dissertation, we intend to explore the integration and interplay between RS, SA, LLM, EDA and data cleaning and pre-processing; and then by using each, we want to develop different variants of a DRecSys focused on user-generated content, evaluation the effectiveness of each based on the techniques used. This approach not only emphasizes methodological rigor while stressing the value of exploring and interpreting data to create effective, context-specific solutions.

To address this need, we introduce DRecSys-SUSA, a system explicitly designed to leverage user feedback as a foundational element. The naming of DRecSys-SUSA emphasizes its reliance on user-generated content rather than official medical data. This distinction is crucial, as

it highlights the system's focus on real-world experiences and user perspectives, while also addressing the unique challenges associated with interpreting and utilizing such unstructured, diverse data effectively.

1.2 Overview of the Proposed Solution

The development of DRecSys follows a structured approach, ensuring the effective integration of key components such as dataset exploration and preparation, computational optimization, and workflow design. A critical aspect of this work is the implementation of the LLM Large Language Model Meta AI (Llama)2 for text generation, enabling the system to generate ranked drug recommendations and concise, informative summaries based on user reviews and the model's base knowledge. These components collectively enhance healthcare decision-making by providing personalized recommendations and valuable insights into drug efficacy, potential side effects, and overall patient satisfaction.

This research utilizes the UCI ML Drug Review dataset [59], which comprises user-generated reviews of various drugs, their associated conditions, and a 10-star rating system reflecting user satisfaction. By analyzing this dataset, the system extracts meaningful insights from real-world experiences to improve the accuracy and relevance of drug recommendations.

To achieve the objectives of this dissertation, the proposed solution is structured into four steps, each contributing to the final development of the Drug Recommendation System (DRecSys):

Exploratory Data Analysis The first step focuses on thoroughly exploring the dataset to understand its structure and key attributes. This step involves statistical and visual analysis of the data to identify trends, distributions, and potential relationships. By doing EDA, we establish a strong foundation for the subsequent steps and phases, ensuring that we are working with well-understood data.

Data Cleaning and Pre-processing After the EDA, the dataset undergoes cleaning and pre-processing. This step includes handling missing values, removing irrelevant attributes, and applying Natural Language Processing (NLP) techniques. These step ensures that the data is prepared for analysis, allowing for the extraction of relevant information that can enhance the performance of the Recommendation System (RS).

Sentiment Analysis The third step focuses on analyzing the sentiment of user-generated reviews. Using TextBlob and VADER, the system classifies reviews as negative, neutral, or positive on a scale of 0 to 10. This SA process helps capture user sentiment toward specific drugs, playing a critical role in assessing the relevance and reliability of the reviews. By integrating SA into the RS, we aim to evaluate its impact on enhancing the system's ability to provide contextually relevant and user-aligned recommendations. This step, along with the previous two, formed the basis of the paper "Enhancing Drug Reviews Insights through Exploratory Data Analysis and Sentiment Analysis"[88], which was presented at the 28^{th} International Conference on Information Visualization.

Text Generation for Drug Recommendation The final step of the process utilizes an LLM, specifically Llama2, which we fine-tuned to generate personalized drug recommendations. Through prompt engineering and model customization, we designed multiple variants of DRecSys by integrating a structured workflow and training the Llama2 model on our dataset. This allowed the system to process user-specific inputs—such as age, sex, symptoms, current medications, and other relevant medical details—while synthesizing insights from user reviews and medical data. Following the implementation, we conducted the evaluation of different DRecSys-SUSA variants. Using an LLM-as-judge process with AI-generated real-world scenarios, we assessed various configurations. This approach allowed us to compare their effectiveness in delivering clinically appropriate and user-friendly drug recommendations, ultimately identifying the most optimal system variant for healthcare applications.

This research will not only contribute to the growing body of knowledge on AI in the healthcare domain but will also provide valuable practical insights for future development of DRecSys and Health Recommendation Systems (HRS). By evaluating what works and why, we aim to inform best practices for the creation of systems that can meaningfully assist both healthcare professionals and patients in navigating the complexities of drug management.

1.3 Contributions

Firstly, we conducted comprehensive EDA and SA that resulted in a paper entitled **"Enhancing Drug Reviews Insights through Exploratory Data Analysis and Sentiment Analysis"** [88], presented at the 28^{th} International Conference Information Visualization, 2024.

Secondly, we design and propose a RS framework that provides a structured approach for developing DRecSys integrating user-generated content with modern AI techniques.

Thirdly, we implemented DRecSys-SUSA, a personalized DRecSys that demonstrates the practical application of our framework. The system suggests drugs based on user-specific inputs such as age, sex, current medications, and symptoms, while reinforcing the importance of fine-tuning LLMs for domain-specific applications, particularly in healthcare.

Fourthly, we conducted a systematic evaluation of different DRecSys-SUSA variants. Through an iterative process of implementation and evaluation using an LLM-as-judge methodology with AI-generated real-world scenarios, we explored and compared different configurations and models to determine the most effective approaches for providing clinically appropriate and user-friendly drug recommendations.

The code used and developed is open-sourced, making it available for further academic exploration and refinement, as this dissertation focuses on demonstrating the system's feasibility as a research project, not for commercial use. The code and dataset for the first three steps are available at `https://github.com/matpato/EDRISA.git` and the fourth and final step is available at `https://github.com/matpato/DRecSys-SUSA.git`.

1.4 Document Structure

The document is structured as follows:

- **Chapter 2: Background** provides an overview of the fundamental concepts and methodologies that form the foundation of this research. It covers topics such as the evolution of RS, SA techniques, and LLMs. Additionally, it explores key algorithms and methodologies relevant to drug recommendation and text generation.
- **Chapter 3: Related Work** reviews existing literature and previous research in RS, particularly in healthcare applications. It discusses SA in medical contexts, the role of EDA, and the application of LLMs in healthcare. This chapter also highlights current challenges and opportunities in developing AI-driven DRecSys.
- **Chapter 4: Proposed Solution** details the methodology of this research. It describes the dataset, its structure, and key attributes, followed by an in-depth explanation of the four-step framework used for developing the DRecSys-SUSA. This includes EDA, data cleaning and pre-processing, SA, and the integration of an LLM on the workflow of the variants of the DRecSys for personalized drug recommendations.
- **Chapter 5: Implementation** focuses on the practical development of the system. It explains the step-by-step process of implementing each component of DRecSys-SUSA, including dataset construction, fine-tuning of Llama2, prompt refinement and the overall workflow od the DRecSys. This chapter also discusses the technical challenges encountered and the solutions applied.
- **Chapter 6: Evaluation & Results Analysis** presents a systematic evaluation of different DRecSys-SUSA variants. It describes the evaluation methodology, including the use of an LLM-as-judge approach and AI-generated real-world scenarios. The chapter provides detailed performance analysis based on accuracy metrics, ranking-based assessments, and user-friendly drug recommendation criteria.
- **Chapter 7: Final Considerations** summarizes the key findings and contributions of this dissertation. It reflects on the broader implications of the research in healthcare and AI, particularly in improving DRecSys. Additionally, it outlines potential future research directions and improvements to enhance the effectiveness of DRecSys-SUSA.

2 Background

In this chapter, we explore the concepts and methods that are relevant having in mind the domain of this dissertation. It covers important topics such as Recommendation System (RS), Sentiment Analysis (SA) and Large Language Model (LLM), all of which are integral to the development of the proposed solution of this dissertation.

2.1 Historical Context

The development of RS, SA, and LLM has evolved significantly, with each technological leap bringing new tools and methodologies that expand their applications in various domains, including the domain of this dissertation–healthcare. Practices like EDA have supported these advancements from the beginning, by facilitating a foundational understanding of datasets and identifying patterns and ensuring the reliability of downstream applications, enabling personalized systems to address real-world challenges, [50, 55, 74, 104, 126].

RS have their roots in the 1950s and 1960s, beginning with the creation of information retrieval systems that could organize and retrieve relevant documents from vast collections. Elaine Rich created the first RS in 1979, called Grundy, where she looked for a way to recommend users books they might like. Her idea was to create a system that asks users specific questions and classifies them into classes of preferences, or "stereotypes", [19]. By the 1980s and 1990s, RS evolved into Content-based Filtering (CBF), considered one of the most widely used and researched recommendation approaches and Collaborative Filtering (CF) methods, exemplified by systems like Tapestry, which matched users with items based on shared preferences [19, 46]. These early systems laid the groundwork for personalized recommendations, which became widely adopted in the 2000s with platforms like YouTube®, Facebook®, Netflix®etc., many using hybrid techniques to refine user experiences. In healthcare, this progression was mirrored by the emergence of Health Recommendation Systems (HRS) which used Machine Learning (ML) to provide personalized recommendations for dietary choices [56], a healthy lifestyle [89, 122], fitness [104, 117], decision-making for patients and physicians [29, 77, 109], and disease prediction [29, 54, 67, 115, 120].

The rise of SA in the 2000s, driven by the exponential development of social media, brought new opportunities for understanding user opinions. Early SA started with the application of lexicon-based approaches, that can still be found useful nowadays, where predefined word lists were used to assess the polarity of text [98]. However, these methods can struggle with nuance and domain-specific language. Advances in ML in the 2010s introduced more sophisticated techniques, such as vectorized word embeddings, allowing for a deeper understanding of context [76]. In healthcare, SA has been applied to patient/user reviews and feedback, providing insights into drug efficacy, side effects, and overall patient satisfaction, further explored in the Chapter 3 of this dissertation.

Transformer architecture [107], introduced in 2017, revolutionized NLP and led to the development of LLM such as BERT [66], GPT [2], and Llama [6], etc. These models leverage bidirectional context and massive datasets to achieve unprecedented performance in language understanding and generation. The architecture's modularity and scalability have made it adaptable to a wide range of NLP tasks, including language modeling, text generation, and question answering. Its ability to efficiently process long sequences and understand contextual relationships has transformed not only NLP but also other domains like vision [39] and speech processing [64, 124]. More recently, foundation models like Llama have shown remarkable potential in healthcare applications, from analyzing clinical notes to generating personalized recommendations [111, 118] or even to do DNA analysis [123].

The integration of RS, SA, and LLM has been used to address different challenges. A recent example of it in healthcare was during the COVID-19 pandemic (starting at the end of 2019) RS proved invaluable, providing data-driven recommendations and insights from vast, often unstructured datasets. EDA was a extremely important step during this time, helping researchers pre-process and interpret pandemic-related data, such as patient feedback and treatment outcomes, which informed the development of reliable RS [44]. As datasets grew in size and complexity, EDA emerged as a critical step in the development of RS, SA, and LLM. EDA allows researchers to uncover patterns, detect anomalies, and identify biases within datasets. When it comes to RS, EDA can help understand user-item interactions, preferences, and sparsity issues, enabling the design of effective models that accurately reflect user behaviors. In healthcare RS, EDA is indispensable for analyzing all types of data, ensuring that RS are based on reliable and meaningful inputs. Furthermore, it enhances the interpretability of outputs by providing insights into the underlying data upon which the RS is built, enabling a certain degree of explainability. This is even more important with the use of RS that use LLMs, allowing researchers to assess the quality and diversity of training data, understand model outputs, and ensure alignment with specific domain requirements. Without this foundational step, insights derived from RS may lead to undesired outcomes and limiting the effectiveness of these systems in real-world applications. This is also further explored in Chapter 3 of this dissertation.

2.2 Recommendation Systems

On the Internet, where the number of choices is overwhelming, there is a need to filter, prioritize and efficiently deliver relevant information in order to help alleviate the problem of

information overload, which created a potential problem to many Internet users. This problem of information overload challenges users in various domains, from casual browsing to critical decision-making, contributing to many fields such as e-commerce, education, and the domain at hand in this dissertation, healthcare. RS address this challenge by analyzing vast amounts of dynamically generated information and providing users with personalized suggestions for content, products, or services. By using user preferences, behaviors, and contextual factors, RS have transformed how users interact with digital platforms and access information [55].

In e-commerce, platforms like Amazon® and eBay® use RS to recommend products based on a user's browsing and purchasing history, driving sales and improving customer experience. In the entertainment industry, services like Netflix®, Spotify®, and YouTube® rely on RS to suggest movies, music, and videos personalized to individual tastes, fostering deeper user engagement, [10, 47]. Education technology platforms leverage RS to recommend courses, resources, and learning paths based on user performance and preferences, enabling more effective learning journeys. In healthcare, RS are increasingly used to provide personalized diet, fitness and drugs suggestions, lifestyle recommendations, assist in clinical decision-making by analyzing user histories and real-time data [117] or helping with the extraction of valuable insights in biomedical fields [85].

At the base of RS is data, which plays a crucial role in enabling accurate and relevant predictions. This data encompasses multiple dimensions, including user information, such as past interactions, demographic details, and explicit preferences; item characteristics, such as categories, tags, and attributes; and contextual information, including time, location, or the device being used, [55].

A defining characteristic of RS is their ability to enhance user experiences by offering personalization, efficiency, and opportunities for exploration. Users benefit from reduced search time, as RS surface relevant options quickly and precisely. Additionally, by tailoring content delivery to individual preferences, RS increase satisfaction while encouraging users to discover new products or ideas that align with their interests. Despite these benefits, RS face significant challenges, including the cold start problem, where systems struggle to recommend items for new users or products with little historical data. Data sparsity is another common issue, as limited interaction data within large datasets can hinder model performance, [55].

Furthermore, ensuring fairness and minimizing bias in recommendations remains an ongoing concern, particularly in sensitive domains like healthcare. The relevance of RS, the challenges and the application of this in the healthcare domain are better explored in Chapter 3.

There are several types of RS, each using different solutions to tailor recommendations to individual user preferences. Here are the primary types [10, 19, 20]:

- **Collaborative Filtering Method** recommends items by identifying patterns of user interaction. It assumes that users who had similar preferences in the past, are likely to agree again in the future. CF can be divided into: (a) **User-based** recommends items based on the preferences of similar users (User-based). It finds users with similar tastes and recommends items they have liked; and, (b) **Item-based** recommends items similar

to those a user has liked before (Item-based), based on similarity attributes between items.

- **Content-Based Filtering Method** relies on the characteristics and attributes of the items themselves (such as movie genres, article keywords, tags, categories, filters) to recommend similar items based on a user's history of preferences.
- **Hybrid Method** combine multiple recommendation techniques to improve accuracy and cover the limitations of using any single method. For example, Netflix uses a hybrid model combining collaborative filtering, content-based filtering, and other algorithms to personalize content offerings [47].
- **Knowledge-Based Recommendation Systems** use domain knowledge about how specific items meet user needs. Instead of relying on user interactions or item features, they apply rules or logic based on predefined knowledge. This method is especially effective when there is no user history.
- **Deep Learning (DL)-Based Recommendation Systems** identify complex patterns in user behavior and item attributes. They are valuable and used in sophisticated systems like Google recommendations.

Each type of RS has its strengths and limitations, and our DRecSys will be chosen based on the specific necessities of it is final application, the data available, and the desired user experience. In Table 2.1, we present examples of well-known RS categorized by their underlying approach and application type.

Table 2.1: Examples of Recommendation Systems by Type and Category

Recommendation Systems	Categories	Example Platforms
Collaborative Filtering	User and Item-Based	Amazon®, Netflix®, eBay®, Last.fm®, Goodreads®
Content-Based Filtering	Item Characteristics	YouTube®, Pandora®, LinkedIn®, Medium®
Hybrid Method	Mixed Methods	Netflix®, Alibaba®, Facebook®, Instagram®, YouTube®
Knowledge-Based RS	Domain Knowledge	Movielens®, Google News, Zillow®, TripAdvisor®
Deep Learning-Based RS	Complex Patterns	TikTok®, Twitter®, Spotify®, Pinterest®

2.3 Sentiment Analysis

Sentiment Analysis (SA), also referred as *opinion mining*, is the field of study that analyses (with the use of NLP) people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes [72]. SA methods can be classified into three main approaches, as seen in Figure 2.1: lexicon-based, ML-based and hybrid approaches [93].

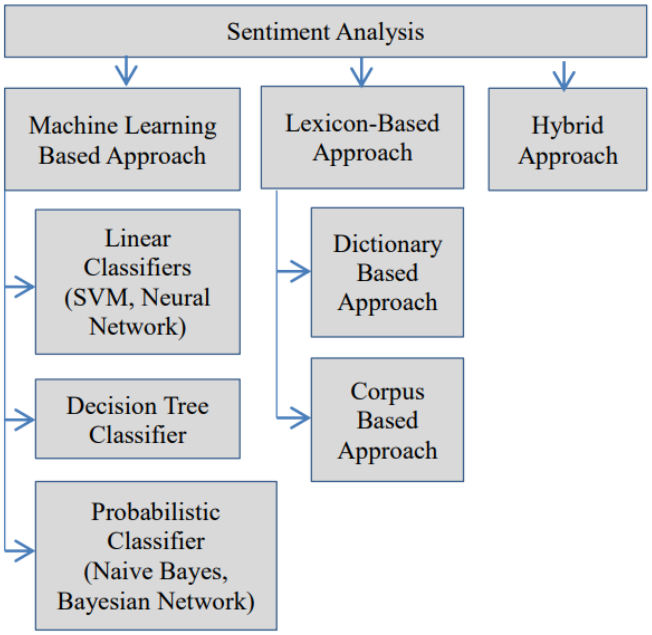


Figure 2.1: *Sentiment analysis methods [93]*

Lexicon-based methods in SA are approaches that rely on a pre-defined list of words—Sentiment Lexicon, each of which is associated with a specific sentiment value – Polarity Score. These methods are based on the idea that the overall sentiment of a text can be determined by the individual sentiments of the words or phrases it contains – Scoring Rules. Lexicon based approach can further be divided into two categories: Dictionary based approach (based on dictionary words i.e. WordNet or other entries) and Corpus based approach (using corpus data, can further be divided into Statistical and Semantic approaches). While basic lexicon-based methods might simply sum up sentiment scores, more advanced approaches consider the context in which words are used. This might involve adjusting scores based on word order, sentence structure, or the presence of negations or modifiers. One key advantage of lexicon-based methods is their simplicity and transparency. They don't require training on large datasets, unlike ML-based approaches. However, their main limitation is that they might not accurately capture the nuances of sentiment in texts with complex structures, sarcasm, context or domain-specific language. ML-based methods, on the other hand, rely on training models using labeled datasets to classify sentiment. These methods can capture intricate patterns in data but require substantial annotated data and computational resources for effective training. To leverage the strengths of both approaches. The hybrid methods combine both ML and lexicon-based methods. In the Figure 2.1 a simplified schematic is presented, illustrating the primary SA methods as described by [93].

In this dissertation, SA plays a central role in evaluating drug reviews. For ranking drug reviews on a scale different from the "usual" negative, neutral and positive based on sentiment, it is important to choose a tool or library considering the complexity of the task and the level of accuracy it requires. Still the objective will always be that this tool or library can accurately interpret the overall polarity and context of, in the specific case of this dissertation, user reviews and feedback, which, in the context of this research can often include some medical terminology

and mixed sentiments. Here are some of the most popular tools for this task that where taken in consideration:

1. **Lexicon-Based Approaches**

   - **TextBlob** (Python library) is a well-known Python library used for processing textual data [99]. In SA, TextBlob provides a polarity score, which is a floating-point number ranging between -1 and 1. A score of -1 means a negative sentiment, while a score of $+1$ indicates a positive sentiment. This polarity score is a key feature of TextBlob's sentiment analyser, which assesses the overall sentiment of a given input sentence. While TextBlob has proved to be effective for basic applications, its accuracy may be limited in more complex SA scenarios.
   - **Valence Aware Dictionary and sEntiment Reasoner (VADER)** (NLTK library) is particularly effective at handling sentiments expressed in social media, which means it can be effective for user-generated content like reviews [96]. It's adept at understanding text with mixed sentiments, emojis, slang, etc. However, it might not be as effective with specialized medical terminology.
   - **SentiWordNet** is a lexical resource derived from WordNet, where each synset (set of synonyms) is assigned numerical scores for positivity, negativity, and neutrality. It is widely used for assigning polarity scores to words in texts and provides a more structured lexicon compared to other resources, [94].

2. **Machine Learning-Based Approaches**

   - **Scikit-learn** (Python library) is very used for implementing classical ML models [36]; including Support Vector Machines (SVM), decision trees, and ensemble methods like Random Forests. It provides a comprehensive set of tools for pre-processing, training, and evaluating models, making it ideal for SA tasks.
   - **spaCy** is an industrial-strength NLP library. It's known for its speed and efficiency [97]. While not primarily a SA tool, spaCy allows for training custom ML models, making it a flexible option for building classifiers like SVM, neural networks, or probabilistic classifiers.
   - **Transformers** (e.g. accessed from HuggingFace) provides access to state-of-the-art ML models like BERT, GPT, Llama, etc [105]. Advanced models like this can be fine-tuned o the specific dataset. They are capable of understanding complex sentence structures and nuanced language, which makes them a strong candidate for analyzing drug reviews. However, they require more computational resources and a deeper understanding of ML.
   - **TensorFlow** (Google) is a highly scalable and production-ready framework for building and training DL models [37]. It is particularly effective for implementing neural networks such as Convolutional Neural Network (CNN) and Recurrent Neural Networks (RNN), as well as modern architectures like Transformers. TensorFlow's high-level API, simplifies the process of creating, training, and deploying models

for tasks like SA. TensorFlow excels in scalability, making it ideal for large datasets and complex models. However, it is a hard tool to start working with and it needs a lot of computational resources, these are challenges for beginners or smaller-scale projects.

- **PyTorch** (Facebook) is a flexible and research-oriented DL framework [35]. It is particularly valued for its dynamic computation graph, which makes debugging and experimentation easier compared to TensorFlow. PyTorch is well-suited for creating advanced architectures like Transformers and supports fine-tuning pre-trained models for domain-specific tasks.

Finally, it is also possible to consider a **Custom Machine Learning Model**, by having a dataset of drug reviews, building a custom model is considered a advantageous approach. This way, it is possible to train the model specifically on the kind of language and sentiment expressed in drug reviews. This approach requires the most effort but could yield very accurate results. Even though the limited data present in the particular dataset could significantly impact the the training of this time of model, leading to unwanted results.

The choice of the SA tool is primarily based on its observed efficacy and results in a specific work. In our dissertation, given that there is an initial rating available for comparison, we were able to assess the performance of each tool more effectively.

2.4 Algorithms for Text Generation

Given the objectives of this dissertation, we concentrated on understating various algorithmic approaches for text generation. The field of NLP has seen significant advancements, presenting a range of methodologies. It was essential to understand and evaluate these approaches, considering their applicability, strengths, and limitations within the context of our specific dataset and requirements. This helps ensure that the chosen method would be effective for generating accurate and contextually relevant recommendations based of user feedback.

- **Deep Learning Approaches** have significantly advanced text generation by effectively handling complex textual data, with Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks being prominent earlier methods [90]. RNN are designed to process sequences of data, using internal memory to capture information from previous elements in a sequence, making them well-suited for text-based applications. LSTM, a specialized form of RNN, excel in learning long-range dependencies and mitigating the vanishing gradient problem, which enhances their ability to generate text where context from earlier sequences is critical. Despite their strengths, RNN and LSTM require substantial computational resources and extensive datasets for effective training, posing challenges in adoption. Pre-trained RNN offer potential solutions, yet the structured nature of datasets and the limited relevance of earlier context in some cases question their suitability for certain tasks. While LSTM have seen applications in areas like SA [69], these models have largely been superseded by transformer architectures due to the sequential processing constraints inherent in RNN.

- **Attention Mechanisms and Transformer-Based Models**, introduced in [107], marked a very important step in NLP. Its self-attention mechanism enables efficient processing of entire sequences, capturing long-range dependencies and parallelizing computations. Transformers form the foundation for state-of-the-art models lik GPT, Llama and T5, which excel at both understanding and generating text. Their scalability and adaptability make them a very good solution for text generation tasks in this dissertation – generating personalized recommendations from user feedback.

- **Transfer Learning with Pre-Trained Language Models** involves using LLM based on transformer architectures, such as RoBERTa [66], GPT [2], Llama [6] and T5. These models have undergone extensive pre-training on vast amounts of text and can be fine-tuned in a smaller, contextual problem based dataset, for specific tasks like text generation based on the analysis and contents of user comments, allowing them to adapt their broad language understanding to a particular context or style. Notably, they have proven to excel in grasping semantic meanings and contextual nuances in text. This type of model has previously been applied to the medical field with success [17].

- **Hybrid Models** are the integration of traditional ML techniques with DL or transfer learning models, [62]. This hybrid approach might involve initial feature extraction using conventional NLP methods, followed by sentiment classification via a pre-trained language model. While this could potentially yield improved outcomes, it demands considerable effort and understanding, which might not proportionately translate into superior results.

- **Variational Autoencoders (VAE)** are generative models that encode textual data into latent representations and can sample from these representations to create diverse and meaningful text [23, 60]. Their ability to model variability makes them suitable for creative text generation tasks, though they may struggle with maintaining coherence over long sequences which may not be beneficial for the work at hand that my need to interpret long texts from an user.

- **Generative Adversarial Networks (GAN)**, consisting of a generator and a discriminator, are increasingly adapted for text generation. Innovations like SeqGAN [114] address challenges in applying GAN to sequential data, making them promising for creative applications or synthetic data augmentation. However, their training remains computationally intensive and less stable compared to other methods [49].

In the end, the choice of using a Pre-Trained Language Model with the LLM Llama depended on accessibility, cost, and specific work requirements and objectives. Additionally, regardless of the chosen model, careful monitoring and validation are essential, given the sensitivity of the medical domain, generating text in the medical domain comes with significant ethical responsibilities.

2.5 Large Language Models for Recommendation Systems

LLM are a subset of pre-trained language models distinguished by their large size, typically with billions or trillions of parameters, and their ability to handle a very wide range of tasks. LLM represent a significant advancement in NLP, offering the ability to handle complex linguistic tasks and generate contextually relevant text. These models are pre-trained on vast datasets and can be fine-tuned for specific tasks, making them powerful tools for generating personalized recommendations, such as the ones of our future drug recommendation system. Below are some of the most notable LLM relevant to this dissertation:

- **Llama** (Meta AI) is a family of transformer-based language models [6], characterized by their range of sizes, from small to ultra-large. These models are designed for efficiency and adaptability, using advanced neural network architectures to achieve high performance in natural language understanding and generation. Llama's scalability makes it suitable for diverse applications, including those with limited computational resources. Choosing a smaller variant of this model may reduce the risk of over-fitting, which is a common concern with smaller datasets.

- **GPT** (OpenAI) models, such as GPT-3 and GPT-4, are state-of-the-art, large-scale language models known for their DL-based transformer architecture [2]. They excel in generating human-like text, understanding context, and adapting to various styles and formats. Pre-trained on vast datasets, GPT models are highly effective in a wide range of NLP tasks, including text generation, translation, and question answering. GPT models, particularly the latest versions like GPT-3 or GPT-4, have a proven track record in various text generation tasks, including nuanced and context-heavy domains, known for their exceptional fine-tuning capabilities. They can adapt well to the specific linguistic style and content of a small dataset [24].

- **BERT** (Google) is a bidirectional transformer model that captures context from both directions of a text sequence. It is particularly effective for tasks such as SA, named entity recognition, and question answering. While not designed for text generation, BERT is an essential precursor to generative models and excels in understanding language [38].

Pre-trained LLM have gained significant popularity when building RS due to their ability to handle complex language tasks and leverage vast external knowledge. They can interpret nuanced relationships between user inputs and generate contextually relevant outputs, making them ideal for systems requiring deep understanding and interaction, such as drug recommendations. In the context of LLM4Rec [111], the capabilities of a LLM are leveraged to build a RS. Both fine-tuning and prompt tuning techniques can be applied to adapt the pre-trained model to the unique requirements of recommendation tasks. These methods enable the model to better understand user-specific needs and generate more relevant, context-aware recommendations:

- **Fine-Tuning** involves tuning the pre-trained language model with data specific to the downstream task. For RS, this data typically includes `<user, item>` interactions, item

descriptions, user profiles, and other contextual information. By fine-tuning, the model learns to incorporate and leverage these specific patterns and relationships to enhance its recommendation capabilities. In this method is important to explain the two following Fine-Tuning Techniques as they are later used in this dissertation:

  - **Low-Rank Adaptation (LoRA)** enables efficient adaptation of the Llama model by freezing its existing parameters and introducing new, low-rank matrices. This approach allows the model to incorporate new, domain-specific information without modifying the core knowledge base. As a result, the model retains its general-purpose capabilities while specializing in the target domain, [34].
  - **Quantized Low-Rank Adaptation (QLoRA)** extends the LoRA technique by incorporating quantization. It compresses the model's weight parameters from the standard 32-bit to a 4-bit format, significantly reducing the memory footprint and computational requirements. This allows for the efficient fine-tuning of large models like Llama on limited hardware resources, such as those provided by Google Colab. By employing QLoRA, we were able to perform the fine-tuning process within the constraints of accessible computational environments, making it both cost-effective and practical [34].

- **Prompt Tuning** task is aligned with the pre-training loss by crafting prompts that guide the model towards the desired output. This technique helps in utilizing the knowledge embedded in the Llama model during its pre-training phase, allowing it to generate outputs or recommendations that are more aligned with the specific context of the task [24].

These methods effectively leverage the strengths of an LLM, enabling it to provide more accurate and contextually relevant recommendations.

3 Related Work

In this chapter, we review the existing literature and previous research relevant to our dissertation domain. This chapter aims to position our work within the academic context by discussing the application and challenges of RS, the use of SA in healthcare, the opportunities and challenges of using RS based on user sentiment, EDA in healthcare datasets, the role of LLM in healthcare applications, and future trends in DRecSys.

3.1 Application and Challenges of Recommendation Systems

RS is integral in managing information overload across various domains, including e-commerce, news, entertainment, healthcare, etc. (see Table 2.1 for examples of platforms that use different RS). RS uses algorithms to filter, prioritize, and deliver relevant information to the user, making them a powerful tool for personalizing user experiences. [47, 55, 104]

RS has evolved through various types, mainly CF, CBF, and hybrid models. CF, for instance, leverages past interactions to recommend items, while CBF focuses on item attributes. These methods are increasingly supplemented by DL and knowledge-based approaches, which offer advanced capabilities for complex tasks, making then useful for a amount of platforms and applications as seen in Table 2.1. The RECOMED system, for example, integrates knowledge-based components and various features like user conditions, drug side effects, and demographic data to personalize recommendations, showcasing a comprehensive pharmaceutical recommendation approach [125]. Similarly, advancements in pre-trained language models have shown potential in other domains, such as toxic comment classification, as demonstrated by [121], highlighting the versatility and effectiveness of these models in diverse applications.

DRecSys are increasingly important to the healthcare domain, where they can support critical decision-making for both patients and professionals and help reduce drugs errors–a significant global health issue, [12]. During the COVID-19 pandemic, as highlighted in [44] many patients self-medicated without proper guidance, intensifying health risks. In such situations, RS can mitigate these risks by providing reliable, data-driven drugs recommendations, ensuring that users receive informed and appropriate options [12, 44].

As healthcare data and user-generated content continue to grow, the role of RS becomes

even more essential, as these systems transform vast amounts of raw data into actionable insights, enhancing access to crucial information and can even help individuals to make well-informed health choices. By integrating personalized recommendations, RS have the potential to profoundly impact patient outcomes and improve healthcare delivery [117, 125].

The fast pace of modern life, along with increased stress and unhealthy habits, has left many people in a state of poor health that, if not addressed, can lead to serious diseases. RS have been recognized as tools that can aid in maintaining and improving health by offering personalized lifestyle and healthcare recommendations [26]. HRS have emerged as valuable tools to provide personalized health information, integrating with PHR to deliver patient-specific guidance. HRS extend PHR systems by offering easily understandable health information to patients and providing healthcare providers with context-specific medical resources, thereby supporting informed decision-making [109]. Also, it highlights key technical challenges, including handling complex medical terminology, information overload, and the need for personalization, which are crucial considerations in developing reliable health-focused RS.

The COVID-19 pandemic further emphasized the need for effective health-oriented systems, as it disrupted daily life and the healthcare sector worldwide, accelerating the development of digital health solutions, including health-focused RS that support both clinical and non-clinical applications [44]. HRS have shown promise in areas like dietary choices [56], a healthy lifestyle [89, 122], fitness [104, 117], decision-making for patients and physicians [29, 77, 109], and disease prediction [29, 54, 67, 115, 120]. While early health HRS focused on basic techniques, recent advancements in algorithms and models have expanded their uses in healthcare [15, 58, 86, 106]. Key applications in healthcare RS include:

Dietary Recommendations: Personalized dietary RS use various methods, such as CBF and knowledge-based RS, to provide suggestions for healthier food choices. These systems consider multiple factors like user preferences, demographics, and nutritional needs. [1] proposes a method for healthier dietary recommendations through food substitutions, using positive pointwise mutual information and truncated Singular Value Decomposition (SVD) to analyze food attributes and contextual relationships. This approach aims to identify similar yet healthier food alternatives that meet user preferences and nutritional needs. Another example of this use case in the paper [48], that discusses a system designed to offer personalized nutrition recommendations based on individual dietary needs, health goals, and lifestyle factors. The authors highlight the use of ML techniques and data from food intake, user preferences, and health conditions to generate personalized dietary suggestions. By focusing on personalization, the system aims to support users in making healthier food choices aligned with their wellness goals, thereby promoting better health outcomes through improved diet management. [9] proposes a context-aware RS that considers the specific context in which foods are consumed to offer personalized dietary recommendations. By analyzing the circumstances under which certain foods are chosen, the system can suggest contextually relevant and healthier alternatives, allowing for more personalized recommendations that align with users' lifestyles and preferences, rather than relying solely on general dietary advice. This approach enhances the personalization of dietary recommendations, supporting users in making health-conscious choices that fit into their daily routines. The DIETOS (DIET Organizer System) RS, [4] is designed for health

profiling and diet management, particularly for individuals with chronic diseases. Utilizing a CBF approach, it provides personalized dietary recommendations by analyzing food consumption data from both healthy individuals and patients with diet-related chronic conditions. This system aims to support users in managing their dietary needs effectively, promoting better health outcomes through personalized food choices that account to their specific health profiles. Lastly, [106] evaluated recommendation algorithms for online recipe portals, emphasizing that the choice of algorithm greatly affects the quality of personalized dietary recommendations. Their work highlights the importance of selecting and fine-tuning algorithms to best support users' health and nutrition goals in different contexts.

Healthy Lifestyle Recommendations: encouraging a healthy lifestyle by recommending activities and choices that promote physical and mental well-being and improve adherence to positive lifestyle changes. Smartphone-based RS have appeared as effective tools for promoting physical activity through engaging approaches. The *Cypress* system recommends enjoyable *exergames* (name given by the authors, representing exercises + games) by combining sensor data with user preferences to create personalized activity suggestions [5]. This approach exemplifies how RS can support healthier lifestyle choices by aligning physical activities with user enjoyment, thereby enhancing long-term exercise adherence. Additionally, [95] introduced a project involving a mobile app designed to help users enhance their well-being by recording three positive experiences each day – the "Three-good-things" exercise. The app uses CF to analyze user behavior patterns and identify individuals with similar interests. Based on the preferences and activities of these similar users, the app recommends nearby activities and places likely to foster positive experiences, encouraging users to engage in enjoyable activities and ultimately promoting a greater sense of daily happiness. Mental health management is another crucial aspect of maintaining a healthy lifestyle. Modern RS have demonstrated significant potential in supporting mental healthcare delivery. For example, the Ginger mental health platform employs knowledge-based recommendation algorithms to enhance user engagement with self-guided mental health content [28].

Training Recommendations: RS in fitness and training provide personalized exercise plans by analyzing user behavior and preferences. These systems can predict motivation drops, identify areas for improvement in training plans, and alert coaches or users, thereby increasing the likelihood of achieving fitness goals. For example, [87] demonstrates a RS within a healthcare platform that identifies users losing motivation to exercise. By spotting these patterns early, the system offers personalized recommendations to re-engage users and maintain their commitment to fitness routines, supporting long-term adherence to physical activity. Another relevant example is [22] that presents a system designed to predict workout quality, enabling coaches to provide timely and targeted support to athletes. By analyzing performance data, the system helps identify areas for improvement and makes sure that training plans are optimized for each individual, enhancing the overall effectiveness of fitness programs. Finally, [41] developed a prototype model for recommending workout videos that fit users' fitness levels and goals. By tailoring recommendations to individual needs, the system supports users in maintaining their exercise routines and staying motivated.

Decision-Making for Patients and Physicians: RS in clinical decision support help patients

manage chronic conditions and provide physicians with data-driven insights for treatment plans. For instance, [43] proposes argumentation-based recommendations to enhance patient empowerment by offering explanations and personalized suggestions that align with individual health needs and preferences, enabling patients to make informed decisions about their care. Systems like ArgoRec ([43]) by having argumentation-based reasoning and natural language interaction to deliver personalized recommendations, improve patient experience and recommendation quality. Additionally, [101] introduces an engagement scoring model designed to optimize care-gap interventions. This system assesses patient behavior and identifies opportunities for timely interventions, ensuring that patients receive necessary support when most needed. Such approaches highlight the potential of RS to improve healthcare delivery by addressing gaps in care and fostering better patient outcomes. Furthermore, [58] developed a hybrid health journey RS using Electronic Medical Records (EMR). By integrating EMR with personalized recommendations, this system aids in guiding patients through their healthcare journeys, offering contextually relevant advice and enhancing the decision-making process for both patients and physicians. The ability to combine clinical data with personalized recommendations underscores the potential of RS in delivering personalized healthcare solutions.

Disease Prediction: RS is increasingly applied in disease-related prediction tasks, using patient data to identify correlations with disease markers. These systems enhance early detection and enable personalized treatment strategies. For example, [11] developed a multimodal sensing framework for behavioral analysis, aimed at improving care for Parkinson's and Alzheimer's patients; By integrating multiple data sources, this system identifies patterns that assist in disease monitoring and early intervention. Similarly, [67] proposed an improved CF system to predict circRNA-disease associations, demonstrating the potential of RS in identifying complex biological correlations to advance precision medicine. Further, significant progress has been made in predicting baseline data for rare diseases. [115] introduced a CF approach that optimally combines user- and item-based predictions to support Friedreich's ataxia patients, enabling personalized care strategies based on predicted baseline data. Expanding on this, [116] developed a hybrid CF model that integrates memory-based and model-based techniques to improve the accuracy of baseline data predictions, ensuring more reliable insights for disease management. These applications illustrate the critical role of RS in using diverse datasets to predict disease progression, optimize care plans, and support personalized healthcare.

Our DRecSys primarily supports decision-making for both patients and physicians, with a secondary focus on disease prediction. By leveraging user data and information from other patients, it delivers personalized drug recommendations to assist in making informed treatment decisions. Additionally, it can predict diseases based on user-reported symptoms. By integrating advanced techniques, mainly Deep Learning (DL)-based RSs. The DRecSys-SUSA identifies patterns and correlations of symptoms with conditions to suggest the most suitable drugs, supporting disease management and optimizing healthcare outcomes.

In addition to these main applications, HRS are also used in areas such as sleep improvement, smoking cessation, and collaborative medical research. Context-aware RS have been developed to enhance lifestyle and sleep quality, as demonstrated in [81], who proposed a system that uses personalized user modeling to recommend context-specific actions for better sleep; These

recommendations adapt to individual habits and environments, offering personalized advice to promote healthier sleep patterns. Similarly, RS can play a significant role in smoking cessation efforts. Personalized health interventions have appeared across various medical domains. For instance, advanced smoking cessation applications now integrate RS to deliver personalized motivational messages based on user preferences and behavioral data, helping users maintain their commitment to quitting [52]. The application of HRS has also revolutionized assistive technologies, particularly in hearing aid devices. Modern hearing aids now function as sophisticated RS, automatically adjusting settings based on user preferences and environmental contexts to optimize audio quality and user experience [83]. All these applications showcase the significant impact RS can have in healthcare by offering personalized recommendations that enhance patient outcomes, improve lifestyle choices, and support medical professionals in making informed decisions. As recommendation techniques continue to evolve, their role in healthcare will expand, further supporting both preventive care and personalized treatment strategies.

Health Recommendation Systems (HRS) present unique challenges compared to other domains. A key issue is the potential harm caused by incorrect recommendations. In e-commerce, a bad recommendation might lead to a poor shopping experience, but in HRS, it can affect physical or mental health [57].

For instance, a poor recommendation for a training plan could cause injury, while inaccurate disease-related recommendations might lead to wrong treatments or excessive stress. In dietary recommendations, inappropriate suggestions might fail to consider individual preferences, costs, or food restrictions, potentially causing frustration or adverse effects [40]. These issues highlight the importance of achieving high recommendation accuracy, which may require collaboration among experts and users, high-quality data, and personalized algorithms.

Another challenge is data collection and integration. While health platforms can still generate available vasts amounts of data, much of it lacks quality and consistency. Issues like data loss, duplication, and input errors make cleaning and extracting meaningful information time-consuming. Additionally, many relevant health datasets are difficult to access due to the sensitive nature of the information or require extensive effort to extract relevant insights. [53] explores these challenges, highlighting the complexity of integrating heterogeneous healthcare data and emphasizing the need for standardized approaches to improve data quality and interoperability.

Interpretability is another critical concern for health RS. Transparent algorithms and clear explanations not only enhance user trust but also empower users to evaluate whether to follow a recommendation. For example, [16] highlight the conflicting goals of explanations in RS, emphasizing that while explanations should clarify the reasoning behind recommendations, they must also be concise and comprehensible to users. Similarly, [18] proposes a CF model that generates explainable recommendations by addressing the diverse preferences of users, making recommendations more relatable and actionable. Beyond accuracy, health RS must also account for factors like novelty, richness, cost, and user acceptability. For instance, [42] demonstrate how explainable RS can enhance the user experience by providing personalized race-time predictions and personalized training plans for marathon runners, showcasing the value of explainable

recommendations in supporting user goals. However, accurate recommendations alone may not always be practical in health RS, such as treatments that are prohibitively expensive or overly painful. [119] addressed this challenge by introducing a posterior network for generating recommendation reasons inspired by user preferences, allowing users to better understand certain recommendations. These advancements in explainability underscore the importance of balancing accuracy with interpretability, ensuring that health RS deliver recommendations that are not only precise but also understandable, practical, and aligned with users' needs.

Lastly, privacy preservation is extremely important, [3]. HRS handle sensitive data, including personal health records and real-time health metrics. Protecting user privacy is essential not only for user safety but also for the long-term success of healthcare technologies such as a DRecSys.

3.2 Sentiment Analysis in Healthcare

Sentiment Analysis (SA) is a powerful tool for understanding user-generated content in healthcare. By analyzing textual data such as reviews and feedback, RS provides insights into patient emotions, satisfaction, and concerns, which are essential for improving healthcare services and enabling personalized recommendations. These insights support various applications, including patient experience monitoring, health policy adjustments, and treatment optimization [91, 126].

Lexicon-based approaches like TextBlob [99] and VADER [96] are frequently used due to their simplicity and interpretability, but they can often struggle with the nuanced and domain-specific language of healthcare. Advanced Deep Learning (DL) techniques, such as Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) models, offer improved capabilities for extracting context and handling complex sentence structures. For example, [65] integrated TextBlob with a CNN-LSTM model to analyze sentiments in healthcare reviews, capturing both polarity and contextual nuances. Similarly, [112] build a bidirectional LSTM (BiLSTM) to better understand sentiment in healthcare data, using bidirectional context for improved representation. Despite these advancements, the application of SA in healthcare often requires extensive computational resources and domain-specific fine-tuning.

One of the areas for SA in healthcare is the analysis of drug reviews, relevant for the context of this dissertation. Drug reviews often contain valuable information about patient experiences, side effects, and drugs efficacy, making them a critical resource for healthcare providers and patients alike. [44] applied SA to predict sentiment scores from patient feedback, aiding in drug safety monitoring and the recommendation of safer alternatives. These applications not only help identify Adverse Drug Reactions (ADRs) but also assist in recommending drugs based on aggregated patient feedback, enabling more informed decisions by both patients and physicians. While challenges such as specialized medical terminology, data sparsity, and the interpretability of advanced models remain, innovations like domain-specific embeddings such as BioBERT [66] and multimodal analysis combining text with other data types promise to enhance the impact of SA in healthcare.

3.3 Opportunities and Challenges in Recommendation Systems-Based on User Sentiment

User sentiment represents a valuable data source for enhancing the capabilities of RS, including DRecSys, offering insights derived directly from patient experiences. By incorporating SA into RS, these platforms can effectively capture user emotions, preferences, and feedback regarding drug efficacy and side effects. This integration allows the creation of more user-centric and data-driven recommendations, providing value not only to individual users but also to healthcare professionals seeking to make informed decisions based on this knowledge, [32, 74, 80, 126].

The analysis of user sentiment can uncover patterns and trends in drug usage, adverse reactions, and treatment satisfaction, which are difficult to identify through traditional clinical studies alone. [65] showed the potential of SA by analyzing drug review, their study extracted nuanced sentiments from user feedback, offering actionable insights into patient experiences that could enhance the drug recommendation process. Similarly, [112] used BiLSTM models for SA, showcasing the ability to capture contextual details in user comments, which is particularly useful for understanding complex healthcare data. [44] extended this concept by using ML techniques to assess patient sentiment on various drugs, identifying trends in drug safety and efficacy. This approach demonstrated the feasibility of predicting drug outcomes based on real-world patient experiences, enabling the recommendation of alternative treatments for those reporting negative side effects or inadequate results. Such systems can also provide healthcare providers with aggregated sentiment data, facilitating more personalized and informed prescribing practices. It is also important to have in mind multimodal SA, which combines textual data with other sources such as audio or video, also holds promise for capturing a more comprehensive picture of patient sentiment and behavior [63].

In summary, integrating user sentiment into drug RSs unlocks significant potential for improving patient outcomes and personalizing healthcare. By capturing real-world experiences and emotions, these systems can enhance the relevance and accuracy of drug recommendations, bridging the gap between clinical evidence and individual patient needs [32].

However, using sentiment-based drug recommendations presents several challenges. SA in healthcare must address the complexity of medical language, mixed sentiments within reviews, and the need for accuracy due to health implications. Additionally, sentiment-based recommendations must maintain a balance between user feedback and clinically validated information to avoid misinformation [65, 112].

So, one of the challenges lies in the integration of SA with clinical decision-making. [18] proposes an explainable framework for recommendations, emphasizing the need for interpretability in health applications. Without clear explanations, sentiment-driven recommendations may not gain the trust of healthcare professionals, who need to understand the rationale behind suggestions to adopt them in clinical practice. A clear understanding of the information behind the RS is critical.

Data-related challenges also play a significant challenge in this area. High-quality datasets specific to healthcare SA are often scarce, and existing datasets may suffer from issues like

limited scope, noise, or lack of diversity (in drugs, different opinions so on). [119] emphasizes the importance of building robust models that can generalize well across diverse datasets while maintaining interpretability.

Finally, adapting sentiment-based systems to the dynamic nature of healthcare is an ongoing challenge. Medical knowledge evolves rapidly, and RS must remain up-to-date to provide relevant insights. Advances in domain-specific NLP tools like BioBERT [66] show promise in addressing some of these issues by enabling more accurate and context-aware analysis of healthcare text. However, the integration of such advanced tools into scalable, real-world applications remains a significant technical and logistical barrier.

3.4 Exploratory Data Analysis in Healthcare Datasets

EDA is a fundamental step when dealing with a healthcare RS, offering critical insights into data trends, anomalies, and distribution patterns. These insights are essential for developing accurate, interpretable, and reliable models that can address the dynamic nature of healthcare. For example, [80] demonstrated the significance of EDA by analyzing drug reviews to uncover patterns in drug efficacy and patient satisfaction, which subsequently informed the development of SA techniques and recommendation models. This foundational analysis makes sure that a possible resulting RS can better reflect real-world patient experiences and guide informed healthcare decisions.

Similarly, other studies have shown that EDA can uncover biases, inconsistencies, and missing data, all of which are critical factors in ensuring the reliability of healthcare RS, [50]. Moreover, the interpretability of models built upon healthcare data heavily depends on EDA. Insights gained through EDA can guide feature selection and ensure that the resulting models align with clinical relevance and decision-making processes. Transparent and explainable the data provided for the RS frameworks rely on the foundational understanding provided by EDA to gain the trust of both healthcare professionals and patients, [50, 80].

In summary, EDA provides the foundation for building responsive and reliable healthcare RS, uncovering insights that inform model development and decision-making. Overcoming data-related challenges and ensuring the interpretability and scalability of RS are key to using EDA effectively in healthcare applications.

3.5 Large Language Models in Healthcare Applications

The fast and recent advancement of LLM has expanded the capabilities of RS in all sectors, including healthcare. LLM such as BERT [66], GPT [2] and Llama [6] have shown potential in interpreting complex language and generating contextually relevant recommendations (outputs) based on user inputs. Especially GPT and Llama models; [2, 24, 102, 103] demonstrated that this models show significant improvements on benchmark datasets and human evaluation, and therefore stand as state-of-the-art language models. In this dissertation, we leverage Llama 2 as our base model and created a workflow on an implementation of a DRecSys with fine-tuning of this model.

Such advancements in LLM can be attributed to two main factors: (1) scaling up the size of language models; and (2) expanding text corpora in the pre-training stage [24, 31, 100, 108, 110]. Their ability to process and interpret vast amounts of text data makes them uniquely suited for addressing the nuanced and dynamic needs of healthcare, where precise and context-aware responses are critical. Their adaptability allows them to handle diverse data inputs, from textual descriptions of symptoms to patient-generated queries and enabling more human-like interactions. LLM are applied as RS to understand item text features and improve recommendation performance with natural language output, having in mind the use of fine-tuning and prompt tuning [45, 70, 71].

A notable example of the integration of LLM in healthcare in the LargE languAge moDel distilling drugs Recommendation – LEADER, [74], it proposes the powerful semantic comprehension and input agnostic characteristics of LLM for text generation in the domain of healthcare, creating appropriate prompt templates that enable LLM to suggest drugs effectively. Its success underscores the potential for LLM to address complex healthcare queries and provide recommendations, aligning with the goals of precision medicine. However, challenges remain, particularly in addressing out-of-corpus issues, where models encounter terms or contexts not present in their training data. This limitation can reduce the accuracy and reliability of recommendations in highly specialized medical domains. This dissertation is a important to show a connection of our domain to the phase of text generation using an LLM and its challenges. In LLM4Rec using Llama [111], fine-tuning and prompt tuning are used to create a RS. Fine-tuning adapts the pre-trained model with task-specific data such as <user, item> interactions and item descriptions. Prompt tuning aligns the downstream task with the pre-training objective, guiding the model to produce more contextually relevant recommendations. LlamaRec [118] introduces a two-stage RS using LLMs for ranking. In the first stage, it uses a candidate generation process to narrow down potential items. In the second stage, Llama is used to rank these candidates by fine-tuning the model on specific interaction data, improving the relevance and quality of the recommendations. This approach enhances the system's ability to understand user preferences and item context, leading to more accurate and personalized recommendations.

The potential for LLM in healthcare represent a significant step forward in enhancing the functionality of healthcare RS. Their ability to scale across languages and medical subdomains, coupled with their proficiency in understanding context and generating human-like text, positions them as great tools for healthcare innovation. As models like BioBERT [66] and MedGPT [17], which are specifically fine-tuned on biomedical datasets, continue to evolve, the applicability of LLM in healthcare RS is expected to grow, further bridging the gap between patient needs and clinical expertise, [74, 78].

The evaluation of RS built with LLM requires a multidimensional approach to ensure accuracy, reliability, and applicability, particularly in the domain of healthcare.

These systems must be assessed using a combination of traditional metrics, such as accuracy, precision, recall, and F1 score, which evaluate the correctness and completeness of recommendations, alongside ranking focused metrics like Hit Rate (HR)and Mean Reciprocal Rank

(MRR) [74, 118]. Beyond quantitative metrics, the explainability and trustworthiness of recommendations are crucial, as healthcare professionals and patients need to understand the reasoning behind the suggestions for clinical decision-making [30].

Domain-specific challenges also play a significant role in the evaluation process. LLM-based RS must demonstrate an ability to effectively interpret complex medical terminology and patient-specific contexts [66], while handling data imbalances, such as rare diseases or uncommon drug interactions [74]. To address these challenges, human-in-the-loop evaluation methodologies are essential, incorporating feedback from clinicians and patients to ensure that recommendations are both practical and clinically relevant. For example, "Towards Interactive Recommender Systems with the Doctor in Loop" [51] emphasizes the role of medical professionals in refining Recommendation System (RS) through interactive collaboration. Real-world scenario testing and usability studies further enhance the reliability of these systems by simulating actual healthcare environments [14, 27, 82, 104, 113].

Explainability and ethical compliance are critical aspects of evaluation, particularly in healthcare settings where trust and accountability are paramount. Metrics that assess the clarity of recommendation explanations and adherence to privacy regulations are increasingly important [18], as are evaluations of fairness to ensure equitable recommendations across diverse patient populations [30]. Additionally, LLM-based RS must be robust and adaptable, with continuous learning mechanisms that enable dynamic updates in response to new data or shifts in medical practices [25]. Evaluating these systems for their ability to integrate new knowledge without compromising performance is extremy important, especially in a rapidly evolving field like healthcare.

Benchmarking against established models, such as BioBERT [66], GPT [2], Llama [7], or systems like LEADER [74] and LlamaRec [118], can provide valuable insights into performance standards and optimization techniques. These benchmarks highlight the importance of fine-tuning and prompt engineering in tailoring LLM for specific healthcare applications. In the end, comprehensive evaluation frameworks that encompass performance metrics, domain-specific considerations, ethical compliance, and adaptability will ensure the safe and effective deployment of LLM-based Recommendation System (RS).

Lastly, evaluating the outputs of LLM is still a challenge due to their diversity and complexity. A growing trend in research to continue helping address this issue involves using one LLM to evaluate another outputs, an approach often referred to as LLM-based evaluation or AI-as-a-judge. This method leverages the LLM's advanced capabilities to assess multiple content dimensions, including tone, relevance, coherence, consistency, reliability, and accuracy, as illustrated in Figure 3.1. This approach offers a scalable and cost-effective alternative to traditional human evaluations [27, 68, 82, 113]. However, while this approach has shown promise in various studies it also introduces challenges such as potential bias propagation and vulnerability, still, despite these limitations, LLM-based evaluation is emerging as a practical and increasingly common methodology in assessing the effectiveness of AI-generated outputs. In the study [68] the authors introduced a detailed collection of benchmarks for LLM-as-judge, along with an analysis of current challenges and future directions, offering valuable resources and insights for advancing this emerging field.

Figure 3.1: *LLM are capable of judging various attributes [68]*

3.6 Future Trends in Drug Recommendation Systems

Looking ahead, RS in healthcare are likely to incorporate multimodal data, including clinical records, genetic data, and patient feedback, to enhance personalization and precision [73]. The use of pre-trained LLM for real-time recommendations is expected to grow having in mind the exponential growth of LLM applications as foundation models [21, 24], and with models like Llama being fine-tuned for healthcare and medical domain [30, 78, 111].

Another potential development is the increased focus on interpretability and explainability in RS, ensuring users and professionals can understand the reasoning behind recommendations [30]. Transparent and interpretable RS will not only build trust among healthcare providers but also facilitate regulatory compliance and ethical accountability.

Ethical considerations, including data privacy and bias mitigation, will also play a very important role as RS in healthcare continue to evolve, particularly in applications involving sensitive patient data [30]. Additionally, addressing patient trust, ensuring equitable access, and adapting to evolving regulatory frameworks will remain critical to the successful deployment of these technologies. Future RS may also leverage continuous learning mechanisms, dynamically adapting to new data streams to refine recommendations over time [25].

In conclusion, despite the comprehensive literature reviewed in this chapter, several open challenges can be seen in the development of DRecSys based on user-generated content. In particular, there is a lack of studies that effectively integrate different techniques such SA with LLMs to generate both personalized drug recommendations. Moreover, few works use the use of LLMs as judges ("LLM-as-judge") applied to realistic, AI-generated scenarios. This dissertation distinguishes itself by proposing DRecSys-SUSA—a system that combines EDA, data pre-processing, SA, and text generation using fine-tuned LLMs—to explore and compare different system variants based on personalized user inputs. In doing so, this work aims to address research gaps and contribute a structured methodological approach and practical insights into the application of AI-driven DRecSys in healthcare.

4 Proposed Solution

This chapter begins with a data description, exploring the data collection process and its relevant features. Following this, the proposed methodology for the implementation of the DRecSysis presented in four main steps: EDA, Data Cleaning and Pre-processing, SSA (focusing on review polarity), and Text Generation.

4.1 Data Description

The UCI ML Drug Review dataset [59] provides patient/user reviews on a vast amount of drugs along with related conditions and a 10-star user rating system reflecting the overall user satisfaction, this rating will be referred as the original rating in this dissertation. The dataset is claimed to have been gathered through a process of crawling various online pharmaceutical review platforms. This method involves systematically extracting information from these websites, allowing for the compilation of user reviews and feedback about different pharmaceutical products – the dataset has more that over 200,000 user drug reviews.

We chose this dataset because it aligns with the central idea of building a RS based on user feedback rather than official medical data, emphasizing real-world experiences from other users. The dataset also contains attributes that are very relevant to our goals, including information on side effects and diverse opinions, efficacy, and satisfaction ratings that can help will the effectiveness of the step of SA. Additionally, the dataset presents a large range of conditions and drugs, making it well-suited for developing a system intended to assist diverse healthcare needs.

Table 4.1: Number of instances for the five main attributes from the dataset [59]

Attribute	Count
`drugName`	3671
`condition`	916
`UsefulCount`	602.1980
`uniqueID`	21.5063
`review`	21.5063

Through a preliminary study of the reviews, we observed that the dataset features a significant number of reviews detailing side effects and outcomes of using the specific drug, also many users present reason why they felt it was effective or not, often written in clear and simple language. This allows for nuanced analysis of user sentiment. Additionally, the reviews are contributed by a broad and diverse group of individuals, enhancing the different insights derived from the data. However, it is important to note that it is not possible to verify whether some reviews come from the same person, which may introduce a degree of redundancy or bias. Despite this limitation, the dataset offers a good foundation for exploring the interplay between user sentiment, symptoms, and drug efficacy in a real-world context.

Finally, the dataset is provided in the format `.csv` and it's divided into two files, this being the test and training partitions, 27,64 MB and 82.99 MB respectively, 110.63 MB combined. The dataset is composed by seven columns (features) this being described in Table 4.2 and the number of instances for the five main attributes present in the dataset can be seen in Table 4.1, there are not missing values, the discrepancy seen in the attribute `UsefulCount` is because this number represents the sum of the `UsefulCounts` values.

Table 4.2: Dataset Feature Descriptions

Feature	Description
`uniqueID`	A unique identifier for each review
`drugName`	The name of the drug
`condition`	The medical condition being treated
`review`	The free text of the review
`rating`	A patient-provided rating on a 10-point scale
`date`	The date the review was entered
`UsefulCount`	The number of users who found the review useful

4.2 Proposed Methodology

We begin by exploring and preparing our dataset, two foundational steps that set the tone for the rest of the analysis. Through extensive data exploration and personalized cleaning and pre-processing, we establish a baseline that supports the integrity and quality of subsequent analyses. These initial steps are crucial–they ensure that the data given as base for our models is consistent, reliable and well understood.

Then, the step SA, where we focus on the user reviews present in the dataset to extract emotional insights from the users on specific drugs. This analysis allows us to categorize sentiments and identify aspects of reviews that enhance their usefulness to others. By exploring these elements, we hope to refine our DRecSys to be more aligned with user expectations and needs based on real-world experiences.

The final step of our process is the development of an text generation model. This model, built on the principles of transfer learning and the strengths of pre-trained large language models, is designed to synthesize vast amounts of information to produce personalized drug recommendations. These DRecSys are not just predictive but prescriptive, offering customized

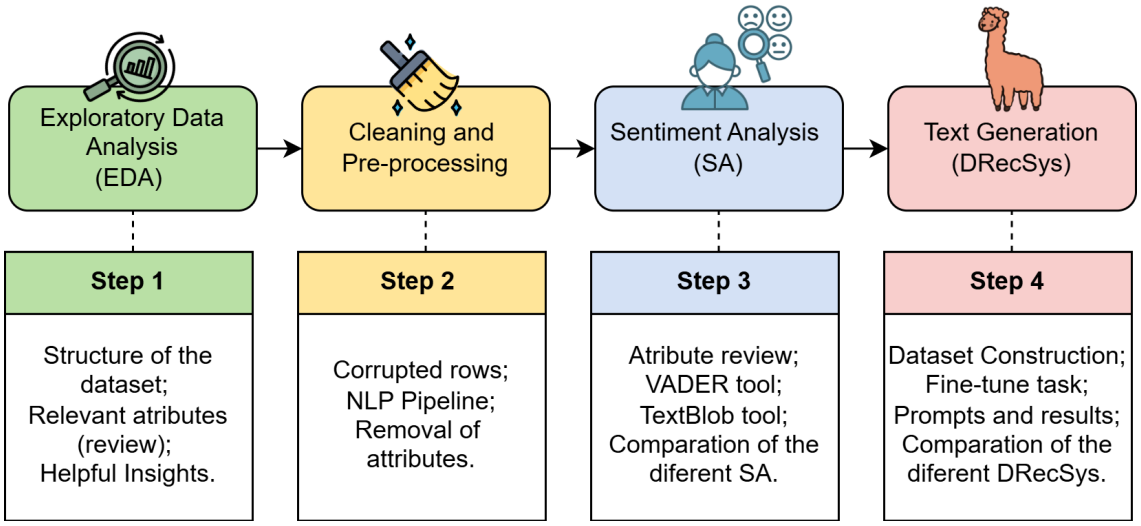


Figure 4.1: *The four steps of the proposed methodology*

advice closely personalized to individual preferences based on user feedback.

As we explore various methods, techniques, and datasets throughout this process, we hope not only to refine a final DRecSys, but also to gain valuable insights into the potential directions and necessities for future developments. The creation of different models thought different techniques and previous steps leads to various iterations of DRecSys, providing us with a clearer understanding of what works and what can be improved. These insights are invaluable as they help our understanding of the dynamic relations between technology and healthcare, highlighting the ongoing need for innovation and adaptation in our approaches. We explain the methods, tools and objectives of each of the main four steps in the next following four subsections. For a clearer understanding of the proposed methodology, Figure 4.1 provides a visual representation outlining the four integral stages of our approach. This figure offers a concise description of the key processes and activities carried out in each stage. This are 4 integral steps of our solution:

Step 1: Exploratory Data Analysis (EDA) In the first step, we focused on thoroughly examining the dataset to understand its structure and contents. The EDA process involved reviewing the dataset attributes and assessing their distribution to gain insights into the dataset's composition. This step is of great importance not only for validating the data format and content but also for evaluating the effectiveness of the future steps, mainly, the data cleaning and pre-processing efforts and the subsequent SA discussions as text generation and the impact of fine-tuning the model, as this step establish a solid foundation for our understanding of the answers provided by DRecSys, ensuring a solid basis for building this models and future models and algorithms.

Step 2: Data Cleaning and Pre-processing To ensure the integrity and quality of our analysis, the dataset undergoes this essential second step – data cleaning and pre-processing. In this step, all the attributes will be checked for missing values and other possible inconsistencies.

Also, the removal of the attribute `rating` will take place in this phase, so future ranking of features is only based on SA. Then the focus will be on the individual pre-processing of the attribute `review`, utilizing the NLTK toolkit to transform raw text into cleaned text—these

steps were based on the observed needs of the reviews present in the dataset. The pre-processing steps can be seen in Table 4.3. These steps ensure a consistent and clean dataset for downstream tasks, such as SA and a base for the training of models for text generation. While this traditional pre-processing approach lays a strong foundation, we recognize that modern LLMs can process raw text effectively without aggressive pre-processing. Thus, the necessity of each step will be further evaluated based on the observed impact on performance. This pre-processing is uniformly applied to both the test and training datasets to ensure consistency in data handling and analysis in the future steps.

Table 4.3: Text Pre-processing Steps and Descriptions

Step	Description
`HTML Entity Removal`	Cleaning up artifacts from online sources, such as replacing encoded characters
`Stopwords Removal`	Eliminating common words that add little semantic value to the analysis
`Lowercase Conversion`	Standardizing text by converting all characters to lowercase
`Punctuation Removal`	Stripping away punctuation marks to focus on textual content
`Tokenization`	Breaking down the text into individual words or tokens
`POS Tagging`	Assigning parts of speech to each word, like nouns, verbs, adjectives, etc.
`Stemming/Lemmatization`	Reducing words to their base or root form, with lemmatization being the preferred method for retaining contextual meaning

Step 3: Sentiment Analysis – polarity of the attribute review In the process of analyzing the sentiment polarity of the attribute *reviews* post pre-processing, we used the lexicon-based tools TextBlob and VADER–from NLTK. Each review is classified based on its sentiment – negative, neutral and positive categories, and sentiments are rescaled to fall within a 0 to 10 range, as transforming the polarity score to a range of 0 to 10 to provide a more nuanced understanding of the sentiments expressed and a better comparison of it is accuracy with the original attribute `rating`.

The two selected lexicon-based tools, TextBlob and VADER, will be employed to assess the sentiment polarity of each review, with a focus on both their cleaned and raw forms. This dual approach is strategic, when applying TextBlob, the effectiveness of the NLP pre-processing will be evaluated, highlighting how the removal of noise and irrelevant information impacts SA. On the other hand, the use of raw data can be particularly advantageous for VADER. Given its design and tuning for understanding nuances in informal or social media text, VADER is said effectively interpret language constructs like emoticons, slang, and sentence capitalization, which might be lost during pre-processing. VADER is specifically tuned for sentiments expressed in social media, making it effective for user-generated content like reviews. It may be better at understanding text with mixed sentiments and mixed polarity, which can be common in drug reviews that often contain some medical terminology, mixed with sentiments from the user or, for example, positive reviews that still demonstrate the presence of side effects. Furthermore,

while TextBlob is known for its ease of use, its performance in complex SA tasks can be limited. Therefore, both TextBlob and VADER is also used in this step to examine if one exhibits better results in interpreting the nuances present in drug reviews. This strategy allows for a comprehensive understanding of how different data forms influence the accuracy and efficacy of Lexicon-based SA tools, ensuring a robust and thorough examination of the sentiment polarity in each review.

Since the initial assessments using TextBlob and VADER proved sufficient, we did not find it necessary to transition to more sophisticated methods, such as transformers, as outlined in the background section. While we were prepared to do so based on the objectives of this work, these simpler approaches provided a solid foundation for the construction of the DRecSys.

Our primary objective for this step is to accurately capture and quantify the sentiments expressed in these reviews, thereby enriching our understanding and interpretation of these important data points and the comprehension of patient feedback that can contribute valuable insights into the dynamics of patient reviews in the healthcare sector. To facilitate this, we will introduce the *polarity* attribute into a dataset. This addition is crucial for the subsequent step of our approach, focused on text generation. Integrating the polarity attribute is anticipated to enhance both the training process and the performance of our text generation model.

Step 4: Text Generation The text generation step, which entails building the DRecSys, is the most important and time-intensive part of this project. This step involves choosing the right model for the task, the fine-tuning of the model, carefully choosing and refining the prompts given to this, the construction of datasets for different tasks and evaluating performance to ensure it delivers accurate and personalized drug recommendations. Additionally, deciding on the design principles requires significant time and consideration during this step.

Due to the complexity of this step, we divided it into the following decisions:

- **The use of Transfer Learning with LLM**: Transfer learning with pre-trained LLMs involves using models such as the ones previously mentioned, which are built on transformer architectures. These models undergo pre-training on vast amounts of text data, capturing semantic meanings and contextual nuances through semi-supervised learning. In our case, we fine-tuned an LLM on a dataset build for drug recommendations with the information from the original dataset mentioned. This allows the model to adapt its general language understanding to the specific nuances of drug information from user feedback, enhancing its performance for this particular context. Furthermore, fine-tuning enables the model to evolve alongside new medical research and user feedback, ensuring continuous improvement in recommendation accuracy [103]. A key reason for choosing an LLM is its ability to generate outputs in natural language, ensuring responses are user-friendly, consistent in structure, and easy to understand. Additionally, LLMs excel in handling inputs with mistakes, such as typos or ambiguous phrasing, making them particularly suited for real-world applications where users may not always provide perfectly structured queries. Based on user-specific details like age, sex, condition, symptoms, and current drugs, the model suggests treatments while considering efficacy and potential

side effects, the output also serves as a benchmark for evaluating the success of fine-tuning, highlighting areas where prompt adjustments or dataset refinements may improve performance.

- **Prompt Template in the workflow of the DRecSys:** The input for the LLM is structured using a designed Prompt Template, [74]. This template serves as a guide, ensuring that the user provides all necessary details (which includes gender, age, specific condition, symptoms, and current drugs) for a comprehensive assessment and the most informative and personalized output. Based on this structured input, the LLM generates a personalized response. This response not only recommends appropriate drugs (based on the user specific necessities and, in some cases, the raking of different drugs present of the specific dataset) but also includes a summary of the drugs with possible information from the fine-tuning task, encompassing real-world experiences and outcomes reported by users and the information on which the LLM was constructed. The goal the DRecSys to provide the user with the best drug recommendations for their specific condition, taking into account their physical and mental well-being, and ensuring that the suggestions are grounded in real-world data and experiences, thereby offering a more informed and personalized recommendation still having a great capacity to adapt and better expose the information based on the amount of information it was build on.

- **Correlating Symptoms to Conditions (SKR Task)**: To improve the accuracy of symptom-to-condition correlation, we used a smaller LLM to analyze user symptoms and cross-references them with conditions in the dataset. This task is mentioned in the dissertation as Semantic Key Retrieval (SKR) task, since it provides the key to then extract the information (the drugs) associated with that key from a dataset created. This task takes into account user-specific details like gender to refine the list of potential conditions. This method, combined with a condition-drug ranking dataset, helps streamline the search for appropriate drugs [118].

- **The Choice of LLMs:** In this dissertation we use fine-tuning, a specific form of transfer learning, using a pre-trained LLM for text generation. Specifically, we selected Llama2 [75], an open-source LLM known for its state-of-the-art text generation capabilities, just liked it was observed in Chapter 2 and 3, where Llama serves as a robust foundation for constructing RS due to its suitability for both research and commercial applications.

  Llama2 was selected for our DRecSys due to its balanced performance, efficiency, and accessibility. Although newer versions like Llama3 may offer more advanced capabilities, Llama2 stands out as a versatile and practical option, especially for applications that require a combination of good performance and cost-effectiveness, [102].

  Llama2 shows strong performance across various natural language processing tasks, offering models that range from 7 billion to 70 billion parameters. This range allows for adaptability based on the needs of specific use cases. In our system, we used the 7B version due to its balance between complexity and computational demands. The smaller model is projected to have sufficient capability in processing simple medical data and

generating drug recommendations, making it a practical starting point for handling the straightforward medical scenarios required in our system.

A key strength of Llama2 is its improved token efficiency, which is very important when working with large datasets like patient records and drug interactions. The model's tokenizer reduces the token count needed to represent complex data, enabling more efficient processing of large volumes of information. This optimization is particularly beneficial in applications requiring real-time or near-real-time responses, such as our DRecSys, where prompt decision-making is essential [103]. Another advantage of Llama2 is its open-source nature, which facilitates easy integration into various platforms, including AWS, Microsoft Azure, and Hugging Face. The availability of the model accelerates the development and deployment process, making it accessible to researchers and developers across different environments. Compared to larger models like Llama3, Llama2 is significantly more resource-efficient. It requires fewer computational resources for both training and inference, which allows for more economical deployment in environments with limited hardware availability. During our development and testing phases, we observed that Llama2 performed well without necessitating high-end infrastructure, making it a more practical choice for healthcare applications where cost and efficiency are essential considerations [103]. Also, a preliminary study on possibly using Llama3, showed problems in the fine-tuning phase (even with the dataset created with the changes from the documentation of Llama3) and the prompt phase, the outputs proved unreliable and confusing at times, displaying not only answers in different languages, as the appearance of random code in the middle on text, even after the prompt reinforced the output wanted. Another important think to have in mind, is that Llama2 aligns with responsible AI practices, particularly in sensitive domains such as healthcare. The model's development includes considerations for ethical compliance, ensuring that it can be deployed safely while adhering to data privacy regulations and maintaining patient confidentiality. This commitment to responsible AI usage was a critical factor in our decision to integrate Llama 2 into our drug recommendation system [8, 13].

In conclusion, for this dissertation, choosing Llama2 provides a well-rounded solution for building a DRecSys, combining high performance, efficiency, and ease of integration; its ability to deliver high-quality recommendations without requiring extensive computational resources or incurring high costs made it the optimal choice for our research.

Finally, we used Phi-3 from Microsoft model [92] for initial symptom analysis, the Semantic Key Retrieval (SKR) task. Microsoft's Phi-3-Mini-128K-Instruct is a lightweight yet powerful model, part of the Phi-3 family. With 3.8 billion parameters and training on 3.3 trillion tokens, it can handle up to 128K tokens in context, making it particularly useful for long-form reasoning and maintaining context. It was designed with safety and alignment in mind through supervised fine-tuning and preference optimization. Despite its smaller size, Phi-3 performs well on benchmarks, matching the capabilities of larger models like GPT-3.5, scoring 69% on Massive Multitask Language Understanding (MMLU) and 8.38 on MT-bench [92].

5 Implementation

This chapter begins with the implementation of EDA, which involves examining the dataset to uncover its structure and key characteristics. This is followed by Data Cleaning and Pre-processing, with a primary focus on the `review` attribute, detailing the steps taken to clean and prepare the data for further analysis. Next, SA is conducted to classify the polarity of the `review` attribute, determining the sentiment expressed in each review using tools such as VADER and TextBlob. Finally, the chapter concludes with an explanation of the construction of the DRecSys, outlining its design and workflow.

Figure 5.1 illustrates the workflow of the implementation for the first three steps. This diagram will be referenced in the following three sections (Figure 5.1. 1 – green, Figure 5.1. 2 – yellow and Figure 5.1. 3 – blue) to enhance visual understanding.

5.1 Exploratory Data Analysis

The dataset for the EDA (Figure 5.1. 1) has a size of (215.063, 7), this being the combination of all the data (train and test dataset together). Table 4.1 showed the number of instances for the five different and relevant attributes for the work to follow, since they will be considered for the fine-tuning phase of the DRecSys. This relationship between this attributes and the insights they may bring is now explored.

Figure 5.2 shows the relation between the percentage of reviews and the rating they have.

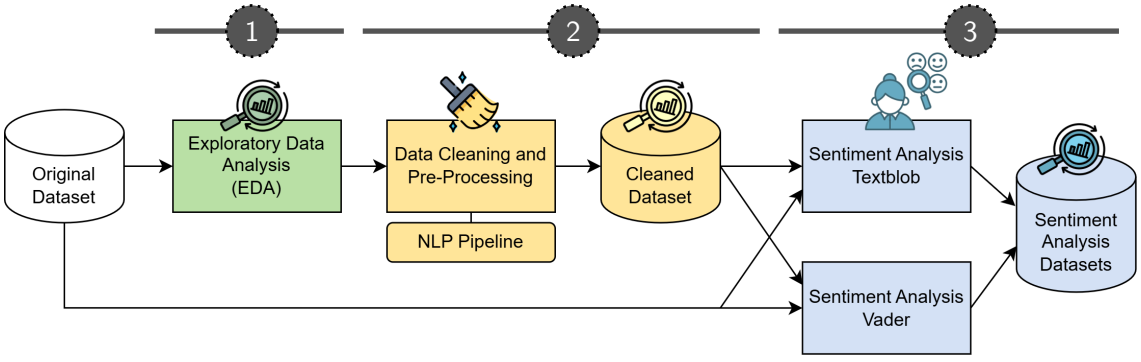


Figure 5.1: *Diagram of the workflow of the implementation of the first three steps*

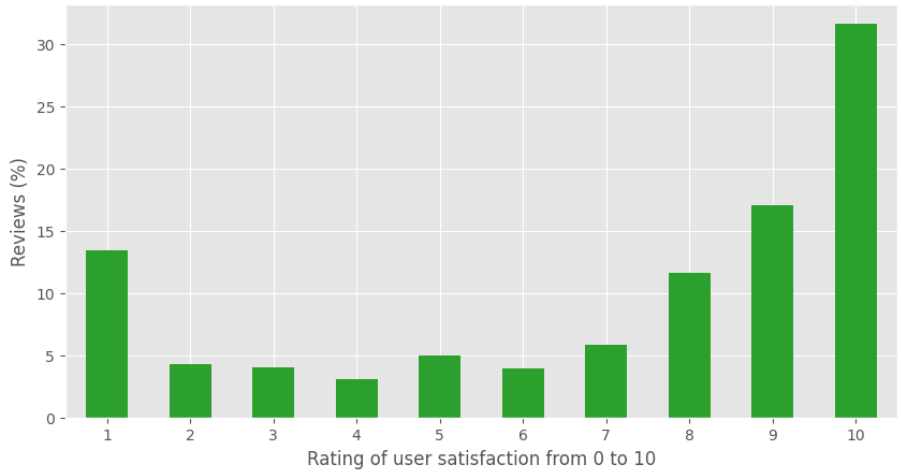


Figure 5.2: *Distribution of Review Counts by Rating (%)*

Drugs with rating 10 are the most reviewed, with a percentage of 31,62%, then 9 and 1 with 17,07% and 13,45%, respectively. This indicates a bias in the dataset toward more positive reviews compared to negative ones. While a higher prevalence of positive reviews can reflect the general satisfaction with certain drugs, it may introduce challenges for fine-tuning a DRecSys, as it may lead the model to overestimate the efficacy or appeal of drugs, potentially under-representing negative feedback or experiences, which are equally important for making balanced recommendations. Figure 5.3 shows the relation between the percentage of the attribute `UsefulCount` and the rating they have. Drugs with rating 10 have the reviews considered most useful, with a percentage of 42,32%, then 9 and 8 with 20,60% and 12,18% respectively, and then 1, with a percentage of 7,53%. While the dataset exhibits a bias toward positive reviews, the fact that users tend to find these higher-rated reviews more helpful lends some justification to this imbalance. Positive reviews likely reflect experiences where treatments were effective and side effects were minimal, which aligns with what most users seek in healthcare solutions. As such, emphasizing these reviews may still provide valuable insights and recommendations for users looking for effective treatments. These visualizations are helpful for understanding how users generally feel about a drug or experience of a user based on the number of the attribute `review` and `UsefulCount`.

Figure 5.4 and Figure 5.5 show the top 10 drugs based on the review count along with their average ratings, and the top 10 drugs ranked by `UsefulCount` alongside with their average ratings, respectively. The average for the top 10 most reviewed drugs is 6,67, whereas for those ranked by `UsefulCount` is 7,52. The top 10 drugs shown in the two figures are different, only sharing the drugs `Sertraline` and `Escitalopram` in different positions and `Phetermine` at number 7. With the intention to choose a reference of the 'Top 10 drugs' for future analysis, we will give more importance to the attribute `UsefulCount` since it has much more data than the attribute reviews (as it has "silent comments") and it has a bigger average mean rating. A higher average mean rating for drugs ranked by `UsefulCount` implies these drugs are associated with positive experiences that users found worth emphasizing. This aligns well with the goal of the RS, which aims to prioritize treatments that have a track record of success and user satisfaction, also, reviews with higher `UsefulCount` scores are more likely to contain

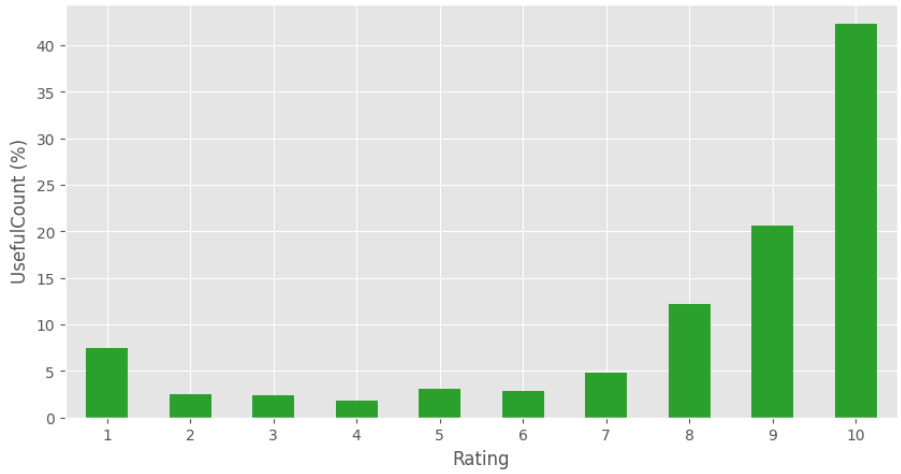


Figure 5.3: *Distribution of the Number of Helpful Votes by Rating (%)*

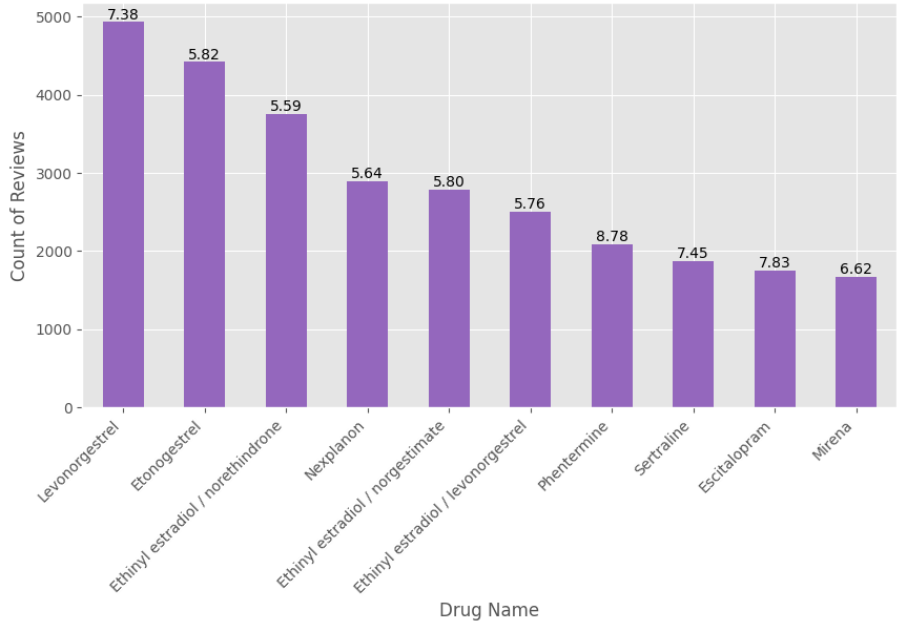


Figure 5.4: *Top 10 Most Reviewed Drugs with Their Mean Ratings*

detailed, actionable feedback and the higher mean rating among `UsefulCount`-ranked drugs can serve as a sign of reliability, as these ratings reflect not just individual experiences but collective validation from the user community ("silent comments"). Now, in Figure 5.6 we can see the top 10 conditions by `UsefulCount`. We can see that there is a combination of both physical and mental health conditions, reflecting the dataset's relevance to diverse areas of healthcare. For example, mental health conditions like Depression, Anxiety, and Bipolar Disorder are prominent alongside physical conditions such as Pain, Weight Loss, and High Blood Pressure.

Figures 5.7 and 5.8 show the number of unique conditions (`condition`) for the top 10 drugs by `UsefulCount` and the number of unique drugs (`drugName`) for the top 10 conditions by `usefulCount`, respectively. Analyzing the Figure 5.7, it shows that from the top 10 drugs by `usefulCount`, `Gabapentin` includes the greater number of conditions, this number being 31, then `Zoloft` with 21 and `Phentermine` with the least number of unique conditions

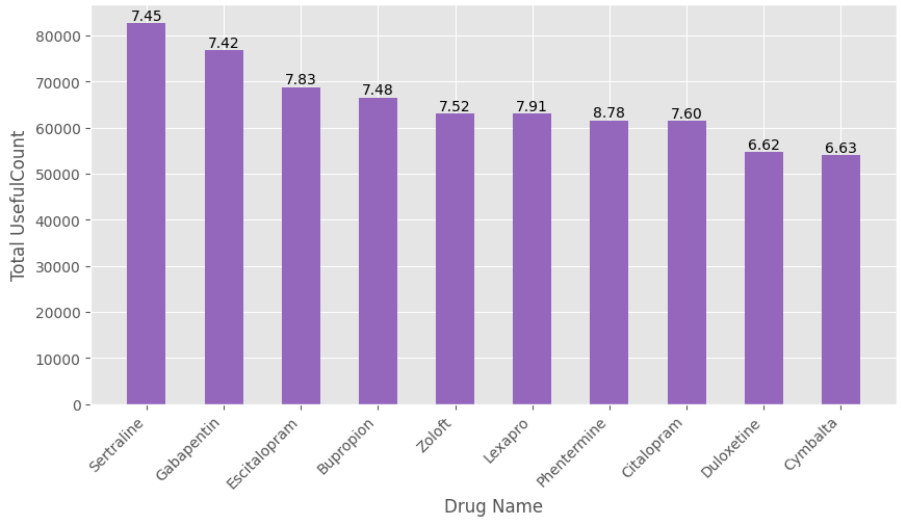


Figure 5.5: *Top 10 Drugs by UsefulCount and Their Mean Ratings*

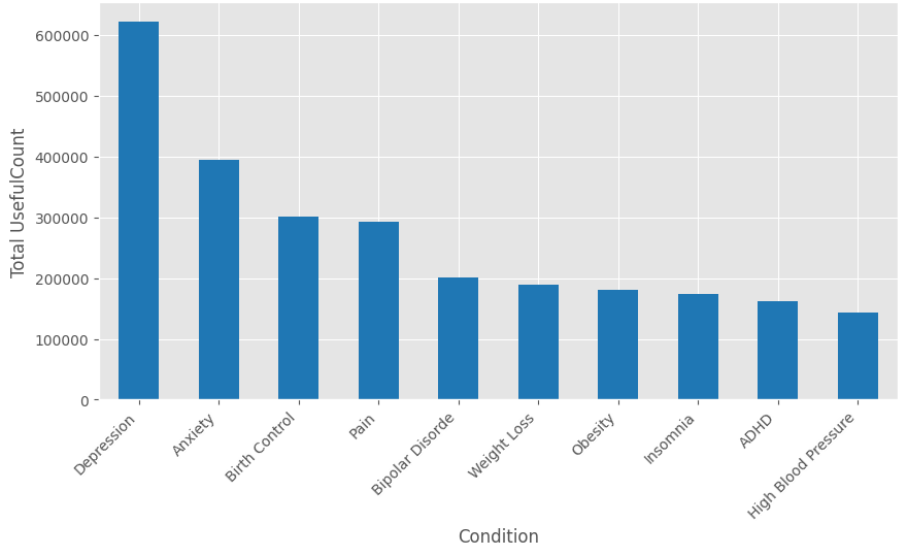


Figure 5.6: *Top 10 Conditions by UsefulCount*

(`condition`), this being only 3. Figure 5.8 shows that Pain, Birth Control and High Blood Pressure are the conditions with the greater number of unique drugs associated to them (from the top 10 conditions (`condition`) by `UsefulCount`) with 219, 181 and 146 unique drugs respectively. Weight Loss is the condition with the least number of unique drugs (`drugName`) only having 22 unique drugs associated. Even if some conditions or drugs have low `UsefulCount` values or fewer records, we chose not to remove them from the dataset. This is because every piece of data, even if less frequent, can still provide important insights or represent unique cases. By keeping all the data, we make sure the dataset stays diverse and reflects real-world situations. This way, the model can learn from all available information and find patterns on its own, without being influenced by removing less common entries. This approach helps create a RS that works for a wide range of conditions and drugs, no matter how common or rare they are in the dataset.

Furthermore, we explored the top 10 prevalent `drug-condition` pairs for the top 10 conditions

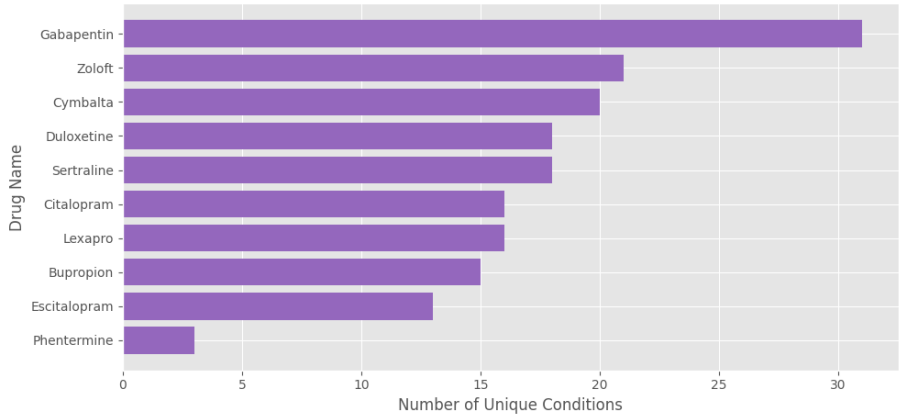


Figure 5.7: *Number of Unique Conditions for the Top 10 Drugs by UsefulCount*

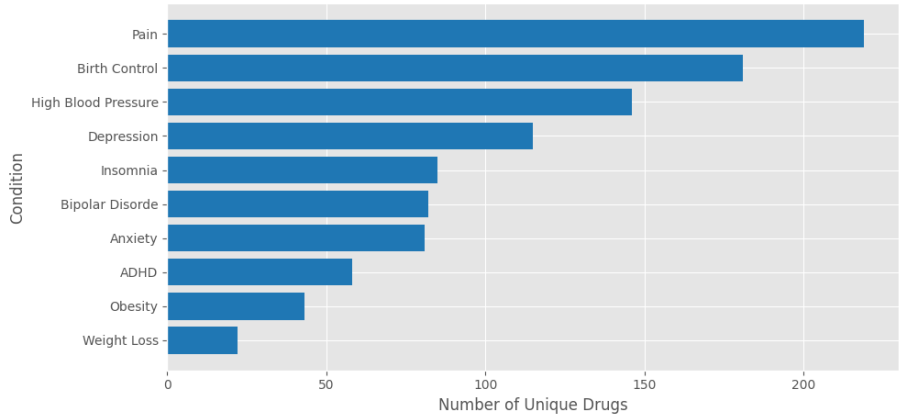


Figure 5.8: *Number of Unique Drugs for the Top 10 Conditions by UsefulCount*

(attribute `condition`) by `UsefulCount`, Figure 5.6, so for each condition we analyzed the top 10 drugs by `UsefulCount` and if these drugs were in the overall 'top 10 drugs by `UsefulCount`', Figure 5.5. A summary of the drugs common to both the top 10 drugs of the `condition` by `UsefulCount` and the overall top 10 drugs by `UsefulCount` can be seen in Table 5.1. It is important to know, that some drugs are common but simply are not presented in the top 10 drugs for the specific condition by `UsefulCount`. From this analysis, we can see that some drugs appear in both the top drugs for specific conditions and the overall top 10 drugs by `UsefulCount`, showing that these drugs are versatile and work for multiple conditions. On the other hand, some drugs in the overall top 10 do not show up in condition-specific lists, which suggests that these drugs might be more useful in general but not as commonly associated with any one specific condition.

The Figure 5.9 shows how useful users found a review compared to the rating it received. The rating with the highest total useful count is 10 (Total useful count: 254,846.4), while the rating with the lowest total useful count is 4 (Total useful count: 110,241). Some reviews stand out as outliers, especially for ratings of 10, where certain reviews received a much higher number of useful votes than others. This suggests that these reviews were particularly meaningful or relevant to users. Even at lower ratings, a few reviews also received high usefulness votes, showing that negative reviews can still be very helpful. These differences show that the content of a review matters a lot, beyond just its rating.

The Figure 5.10 illustrates the relationship between usefulness and review length. The review with the highest usefulness count, at 187 words, demonstrates that shorter but informative reviews tend to be the most appreciated by users. Additionally, very few reviews exceeded 750 words, and those that did did not receive particularly positive user feedback. The Figure 5.10 also highlights the presence of outliers, with some extremely short reviews achieving unusually high usefulness counts. After examining these outliers, it was observed that they often had high ratings and were written in a clear and easy-to-understand manner, making them particularly impactful despite their brevity. This observation was considered when constructing the instructions for the output explanation of the DRecSys. It indicates that shorter and more concise explanations might be better received by users, reflecting their preference for clarity and straightforward communication.

In the Table 5.2 we can see the found to be most useful review by the users for three of the drugs present in the top 10 drugs by `UsefulCount`, this being: (1) `Sertraline` – a review with a rating 10 and UsefulCount of 1291; (2) `Gabapentin` – a review with a rating of 10 and a `UsefulCount` of 500; (3) `Escitalopram` – a review with a rating of 9 and a `UsefulCount` of 458 . In these reviews, we can observe positive feedback highlighted in green and mentions of side effects highlighted in blue. The reviews often have a personal and relatable tone, with some being sentimental and providing personal background about the user's experience before taking the drug, not only reflecting on the practical effects of the drug but also saying the emotional and contextual impact it has on users. Incorporating this information into a DRecSys is important, as some insights are not straightforward or easily quantified. The ability to analyze and interpret such feedback allows for more empathetic and informed healthcare recommendations, making the system a valuable tool for understanding real-world user experiences.

Finally, Figure 5.11 presents the review sentiment over time (2008 until 2018), analyzing closely the sentiment over the years we observed that in 2008 there was only 2% of negative comment, increasing to around 18% in the years on 2009 to 2012. In the years of 2013 and 2014 there was a decrease of negative comments rounding the 12% and then an exponentially increased in the years 2015 to 2017, with the percentages, 25.8%, 34.9% and 36.9%, respectively. This trend

Table 5.1: Prevalent top 10 drug-condition pairs

Condition	Common Drugs
Depression	Bupropion, Sertraline, Escitalopram, Citalopram, Duloxetine, Cymbalta
Anxiety	Escitalopram, Lexapro
Birth Control	None
Pain	Gabapentin
Bipolar Disorder	None
Weight Loss	Phentermine
Obesity	Phentermine
Insomnia	None
ADHD	None
High Blood Pressure	None

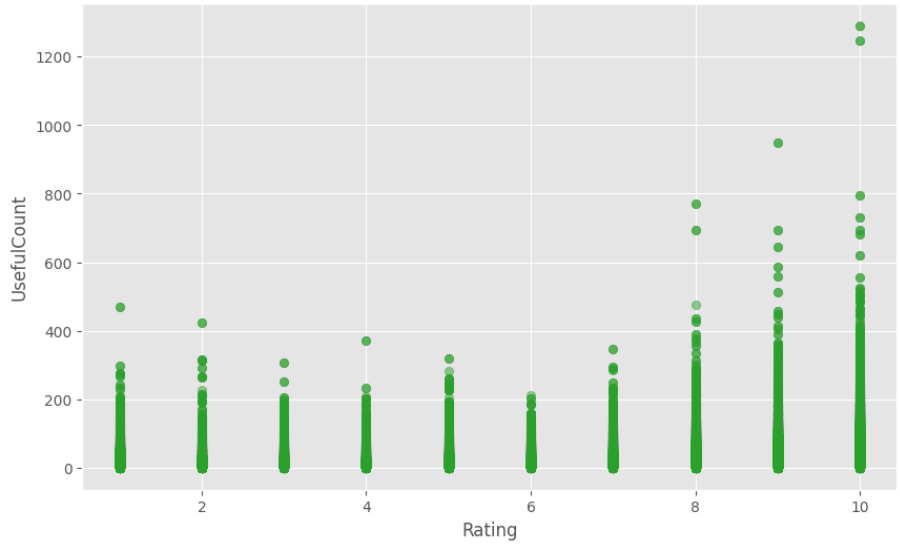


Figure 5.9: *Relationship Between the Number of Helpful Votes and Ratings*

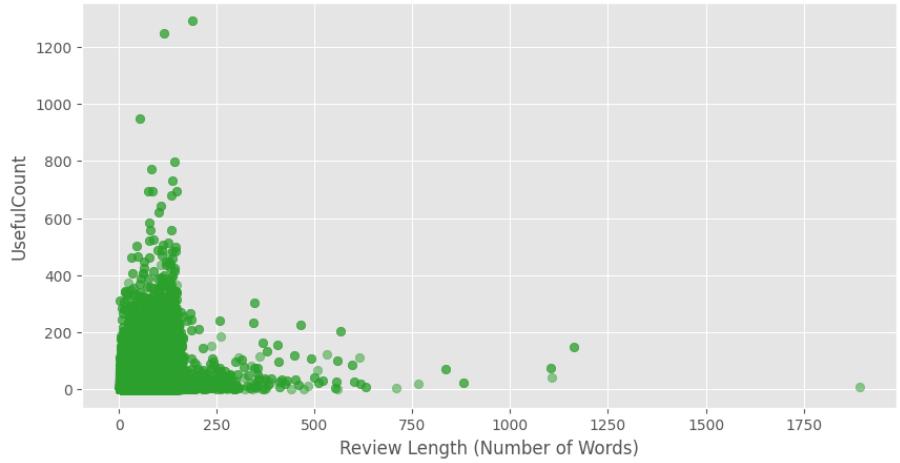


Figure 5.10: *Relationship Between the Number of Helpful Votes and Review Length*

in sentiment over time could be influenced by external factors like changes in drug availability, public opinion, or healthcare policies that affected user experiences and reviews. The sharp rise in negative comments after 2015 might point to growing dissatisfaction with certain drugs, increased awareness of side effects, or simply more users sharing their feedback online.

5.2 Data Cleaning and Pre-processing

A thorough data cleaning and pre-processing step was initiated to prepare the dataset for deeper analysis (Figure 5.1.2) . Initially, we identified and removed corrupted rows based on inaccuracies on the `condition` attribute. These 900 entries (in the training dataset) and 271 entries (in the test dataset) incorrectly contained user feedback counts instead of valid medical conditions, which could have skewed the dataset's reliability.

Following this, we enhanced the textual data (attribute `review`) through a pre-processing pipeline. This involved converting all text to lowercase, stripping `HTML` entities, and removing punctuation. We also filtered out stopwords (present in the library from NLTK) and applied

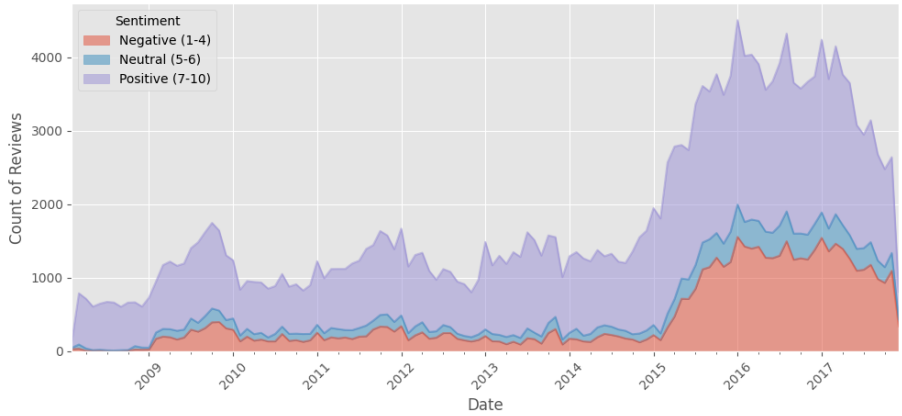


Figure 5.11: *Trends in the Sentiment behind a Review Over Time*

lemmatization to normalize the words, preserving the relevant parts of the review while maintaining the sentiment behind it. Additionally, we removed the attribute `rating` and introduced the new column featuring the processed reviews.

In the cleaning process, we thought about the possibility of removing drugs with a low number of reviews. However, we decided against it because 32.42% of the dataset (both train and test) consists of drugs with just one or two reviews. Removing these would have reduced the variety in the dataset and could have caused us to lose important insights from less commonly reviewed drugs. Our goal is to determine if the pre-processing pipeline adds value when working with a tool like an LLM. Specifically, we want to see if these steps are necessary or if LLM can perform just as well with raw, unstructured text. Pre-processing was applied to all reviews, even those with fewer entries, to create a consistent dataset for evaluation. One advantage of pre-processing is that it significantly reduces the length of the reviews, which can improve efficiency during fine-tuning and testing. This analysis will help us weigh the benefits and drawbacks of using raw versus processed data for NLP tasks with modern language models.

Finally, we saved the cleaned data into two files, one for the training data and one for the test data (Figure 5.1.3), ensuring it was well-structured for the subsequent analysis. In the Table 5.3 we can see a specific review after the pre-processing pipeline for the drug `Abilify`. The highlighted colors help emphasize key aspects of the review: purple marks the drug name, green highlights positive feedback, and blue indicates mentions of side effects.

5.3 Sentiment Analysis

In the SA step (Figure 5.1.3), firstly essential functions were defined for rescaling sentiment scores and labeling sentiment based on these rescaled scores. Following this, we implemented a function designed to process the dataset by doing SA using both TextBlob and VADER lexicon-based tools. This function adjusts the sentiment scores, applies labels to them and in the end saves the results into separate files. We executed this function twice to accommodate both the raw and cleaned versions of the dataset. This workflow resulted in four output files for the train dataset and four files for the test dataset, capturing the SA results – from both TextBlob and VADER for each data type (raw or clean).

Three new columns were then added to each of the eight dataset depending on the tool used for SA – TextBlob or VADER:

(1) `TextBlob_score or VADER_score` – Polarity score from TextBlob or VADER, ranging from -1 (most negative) to 1 (most positive);
(2) `TextBlob_rescaled or VADER_rescaled` – Re-scaled TextBlob or VADER score ranging from 0 to 10;
(3) `TextBlob_label or VADER_label` – Sentiment label for TextBlob or VADER scores (Negative, Neutral, Positive) based on the rescaled score.

Table 5.2: Three most useful reviews of the three drugs with the most UsefulCount

Drug	Review
Sertraline	*I remember reading people's opinions, online, of the drug before I took it and it scared me away from it. Then I finally decided to give it a try and it has been the best choice I have made. I have been on it for over 4 months and I feel great. I'm on 100mg and I don't have any side effects. When I first started I did notice that my hands would tremble but then it subsided. So honestly, don't listen to all the negativity because what doesn't work for some works amazing for others. So go based on yourself and not everyone else. It may be a blessing in disguise. The pill is not meant to make you be all happy go lucky and see "butterflies and roses", it's meant to help put the chemicals in your mind in balance so you can just be who you are and not overly depressed. I still get sad sometimes, but that is normal, that is life, and it's up to people to take control to make a change. I did so by getting on this pill.*
Gabapentin	*Neurontin has changed my life dramatically. I suffered from high anxiety with panic and anxiety attacks. I started this medicine at about 19 and had been off and on it the past 13 years. It literally has made me a new person. When I don't use it I notice a lot of fatigue, worry, anxiety, etc. When I am using this drugs, I can concentrate on the better side of things and am not so negative. In the beginning it did have some side effects such as a "loopy"feeling. Over time that did pass, and now I can also sleep much easier, and my carpal tunnel syndrome is dramatically better. Before I felt very snappy and impatient, now I handle things with a much more calm disposition. Love it.*
Escitalopram	*I wasn't severely depressed but I was always CRAZY sensitive, flipped out and cried at the smallest things, had pretty bad anxiety for no reason and would think negative thoughts over and over until I couldn't sit with myself. I finally decided to see a psychologist who recommended I have a psychiatric evaluation- in which they prescribed Lexapro. They started me on 10mg for two weeks and bumped me up to 20mg, which I have been taking for the last 2 months. I can't even express the immense change I experienced. I have never felt so content in my life. It never changed my personality in a drastic way, but I feel very confident, I can brush things off easily, I don't feel stuck in life, I can actually have fun. I LOVE this medicine.*

Table 5.3: Comparison of raw and processed/Clean review texts for the drug Abilify

Type	Content
Raw Review	"*Abilify changed my life. There is hope. I was on Zoloft and Clonidine when I first started Abilify at the age of 15. Zoloft for depression and Clondine to manage my complete rage. My moods were out of control. I was depressed and hopeless one second and then mean, irrational, and full of rage the next. My Dr. prescribed me 2mg of Abilify and from that point on I feel like I have been cured though I know I'm not. Bi-polar disorder is a constant battle. I know Abilify works for me because I have tried to get off it and lost complete control over my emotions. Went back on it and I was golden again. I am on 5mg 2x daily. I am now 21 and better than I have ever been in the past. Only side effect is I like to eat a lot.*"
Processed Review	"abilify change life hope zoloft clonidine first start abilify age 15 zoloft depression clondine manage complete rage mood control depress hopeless one second mean irrational full rage next dr prescribe 2mg abilify point feel like cure though know im bipolar disorder constant battle know abilify work try get lose complete control emotion go back golden 5mg 2x daily 21 good ever past side effect like eat lot"

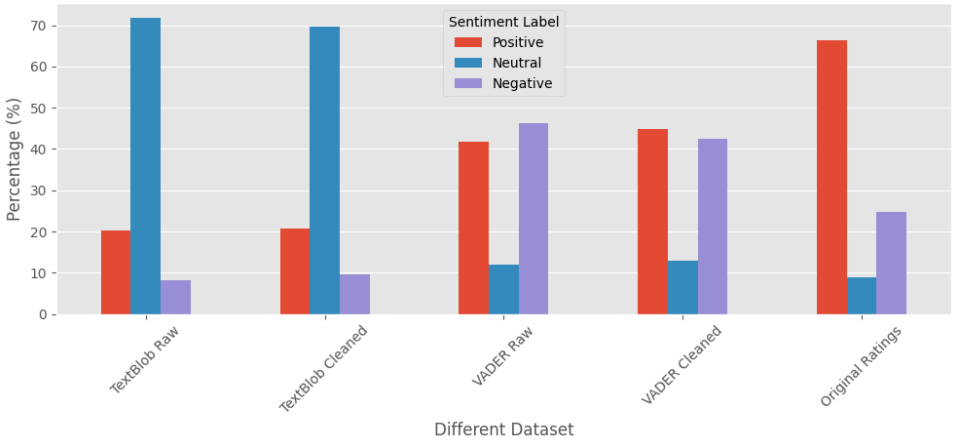


Figure 5.12: *Comparison of Sentiment Classifications and Ratings in Different Datasets*

Through this SA, we observed that VADER provided more distinct and nuanced results compared to TextBlob (Figure 5.12). In particular, TextBlob was found to have a tendency to categorize a significant proportion of ratings as 'neutral'. This has the potential to obfuscate the insights that can be gained from the data (Figure 5.13). VADER's analysis, on the other hand, offered a better differentiation of sentiment, allowing for a more detailed understanding of user perceptions and experiences on the drug corresponding to the specific review (Figure 5.14). Additionally, VADER distinguishes itself at identifying negative reviews, a crucial feature for the healthcare sector where identifying potential issues, such as side effects, is as important as highlighting positive outcomes, and since we found the bias of positive reviews in the EDA, this helps even the ground.

This capability makes VADER a preferable tool in the contexts of this dataset, and this work, where distinguishing between subtle sentiment tones is crucial for data interpretation and decision-making. Both tools gave ratings less positive comparing with the users original rating.

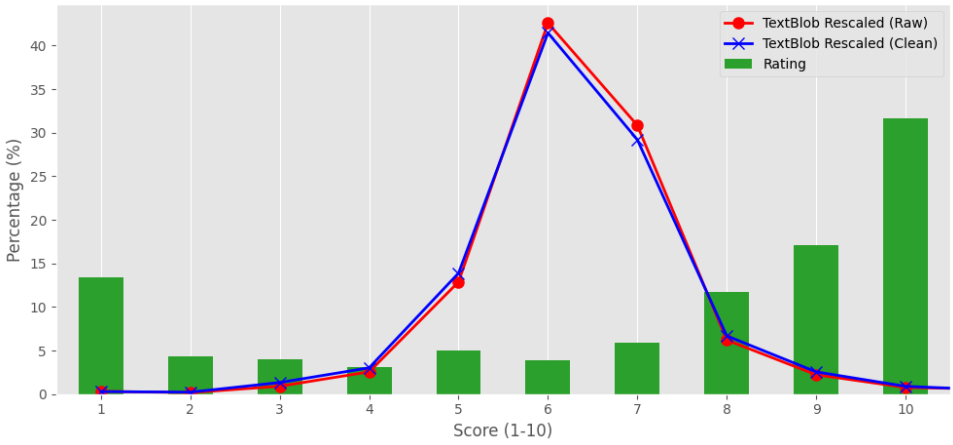


Figure 5.13: *Comparison of Ratings Frequency and TextBlob Rescaled Scores*

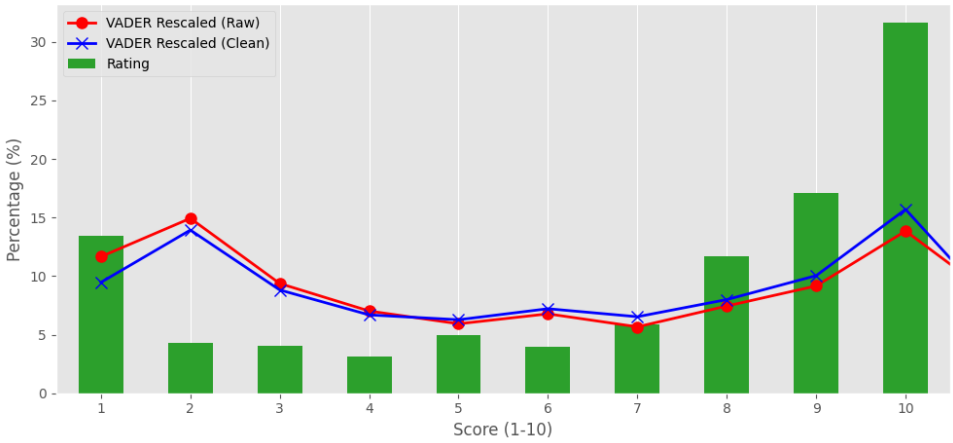


Figure 5.14: *Comparison of Ratings Frequency and VADER Rescaled Scores*

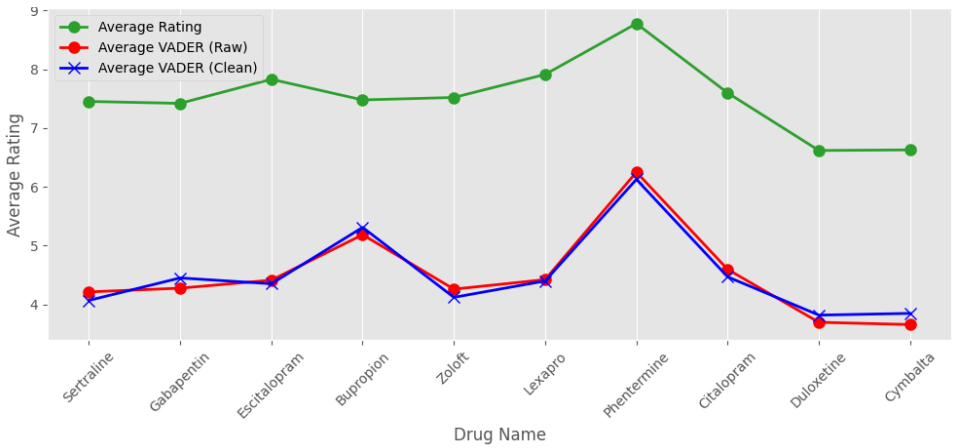


Figure 5.15: *Comparative Analysis of the Top 10 Drugs: Average Ratings vs. VADER Sentiment Scores*

This observed discrepancy between user ratings and SA scores underscores the complexity of interpreting review data. The difference suggests that while users may rate drugs highly overall, their reviews often contain mixed sentiments, such as positive outcomes tempered by side effects, Figure 5.15. This highlights the importance of using tools like VADER to extract meaningful insights beyond the ratings we already had, ensuring that the RS reflects both benefits and challenges of treatments. Integrating these findings with LLM-driven analysis could further enhance the system's ability to interpret nuanced feedback and provide more accurate, context-aware recommendations, connecting between structured SA and advanced language modeling.

5.4 Text Generation

Building upon the previous results, this final phase focuses on implementing the workflow for text generation and evaluating different variants of DRecSys. These models were fine-tuned and tested on the original dataset and VADER ratings, both in clean and raw versions, with and without the SKR task. The decision to use VADER over TextBlob was guided by the SA step, which demonstrated that VADER provided more nuanced sentiment differentiation, particularly in detecting negative feedback—an essential factor for healthcare applications. Before implementing the text generation workflow in DRecSys, several preparation phases were required. These included structuring datasets, securing computational resources, fine-tuning the models, and defining effective prompt templates. These phases were crucial for aligning the system with the proposed methodology.

The first phase of preparation focused on dataset construction. Particular attention was given to structuring the dataset input for the fine-tuning of the LLM, Llama2, to ensure accurate model training. Additionally, a separate dataset was designed specifically for the SKR task to enhance its effectiveness.

Another critical aspect was ensuring access to adequate computational resources. Fine-tuning LLM models is resource-intensive, requiring substantial storage for datasets and model checkpoints. To optimize efficiency, we utilized Llama2 with 7 billion parameters, balancing computational cost and performance. The fine-tuning process was conducted on Google Colab servers using Quantized Low-Rank Adaptation (QLoRA), significantly reducing computational overhead. Additionally, open-source tools—including the Hugging Face Transformers, Parameter-Efficient Fine-Tuning (PEFT) (for LoRA implementation), BitsAndBytes (for model quantization), and Transformers Reinforcement Learning (TRL) (for training loop management)—were integrated to streamline the process.

Finally, implementing the workflow required the design of input-output prompt templates. These templates served as the connection between user queries with user information and the model's generated responses, ensuring recommendations were relevant, personalized, and consistent while maintaining the ranked list of drugs and user-friendly explanation format.

Through these structured phases, the DRecSys variants were optimized to generate personalized drug recommendations, providing valuable insights into system implementation to enhance healthcare decision-making.

5.4.1 Dataset Construction

The dataset construction phase was a critical foundation for the development of DRecSys, ensuring that the data was properly prepared and aligned with the objectives of the fine-tuning and SKR tasks. This phase focused on transforming the raw and preprocessed datasets into formats optimized for use with LLM, in specific Llama2, also considering the unique requirements we wanted for the DRecSys. With the development of this phase two main datasets were built: one for fine-tuning the model and another forSKR to be use during the prompt and results phase of this final step.

For the fine-tuning task, we structured the datasets to align with the `<prompt, completion>` format, commonly used for training language models and aligned with what the Llama2 required. In this case, each data point was divided into a prompt and a completion. The function `format_rows(row)` is used to create prompts and completions for fine-tuning. It generates the following:

1. **Prompt**:

```
"User review on the drug [drugName] for [condition]"
```

2. **Completion**:

```
"User gave it a rating of [rating]:\n[review]. [usefulCount]
people found this review useful."
```

The placeholders `[drugName]`, `[condition]`, `[rating]`, `[review]`, and `[usefulCount]` are filled with values from the corresponding fields in the dataset. The function returns a structured output containing the prompt and completion in a format suitable for fine-tuning tasks.

The original dataset was used, with minimal cleaning–only corrupted rows were removed. The SA tool VADER was then applied to both the clean and raw versions of the dataset, as we want to evaluate the step of cleaning and pre-processing. Structuring the datasets in this format ensures compatibility with Llama2, as it mirrors the structure used during the pre-training phase. This format helps the model associate user queries with the corresponding responses, enhancing its ability to generate accurate, context-aware recommendations. For each dataset, the prompt was formatted as a user review on a drugs for a specific condition, and the completion included the user's `rating`, `review` text and `Usefulcount`. For the VADER datasets, the attribute `rating` was replaced by `VADER_rescaled` to reflect the adjusted sentiment-based ratings. One of the key aspects of this phase was ensuring that the datasets maintained a balance between preserving user feedback and structuring the data to enhance the model's learning efficiency. For instance, reviews had to be paired with other data points such as drug names, conditions, ratings and the attribute `UsefulCount`, without over complicating the input data for the fine-tuning of the LLM. This ensured that Llama2 could learn to provide context-aware drug recommendations. In Table 5.4 we can see the first line of the dataset

`train_ready_fine_tuning_or`, a user Review on `Valsartan` for Left Ventricular Dysfunction. The example highlights how the prompt provides context, while the completion combines sentiment and user generated feedback.

Table 5.4: User Review on Valsartan for Left Ventricular Dysfunction

Prompt	Completion
User review on the drugs Valsartan for Left Ventricular Dysfunction	"'User gave it a rating of 9:""It has no side effect, I take it in combination of Bystolic 5 Mg and Fish Oil"". 27 people found this review useful.'"

This process resulted in six datasets for fine-tuning, each creating different variants of the original and VADER–processed data. By including both cleaned and raw datasets, we aim to evaluate whether pre-processing improves the model's performance or if raw data alone can achieve similar results and in the SA had a positive impact as well.

For the SKR task that occurs during the prompt for text generation and results phase, we constructed datasets based on the `<condition, ranked drugs>` format. It included the `condition` attribute and the list `<ranked the drugs>` based on user ratings in descending order, theSKR task is designed to serve as a possible support to the DRecSys by providing contextually relevant information during text generation and it was designed to ensure that the DRecSys can prioritize drugs effectively, aligning with the drugs present in the dataset and the ranking associated in the same dataset. This task will be better evaluated in the rest of this chapter. In Table 5.5, we display the first line of the `condition_drugs` dataset, showing the ranked drugs associated with the condition Agitation.

Table 5.5: Condition Agitation with the ranked drugs in descending order associated with the specific condition Agitation

Condition	Ranked Drugs
Agitation	['Olanzapine', 'Citalopram', 'Olanzapine', 'Zyprexa Intramuscular', 'Risperidone', 'Loxapine', 'Citalopram', 'Risperidone']

The datasets required for both the fine-tuning and SKR task were generated locally, following the same process used in the previous steps.

5.4.2 Fine-tuning task

For the fine-tuning process to be carried, first we had to acess the model, to access Meta's LLM for this task, we used Hugging Face, where we requested the model after creating and account and were granted access to the following models within three hours:

(1) `"meta-llama/Llama-2-7b-chat-hf"` [79];
(2) `"meta-llama/Meta-Llama-3-8B"` [7].

Once access was secured, we proceeded with the fine-tuning process by loading the base models and the relevant datasets. Following the methodology outlined in our proposal, we implemented

specific parameter configurations (discussed later on in the chapter) and began the supervised fine-tuning process. Additionally, the GPU A100 with 40G VRAM proved necessary for handling the model's complexity, especially for optimizing memory and computation speed. The fine-tuning of each model took approximately 6 hours, with a batch size of 3, which was determined after iterative testing for first, the possibility to fine-tune and then efficiency.

The fine-tuning was done using the QLoRA methodology, which allows for efficient fine-tuning by introducing LoRA within the model's attention mechanism. The LoRA attention dimension (`lora_r = 64`) was set as 64 (balance between efficiency and performance) the alpha parameter for LoRA scaling as 16 (`lora_alpha = 16`) and the dropout porbability for LoRA layers as `0.1` (`lora_dropout = 0.1`). These parameters control the behavior and efficiency of the LoRA adaptation layers added to the original model [33]. We also used a 4-bit quantization strategy, which balances performance and memory usage. Specifically, the base model was loaded in 4-bit precision, using the `nf4` quantization format, known significantly reducing memory requirements while preserving model accuracy, so it provided better accuracy for LLM tasks. We maintained the nested quantization (or double quantization) as `False` (`use_nested_quant = False`), favoring stability and simplicity over maximum memory savings. We opted for bfloat16 precision training (`bf16=True`) due to the compatibility with the A100 GPU and ensuring reasonable precision during calculations despite 4-bit storage. When configuring the training arguments, the fine-tuning parameters were carefully selected to ensure the models were optimized for our specific tasks. Training was conducted over a single epoch (`num_train_epochs=1`), following best practices possible for efficient fine-tuning of large models with large datasets in limited-resource settings [61].

This approach allowed us to adapt a high-capacity model like Llama2 to our domain-specific tasks without requiring large hardware resources, without QLoRA and the minimization of the epoch and even the choice to use Llama2 7b, the fine-tuning would have been very expensive and time consuming.

Throughout the fine-tuning process, several challenges emerged. Initially, we opted to use Google Colab due to its availability of free GPUs and RAM. However, we realized that stable and efficient training required access to higher-performance hardware, specifically an A100 GPU. When attempting to use the T4 GPU, the training time was estimated to exceed 36 hours, and the memory limitations were prohibitive, even with memory-optimization techniques like QLoRA. Another significant challenge was maintaining continuous server availability. Given that each training session lasted several hours, any disruption or loss of connection would terminate the process, requiring a complete restart. This necessitated constant monitoring to avoid unnecessary delays. The fine-tuning task of this dissertation can still be replicated using the free tools available on Google Colab. However, for guaranteed stability and significantly faster performance, we opted to use the A100 GPU, which consumed approximately 11.77 units per hour.

The training resulted in three fine-tuned models, one with the contents of the original dataset, and two with VADER ratings and raw or clear reviews. Each fine-tuned system is composed by three files, the `adapter_config.json`: Contains the configuration for model adapters; the `adapter_model.safetensors`: Stores the fine-tuned model weights; and the file `README.md`:

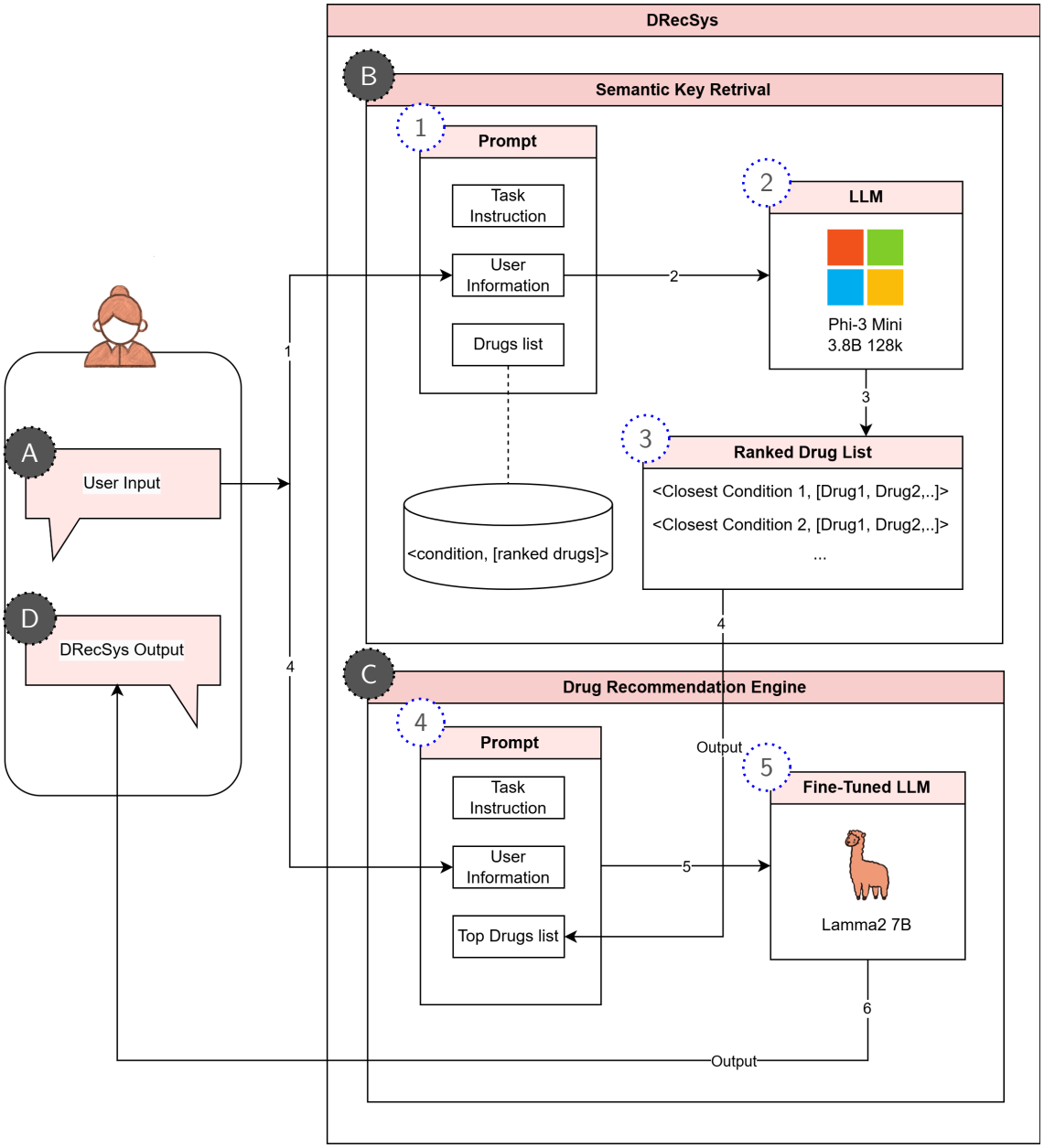


Figure 5.16: *Diagram of the workflow of the implementation of a DRecSys, fourth step*

Documentation related to the specific model and fine-tuning process.

5.4.3 DRecSys Workflow

The final phase of this work focused on developing key prompts and the DRecSys workflow. This phase was consolidated into a single file containing all necessary functions, serving as the core of our DRecSys and guiding the execution of each component to ensure the effective delivery of drug recommendations.

The workflow of the DRecSys implementation can be observed in Figure 5.16, showing the progression from the user's initial input through the system's processing stages to the final output. This diagram highlights how each element in the system connect to generate recommendations.

The process starts with the User Input (Figure 5.16. A), where the user provides personal details including age, sex, current drugs, symptoms, and other relevant medical details it finds useful to say. These inputs are then processed by the first component of the system–the Semantic Key Retrieval (SKR) task (Figure 5.16. B), which uses a custom-designed prompt (Figure 5.16. 1) to interpret and analyze the user input. Its important to note that the efficiency of this task is later discussed. This initial prompt includes three elements: the task instructions, user-specific data extracted from the input (age, sex and symptoms) , and a comprehensive list of drugs from the dataset. The prompt is then fed into the Phi-3-Mini LLM (Figure 5.16. 2), a LLM from Microsoft (3.8B, 128k), which conducts the semantic analysis of the user's symptoms and other details. The model identifies the conditions most closely matching the user's input by analyzing similarities between the provided symptoms and the conditions within the original dataset. The output is a Ranked Drug List (Figure 5.16. 3), highlighting the drugs most relevant to the identified condition. The implementation of this task was effective, and Phi-3-Mini, despite being a much smaller LLM compared to Llama2, demonstrated a solid performance in completing the task successfully. Once the model identifies the most relevant conditions, these are passed along to the next phase for drug recommendations.

After generating the ranked drug list, the system transitions into the Drug Recommendation Engine (Figure 5.16. C). In this stage, there is a second prompt designed specifically for the fine-tuned LLM–Llama2 a LLM from Meta (Figure 5.16. 4). This second prompt contains also three elements: task instructions, all of the user's personal information, and the top-ranked drugs produced by the SKR process. The Llama2 model processes this input, producing the DRecSys Output (Figure 5.16. D), which provides the final list of drug recommendations and a brief explanation personalized to the user's needs.

It is also important to reinforce that the Drug Recommendation Engine (Figure 5.16. C) uses a carefully designed prompt for the fine-tuned LLM (Figure 5.16. 4). This prompt was specifically created to ensure accurate, clinically relevant drug recommendations that directly align with the identified conditions or user systems. The design of this prompt prioritizes personalization, taking into account not only the user's symptoms but also their age, sex, current drugs, and other relevant health details, all of which can be critical factors in recommending appropriate treatments.

The development process of both prompts involved iterative refinement, with continuous evaluations and adjustments based on preliminary results. This iterative approach ensured that the final versions of the prompts were effective and personalized to each task, thus enhancing the system's overall accuracy and reliability. With this tests, the output of the DRecSys (Figure 5.16. D) became increasingly consistent as the prompts and model responses were refined. Achieving a high level of consistency was critical for ensuring that the system's output was reliable and easy to interpret. A uniform output structure also makes it possible for users to efficiently extract and analyze relevant information, without the distractions caused by variations in response formats. This consistency significantly improved the clarity and usability of the system.

Focusing again in the prompt for the SKR task (Figure 5.16. 1), which focused on identifying and extracting key semantic elements from the user's input, it required fewer modifications

due to the relative simplicity of the task, it involved a more straightforward design, allowing for greater flexibility in the prompt while still ensuring that the results were accurate and relevant. There is a contrast in complexity between the SKR Task and the Drug Recommendation Task highlights the varying degrees of refinement needed across the workflow, with more complex tasks like drug recommendation requiring much more careful changes.

The two prompt functions for the DRecSys and their specific prompts for each task just metioned are detailed below:

1. **Prompt Function for the Fine-Tuned LLM (Figure 5.16.4)**:

   a) **Function Definition**:

   ```
   # Prompt function for the fine-tuned LLM
   def prompt_model(prompt):
   ```

   b) **Pipeline Initialization**:

   ```
   pipe = pipeline(task="text-generation", model=model,
   tokenizer=tokenizer, max_new_tokens=800)
   ```

   c) **Prompt Processing**:

   ```
   result = pipe(f"<s>[INST] {prompt} [/INST]")
   ```

   d) **Return Statement**:

   ```
   return result[0]['generated_text']
   ```

2. **Prompt Function for the Sentiment Key Retrieval LLM (Figure 5.16.1)**:

   a) **Function Definition**:

   ```
   # Prompt function for the semantic key task using the
   Phi-3-mini LLM
   def prompt_tiny_model(prompt):
   ```

   b) **Pipeline Initialization**:

   ```
   pipe = pipeline(task="text-generation", model=tiny_model,
   tokenizer=tiny_tokenizer, max_new_tokens=275)
   ```

   c) **Prompt Processing**:

   ```
   result = pipe(f"<s>[INST] {prompt} [/INST]")
   ```

d) **Print Result**:

```
print(result)
```

e) **Return Statement**:

```
return result[0]['generated_text']
```

In both functions above, it was crucial to explicitly define the task as `text generation`, utilizing the format `<s>[INST] {prompt} [/INST]`. Additionally, careful attention was given to correctly setting the maximum token limit to control the length of the generated output.

1. **Prompt for the Semantic Key Retrieval task (Figure 5.16.1)**:

a) **Task Description**:

```
You are a highly knowledgeable medical assistant. Based on
the symptoms provided, identify the medical conditions that
most closely relate to these symptoms.
```

b) **Input Information**:

```
Age: <age>
Sex: <sex>
Symptoms: <symptoms>
```

c) **Condition Constraints**:

```
Your response can only include conditions that are present
in the following list:
<', '.join(condition_medication_dict.keys())>
```

d) **Output Format**:

```
Provide the output in the format of a ranked list of
conditions relevant to the symptoms provided. The output
can only include the ordered list with the following format,
with no additional text before or after:

   i.

   ii.

  iii.
```

2. **Prompt for the Fine-Tuned LLM Task (Figure 5.16. 4)**:

   a) **Task Description**:

   ```
   You are a highly knowledgeable medical assistant. Your task
   is to recommend medications based on the patient's details
   and symptoms. The output must strictly follow the exact
   format below without deviation.
   ```

   b) **Strict Guidelines**:

   ```
   Do not include any extra words, explanations, or greetings.
   The response must contain only two sections:
   ```

   c) **Section 1: Ranked List of Medications**:

   ```
   Provide a simple, numbered list of medication names in the
   following format:

      i. <medication 1>

     ii. <medication 2>

    iii. <medication 3>
   ```

   d) **Section 2: Brief Explanation**:

   ```
   After the medications list, provide a brief and
   standardized explanation for why these drugs were chosen,
   including considerations such as medication interactions,
   patient age, sex, and specific symptoms. Do not forget,
   these medications ensure the patient's physical and mental
   well-being.
   ```

   e) **Patient Details**:

   ```
   Age: <age>
   Sex: <sex>
   Current medications: <current_medications>
   Symptoms: <symptoms>
   Other Relevant Details: <other_details>
   ```

f) **Medication Options to Consider**:

```
Here are medications to consider:
<', '.join(recommended_medications)>
```

1. **Prompt for Fine-Tuned LLM Task in DRecSys**:

a) **Task Description**:

```
You are a highly knowledgeable medical assistant. Your task
is to recommend medications based on the patient's details
and symptoms. The output must strictly follow the exact
format below without deviation.
```

b) **Strict Guidelines**:

```
Do not include any extra words, explanations, or greetings.
The response must contain only two sections:
```

c) **Section 1: Ranked List of Medications**:

```
Provide a simple, numbered list of medication names in the
following format:

  i. <medication 1>

 ii. <medication 2>

iii. <medication 3>
```

d) **Section 2: Brief Explanation**:

```
After the medications list, provide a brief and
standardized explanation for why these drugs were chosen,
including considerations such as medication interactions,
patient age, sex, and specific symptoms. Do not forget,
these medications ensure the patient's physical and mental
well-being.
```

e) **Patient Details**:

```
Age: <age>
Sex: <sex>
Current medications: <current_medications>
Symptoms: <symptoms>
Other Relevant Details: <other_details>
```

f) **Medication Options to Consider**:

```
Here are medications to consider:
<', '.join(recommended_medications)>
```

Looking ahead to the evaluation of the effectiveness of the SKR present in the next chapter, we simple do not use the existing code designed for that task and for the fine-tuned LLM Prompt, the evaluation excludes the list of drugs to consider.

6 Evaluation & Result Analysis

In this chapter, we begin by presenting the DRecSys configurations and the experimental methodology, this is followed by three analyses: the evaluation of the generated drug lists, an analysis of the system's conversational behavior, and a discussion of further observations and insights.

To evaluate the performance of the DRecSys systems, six different configurations were tested, each representing a unique combination of dataset preparation with a workflow methodology for the DRecSys, mentioned and implemented in the earlier steps, this configurations are:

- **Original Dataset with SKR** (Figure I.1): used the semantic key retrieval task for structured mapping.
- **Original Dataset without SKR** (Figure I.2): Omitted the retrieval task.
- **Clean VADER Dataset without SKR** (Figure I.3): Used preprocessed data with VADER sentiment analysis but excluded semantic retrieval.
- **Clean VADER Dataset withSKR** (Figure I.4): Combined cleaned data with the retrieval task.
- **Raw VADER Dataset without SKR** (Figure I.5): Evaluated unprocessed data without applying semantic retrieval.
- **LLama2 with no Fine-Tuning** (Figure I.6): Served as a baseline, using the pre-trained model in its default state.

So, after implementing the different DRecSys having in mind the aspects that we wanted validated, we proceeded with testing and comparing their performance. The objective, in terms of the output of the DRecSys, was to evaluate the efficiency and accuracy of these systems in (1) generating ranked lists of drugs and (2) providing correct and personalize explanations based on individual user input. Additionally, and equally important, we do a (3) comparative analysis on the different configurations of the system, including the use of clean data, SKR task and SA to gain insights into their effectiveness. For this to be possible we implemented the experimental methodology.

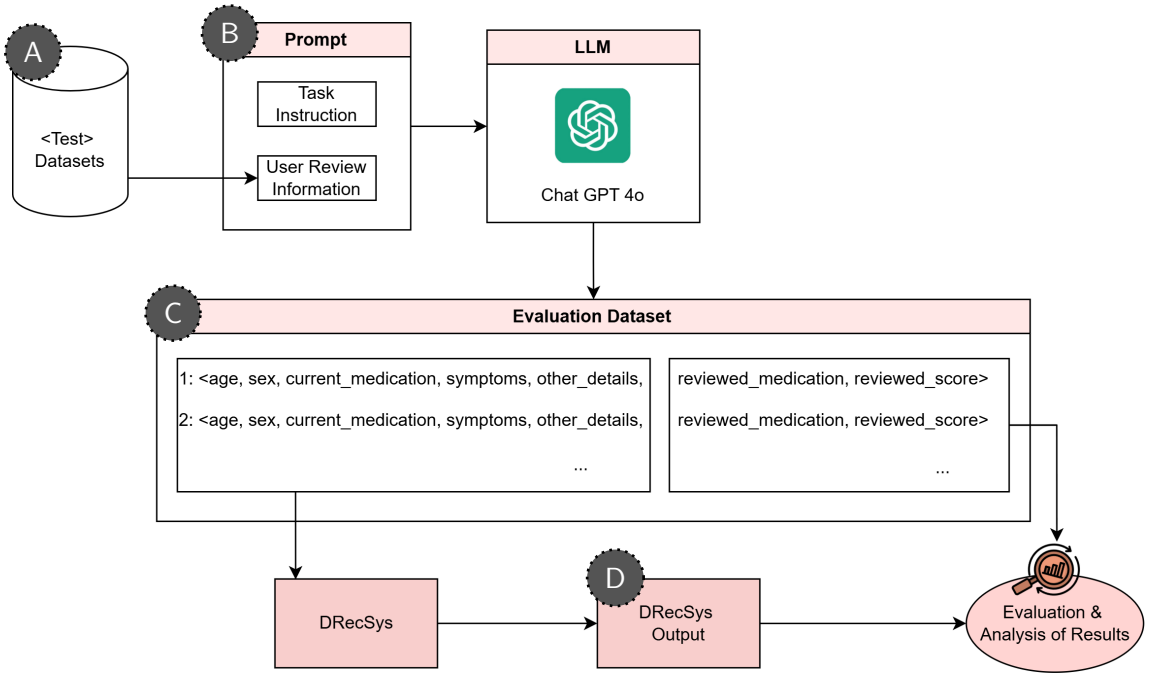


Figure 6.1: *Diagram of the workflow of the testing phase of the fourth and final step*

6.1 Experimental Methodology

Present in the Figure 6.1, we can observe the workflow of the testing process. To assess the system's ability to generate a ranked lists of drugs followed by a brief explanation and to then do a comparative analysis on the different configurations of the system, we tested the system using scenarios generated from the test dataset Figure 6.1.A. This strategy serves to simulate real-world scenario testing, as usability studies further enhance the reliability of these systems by simulating actual healthcare environments, [14, 27, 82, 104, 113]. These scenarios were generated using the AI tool ChatGPT 4.0, one with the original ratings and one with the VADER ratings, and stored in two files to be the model imputs. Figure 6.1.C. Each dataset created contains approximately 1,200 real-world medical scenarios generated using the AI, designed to cover a range of user profiles, medical histories and conditions, as the dataset underwent several adjustments and refinements to ensure it. This method of testing is inspired by relevant studies in the field, [14, 27, 82, 104, 113]. The prompt used for the scenarios generation (Figure 6.1.B) using ChatGPT 4o is as follows:

1. **Prompt for the generation of test scenarios using ChatGPT-4.0**:

   a) **Task Description**:

   ```
   Your goal is to generate a CSV file with the following
   fields:
   ```

   b) **CSV Row Fields**:

age: The age of the user (or inferred from the context).
sex: The gender of the user (or inferred from the context).
current_drugs: Any drugs mentioned that the user is currently taking other than the one being reviewed; if none, write "none".
symptoms: The user's condition or symptoms for which the drug is being taken.
other_details: Any additional relevant details about the user's condition, symptoms, or medical history that are not already covered in the symptoms or current drugs fields; if none, write "none".
reviewed_drugs: The drugs being reviewed by the user.
review_rating: The rating the user gave to the reviewed drug.

c) **Important Rules**:

```
1. The questions you generate should describe the user's
medical condition (including symptoms, age, sex, etc.)
without mentioning the drug that is being reviewed or the
user's reaction/opinion on that drug.
2. Ensure that any field containing commas, semicolons, or
lists (such as current_drugs, symptoms, or other_details)
is properly enclosed in double quotes to maintain CSV
formatting.
```

d) **Expected Output Format**:

```
Each line of the CSV file must have the following
structure:
age, sex, current_drugs, symptoms, other_details,
reviewed_drugs, review_rating
```

This evaluation of each system was around 10 hours with the implementation of the SKR task, and 5 hours without it. Using the GPU A100 with 40G VRAM proved necessary when handling the implementation using the SKR task. As for the outputs of the DRecSys (Figure 6.1. D) , they can be see in five files, for the original and clean VADER ratings with and without the SKR task, and raw VADER rating without the SKR task.

It is important to refer that ages were defined based on details such as life experiences, medical conditions, and references to treatments. Where no clear clues were provided, the age field was left blank. Contextual clues such as the user's medical conditions (e.g., menopause, or postmenopausal symptoms, birth control use, heart associated problems, erectile dysfunctions etc.) and life events (e.g., long-term drugs use or references to childbearing) were used to estimate ages where reasonable. Only personal information, such as user experiences or life details, was included in other_details with no references to the drug's effects or usage

specifics. In Table 6.1, are three of this test scenarios, these test cases provided a controlled environment to evaluate the system's ability to recommend appropriate drugs. And in the Table 6.2 we can see the output provided by the system using the original dataset with the SKR task.

Table 6.1: Three different user scenarios

Attributes	Information/Scenario Generated
Age	45
Sex	Female
Current drugs	Citalopram, Fluoxetine, Amitriptyline
Symptoms	Depression, Insomnia, Anxiety
Other Details	None
Reviewed drugs	Mirtazapine
Review Rating	10
Age	65
Sex	Male
Current drugs	None
Symptoms	Smoking Cessation
Other Details	Quitting after 50 years of smoking
Reviewed drugs	Chantix
Review Rating	10
Age	35
Sex	Female
Current drugs	None
Symptoms	Acne
Other Details	None
Reviewed drugs	Tretinoin
Review Rating	10

6.2 Analysis of the Generated Drug Lists

The first objective of the system is to generate a ranked list of drugs that aligns with the user's symptoms and medical history, while considering the following goals:

1. **Positive Recommendations**: The system aims to recommend drugs that have received positive ratings in the dataset used for fine-tuning, or extremely good drugs known by the LLM. These drugs should be commonly prescribed for the symptoms (or for the `condition`) indicated by the user.

2. **Avoid Negative Recommendations**: The system should avoid recommending drugs that have been negatively ranked in the dataset used for fine-tuning or presented severe side effects, or other contraindications.

3. **Ethical Decision-Making**: The system should refrain from recommending any drugs in cases where the user input is insufficient, incorrect, or unclear.

Table 6.2: Three user scenarios output from the DRecSys orginal with SKR task

Ranked List of Drugs	Brief Explanation:
1. Trazodone 2. Ambien 3. Quetiapine	Trazodone is a sedating antidepressant that can help with sleep, depression, and anxiety. Ambien is a sedating hypnotic that can also help with sleep. Quetiapine is an antipsychotic that can help with sleep, depression, and anxiety. These drugs were chosen because they are generally well tolerated and do not have a lot of side effects, and they are all sedating so they can help with sleep. They are also all relatively inexpensive and available in many pharmacies.
1. Ramipril 2. Sertraline 3. Fluoxetine	Ramipril is a good option for smoking cessation due to its ability to reduce blood pressure, which can make quitting easier. Sertraline is also a good option as it is an antidepressant that can help with cravings. Fluoxetine is also a good option, but it is important to note that it can increase blood pressure.
1. Tretinoin 2. Epiduo 3. Aczone 4. Ziana	Tretinoin is a topical retinoid that works by increasing cell turnover, which helps to reduce the appearance of acne. It is also a good option for patients with sensitive skin, as it can be used in a lower concentration than other retinoids. Epiduo is a combination of tretinoin and benzoyl peroxide, which can help to reduce acne by killing bacteria on the skin and preventing the formation of new acne. Aczone is a topical retinoid that is also effective in reducing acne, but it is milder than tretinoin. Ziana is a combination of tretinoin and clindamycin, which is effective in reducing acne by killing bacteria on the skin and preventing the formation of new acne. All of these drugs are effective in treating acne, but they may have different side effects and may not be suitable for all patients. It is important to consult with a dermatologist to determine the best course of treatment for your specific skin condition.

Having this in mind, and to evaluate the performance of the DRecSys in providing a ranked list of drugs with the outputs from the workflow (Figure 6.1. D), we focused on several key categories that assess how well the drugs suggested by the system align with the drugs reviewed by original users, as well as their corresponding review scores given, these categories can be seen in Table 6.3. These elements provide insights into the effectiveness of the system and highlight both positive and negative aspects of its performance.

This key elements have several strengths in evaluating DRecSys drugs recommendations. First, this approach effectively aligns suggestions with user preferences by comparing the drugs proposed by the system with those reviewed by users, allowing it to measure how well the top-ranked drugs match user satisfaction; as by analyzing user review ratings, the system captures the sentiment behind each review, offering insights into how well DRecSys recommendations resonate with real-world preferences. Additionally, the approach focuses on identifying when a user positively reviews the top recommendation or other suggestions, providing a clear measure of DRecSys's ability to rank its suggestions effectively. It also tracks failure cases, identifying when the DRecSys fails to generate a usable list or deviates from the expected

output, helping to understand areas for improvement or cases where the system refrains from making recommendations due to ethical concerns or insufficient information. Overall, this framework offers a good method for the evaluation of DRecSys's performance, focusing on both accuracy and user satisfaction.

Table 6.3: Insight Categories and Descriptions

Category	Description and Insight Color
`User liked top suggestion`	The system presented a top-ranked drug that received a positive review from the user. (Positive Insight - Dark Green)
`User liked other suggestions`	The system suggested a list of drugs, one of which received a positive review, though not the top-ranked one. (Positive Insight - Dark Green)
`Inconclusive/Neutral`	The system's suggestion received a neutral review, indicating the user had no strong preference. (Positive Insight - Dark Green)
`User disliked`	The system's suggestion led to a negative review of one of the recommended drugs. (Negative Insight - Dark Red)
`Inconclusive (Positive)`	The system suggested drugs, but the user liked one that wasn't highly ranked. (Inconclusive Insight - Dark Yellow)
`Inconclusive (Neutral)`	The system suggested drugs, but the user felt neutral about one that wasn't highly ranked. (Inconclusive Insight - Dark Yellow)
`Not liked by user & not suggested`	The system did not suggest a drug that the user disliked. (Positive Insight - Dark Green)
`Extraction Failed`	Error in extracting the list. (Inconclusive Insight - Dark Yellow)

Despite these strengths, the approach has limitations that could affect the accuracy of its insights. The reliance on numeric review ratings may oversimplify user preferences, as it does not account for detailed feedback, context, or reasoning behind the rating, which can lead to incomplete insights that the LLM was able to capture. Additionally, the focus on the top three drugs may result in suggestions that differ from the drugs the user actually used, even if they remain highly appropriate. By concentrating on comparing top-ranked suggestions with the drugs users chose, this approach might overlook other relevant options offered by the system, potentially underestimating DRecSys's overall performance. This is particularly important when using SKR, where a set of the best drugs is provided to try and help DRecSys's decision-making. Additionally, the method assumes strict adherence to a predefined output format to extract this variables, meaning even minor deviations can trigger an `Extraction Failed` classification, which can also misrepresent DRecSys's capabilities. While the approach provides valuable insights, its rigid and inflexible structure limits the analysis, especially in cases that require more nuanced feedback and adaptability, which is a challenge when evaluating this type of models and validating its outputs. The results of this approach, having in mind the categoried

from Table 6.3 can be visualized in bar charts, highlighting insights for the output of the ranked list of drugs under different DRecSys configurations. we can see then in Figures I.1, I.2, I.3, I.4, I.5, I.6 present in the respective Annex of this dissertation.

6.2.1 Accuracy-based metrics

To further evaluate and compare the system's recommendations, we measured accuracy-based metrics. To extract this metrics we used the data extracted from the simulated real-word scenario, so, it faces the same limitations as mentioned above. The evaluation categorized system outputs based on the defined key labels above mentioned. For the extraction on this metrics: True Positives (TP) included `User liked top suggestion`, where the system correctly identified a top-ranked drugs positively reviewed by the user; `User liked other suggestions`, where a positive review was linked to a non-top-ranked suggestion and `User was neutral to a suggestion`, indicating a neutral yet acceptable recommendation. True Negatives (TN) were identified as `Not liked by user & not suggested`, where the system successfully avoided suggesting disliked drugs. False Positives (FP) were labeled as `User disliked`, capturing instances where the system recommended poorly rated drugs. False Negatives (FN) has the `Inconclusive (Positive)` and `Inconclusive (Neutral)`, where the system failed to align suggestions with user preferences or neutral ratings despite those drugs being liked or accepted. The category `Extraction Failed`, representing inconclusive outputs where no valid recommendations were generated but the DRecSys or where not possible to extract, this was included in the graphs but not considered for traditional metrics. The results are presented in Table 6.4.

Table 6.4: Evaluation Accuracy Metrics for Different DRecSys Configurations

DRecSys Configuration	Accuracy	Precision	Recall	F1 Score
Original Dataset with SKR	0.3147	0.7778	0.1556	0.2592
Original Dataset without SKR	0.3126	**0.8129**	0.1418	0.2423
Clean VADER Dataset with SKR	**0.4995**	0.5269	**0.1948**	**0.2853**
Clean VADER Dataset without SKR	0.4804	0.5314	0.1518	0.2388
Raw VADER Dataset without SKR	0.4854	0.4762	0.1523	0.2334
LLama2 No Fine-Tuning	0.2101	0.7049	0.1237	0.2108

Impact of Raw and Clean Data: When comparing the raw VADER dataset without SKR to the clean VADER dataset without SKR, the results highlight the benefits and limitations of cleaning the dataset. The clean VADER dataset achieved a slightly lower accuracy (0.4804 vs. 0.4854) but a higher F1 score (0.2388 vs. 0.2334), indicating that cleaning the data contributed to better overall performance despite the marginal drop in accuracy. Recall remained comparable (0.1518 for the clean dataset vs. 0.1523 for the raw dataset), but cleaning contributed to better precision (0.5314 for the clean dataset vs. 0.4762 for the raw dataset). These results suggest that cleaning the dataset enhanced the system's ability to make more specific and relevant recommendations.

Performance with Sentiment Analysis: When comparing the original dataset without SKR

to the raw VADER dataset without SKR, the results reveal the impact of incorporating VADER for SA. The raw VADER dataset had higher accuracy (0.4854 vs. 0.3126) and a similar F1 score (0.2334 vs. 0.2423), indicating that SA helped improve the system's overall effectiveness. Recall was nearly identical (0.1523 for VADER vs. 0.1418 for the original), however, precision was lower for the raw VADER dataset (0.4762 vs. 0.8129), reflecting a trade-off between broader coverage and specificity. These results demonstrate that SA with VADER added value to the system by improving accuracy, though at the cost of reduced precision.

Effectiveness of SKR: When comparing the clean VADER dataset with SKR to the clean VADER dataset without SKR, the results highlight the impact of the SKR task. The inclusion of SKR improved accuracy (0.4995 vs. 0.4804) and F1 score (0.2853 vs. 0.2388), demonstrating its ability to enhance overall performance. Recall was also higher for the SKR configuration (0.1948 vs. 0.1518), showing that SKR broadened the range of relevant recommendations. However, this improvement came at the expense of slightly lower precision (0.5269 for SKR vs. 0.5314 without SKR), as the system retrieved a wider set of recommendations, some of which were less specific. These findings suggest that SKR is effective in improving recall and overall performance while maintaining a balance with precision.

Comparison with the System Without Fine-Tuning: When comparing the original dataset without SKR to the Llama2 configuration without fine-tuning, the results clearly demonstrate the importance of fine-tuning in improving system performance. The original dataset without SKR achieved significantly higher accuracy (0.3126 vs. 0.2101) and F1 score (0.2423 vs. 0.2108), showing that fine-tuning enhances the system's ability to align its recommendations with the dataset. Recall was also slightly better in the fine-tuned system (0.1418 vs. 0.1237), indicating that it retrieved more relevant options. Precision, while higher in the no fine-tuning configuration (0.7049 vs. 0.8129), reflects the inability of the non-fine-tuned system to adapt to domain-specific nuances effectively, relying instead on overly generic recommendations. This result helps to highlight the importance and impact of fine-tuning for domain-specific tasks, as a general-purpose language model struggles to provide accurate or relevant drugs recommendations that where used by users and saved in a dataset, so, the fine-tuning helped make the system more aware this specific drugs.

Trade-off Between Precision and Recall: Across all configurations, there is a noticeable trade-off between precision and recall. High precision in original datasets reflects a better alignment with highly relevant recommendations, while higher recall in SKR-based models suggests a broader retrieval of potentially useful options, but reducing specificity.

The evaluation of this accuracy-based metrics reveals that **fine-tuned models with clean VADER datasets and SKR** provide the most balanced performance, excelling in accuracy and F1 score. The analysis also demonstrates the critical role of fine-tuning and task-specific optimizations, such as SKR and SA, in improving the recommendation capabilities of the DRec-Sys. Also, this evaluation suffered from the large amount of `Inconclusive (Positive)` of extractions, still, this lies with the challenges mentioned above, since we are only looking at a ranking of three drug recommendations, and working with a knowledgeable model, with much available information and alternatives. The SKR task can also be underestimated with this metrics, as it chooses drugs that have automatically the best rating, but maybe it was not the

one the user used in this scenario, still we cannot forget that this method may take from the personalized aspect of using and creating the DRecSys.

6.2.2 Ranking-based Metrics

To evaluate the system's ability to prioritize relevant drugs effectively, we used ranking-based metrics. To extract this metrics we used the data extracted from the simulated real-word scenarios, so, it faces the same limitations as mentioned above. We focused on the metrics: HR to measure how often the top suggestion matched the user's preferred drugs, MRR to evaluate how early relevant drugs appeared in the ranking, and NDCG to prioritize relevant suggestions at the top of the list. The results are presented in Table 6.5.

Table 6.5: Evaluation Ranking Metrics for Different DRecSys Configurations

DRecSys Configuration	HR	MRR	NDCG
Original Dataset with SKR	**0.0408**	**0.0669**	**0.0849**
Original Dataset without SKR	0.0362	0.0587	0.0715
Clean VADER Dataset with SKR	0.0325	0.0510	0.0594
Clean VADER Dataset without SKR	0.0233	0.0371	0.0417
Raw VADER Dataset without SKR	0.0192	0.0326	0.0374
LLama2 No Fine-Tuning	0.0102	0.0225	0.0315

Impact of Clean and Raw Data: Models using the clean VADER dataset show an improvement in ranking-based metrics compared to raw VADER data, particularly in HR and MRR. This suggests that cleaning and standardizing the input data enhances the model's ability to make more informed recommendations. However, the improvement remains moderate, so it should not be taken as a definite improvement, as it needs future work and evaluation.

Effectiveness of Sentiment Analysis: When comparing the original dataset without SKR to the raw VADER dataset without SKR, this evaluation focuses solely on the quantitative value of the rating attribute rather than the sentiment present in the reviews. The raw VADER dataset shows a decline in ranking-based metrics such as HR (0.0192 vs. 0.0362), MRR (0.0326 vs. 0.0587), and NDCG (0.0374 vs. 0.0715). These results suggest that while VADER's SA adjusts the ratings based on the sentiment of the reviews, it may not translate into better ranking-based performance when evaluated solely on numerical ratings. However, this does not diminish the value of SA, as it provides additional context to the ratings that could be beneficial in other aspects of recommendation generation, such as improving the richness of explanations or supporting more personalized results. Future work could explore ways to combine sentiment insights with numerical ratings to maximize their impact on ranking-based performance.

Effectiveness of SKR: SKR consistently helps to have better results in MRR and NDCG scores across configurations, suggesting that it improves the ranking quality by prioritizing highly rated drugs. Nonetheless, the downside is evident, as SKR reduces the personalized nature of recommendations, focusing on general high ratings rather than specific user choices.

Comparison with the System Without Fine-Tuning: The no fine-tuning configuration

performs the worst across all ranking metrics. This result further helps understand the positive impact of fine-tuning, as general-purpose models fail to provide domain-specific ranking relevance, making their output less useful for real-world scenarios.

The configurations using the **original dataset with SKR** achieved the highest scores across all metrics (HR, MRR, and NDCG). This indicates that the SKR task enhances the model's ability to present the most relevant drugs early in the ranking. However, the overall HR value remains low due to the limited alignment of the top-ranked drugs with the user's actual preferred choice, reflecting the challenges of personalization in the recommendations.

The evaluation of this ranking-based metrics reveals that while SKR improves ranking structure, challenges exist in aligning recommendation with user specific usage, still, this can be a sign of better personalized recommendations. Fine-tuned models with clean datasets and the use of semantic retrieval show promise. The consistently low HR values highlight the difficulty in aligning the top suggestion with user preferences. While MRR and NDCG show better performance, the rankings often fail to reflect the specific drugs that users positively reviewed, which may not the a bad insight as it can be a sign of better alignment with personalized recommendations, just as mentioned above.

6.3 Conversational Behavior Analysis

The second objective was for the DRecSys to provide a brief and comprehensive explanation for each recommendation. This output should ensure the following:

1. **Consistency:** The system consistently follows the structured prompt format provided, generating uniform and predictable outputs across all user scenarios. while maintaining personalization in the content.

2. **User-Friendly Explanations**: The explanations generated should be easy to understand, providing clear reasoning behind the recommended drugs. Medical terminology should be simplified to improve accessibility for non-expert users.

3. **Logical Details**: Wherever possible, the system incorporates logical reasoning in the explanations, detailing how specific symptoms, drugs, or potential side effects influenced the recommendation.

4. **Extra encouragement**: The system also encourages users to consult a healthcare professional, especially when the recommendation involves serious conditions or complex drug interactions—encourages the user to search for more information.

To evaluate the effectiveness of the DRecSyss in generating this output, we tested the system using the same AI tool as for the previous test to evaluate the tone, consistency and contents of the explanation given to the users in the tests from the workflow previously done. This method (LLM-as-judge) was also observed on the studies mentioned above in Chapter 3. We tested all DRecSyss configuration, and the differences of the outputs with be highlighted below, still, since they where fairly similar, first we discuss the Consistency, the User-Friendly Explanations,

the Logical Details and the Extra encouragement seen in the results on the clean VADER dataset without SKR task.

Consistency: The system consistently followed the structured prompt format provided, generating uniform and predictable outputs across most user scenarios. After the drug list, a brief explanation is provided, typically in 2-3 sentences, covering why the drugs were chosen, often mentioning their effects and relevance to the user's symptoms or conditions. In Figure 6.2 we can see two examples of consistent drug explanations generated by the DRecSys, in both cases, the system provides a concise explanation regarding each drugs, its function, and why it was recommended. The structure is clear and predictable, showing consistent adherence to the format required for delivering both drugs lists and explanations.

User-Friendly Explanations: The system simplifies the explanation of drugs, avoiding overly technical terminology, making it easier for a non-expert to understand. In Figure 6.2, we can see two examples of user-friendly drug explanations generated by the DRecSys, in the first example, the explanation uses accessible language, such as "help with sleep, depression, and anxiety", making it easy for a non-medical user to understand the purpose of the drugs without delving into complex pharmacology. In the second example he term "long-acting reversible contraceptive" (LARC) is explained in simple terms, with the duration of effectiveness (3 years) made clear, along with how the method works. This is an example of simplifying medical terms for the average user.

Logical Details: The system connects symptoms, current drugs, or other user details with the recommendation, providing a logical flow of reasoning. In Figure 6.2, we have two examples of logical details in drugs explanation generated by the DRecSys. In the fist example we can see that the explanation provides logical details, tying the user's symptoms (sleep issues, depression, and anxiety) to the sedating effects of the recommended drugs, explaining why each was chosen. In the second example, the explanation logically connects the goal of contraception with the effectiveness of the suggested drugs (both long-acting reversible contraceptives). It explains the method of insertion and the duration, providing a reason for their choice.

Extra encouragement: The system encourages the user to consult a healthcare professional or seek more information, particularly in complex cases or for serious conditions. In Figure 6.3, we can see two examples of Extra Encouragement for Professional Consultation generated by the DRecSys, in the first example, we can see that the DRecSys encourages further consultation with a healthcare professional, ensuring that the user does not rely solely on the recommendation and seeks personalized advice and the second example provides an extra nudge for users to discuss the recommendations with a doctor to check for possible drug interactions, promoting a cautious approach.

In comparing the results from the various DRecSys configurations outputs and we evaluated the key factors just mentioned. Although there are differences between the DRecSys, their overall outputs are fairly similar.

When in comes to the consistency in the outputs, across all datasets, the systems generally produce consistent results, but the **clean datasets** stand out for its superior consistency in formatting and structure. Each output is clearly organized, starting with user details (age, sex,

"Trazodone is a sedating antidepressant that can help with sleep, depression, and anxiety. Ambien is a sedating hypnotic that can also help with sleep. Quetiapine is an antipsychotic that can help with sleep, depression, and anxiety. These drugs were chosen because they are generally well tolerated and do not have a lot of side effects, and they are all sedating so they can help with sleep."

"Etonogestrel is a long-acting reversible contraceptive (LARC) that is inserted under the skin of the upper arm. It is effective for 3 years and is very effective at preventing pregnancy. Levonorgestrel is also a LARC that is inserted in the uterus and is very effective at preventing pregnancy."

Figure 6.2: *Example of Consistent, User-Friendly and Logical Details in drugs explanations generated by the DRecSys*

"I recommend you discuss these options with your healthcare provider, as they can help determine the best option for your specific needs and health conditions."

"Before starting any of these drugs, it is important to discuss with your healthcare provider to ensure there are no interactions with your current drugs."

Figure 6.3: *Examples of Extra Encouragement for professional consultation in drugs explanations generated by the DRecSys*

current drugs, symptoms), followed by a ranked list of recommended drugs, a brief explanation with a similar size, and frequent encouragement to consult a healthcare provider. In the clean dataset almost every row follows this structure, making it highly readable and easy to interpret. In contrast, the raw VADER dataset suffers from some formatting issues such as repeated headers and inconsistent spacing, which affects its readability and overall coherence. The original dataset systems maintain a reasonable structure, but they lack the refined clarity and uniformity that the clean dataset has. The clean VADER based system consistently also offers more clear, well-structured reasoning behind drugs recommendations. Each recommendation is accompanied by an explanation of the drug's effects, its relevance to the user's symptoms, and a reminder to seek professional healthcare advice. This uniformity improves readability trust in the system. The other DRecSys, sometimes present varying explanation quality and dimension, in particular the raw VADER, some explanations lack depth or are far to repetitive, which reduces the overall clarity and value of the recommendations. As a result, users may find it harder to interpret and trust the recommendations. The original dataset performs better than the raw version in terms of consistency, but it still doesn't reach the level of clarity seen in the clean dataset. In terms of language consistency, the clean dataset again stands out. It shows standardized phrasing and sentence structures throughout, frequently using language like "This drugs is effective for ..." and emphasizing the importance of consulting a healthcare professional consistently. This level of consistency makes the recommendations feel more trustworthy and easier to follow. By contrast, the raw dataset shows inconsistent language use,

with some fluctuations in tone and repetitive phrasing. The original dataset fares better than the raw version but still doesn't match the clean dataset's consistency in clear terminology and phrasing.

User-friendly explanations are crucial for ensuring that non-experts can understand and act on the recommendations. The clean dataset shows very good results in this regard, offering explanations that avoid medical terminology and provide clear, straightforward descriptions, such as "helps with sleep" or "reduces inflammation". This makes the information accessible to a broader audience, particularly those without medical training. The raw VADER is less consistent in this area, with some explanations being overly repetitive, making them harder to follow. The original dataset performs better than the raw version, but its explanations are still slightly lacking when compared with the clean dataset.

When it comes to logical reasoning, the clean dataset provides well-connected and logical explanations that clearly tie together symptoms, recommended drugs, and their effects. Each recommendation includes a reasoning chain that explains why a particular drugs is suggested and how it addresses the user's specific condition. This makes the recommendations more trustworthy. In contrast, the raw VADER system often shows weaker logical reasoning, with less thorough connections between the user's symptoms and the recommended drugs. The original dataset includes logical reasoning, but it doesn't provide the same detailed connections that are consistently present in the clean dataset.

An important aspect of a reliable DRecSys is the inclusion of extra encouragement to consult healthcare professionals, especially when dealing with complex conditions or potential drug interactions. The clean dataset frequently and consistently includes reminders for users to seek professional guidance, ensuring that users are encouraged to take responsible actions based on the recommendations provided. The raw VADER system, on the other hand, is less consistent in providing these crucial reminders. The original dataset includes more of this reminders but less consistently than the clean dataset, where this advice is a standard practice across all outputs.

In conclusion, while the outputs across the datasets exhibit overall similarities, the **clean dataset** stands out for its clarity, structured recommendations, and attention to detail, making it the most effective and user-friendly configuration. However, its impact on significantly improving system metrics remains limited.

The primary advantage of the clean dataset lies in enhancing the actionability of the system's recommendations and fostering greater user trust, rather than delivering substantial performance gains. It is important to note that all aspects of the system can be further improved by refining the introductions in the prompt. Testing revealed that the system can consistently include specific phrases, such as, *"Please always consult a medical professional before starting any medication."* Furthermore, the system demonstrated the ability to incorporate essential disclaimers, particularly in contexts that align with the user's symptoms and the medication being reviewed. This capability underscores a key advantage of employing this technology: its potential for delivering personalized and contextually appropriate recommendations.

However, it was also observed that overly long prompts with excessive instructions tend to

diminish the system's ability to provide personalized and insightful commentary on medications, personalized to the user's symptoms and information. This observation highlights the importance of balancing system design priorities – whether to maximize the technology's potential for personalization or to enforce specific, predefined outputs. Striking the right balance is crucial when building systems that aim to leverage technology effectively while addressing user needs and concerns.

6.4 Further Observations and Insights

Another insights we can retrieve from the outputs from the DRecSys, is a comparative with the insights from the EDA previously done. In the Table 6.6 we can see a compare the top 10 most suggested drugs from the DRecSys with the original ratings and the VADER ratings, and observe which of these where present and not present in the original dataset.

For example, drugs such as `Venlafaxine`, `Sertraline` and `Escitalopram` and are present in both rating systems and are frequently suggested, demonstrating they are highly recommended medications.

`Sertraline` is one of the most prominent drugs across both the VADER and Original rating DRecSys recommendations appearing in Figure 5.4, 5.5, 5.8. This indicates that `Sertraline` is highly regarded and widely used. Its inclusion in the "Top 10 Drugs by UsefulCount" (Figure 5.5) shows that it has received a significant number of useful reviews, suggesting that users find it effective. Additionally, its presence in the "Top 10 Most Reviewed Drugs" (Figure 5.4) indicates that it is frequently prescribed and widely discussed among users. Furthermore, it is associated with multiple conditions in the "Number of Unique Conditions for the Top 10 Drugs by UsefulCount" (Figure 5.8) suggests that `Sertraline` is versatile in its application, making it a crucial drug for treating a variety of mental health conditions. Overall, the drug's strong presence across all metrics explains its high number of recommendations and positive reception by the users.

`Escitalopram` is another important drug, appearing in the clean VADER top 10 list (Table 6.6) and showing up also in Figure 5.4, 5.5, 5.8. Like `Sertraline`, it also appears in the "Top 10 Drugs by UsefulCount" (Figure 5.5) and "Top 10 Most Reviewed Drugs" (Figure 5.4) indicating that it is frequently prescribed and widely discussed by users. Its presence in the "Number of Unique Conditions for the Top 10 Drugs by UsefulCount" (Figure 5.8) demonstrates that `Escitalopram` is used for treating a wide variety of conditions. Its consistent presence across these metrics shows it is both effective and well-received by the user base.

`Fluoxetine` appears in both the clean VADER and original rating System lists (Table 6.6) and is featured in Figure 5.5, 5.8. It ranks in the "Top 10 Drugs by UsefulCount" (Figure 5.5) meaning that users consider it to be an effective treatment, though it is slightly less represented compared to `Sertraline` or `Escitalopram`. `Fluoxetine` is also notable for its versatility, as indicated by its inclusion in the "Number of Unique Conditions for the Top 10 Drugs by UsefulCount" (Figure 5.8). This implies that while it is widely used, its range of applications may be somewhat narrower compared to the other drugs like `Sertraline` and `Escitalopram`.

Levonorgestrel is included in the Original rating system list (Table 6.6) and appears in Figure 5.4. This suggests that Levonorgestrel was frequently reviewed and recommended, likely because of its common use as a birth control drug. Its high review count reflects its broad use in reproductive health and contraceptive treatments – Birth Control condition. The volume of reviews indicates its widespread prescription and general familiarity among users. Ethinyl also appears in both the clean VADER rating system list, also showing up in Figure 5.4. Similar to Levonorgestrel, its appearance here is likely due to its prominent use as a birth control drug.

The presence of these drugs across multiple graphs outlines their significance in treating a wide range of conditions. Drugs like Sertraline, Escitalopram, Fluoxetine and Venlafaxine play a crucial role in mental health treatment Figure 5.8, consistently receiving positive feedback and high usefulness ratings from users. On the other hand, Levonorgestrel and Ethinyl are important in reproductive health, particularly for Birth Control (Figure 5.8), as evidenced by their large number of reviews.

However, it is particularly interesting to analyze the drugs marked with a cross ($\times$), which were recommended by the DRecSys but are absent from the original dataset. For instance, in the VADER drug list, Ortho and Microgestin were recommended but are not present in the original dataset. Similarly, in the original drug Junel was suggested without being part of the VADER ratings' recommendations. These drugs are well-known medications; for example: Ortho is widely known oral contraceptives, which suggests the system is identifying these as relevant based on broader context or patterns in related data, even though they were not included in the original dataset. Microgestin and Junel are also common components of contraceptives, potentially highlighting the system's ability to generalize recommendations for this category (Birth Control).

The presence of these drugs in recommendations despite their absence in the dataset shows the system's potential to suggest widely known and relevant medications. This is attributed to the model's pre-training knowledge or the sentiment/context derived from related reviews. The combination of recommending drugs present in the dataset alongside widely known medications demonstrates the effectiveness of fine-tuning and the system's potential to suggest well-established drugs. This highlights the capability of using an LLM for a DRecSys, showcasing its ability to balance dataset-specific insights with broader, generalizable medical knowledge.

This evaluation of the DRecSys highlights a key challenge: balancing personalized recommendations with consistent, standardized outcomes. Personalization focuses on tailoring suggestions to the unique needs and preferences of individual users, using the system's ability to understand the details provided by the user. However, this can sometimes shift the focus away from producing reliable and consistent outputs, especially when systems like SKR prioritize highly rated drugs over user-specific needs. On the other hand, controlled outcomes focus on consistency and accuracy, ensuring that the recommendations meet clear and measurable standards. Finding the right balance between these goals is essential and depends on the specific purpose of the DRecSys. Personalized recommendations make the system more engaging and useful for individuals, while controlled outcomes can improve reliability and trust. In the end, the success of a DRecSys depends on its ability to combine both, delivering recommendations that

are reliable, consistent, and still feel personal to the user.

Table 6.6: Top 10 Most Suggested Drugs with VADER and Original Rating Systems with Tick/Cross Status of their presence in the original dataset

Vader Drug	**Suggestions**	✓/×	**Original Drug**	**Suggestions**	✓/×
Ortho	168	×	Yaz	136	✓
Seasonique	165	✓	Ortho	114	×
Venlafaxine	105	✓	Venlafaxine	98	✓
Sertraline	99	✓	Levonorgestrel	82	✓
Microgestin	88	×	Sertraline	77	✓
Fluoxetine	83	✓	Gabapentin	72	✓
Lamotrigine	56	✓	Escitalopram	64	✓
Escitalopram	55	✓	Zoloft	47	✓
Seasonale	53	✓	Lamotrigine	41	✓
Ethinyl	47	✓	Junel	39	×

7 Final Considerations

7.1 Conclusions

In this dissertation, we developed and evaluated various variants of a Drug Recommendation System (DRecSys) to understand the impact of different processes on it's performance. By fine-tuning a LLM (Llama2) using the UCI ML Drug Review dataset, our objectives were to generate a personalized recommendations based on user-generated feedback and medical histories while evaluating the significance of distinct development steps.

The proposed solution was organized into four main steps: (1) EDA, (2) data cleaning and pre-processing, (3) SA, and (4) the implementation of the phases associated with the workflow and evaluation of the DRecSys. These stages provided a structured approach to building the RS while also facilitating the evaluation of various configurations, particularly the impact of data cleaning, SA, SKR, and fine-tuning.

The baseline model, represented by Llama2 without fine-tuning, exhibited the lowest performance across all metrics, with an accuracy of 0.2101 and an F1 score of 0.2108. This underscores the critical importance of fine-tuning in adapting general-purpose models for domain-specific tasks, such as healthcare. Without fine-tuning, the model struggled to generate meaningful or relevant drug recommendations, demonstrating its limited ability to understand the nuances of specific drugs and their correlations to medical conditions present in the dataset.

Regarding data cleaning and pre-processing, while the cleaned datasets did not achieve higher accuracy compared to the raw datasets (e.g., 0.4804 vs. 0.4854 for the clean and raw VADER datasets without SKR), they did improve other performance metrics, such as F1 score (0.2388 vs. 0.2334) and precision (0.5314 vs. 0.4762). This indicates that while cleaning the dataset may not significantly impact accuracy, it enhances the system's ability to provide more specific and relevant recommendations. Additionally, cleaning improved the clarity and usability of the outputs, making the recommendations more structured and user-friendly, which is crucial for enhancing user trust and engagement.

In terms of SA, the comparison between VADER and TextBlob showed that VADER performed better in capturing sentiment nuances related to drug efficacy and side effects, particularly when dealing with informal or domain-specific user reviews, so the ratings used of a variant of

the DRecSys where VADER's; And SA using VADER played an important role in improving the system's contextual understanding of user reviews. When comparing the raw VADER dataset (accuracy 0.4854, F1 score 0.2334) with the original dataset without sentiment analysis (accuracy 0.3126, F1 score 0.2423), the benefits of sentiment analysis in aligning recommendations with user satisfaction are evident. However, the lower precision in VADER-based configurations (e.g., 0.4762 for raw VADER vs. 0.8129 for the original dataset) highlights a trade-off between broader recommendation coverage and specificity.

Fine-tuned models using the **clean VADER dataset with SKR demonstrated the best overall performance**, achieving the highest accuracy (0.4995) and F1 score (0.2853). Despite these gains, a significant proportion of outputs fell under the "Inconclusive (Positive)" category, revealing challenges in aligning recommendations confidently with user preferences. This suggests further improvements are needed in personalization and decision-making processes to better capture user-specific nuances.

When it comes to ranking-based metrics, including HR (0.0408), MRR (0.0669), and NDCG (0.0849) for the original dataset with SKR, highlighted the system's strengths in prioritizing relevant drugs but also revealed limitations in personalization. While SKR improved the ranking structure, it can make the DRecSyss fail to prioritize specific drugs that users positively reviewed, or are a better fit having in mind the user's provided medical history, reflecting the inherent trade-off between optimizing rankings and maintaining personalization.

Overall, fine-tuning the 7B Llama2 model significantly improved domain-specific performance. However, the model's limited parameter size introduced challenges in handling complex relationships, leading to a notable prevalence of inconclusive results. Addressing these limitations could involve fine-tuning larger models, incorporating domain-specific medical datasets, and enhancing training processes to improve the system's ability to understand nuanced relationships between drugs, conditions, and symptoms, all options discussed in the next section.

By integrating EDA, data cleaning, SA tools, fine-tuning, and techniques like SKR, this research contributes to the growing field of AI-driven healthcare. While limitations remain, these findings demonstrate the potential of combining modern AI techniques with traditional methodologies to deliver actionable, contextually relevant insights. These results lay the groundwork for future efforts to create more accurate, personalized, and user-friendly DRecSys, advancing the role of AI in healthcare decision-making, while challenges remain, the findings of this dissertation help with future efforts to create more accurate, personalized, and user-friendly DRecSys, enhancing healthcare decision-making for both patients and healthcare professionals.

The evaluation of the results also highlighted areas where the DRecSys encountered limitations, as well as challenges within the evaluation process itself, which are now addressed in the Future Work section.

7.2 Future Work

Identifying areas for refinement, potential enhancements, and further exploration is essential for advancing the DRecSys's performance and reliability.

- **Fine-Tuning on Healthcare-Specific Data**: Future iterations of DRecSys could benefit from fine-tuning on larger and more domain-specific healthcare datasets. This would enable the model to better understand clinical language, drug interactions, and nuanced user feedback, making the recommendations more accurate and medically appropriate, making it easier for the DRecSys to also understand the relationships between drugs, conditions, and symptoms. A valuable source of healthcare-specific data for this purpose is the Patient Information Leaflet (PIL) of various medications. PILs provide structured and detailed information about drug indications, contraindications, side effects, and usage instructions. For example, contraindications and adverse effects listed in the PIL could be interpreted as negative reviews, while indications and therapeutic benefits could be treated as positive feedback. By incorporating this rich dataset into fine-tuning, the DRecSys could achieve a deeper understanding of the clinical context, ensuring that its recommendations align closely with medical standards and user needs.

- **Leveraging Larger Language Models**: The current DRecSys utilized Llama2 with 7 billion parameters, which demonstrated significant improvements through fine-tuning but revealed limitations in handling more complex relationships within the data. Future iterations could benefit from leveraging larger models, such as Llama2 with 13B or 65B parameters, which possess greater capacity to capture nuanced patterns and intricate connections between drugs, symptoms, and conditions. These larger models, with their enhanced contextual understanding and broader generalization abilities, could enable the DRecSys to deliver more precise and personalized drug recommendations while better addressing the diversity and complexity inherent in medical datasets. This scalability would also allow for more effective integration of additional features, such as advanced sentiment analysis or improved handling of rare conditions, ultimately enhancing the system's reliability and user trust.

- **Enhancements in Data Cleaning and Pre-Processing**: While the performance of the DRecSys based on clean datasets was only moderately better than that of raw datasets, future work could explore more sophisticated pre-processing techniques to unlock greater potential. Advanced methods such as semantic parsing and NER could help better standardize input data by identifying and structuring key entities like drug names, conditions, and symptoms. These techniques would enable the DRecSys to interpret user inputs more accurately, reducing ambiguities and inconsistencies in the data. This improvement could lead to more precise recommendations and greater personalization, particularly when dealing with complex or nuanced medical scenarios.

- **Advanced Sentiment Analysis Techniques**: While VADER demonstrated its effectiveness in extracting sentiment from user reviews, future iterations of the DRecSys could benefit from integrating more advanced SA techniques, such as transformer-based models

like BERT or RoBERTa. These models, pre-trained on vast and diverse datasets, have the potential to capture subtle nuances and contextual dependencies in user feedback that simpler tools like VADER or TextBlob might overlook. By using these advanced techniques, the DRecSys could gain a more precise understanding of user sentiment, enabling it to make recommendations that better align with user preferences and expectations.

- **Optimization of the Semantic Key Retrieval Task**: The semantic key retrieval task, although helpful in structuring recommendations, diminished personalization when predefined lists were used. Improving this task by integrating more sophisticated algorithms could help the system personalize suggestions more effectively to individual user needs/ inputs.

- **More Evaluation Scenarios:** We conducted approximately 1,200 scenarios to test DRecSys. However, increasing the number of scenarios could potentially lead to more comprehensive insights and improved results. Additionally, further exploration of diverse and edge-case scenarios might help better evaluate the system's robustness and adaptability.

- **Ethical Decision-Making and Professional Guidance**: Upholding ethical standards in drug recommendation systems is essential, particularly when addressing complex medical scenarios. Future iterations of the DRecSys could incorporate more robust mechanisms to verify the safety and appropriateness of its suggestions, ensuring strict alignment with established medical guidelines. Furthermore, consistent emphasis should be placed on encouraging users to consult healthcare professionals before acting on recommendations, especially for cases involving drug interactions or high-risk conditions. This can be achieved through the development and testing of ethically designed prompts, ensuring the system reinforces the importance of professional medical advice while maintaining user trust and safety.

In conclusion, while the DRecSys demonstrated the potential to provide personalized, data-driven drug recommendations, future improvements – particularly in the areas of fine-tuning, sentiment analysis, and model complexity – are essential to advancing the system's capability and ensuring it becomes a trusted tool in healthcare decision-making.

Bibliography

[1] P. Achananuparp and I. Weber. "Extracting food substitutes from food diary via distributional similarity". In: *arXiv preprint arXiv:1607.08807* (2016) (cit. on p. 18).

[2] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. "Gpt-4 technical report". In: *arXiv preprint arXiv:2303.08774* (2023) (cit. on pp. 8, 14, 15, 24, 26).

[3] I. T. Agaku, A. O. Adisa, O. A. Ayo-Yusuf, and G. N. Connolly. "Concern about security and privacy, and perceived control over collection and use of health information are related to withholding of health information from healthcare providers". In: *Journal of the American Medical Informatics Association* 21.2 (2014), pp. 374–378 (cit. on p. 22).

[4] G. Agapito, M. Simeoni, B. Calabrese, P. H. Guzzi, G. Fuiano, and M. Cannataro. "DIETOS: A Recommender System for Health Profiling and Diet Management in Chronic Diseases." In: *HealthRecSys@ RecSys*. 2017, pp. 32–35 (cit. on p. 18).

[5] E. Agu and M. Claypool. "Cypress: A cyber-physical recommender system to discover smartphone exergame enjoyment". In: *Proceedings of the ACM workshop on engendering health with recommender systems*. 2016 (cit. on p. 19).

[6] M. AI. *Introducing LLaMA: A foundational, 65-billion-parameter language model*. https://ai.meta.com/blog/large-language-model-llama-meta-ai/. Accessed: 2024-02-23. 2023 (cit. on pp. 8, 14, 15, 24).

[7] M. AI. *Meta-Llama-3-8B*. https://huggingface.co/meta-llama/Meta-Llama-3-8B. 2023 (cit. on pp. 26, 50).

[8] M. AI. *Responsible AI*. https://ai.meta.com/responsible-ai/. 2023 (cit. on p. 35).

[9] S. Akkoyunlu, C. Manfredotti, A. Cornuéjols, N. Darcel, and F. Delaere. "Investigating substitutability of food items in consumption data". In: *Second International Workshop on Health Recommender Systems co-located with ACM RecSys*. Vol. 5. 2017 (cit. on p. 18).

[10] A. Al-Wahab, A. Elhabbash, and A. Fayyoumi. "A Survey on Modern Recommendation System based on Big Data". In: *arXiv preprint arXiv:2206.02631v4* (2022). URL: https://ar5iv.labs.arxiv.org/html/2206.02631v4 (cit. on p. 9).

[11] F. Alvarez, M. Popa, V. Solachidis, G. Hernandez-Penaloza, A. Belmonte-Hernandez, S. Asteriadis, N. Vretos, M. Quintana, T. Theodoridis, D. Dotti, et al. "Behavior analysis through multimodal sensing for care of Parkinson's and Alzheimer's patients". In: *Ieee Multimedia* 25.1 (2018), pp. 14–25 (cit. on p. 20).

[12] J. G. Anderson and K. Abrahamson. "Your health care may kill you: Medical errors." In: *ITCH* 234 (2017), pp. 13–17 (cit. on p. 17).

[13] Anonymous. "LoRA Fine-tuning Efficiently Undoes Safety Training in Llama 2-Chat 70B". In: *arXiv preprint arXiv:2310.20624* (2023). URL: https://ar5iv.labs.arxiv.org/html/2310.20624 (cit. on p. 35).

[14] Arthur AI Team. *LLM-Guided Evaluation: Using LLMs to Evaluate LLMs*. Arthur Blog. 2023. URL: https://www.arthur.ai/blog/llm-guided-evaluation-using-llms-to-evaluate-llms (cit. on pp. 26, 60).

[15] J. Aswal and N. Srivastava. "A Recommender System for Informal Bibliotherapy". In: *Proceedings of the Workshop on Health Recommender Systems at RecSys*. 2020, pp. 23–27 (cit. on p. 18).

[16] K. Balog and F. Radlinski. "Measuring recommendation explanation quality: The conflicting goals of explanations". In: *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*. 2020, pp. 329–338 (cit. on p. 21).

[17] Z. Bao, W. Chen, S. Xiao, K. Ren, J. Wu, C. Zhong, J. Peng, X. Huang, and Z. Wei. "Disc-medllm: Bridging general large language models and real-world medical consultation". In: *arXiv preprint arXiv:2308.14346* (2023) (cit. on pp. 14, 25).

[18] O. Barkan, Y. Fuchs, A. Caciularu, and N. Koenigstein. "Explainable recommendations via attentive multi-persona collaborative filtering". In: *Proceedings of the 14th ACM Conference on Recommender Systems*. 2020, pp. 468–473 (cit. on pp. 21, 23, 26).

[19] J. Beel, B. Gipp, S. Langer, and C. Breitinger. "Paper recommender systems: a literature survey". In: *International Journal on Digital Libraries* 17 (2016), pp. 305–338 (cit. on pp. 7, 9).

[20] J. Bobadilla, S. Alonso, and A. Hernando. "Deep Learning Architecture for Collaborative Filtering Recommender Systems". In: *Applied Sciences* 10.7 (2020), p. 2441. DOI: 10.3390/app10072441. URL: https://www.mdpi.com/2076-3417/10/7/2441 (cit. on p. 9).

[21] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. "On the opportunities and risks of foundation models". In: *arXiv preprint arXiv:2108.07258* (2021) (cit. on p. 27).

[22] L. Boratto, S. Carta, W. Iguider, F. Mulas, P. Pilloni, et al. "Predicting workout quality to help coaches support sportspeople". In: *CEUR Workshop Proceedings*. Vol. 2216. CEUR-WS. 2018, pp. 8–12 (cit. on p. 19).

[23] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio. "Generating sentences from a continuous space". In: *arXiv preprint arXiv:1511.06349* (2015) (cit. on p. 14).

[24] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf (cit. on pp. 15, 16, 24, 25, 27).

[25] G. Cai, J. Zhu, Q. Dai, Z. Dong, X. He, R. Tang, and R. Zhang. "Reloop: A self-correction continual learning loop for recommender systems". In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2022, pp. 2692–2697 (cit. on pp. 26, 27).

[26] A. Calero Valdez, M. Ziefle, K. Verbert, A. Felfernig, and A. Holzinger. "Recommender systems for health informatics: state-of-the-art and future perspectives". In: *Machine Learning for Health Informatics: State-of-the-Art and Future Challenges* (2016), pp. 391–414 (cit. on p. 18).

[27] M. Cascella, J. Montomoli, V. Bellini, and E. Bignami. "Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios". In: *Journal of medical systems* 47.1 (2023), p. 33 (cit. on pp. 26, 60).

[28] A. Chaturvedi, B. Aylward, S. Shah, G. Graziani, J. Zhang, B. Manuel, E. Telewa, S. Froelich, O. Baruwa, P. P. Kulkarni, et al. "Content Recommendation Systems in Web-Based Mental Health Care: Real-world Application and Formative Evaluation". In: *JMIR formative research* (2023) (cit. on p. 19).

[29] E. Choi, A. Schuetz, W. F. Stewart, and J. Sun. "Using recurrent neural network models for early detection of heart failure onset". In: *Journal of the American Medical Informatics Association* 24.2 (2017), pp. 361–370 (cit. on pp. 7, 18).

[30] J. C. Chow, V. Wong, and K. Li. "Generative Pre-Trained Transformer-Empowered Healthcare Conversations: Current Trends, Challenges, and Future Directions in Large Language Model-Enabled Medical Chatbots". In: *BioMedInformatics* 4.1 (2024), pp. 837–852 (cit. on pp. 26, 27).

[31] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al. "Palm: Scaling language modeling with pathways. arXiv 2022". In: *arXiv preprint arXiv:2204.02311* 10 (2022) (cit. on p. 25).

[32] C. N. Dang, M. N. Moreno-García, and F. D. l. Prieta. "An approach to integrating sentiment analysis into recommender systems". In: *Sensors* 21.16 (2021), p. 5666 (cit. on p. 23).

[33] T. Dettmers, M. Lewis, S. Shleifer, and L. Zettlemoyer. "LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale". In: *Advances in Neural Information Processing Systems*. 2022 (cit. on p. 51).

[34] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. "Qlora: Efficient finetuning of quantized llms". In: *Advances in Neural Information Processing Systems* 36 (2024) (cit. on p. 16).

[35] P. Developers. *PyTorch: An open-source machine learning framework*. https://pytorch.org. Accessed: 2024-11-20 (cit. on p. 13).

[36] S. learn Developers. *Scikit-learn: Machine Learning in Python*. https://scikit-learn.org. Accessed: 2024-11-20 (cit. on p. 12).

[37] T. Developers. *TensorFlow: An end-to-end open-source machine learning platform*. https://www.tensorflow.org. Accessed: 2024-11-20 (cit. on p. 12).

[38] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *arXiv preprint arXiv:1810.04805* (2018) (cit. on p. 15).

[39] A. Dosovitskiy. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020) (cit. on p. 8).

[40] J. D. Ekstrand and M. D. Ekstrand. "First do no harm: Considering and minimizing harm in recommender systems designed for engendering health". In: *Engendering Health Workshop at the RecSys 2016 Conference*. ACM. 2016, pp. 1–2 (cit. on p. 21).

[41] E. Ezin, E. Kim, and I. P. Carrascosa. "'Fitness that Fits':-A Prototype Model for Workout Video Recommendation". In: *12th ACM Conference on Recommender Systems*. 2018 (cit. on p. 19).

[42] C. Feely, B. Caulfield, A. Lawlor, and B. Smyth. "Providing explainable race-time predictions and training plan recommendations to marathon runners". In: *Proceedings of the 14th ACM Conference on Recommender Systems*. 2020, pp. 539–544 (cit. on p. 21).

[43] J. M. Fernández, M. Mamei, S. Mariani, F. Miralles, A. Steblin, E. Vargiu, F. Zambonelli, et al. "Towards argumentation-based recommendations for personalised patient empowerment". In: *Proceedings of the Eleventh ACM Conference on Recommender Systems*. 2017 (cit. on p. 20).

[44] S. Garg. "Drug recommendation system based on sentiment analysis of drug reviews using machine learning". In: *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. IEEE. 2021, pp. 175–181 (cit. on pp. 2, 8, 17, 18, 22, 23).

[45] S. Geng, S. Liu, Z. Fu, Y. Ge, and Y. Zhang. "Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5)". In: *Proceedings of the 16th ACM Conference on Recommender Systems*. 2022, pp. 299–315 (cit. on p. 25).

[46] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry. "Using collaborative filtering to weave an information tapestry". In: *Communications of the ACM* 35.12 (1992), pp. 61–70 (cit. on p. 7).

[47] C. A. Gomez-Uribe and N. Hunt. "The Netflix Recommender System: Algorithms, Business Value, and Innovation". In: *ACM Transactions on Management Information Systems (TMIS)*. Vol. 6. 4. ACM, 2015, pp. 1–19 (cit. on pp. 9, 10, 17).

[48] F. Gutiérrez, K. Verbert, and N. N. Htun. "PHARA: an augmented reality grocery store assistant". In: *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*. 2018, pp. 339–345 (cit. on p. 18).

[49] M. A. Haidar and M. Rezagholizadeh. "Textkd-gan: Text generation using knowledge distillation and generative adversarial networks". In: *Advances in Artificial Intelligence: 32nd Canadian Conference on Artificial Intelligence, Canadian AI 2019, Kingston, ON, Canada, May 28–31, 2019, Proceedings 32*. Springer. 2019, pp. 107–118 (cit. on p. 14).

[50] S. Haneuse, D. Arterburn, and M. J. Daniels. "Assessing missing data assumptions in EHR-based studies: a complex and underappreciated task". In: *JAMA Network Open* 4.2 (2021), e210184–e210184 (cit. on pp. 7, 24).

[51] A. Holzinger, A. C. Valdez, and M. Ziefle. "Towards interactive recommender systems with the doctor-in-the-loop". In: (2016) (cit. on pp. 2, 26).

[52] S. Hors-Fraile, F. J. N. Benjumea, L. C. Hernández, F. O. Ruiz, and L. Fernandez-Luque. "Design of two combined health recommender systems for tailoring messages in a smoking cessation app". In: *arXiv preprint arXiv:1608.07192* (2016) (cit. on p. 21).

[53] N. Huba and Y. Zhang. "Designing patient-centered personal health records (PHRs): health care professionals' perspective on patient-generated data". In: *Journal of medical systems* 36 (2012), pp. 3893–3905 (cit. on p. 21).

[54] A. S. Hussein, W. M. Omar, X. Li, and M. Ati. "Efficient chronic disease diagnosis prediction and recommendation system". In: *2012 IEEE-EMBS conference on biomedical engineering and sciences*. IEEE. 2012, pp. 209–214 (cit. on pp. 7, 18).

[55] F. O. Isinkaye, Y. O. Folajimi, and B. A. Ojokoh. "Recommendation systems: Principles, methods and evaluation". In: *Egyptian informatics journal* 16.3 (2015), pp. 261–273 (cit. on pp. 2, 7, 9, 17).

[56] F. Jabeen, M. Maqsood, M. A. Ghazanfar, F. Aadil, S. Khan, M. F. Khan, and I. Mehmood. "An IoT based efficient hybrid recommender system for cardiovascular disease". In: *Peer-to-Peer Networking and Applications* 12 (2019), pp. 1263–1276 (cit. on pp. 7, 18).

[57] A. Jameson. "A Tool That Supports the Psychologically Based Design of Health-Related Interventions." In: *HealthRecSys@ RecSys*. 2017, pp. 39–42 (cit. on p. 21).

[58] S. Jamshidi, A. Torkamani, and J. Mellen. "A hybrid health journey recommender system using electronic medical records". In: *Proceedings of the 3rd International Workshop on Health Recommender Systems co-located with the 12th ACM Conference on Recommender Systems*. Vancouver, Canada, 2018, pp. 57–62 (cit. on pp. 18, 20).

[59] S. Kallumadi and F. Grer. *Drug Reviews (Drugs.com)*. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5SK5S. 2018 (cit. on pp. 3, 29).

[60] D. P. Kingma. "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114* (2013) (cit. on p. 14).

[61] A. Komatsuzaki. "One epoch is all you need". In: *arXiv preprint arXiv:1906.06669* (2019) (cit. on p. 51).

[62] A. Kumar and S. Shekhar. "Original Research Article Hybrid model of unsupervised and supervised learning for multiclass sentiment analysis based on users' reviews on healthcare web forums". In: *Journal of Autonomous Intelligence* 7.4 (2024) (cit. on p. 14).

[63] S. Lai, X. Hu, H. Xu, Z. Ren, and Z. Liu. "Multimodal sentiment analysis: A survey". In: *Displays* (2023), p. 102563 (cit. on p. 23).

[64] S. Latif, A. Zaidi, H. Cuayahuitl, F. Shamshad, M. Shoukat, and J. Qadir. "Transformers in speech processing: A survey". In: *arXiv preprint arXiv:2303.11607* (2023) (cit. on p. 8).

[65] E. Lee, F. Rustam, H. F. Shahzad, P. B. Washington, A. Ishaq, and I. Ashraf. "Drug Usage Safety from Drug Reviews with Hybrid Machine Learning Approach." In: *Computer Systems Science & Engineering* 46.1 (2023) (cit. on pp. 22, 23).

[66] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining". In: *Bioinformatics* 36.4 (2020), pp. 1234–1240 (cit. on pp. 8, 14, 22, 24–26).

[67] X. Lei, Z. Fang, and L. Guo. "Predicting circRNA–disease associations based on improved collaboration filtering recommendation system with multiple data". In: *Frontiers in genetics* 10 (2019), p. 897 (cit. on pp. 7, 18, 20).

[68] D. Li, B. Jiang, L. Huang, A. Beigi, C. Zhao, Z. Tan, A. Bhattacharjee, Y. Jiang, C. Chen, T. Wu, et al. "From Generation to Judgment: Opportunities and Challenges of LLM-as-a-judge". In: *arXiv preprint arXiv:2411.16594* (2024) (cit. on pp. 26, 27).

[69] F. Li, C. Cui, Y. Hu, and L. Wang. "Sentiment Analysis of User Comment Text based on LSTM". In: *WSEAS Transactions on Signal Processing* 19 (2023), pp. 19–31 (cit. on p. 13).

[70] J. Li, M. Wang, J. Li, J. Fu, X. Shen, J. Shang, and J. McAuley. "Text is all you need: Learning language representations for sequential recommendation". In: *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2023, pp. 1258–1267 (cit. on p. 25).

[71] L. Li, Y. Zhang, and L. Chen. "Prompt distillation for efficient llm-based recommendation". In: *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 2023, pp. 1348–1357 (cit. on p. 25).

[72] B. Liu. "Sentiment analysis: A fascinating problem". In: *Sentiment Analysis and Opinion Mining*. Springer, 2012, pp. 1–8 (cit. on p. 10).

[73] Q. Liu, J. Hu, Y. Xiao, X. Zhao, J. Gao, W. Wang, Q. Li, and J. Tang. "Multimodal recommender systems: A survey". In: *ACM Computing Surveys* 57.2 (2024), pp. 1–17 (cit. on p. 27).

[74] Q. Liu, X. Wu, X. Zhao, Y. Zhu, Z. Zhang, F. Tian, and Y. Zheng. "Large Language Model Distilling Medication Recommendation Model". In: *arXiv preprint arXiv:2402.02803* (2024) (cit. on pp. 2, 7, 23, 25, 26, 34).

[75] A. Meta. "Introducing LLaMA: A foundational, 65-billion-parameter large language model". In: *Meta AI. https://ai. facebook. com/blog/large-language-model-llama-meta-ai* (2023) (cit. on p. 34).

[76] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean. "Efficient Estimation of Word Representations in Vector Space". In: *International Conference on Learning Representations*. 2013. URL: https://api.semanticscholar.org/CorpusID:5959482 (cit. on p. 8).

[77] M. Nasiri, B. Minaei, and A. Kiani. "Dynamic recommendation: Disease prediction and prevention using recommender system". In: *International Journal of Basic Science in Medicine* 1.1 (2016), pp. 13–17 (cit. on pp. 7, 18).

[78] Z. A. Nazi and W. Peng. "Large language models in healthcare and medical domain: A review". In: *Informatics*. Vol. 11. 3. MDPI. 2024, p. 57 (cit. on pp. 25, 27).

[79] NousResearch. *Llama-2-7b-chat-hf*. https://huggingface.co/NousResearch/Llama-2-7b-chat-hf. 2023 (cit. on p. 50).

[80] B. Panda, C. R. Panigrahi, and B. Pati. "Exploratory data analysis and sentiment analysis of drug reviews". In: *Computación y Sistemas* 26.3 (2022), pp. 1191–1199 (cit. on pp. 1, 2, 23, 24).

[81] V. Pandey, D. D. Upadhyay, N. Nag, and R. C. Jain. "Personalized User Modelling for Context-Aware Lifestyle Recommendations to Improve Sleep." In: *HealthRecSys@ RecSys*. 2020, pp. 8–14 (cit. on p. 20).

[82] Y.-J. Park, A. Pillai, J. Deng, E. Guo, M. Gupta, M. Paget, and C. Naugler. "Assessing the research landscape and clinical utility of large language models: A scoping review". In: *BMC Medical Informatics and Decision Making* 24.1 (2024), p. 72 (cit. on pp. 26, 60).

[83] A. Pasta, M. K. Petersen, K. J. Jensen, and J. E. Larsen. "Rethinking hearing aids as recommender systems". In: *CEUR Workshop Proceedings*. Vol. 2439. CEUR-WS. 2019, pp. 11–17 (cit. on p. 21).

[84] M. Pato. *The ISELthesis LaTeX Template's Manual*. Instituto Superior de Engenharia de Lisboa (ISEL-IPL). 2024. URL: https://github.com/matpato/iselthesis (cit. on p. viii).

[85] M. Pato, M. Barros, and F. M. Couto. "Survey on Recommender Systems for Biomedical Items in Life and Health Sciences". In: *ACM Computing Surveys* 56.6 (2024), pp. 1–32 (cit. on p. 9).

[86] F. Pecune, L. Callebert, and S. Marsella. "A Recommender System for Healthy and Personalized Recipes Recommendations". In: *Proceedings of the Workshop on Health Recommender Systems at RecSys*. 2020, pp. 15–20 (cit. on p. 18).

[87] P. Pilloni, L. Piras, L. Boratto, S. Carta, G. Fenu, F. Mulas, et al. "Recommendation in persuasive eHealth systems: An effective strategy to spot users' losing motivation to exercise". In: *CEUR Workshop Proceedings*. Vol. 1953. 2017, pp. 6–9 (cit. on p. 19).

[88] A. S. Pinto, M. Pato, and N. Datia. "Explainable Feature Ranking using Interactive Dashboards". In: *2024 28th International Conference Information Visualisation (IV)*. IEEE. 2024. DOI: 10.1109/IV64223.2024.00042 (cit. on pp. 3, 4).

[89] J. J. Prochaska and N. L. Benowitz. "The past, present, and future of nicotine addiction therapy". In: *Annual review of medicine* 67.1 (2016), pp. 467–486 (cit. on pp. 7, 18).

[90] *Recurrent Neural Networks - IBM*. https://www.ibm.com/topics/recurrent-neural-networks. Accessed: 2024-02-23 (cit. on p. 13).

[91] Repugen. *Sentiment Analysis in Healthcare: Enhancing Patient Experience and Healthcare Delivery*. Accessed: 2024-11-12. 2023. URL: https://www.repugen.com/blog/sentiment-analysis-in-healthcare (cit. on p. 22).

[92] M. Research. *Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone*. Accessed: September 9, 2024. 2023. URL: https://www.microsoft.com/en-us/research/publication/phi-3-technical-report-a-highly-capable-language-model-locally-on-your-phone/ (cit. on p. 35).

[93] A. Sadia, F. Khan, and F. Bashir. "An overview of lexicon-based approach for sentiment analysis". In: *2018 3rd International Electrical Engineering Conference (IEEC 2018)*. 2018, pp. 1–6 (cit. on pp. 10, 11).

[94] F. Sebastiani and A. Esuli. "Sentiwordnet: A publicly available lexical resource for opinion mining". In: *Proceedings of the 5th international conference on language resources and evaluation*. European Language Resources Association (ELRA) Genoa, Italy. 2006, pp. 417–422 (cit. on p. 12).

[95] P. Siriaraya, K. Suzuki, and S. Nakajima. "Utilizing Collaborative Filtering to Recommend Opportunities for Positive Affect in daily life." In: *HealthRecSys@ RecSys*. 2019, pp. 2–3 (cit. on p. 19).

[96] Skillcate. *Sentiment Analysis Using NLTK Vader*. https://medium.com/@skillcate/sentiment-analysis-using-nltk-vader-98f67f2e6130. Accessed: 2024-02-23. 2020 (cit. on pp. 12, 22).

[97] *spaCy: Industrial-Strength Natural Language Processing*. https://spacy.io/. Accessed: 2024-02-23 (cit. on p. 12).

[98] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. "Lexicon-based methods for sentiment analysis". In: *Computational linguistics* 37.2 (2011), pp. 267–307 (cit. on p. 8).

[99] *TextBlob: Simplified Text Processing*. https://textblob.readthedocs.io/en/dev/. Accessed: 2024-02-23 (cit. on pp. 12, 22).

[100] R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, et al. "Lamda: Language models for dialog applications". In: *arXiv preprint arXiv:2201.08239* (2022) (cit. on p. 25).

[101] M. A. Torkamani, M. Jhaveri, J. Mellen, M. Brown-Hayes, J. Chung, B. Pan, and H. Kardes. "Engagement Scoring for Care-gap Intervention Optimization." In: *HealthRecSys@ RecSys*. 2018, pp. 53–56 (cit. on p. 20).

[102] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. "Llama: Open and efficient foundation language models". In: *arXiv preprint arXiv:2302.13971* (2023) (cit. on pp. 24, 34).

[103] H. Touvron et al. "Llama 2: Open Foundation and Fine-Tuned Chat Models". In: *arXiv preprint arXiv:2307.09288* (2023). URL: https://ar5iv.labs.arxiv.org/html/2307.09288 (cit. on pp. 24, 33, 35).

[104] T. N. T. Tran, A. Felfernig, C. Trattner, and A. Holzinger. "Recommender systems in the healthcare domain: state-of-the-art and research issues". In: *Journal of Intelligent Information Systems* 57.1 (2021), pp. 171–201 (cit. on pp. 1, 2, 7, 17, 18, 26, 60).

[105] *Transformers: State-of-the-art Natural Language Processing*. https://huggingface.co/docs/transformers/en/index. Accessed: 2024-02-23 (cit. on p. 12).

[106] C. Trattner and D. Elsweiler. "An evaluation of recommendation algorithms for online recipe portals". In: *Proceedings of the CEUR Workshop on Health Recommender Systems*. CEUR Workshop Proceedings. 2019 (cit. on pp. 18, 19).

[107] A Vaswani. "Attention is all you need". In: *Advances in Neural Information Processing Systems* (2017) (cit. on pp. 8, 14).

[108] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le. "Finetuned language models are zero-shot learners". In: *arXiv preprint arXiv:2109.01652* (2021) (cit. on p. 25).

[109] M. Wiesner and D. Pfeifer. "Health recommender systems: concepts, requirements, technical basics and challenges". In: *International journal of environmental research and public health* 11.3 (2014), pp. 2580–2607 (cit. on pp. 7, 18).

[110] B. Workshop, T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, et al. "Bloom: A 176b-parameter open-access multilingual language model". In: *arXiv preprint arXiv:2211.05100* (2022) (cit. on p. 25).

[111] L. Wu, Z. Zheng, Z. Qiu, H. Wang, H. Gu, T. Shen, C. Qin, C. Zhu, H. Zhu, Q. Liu, et al. "A survey on large language models for recommendation". In: *arXiv preprint arXiv:2305.19860* (2023) (cit. on pp. 8, 15, 25, 27).

[112] G. Xu, Y. Meng, X. Qiu, Z. Yu, and X. Wu. "Sentiment analysis of comment texts based on BiLSTM". In: *Ieee Access* 7 (2019), pp. 51522–51532 (cit. on pp. 22, 23).

[113] Z. Yan. "Evaluating the Effectiveness of LLM-Evaluators (aka LLM-as-Judge)". In: *eugeneyan.com* (2024). URL: https://eugeneyan.com/writing/llm-evaluators/ (cit. on pp. 26, 60).

[114] L. Yu, W. Zhang, J. Wang, and Y. Yu. "Seqgan: Sequence generative adversarial nets with policy gradient". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 31. 1. 2017 (cit. on p. 14).

[115] W. Yue, Z. Wang, W. Liu, B. Tian, S. Lauria, and X. Liu. "An optimally weighted user-and item-based collaborative filtering approach to predicting baseline data for Friedreich's Ataxia patients". In: *Neurocomputing* 419 (2021), pp. 287–294 (cit. on pp. 7, 18, 20).

[116] W. Yue, Z. Wang, B. Tian, M. Pook, and X. Liu. "A hybrid model-and memory-based collaborative filtering algorithm for baseline data prediction of Friedreich's ataxia patients". In: *IEEE Transactions on Industrial Informatics* 17.2 (2020), pp. 1428–1437 (cit. on p. 20).

[117] W. Yue, Z. Wang, J. Zhang, and X. Liu. "An overview of recommendation techniques and their applications in healthcare". In: *IEEE/CAA Journal of Automatica Sinica* 8.4 (2021), pp. 701–717 (cit. on pp. 7, 9, 18).

[118] Z. Yue, S. Rabhi, G. d. S. P. Moreira, D. Wang, and E. Oldridge. "LlamaRec: Two-stage recommendation using large language models for ranking". In: *arXiv preprint arXiv:2311.02089* (2023) (cit. on pp. 8, 25, 26, 34).

[119] H. Zhan, H. Zhang, H. Chen, L. Shen, Y. Lan, Z. Ding, and D. Yin. "User-inspired posterior network for recommendation reason generation". In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2020, pp. 1937–1940 (cit. on pp. 22, 24).

[120] L. Zhang, X. Chen, N.-N. Guan, H. Liu, and J.-Q. Li. "A hybrid interpolation weighted collaborative filtering method for anti-cancer drug response prediction". In: *Frontiers in pharmacology* 9 (2018), p. 1017 (cit. on pp. 7, 18).

[121] Z. Zhao. "Using Pre-trained Language Models for Toxic Comment Classification". PhD thesis. University of Sheffield, 2022 (cit. on p. 17).

[122] X. Zhou, X. Wei, A. Cheng, Z. Liu, Z. Su, J. Li, R. Qin, L. Zhao, Y. Xie, Z. Huang, et al. "Mobile Phone–Based Interventions for Smoking Cessation Among Young People: Systematic Review and Meta-Analysis". In: *JMIR mHealth and uHealth* 11.1 (2023), e48253 (cit. on pp. 7, 18).

[123] Z. Zhou, Y. Ji, W. Li, P. Dutta, R. Davuluri, and H. Liu. "Dnabert-2: Efficient foundation model and benchmark for multi-species genome". In: *arXiv preprint arXiv:2306.15006* (2023) (cit. on p. 8).

[124] Y. Zhu, D. Su, L. He, L. Xu, and D. Yu. "Generative Pre-trained Speech Language Model with Efficient Hierarchical Transformer". In: *arXiv preprint arXiv:2406.00976* (2024) (cit. on p. 8).

[125] M. Zomorodi, I. Ghodsollahee, P. Plawiak, and U. R. Acharya. "RECOMMED: A Comprehensive Pharmaceutical Recommendation System". In: *arXiv preprint arXiv:2301.00280* (2022) (cit. on pp. 17, 18).

[126] A. Zunic, P. Corcoran, and I. Spasic. "Sentiment analysis in health and well-being: systematic review". In: *JMIR medical informatics* 8.1 (2020), e16023 (cit. on pp. 1, 2, 7, 22, 23).

I Results of the DRecSys

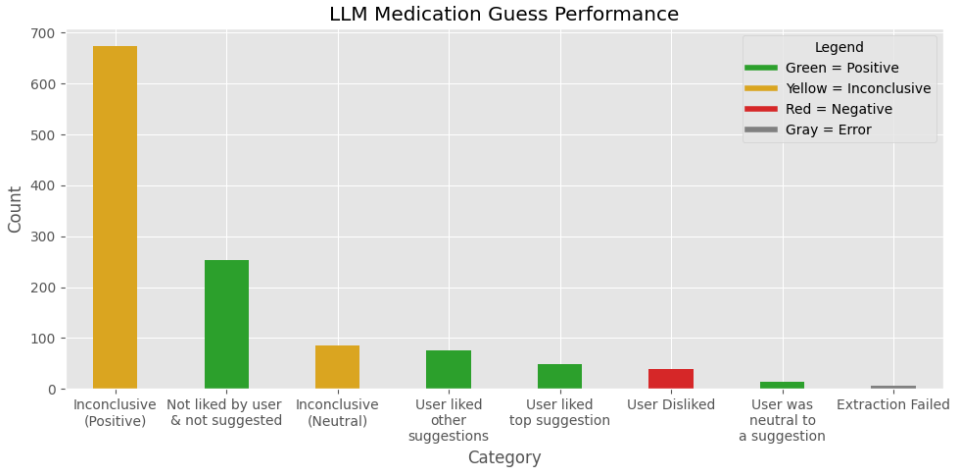


Figure I.1: *Performance of the DRecSys with the original dataset with SKR task*

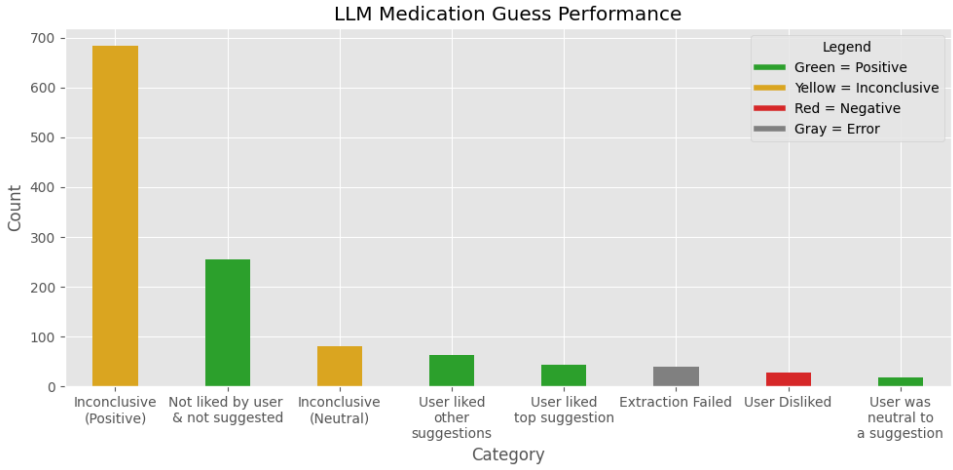


Figure I.2: *Performance of the DRecSys with the original dataset without SKR task*

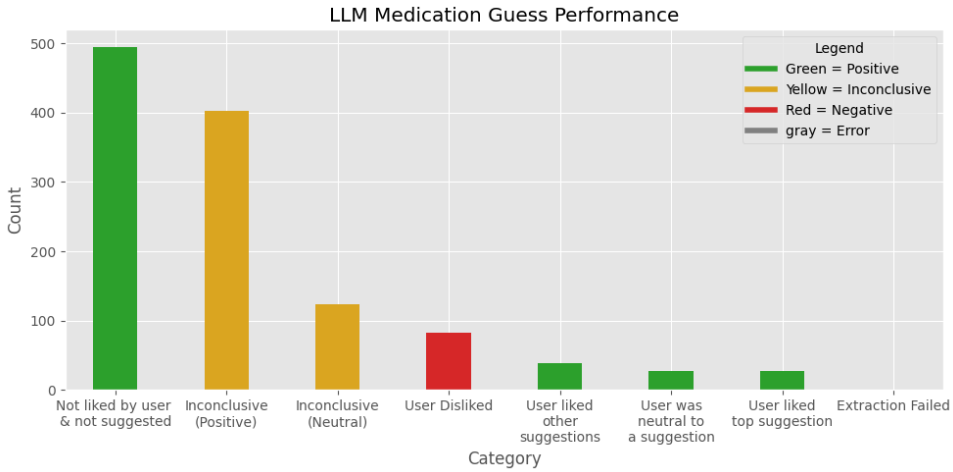


Figure I.3: *Performance of the DRecSys with the clean Vader dataset without SKRtask*

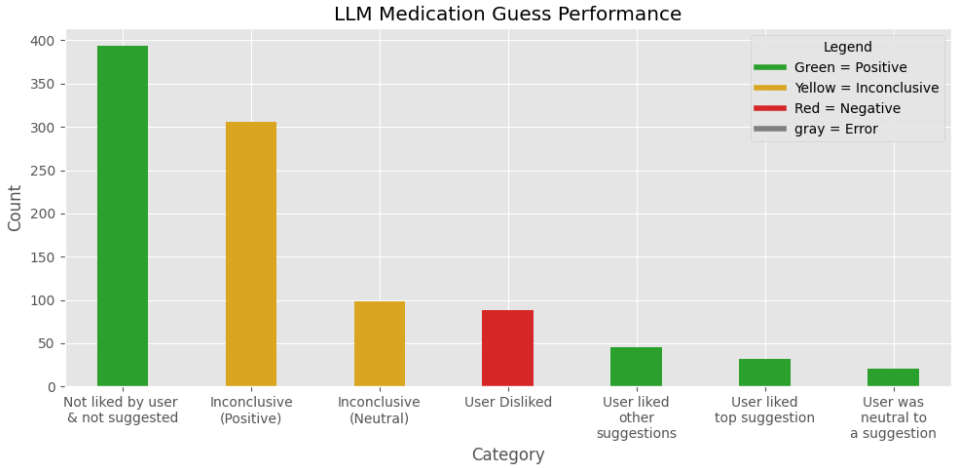


Figure I.4: *Performance of the DRecSys with the clean Vader dataset with SKR task*

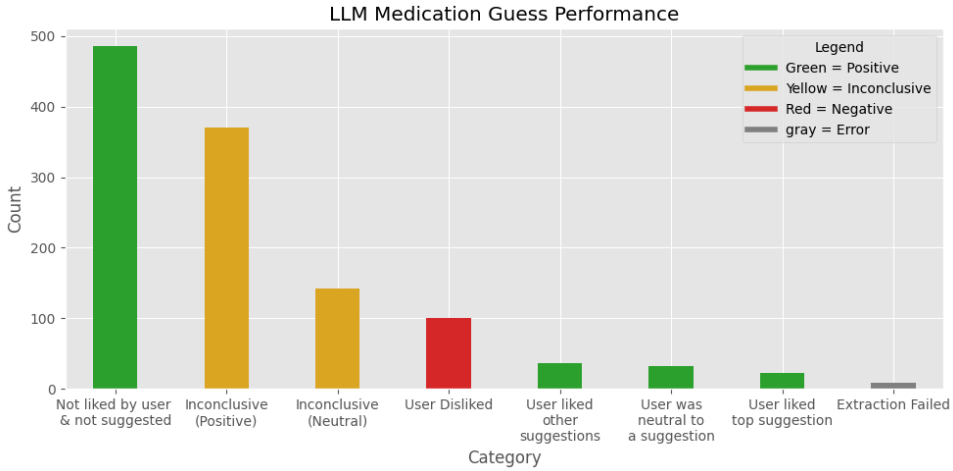


Figure I.5: *Performance of the DRecSys with the raw Vader dataset without SKR task*

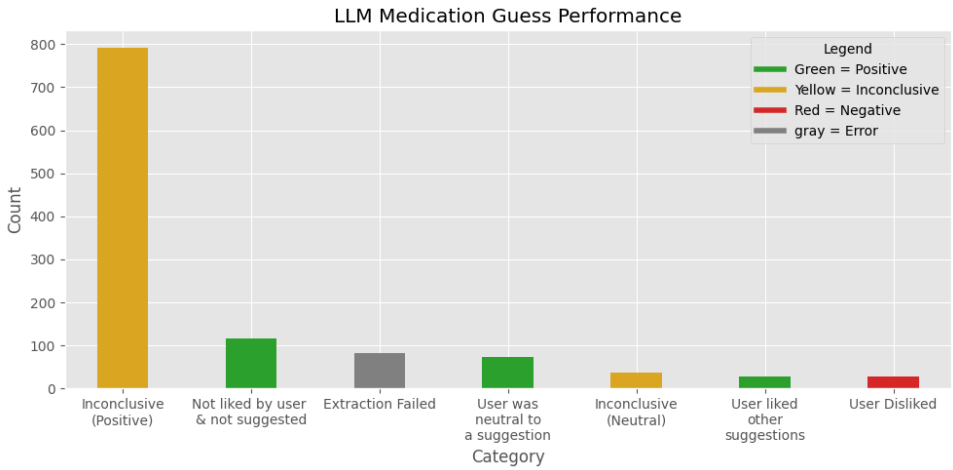


Figure I.6: *Performance of the DRecSys using LLama2 with no fine-tuning*

@is@a@figure