ISEL
INSTITUTO SUPERIOR DE ENGENHARIA DE LISBOA

A Topic Modelling-Based Recommender System for Drugs Using User Experience Reviews [TopicDrugRec]

RAFAEL REIS DE CARVALHO

(Licenciado em Engenharia Eletrónica e de Telecomunicações e Computadores)

Dissertação para obtenção do Grau de Mestre em

Engenharia Informática e de Computadores

Orientadores: Doutora Matilde Pós-de-Mina Pato

Doutor Nuno Miguel Soares Datia

Júri:

Presidente: Doutor José Manuel de Campos Lages Garcia Simão

Vogais: Doutora Vânia Patrícia Padrão Mendonça

Doutora Matilde Pós-de-Mina Pato

Novembro 2025

A Topic Modelling-Based Recommender System for Drugs Using User Experience Reviews [TopicDrugRec]

RAFAEL REIS DE CARVALHO

(Licenciado em Engenharia Eletrónica e de Telecomunicações e Computadores)

Dissertação para obtenção do Grau de Mestre em

Engenharia Informática e de Computadores

Orientadores: Doutora Matilde Pós-de-Mina Pato, ISEL-IPL

Doutor Nuno Miguel Soares Datia, ISEL-IPL

Júri:

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

Novembro 2025

Acknowledgements

I would like to begin the acknowledgements by expressing my gratitude **towards my advisors, Prof. Matilde Pato and Prof. Nuno Datia**. Your guidance, support, and expertise made this dissertation possible. The ongoing effort you put into guiding me and the way you constantly pushed me to challenge myself not only shaped my academic journey but also changed my approach in my daily work.

I am also thankful for the opportunity to study at **Instituto Superior de Engenharia de Lisboa**. During my six-year stay I got to meet great professionals and carry with me great friends. Additionally, it brought to me a sharper work ethic, by striving for great results, which made me grow as a professional, all while benefiting from the great environment that supported my studies.

To my parents: This would not have been possible without your ongoing support and encouragement, and the way you always believed in me. I will always remember how, during the writing of this dissertation, you kept cheering me up, bringing me food to keep me going, or even asking when would the dissertation finally be finished. Additionally, I hope you don't forget the basics of VS Code that I taught you so you could run my code while I was away in class. Thank you for supporting me in every decision and for always being there.

To my girlfriend: Thank you for motivating me every single day throughout this year-long journey. Your patience, love, and comprehension gave me strength to keep moving forward, especially on the toughest days. I will always be grateful for understanding my time slots for working on the dissertation, which may have interfered with some dates we had planned. You have always and will always be a major source of motivation, and I am truly lucky to have had your support by my side.

To my friends, especially those I met at ISEL: I will never forget the party nights, study sessions, the labs, and the afternoons at Firmino playing pool, hitting lucky shots and saying it was all "vector analysis". Those moments will carry with me and I am grateful that you were part of this chapter of my life, and I look forward to creating more memories with you.

Finally, **to everyone** who has been part of my time at ISEL, you have my deepest **thank you**, I will carry the memories of this journey with me always.

Thank you, everyone, I will see you in the next chapter, whatever it will be.

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

Rafael Reis de Carvalho

Lisbon, September , 2025

A Topic Modelling-Based Recommender System for Drugs Using User Experience Review [TopicDrugRec]

This document was created using the (pdf)LaTeX processor, based in the "iselthesis" template [71], developed at the DEETC of ISEL-IPL.

Abstract

The increasing volume of patient-reported data, alongside the rise of personalised medicine has made it challenging for healthcare professionals to incorporate patient experiences into their clinical decision-making due to information overload and demanding working shifts. Much of this data is available in the form of numerical drug ratings, which often fail to capture the complexity of user experiences by lacking contextual information and response bias. To address this, this dissertation proposes **TopicDrugRec**, a drug recommender system based on topic modelling and trained on the UCI ML Drug Review dataset, designed to support clinicians in providing safer and more personalised drug prescriptions.

It follows a six step methodology: first, exploratory data analysis, data cleaning, followed by sentiment analysis to mitigate rating bias, topic modelling to extract latent themes from patient reports in the form of free text, integration of medical knowledge (drug-drug interactions, side effects and contraindications) to enhance patient safety, and the implementation of a web application and performance evaluation.

The recommendation algorithm was designed to incorporate topic similarity, user sentiment, and perceived usefulness, allowing for tunable hyperparameters to generate the recommendations. Three topic modelling approaches were evaluated: Latent Dirichlet Allocation, Non-negative Matrix Factorization, and BERTopic. The evaluation showed semantic similarity, derived from topic modelling, to be the most influential factor in recommendation quality. Additionally, grouping medical conditions into ICD-11 categories mitigated dataset imbalanced and improved coverage, with the NMF-based model achieving the best performance on this setup, with a Precision@10 of 0.513 and Mean Reciprocal Rank @10 of 0.676.

Despite being a proof-of-concept, these findings demonstrate TopicDrugRec's potential in reducing information overload, enhancing medic-patient interaction and integrating patient feedback into data-driven decision-making. Additionally, it lays foundation for future work, including real world validation, curating more complex datasets with patient information, and providing explainable recommendations.

Keywords: Recommender System, Drug Recommender System, Topic Modelling, Sentiment Analysis, Patient Reported Outcomes

Resumo

O aumento do volume de dados reportado por pacientes face a experiências passadas, aliado ao avanço da medicina personalizada, tem tornado desafiante para os profissionais de saúde incorporar a experiência dos doentes na tomada de decisão clínica devido ao excesso de informação e turnos prolongados. Muita desta informação provém das classificações numéricas, que não captam a complexidade da experiência do paciente dado que carecem de contexto pessoal, estando sujeito ao enviesamento. Com o objetivo de colmatar estas limitações, esta dissertação propõe o **TopicDrugRec**, um sistema de recomendação de medicamentos baseado em *topic modelling*, treinado no conjunto de dados UCI ML Drug Review.

A sua implementação seguiu uma metodologia em seis etapas: análise exploratória e limpeza dos dados, análise de sentimento para mitigar o enviesamento, *topic modelling* para extração de temas latentes das *reviews*, integração de conhecimento biomédico para reforçar a segurança do paciente, e por fim, a implementação numa aplicação web e avaliação de performance.

O algoritmo de recomendação incorpora como hiperparâmetros a semelhança de tópicos, o sentimento do utilizador e concordância com a avaliação, que podem ser ajustados para gerar as recomendações. Foram avaliadas três abordagens de *topic modelling*: Latent Dirichlet Allocation, Non-Negative Matrix Factorization e BERTopic. Os resultados demonstram que a semelhança de tópicos, derivada do *topic modelling*, é o fator mais influente na qualidade das recomendações. Em simultâneo, a agregação das condições médicas em categorias ICD-11 demonstrou mitigar o desbalanceamento do conjunto de dados, e melhorou a cobertura, sendo que o modelo NMF apresentou o melhor desempenho neste cenário, com uma Precisão@10 de 0.513, e *Mean Reciprocal Rank*@10 de 0.676.

Apesar de se tratar de uma prova de conceito, os resultados demonstram o potencial do TopicDrugRec em mitigar os efeitos do excesso de informação, integrar o *feedback* dos pacientes, e de potenciar maior interação entre o médico e o paciente. Para além disso, estabelece também base para trabalho futuro, incluindo a validação do algoritmo em cenários clinicos reais, a criação de conjuntos de dados mais completos com informação do paciente e de disponibilizar uma explicação informada das recomendações.

Palavras-chave: Sistema de Recomendação, Sistema de Recomendação de medicamentos, Topic Modelling, Análise de Sentimento, Relato de Pacientes

Contents

List of Figures

List of Tables

Acronyms

AI	Artificial Intelligence 2, 26
API	Application Programming Interface 30
BERT	Bidirectional Encoder Representations from Transformers xvi, 3, 9, 11, 12, 13, 19, 20, 32, 33, 34, 35, 36, 55, 56, 57
BoW	Bag of Words xvii, 9, 10, 12, 18
CB	Content-based Filtering 3, 5, 7, 8, 17, 18, 21, 22, 37, 75
CF	Collaborative Filtering 5, 7, 8, 17, 18, 21, 22, 23, 24, 37
CHAID	Chi-Squared Automatic Interaction Detection 20
CNN	Convolutional Neural Network 19
DBSCAN	Density-Based Spatial Clustering of Applications with Noise 23
DDI	Drug-Drug Interaction 37, 38, 68, 69, 70, 73
DL	Deep Learning 11
EDA	Exploratory Data Analysis 4, 5, 23, 24, 27, 28, 29, 41, 42, 46, 83, 84
EHR	Electronic Health Record 24, 26
GA	Genetic Algorithm 18
GA-KM	Genetic Algorithm-optimized K-means clustering 18
HDBSCAN	Hierarchichal Density-Based Spatial Clustering of Applications with Noise 12, 13
HR	Hit Rate 22, 25
HRS	Health Recommender System 5, 21
ICD11	International Classification of Diseases 4, 5, 6, 30, 42, 44, 49, 73, 74, 77, 78, 79, 84, 85, 87, 88
KB	Knowledge-based 3, 4, 5, 8, 21, 37
KG	Knowledge Graph 3, 21
LDA	Latent Dirichlet Allocation xiv, xvi, xvii, 5, 9, 10, 18, 35, 36, 38, 39, 48, 50, 56, 57, 58, 59, 60, 61, 65, 66, 67, 68, 71, 74, 75, 76, 78, 79, 80, 81, 84
LLM	Large Language Model 24, 25, 26

MAE Mean Absolute Error 18
MAP Mean Average Precision 24, 39, 74, 75, 76, 77, 78, 79, 80, 81, 84, 85
MAR Mean Average Recall 39, 74, 75, 76, 78, 79, 80, 81, 84, 85
ML Machine Learning 2, 4, 8, 14, 15, 20, 49, 73, 83, 84
MRR Mean Reciprocal Rank 25, 39, 74, 75, 76, 77, 78, 79, 80, 81, 84, 85
MSE Mean Square Error 23

nDCG normalized Discounted Cumulative Gain 25
NER Named Entity Recognition 68
NLP Natural Language Processing 1, 2, 5, 11, 13, 14, 17, 18, 25, 26, 30, 41, 84
NLTK Natural Language Toolkit 30, 32, 34
NMF Non-negative Matrix Factorization xiv, xv, xvi, xvii, 5, 9, 11, 12, 35, 36, 38, 39, 48, 50, 56, 60, 61, 62, 63, 65, 66, 67, 68, 71, 74, 75, 76, 77, 78, 79, 80, 81, 84, 85

PCA Principal Component Analysis 18, 23
PoS Part-of-Speech 19, 20
PRO Patient-Reported Outcomes 1, 2, 4, 35, 41, 83, 84, 86

RMSE Root Mean Square Error 24
RS Recommender System 1, 2, 3, 4, 5, 7, 8, 17, 18, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 35, 36, 37, 38, 39, 40, 41, 43, 44, 45, 49, 71, 74, 77, 81, 83, 84, 85, 86, 88

SA Sentiment Analysis xv, 2, 4, 5, 6, 7, 13, 14, 17, 19, 20, 24, 25, 27, 29, 31, 32, 33, 34, 35, 38, 41, 50, 51, 52, 53, 83, 84
SVM Support Vector Machine 23

TF-IDF Term Frequency Inverse Document Frequency 12, 13
TM Topic Modelling xv, 1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 18, 25, 27, 29, 31, 35, 36, 38, 39, 40, 41, 43, 46, 47, 50, 51, 56, 63, 65, 67, 68, 71, 75, 83, 84

UMAP Uniform Manifold Approximation and Projection 12, 13

VADER Valence Aware Dictionary and sEntiment Reasoner xv, 33, 34, 53, 54

1 Introduction

This chapter serves to frame the research domain, establish its relevance, and highlight opportunities for incorporating a Recommender System (RS) within the healthcare sector. It outlines the motivation behind this research and provides an overview of the proposed solution, which explores the use of Topic Modelling (TM) to build a RS that incorporates user sentiment. This approach aims to improve the quality of drug prescriptions, reducing information overload on health professionals, while also addressing the medical implications associated with such technology. Furthermore, key objectives are presented, along with the structure of the document to guide the reader through this dissertation.

In parallel with technological advancements, increasingly vast amounts of data are being collected, interpreted, and contextualized, yielding valuable insights into patient health, treatment effectiveness, and healthcare delivery. Patient-Reported Outcomes (PRO), such as user reviews of drugs, represent an under-explored source of information [76, 95]. These qualitative reviews offer insight into patient experiences, treatment efficacy, and the occurrence of adverse effects, which can inform more personalised medical approaches.

In particular, the application of Natural Language Processing (NLP), and more specifically, TM techniques to leverage PRO for personalized drug recommendation is a relatively unexplored subject. Incorporating such user-generated content into clinical decision-making has shown promise in enhancing treatment outcomes. For instance, [37] demonstrated that including PRO in routine cancer care improved communication between patients and healthcare providers, leading to increased satisfaction and more effective care.

1.1 Motivation

Given the rise in data collection, and the increasing emphasis on patient-oriented medical decision-making, healthcare professionals face significant cognitive challenges due to information overload. More often, physicians struggle to effectively interpret and integrate qualitative patient-generated information, such as PROs, into their clinical decision-making, especially given their demanding work environments, and increasingly busy and long working shifts.

Recent research has demonstrated a strong correlation between healthcare professionals' workload and increased error rates. In [47], 1.6 million prescriptions made by 1,066 physicians are analysed, revealing that those working busy shifts were 8.2 **times more likely to make prescription errors** compared to lighter shifts. This had a big impact on the quality of treatment. Likewise, working multiple shifts in a row also lead to increased error rates, with a **third consecutive shift** having an error rate of approximately 2.1% compared to 0.88% during the first shift. Recommender Systems can be used as a tool to address these challenges by condensing patient-generated content, including personal experiences and reported side effects, into accessible and user-friendly applications. These help refine diagnostic and prescription processes, alleviating the heavy cognitive workload faced by healthcare professionals, and improving the speed of clinical decisions.

Although PRO has the potential to improve clinical decision-making, it is important to recognize that user-generated content often has reliability problems and can be compromised by biases. The design and selection of items have been shown to have a significant impact on data integrity [13]. Specific wording, scale choices, and format can affect the quality of the collected data.

One notable issue in this context is the **Extreme Response Bias**, also known as extreme response style. This occurs when respondents overly select extreme ratings on Likert scales (e.g., consistently choosing "1" or "10" on a 0–10 scale, or repeatedly selecting "Strongly Agree" or "Strongly Disagree"), that do not match their true opinions or experiences. The number of response options provided can also contribute to this bias; for example, an odd number of choices may encourage respondents to select a neutral response, while an even number of choices could unintentionally bias responses toward one side, influencing data quality and interpretation [44].

Machine Learning (ML) and Artificial Intelligence (AI) techniques, particularly within RS offer possible solutions to mitigate these reliability and interpretations issues. By leveraging those algorithms, RSs facilitate the analysis of patient-generated data, enabling healthcare professionals to identify meaningful patterns and further comprehend complex data [63]. However, these technologies do not seek to replace healthcare providers but rather aim to complement their expertise, reduce cognitive overloads and ideally enhance clinical decision-making, allowing professionals more time to engage in direct patient interactions.

Aligned with the growth of ML and AI and the current challenges clinicians face due to information overload, this dissertation aims to investigate and present a proof of concept on how TM and Sentiment Analysis (SA) algorithms can be leveraged and adapted to the medical scenario to build a drug RS.

1.2 Proposed Solution

Considering these challenges and technological advancements, this research proposes a **Topic Modelling-based Drug Recommender System**, named ***TopicDrugRec***. This system aims to improve medical prescription process by integrating **user-generated drug reviews**, ML, NLP, and SA techniques as well as considering patient concerns using a

Knowledge-based (KB) approach. Using real-world patient experiences captured in reviews, **TopicDrugRec** generates recommendations that provide a **patient-centred perspective**, aiming not only to improve drug prescription effectiveness, but also to align then the user experience with individual patient needs.

To address challenges like **extreme response bias** in patient ratings, the system employs a **Bidirectional Encoder Representations from Transformers (BERT)-based Sentiment Analysis approach**. By analysing the textual content of reviews, the system detects discrepancies between the sentiment expressed in the text and the corresponding numerical rating, performing corrections. This reduces the impact of skewed data distributions, further enhancing the quality of recommendations. Furthermore, TM is integrated to extract qualitative insights from patient-reported outcomes, allowing RS to focus on expressive feedback. This approach not only minimises the reliance on potentially biased numerical data but also enables more accurate and balanced recommendations. Finally, in terms of patient safety, a Knowledge Graph (KG) is used to consider known negative interactions between drugs, while also providing additional information such as side effects and warnings or considerations when taking a recommended medicine.

The UCI ML Drug Review dataset [85] serves as our foundation, providing rich medication review data that include patient experiences, associated conditions, satisfaction ratings on a 10-star scale, and helpfulness metrics based on user votes. This comprehensive dataset enables **TopicDrugRec** to analyse patient feedback patterns and sentiment, facilitating the extraction of meaningful topics while informing personalized drug recommendations.

The patient safety component is assessed using **DDInter** [102], which is a professional and open-access database specific to drug-drug interactions, providing annotations between each association including the mechanism description, risk levels, and alternative medications.

Recommendations are generated using a hybrid approach, that combines Content-based Filtering (CB) and KB approaches. The recommendation matrix is constructed by grouping similar reviews based on topic distributions and incorporating user sentiment as tunable hyper parameters in the ranking process. This ranking is determined using a **weighted scoring function**, providing a flexible and adaptable approach to providing recommendations.

TopicDrugRec is deployed as a **Flask-based web application**, which handles **data processing, model inference, and recommendation ranking** offering a structured approach of how RS can be integrated into the medical prescription workflow. However, it is important to highlight that this implementation remains a proof of concept, where its usability and effectiveness must be further assessed through user studies and by considering questionnaires.

1.3 Research Questions

This research aims to answer three key questions that reflect the challenges of leveraging user reviews, addressing data biases, and incorporating safety considerations into the recommendation process:

Q1. How effective are embedding-based ML libraries in mitigating extreme response bias in Likert-Scale ratings by extracting sentiment from textual user reviews?

Extreme response bias occurs when users disproportionately select extreme values on a Likert Scale (e.g., always giving '1' or '10'), resulting in skewed data distributions and consequently reducing informational features. To address this issue, I propose a **BERT-based SA** approach. This method analyses the context of a review to extract the sentiment expressed in the text. If differences between the text and the numerical rating are detected, a correction is performed so the rating better aligns with the content of the review.

Q2. How can unstructured PRO be leveraged to enhance drug recommendations and improve medical treatment?

In addition to SA, TM is also integrated into the process, enabling the RS to focus on qualitative insights extracted from patient reviews. By incorporating thematic information, the system reduces reliance on potentially biased quantitative data, resulting in more accurate and balanced recommendations, while focusing on more expressive patient feedback.

Q3. In the context of medical recommendations based on a RS, how can external knowledge be integrated to ensure the safety of the recommendations?

To address this issue this research takes a **safety-oriented approach** by incorporating a KB system that integrates information of known drug-drug interactions and known side effects. This system excludes negative combinations from the recommendations while suggesting safe alternatives when available.

The proposed recommendation system **ensures personalized suggestions** while maintaining safety standards, thereby enhancing both trustworthiness and reliability in drug recommendations.

1.4 Contributions

The main contributions of this work are:

- **The demonstration of the effectiveness of ICD11** groupings in mitigating class imbalance across medical conditions. By decreasing granularity, we showed substantially higher coverage of the dataset within the top 10 most represented target features. This approach proved successful in both the EDA, where coverage nearly doubled from 44.5% at the condition level to 83% at the disease-group level, as well as during the evaluation of recommendation performance, where ICD11-based models consistently yielded better precision and ranking quality metrics.
- **The practical demonstration of a Drug Recommender System through a functional web-based application.** This dissertation also contributed with a deployable and containerised application, enabling reproducible and scalable testing in real-world scenarios. It integrates the recommendation algorithm and external

biomedical knowledge, allowing the users to consult drug side effects, adverse reactions and drug interactions.

- **Academic contributions.** This dissertation also served for producing a scientific article, intended for submission to the *ACM Transactions on Recommender Systems* or *Bioinformatics*. In addition, all the developed code is publicly available to support future research, accessible in `https://github.com/matpato/TopicDrugRec`. Furthermore, the core datasets that power TopicDrugRec are also available in Zenodo, via `https://zenodo.org/records/17188352` [12].

1.5 Document Structure

The document is structured as follows:

- **Chapter 2: Technical Background** covers theoretical foundations that support the dissertation. It provides an overview of Recommender Systems, the most used algorithms, such as Collaborative Filtering (CF) and Content-based Filtering (CB), extending to Knowledge-based (KB). Then, Topic Modelling is discussed, explaining its purpose and further diving into three models: LDA, Non-negative Matrix Factorization (NMF) and BERT, highlighting their underlying algorithmic differences. Finally, it covers Sentiment Analysis, its applications, a typical data processing pipeline for these use cases, and the three categories covered in literature.

- **Chapter 3: Related Work** provides an overview of existing research on Recommender System, with a particular focus on their application in the healthcare sector. It begins by examining foundational approaches to RS and their applications in commercial domains, followed by a review of research on NLP and SA. Finally, it explores Health Recommender System and **Drug Recommender Systems**, analysing studies that integrate machine learning and knowledge-based approaches to enhance clinical decision-making.

- **Chapter 4: The TopicDrugRec System Methodologies** presents the methodologies, tools, and frameworks adopted in the development of **TopicDrugRec**, while addressing the research questions defined. The chapter begins by introducing the six step developing pipeline, beginning with Exploratory Data Analysis (EDA) to characterize the data, and **Data Cleaning and Preprocessing** to introduce International Classification of Diseases (ICD11) groupings and standardize the dataset for the further SA stage, where sentiment is extracted from the data. Continuing with TM where latent topics are extracted and evaluated through the presented metrics. An additional stage of **External Knowledge** is introduced and shown in the final stage of **Implementation and Evaluation**, which describes the algorithm and evaluates the **TopicDrugRec's** performance.

- **Chapter 5: Implementation and Evaluation of Core Components** presents the development stages of **TopicDrugRec**, highlighting the experimental setups

and interpreting results. It begins by evaluating the impact of ICD11 groupings into dataset coverage, followed by presenting the steps taken in **Data Cleaning** and their impact on original data. It extends to presenting the approach taken for SA and comparing the proposed models and its results. It extends to interpreting the results of each TM model, highlighting the best parameter configuration for each. Then, presents the results of integrating external knowledge and covers the web application and its use. The chapter ends with an extensive evaluation and interpretation of results of the recommendation performance throughout multiple test setups.

- **Chapter 6 Final Considerations** provides the final considerations of the dissertation. It summarises the findings through **Conclusions** and reflects on the limitations found during the development of **TopicDrugRec** in the **Discussion** subsection. Finally, it outlines potential **Future Work** to refine and improve upon the presented proof-of-concept.

2 Technical Background

This chapter presents the fundamental theoretical framework underlying **TopicDrugRec**, providing comprehensive coverage of Recommender System (RS), Sentiment Analysis (SA), and Topic Modelling (TM) methodologies. Through systematic examination of these interconnected disciplines, readers gain deeper insight into the theoretical foundations that drive our research methodology.

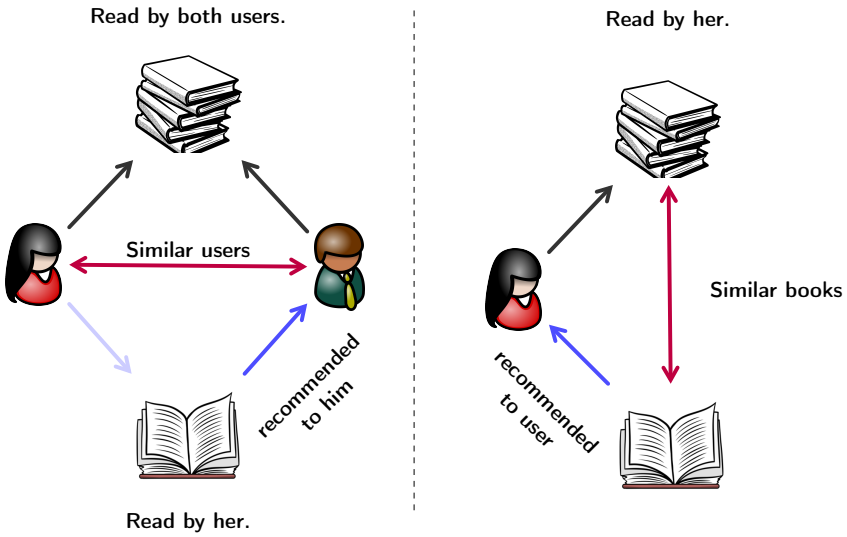
2.1 Recommender Systems

Recommendation systems represent a rapidly evolving field of research that helps users navigate increasingly vast amounts of information. As defined by [61], these systems play a crucial role in modern digital platforms:

" **A subclass of information filtering systems that aim to predict the "rating" or "preference" a user would give to a certain item. For vast amounts of data, these systems are critical in helping users find relevant content, suited to their likes.** "

They demonstrate significant advantages in sectors such as e-commerce (Amazon, Instacart), entertainment (Netflix, Spotify), social media (Facebook, TikTok), news (Wall Streets), and professional networking (LinkedIn), where they help users in filtering large amounts of content within a limited time, while also keeping them engaged and connected to the companies' products, thus boosting sales.

RS employs diverse methodologies to generate personalized recommendations. According to [2], **Collaborative Filtering** represents a well-known strategy, utilizing user-provided ratings to create targeted suggestions. However, this technique faces a significant limitation: the resulting ratings matrices typically exhibit sparsity, as most users have not rated the majority of available items. The main idea of this approach is that unspecified ratings can be imputed based on the observed ratings being highly correlated across different users and items. User ratings can take different forms, they can be explicit, such as Likert Scale [8] allowing for 1-10 or Negative, Neutral or Positive opinions; or implicit, inferred from user behaviour such as interactions or purchases. **Content-based Filtering (CB)**



(a) *Collaborative Filtering model* (b) *Content-based Filtering model*

Figure 2.1: *Techniques of recommender systems: Collaborative (left) and Content-based (right) Filtering. Hybrid approach represents the combination of both.*

uses the items' attributes to make recommendations. In these, the ratings and interaction behaviour are combined with the content information on the item; they are suitable when the ratings of other users are not available, but the items' descriptions are. They are advantageous in situations where new items appear and have insufficient ratings due to the comparison of attributes that might have been rated by the user in other items. The integration of multiple recommendation techniques through **Hybrid Filtering Method** offers a powerful solution for mitigating the individual limitations of each approach. By strategically combining different methodologies, hybrid systems can leverage the strengths of each component while compensating for their respective weaknesses, ultimately leading to more robust and effective recommendation systems.

Before exploring variations of these methods, **Knowledge-based (KB) Recommender System** are particularly useful in the context of items that are not purchased frequently [93]. In these cases, there may not be sufficient ratings, and the items have detailed options, for example, medicine/healthcare domain [84]. In this context, the pathology of the users may differ and specific properties must be considered.

The relevance of Knowledge-based RS to this dissertation lies in their capacity to navigate datasets and extract meaningful insights from user reviews via Topic Modelling techniques, while simultaneously incorporating individual patient characteristics, such as current medication regimens, to avoid potentially harmful drug interactions in recommendations.

2.2 Topic Modelling

Topic Modelling is an unsupervised Machine Learning (ML) technique used to uncover hidden topics or themes within a collection of documents [1]. In this research, TM is applied to automatically group together words that frequently co-occur across multiple reviews. The goal is to, given a set of documents, identify groups of words that represent distinct

topics, uncovering underlying themes within the corpora, as suggested in Figure 2.2. TM methods uncover latent themes in textual data by grouping words that frequently appear together. These methods can be categorised into three main types:

- **Probabilistic models**, such as Latent Dirichlet Allocation (LDA) [11], extract topics by assigning a probability distribution of topics to each document. These models operate under the assumption that documents are mixtures of topics, with each topic being a distribution over words.
- **Deterministic models**, like NMF [51], decompose the document-term matrix into lower-rank matrices, where each document and topic is represented using non-negative weights. These models focus on finding a direct representation of topics without incorporating uncertainty or randomness.
- **Contextual models**, such as BERT-based approaches like BERTopic [29], leverage deep learning to capture the semantic relationships between words, enabling a more nuanced topic representation.

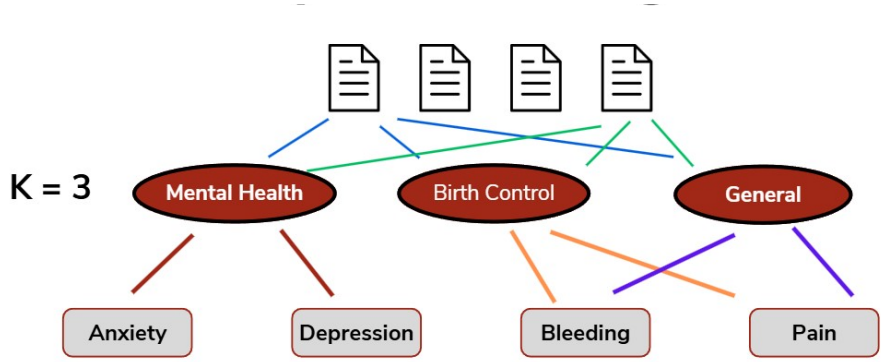


Figure 2.2: *Example of Topic Modelling for 4 documents, uncovering 3 hidden topics*

Latent Dirichlet Allocation

The LDA algorithm analyses word frequencies and co-occurrence patterns within individual documents which are modelled into a document-term matrix using a Bag of Words (BoW) representation [103], where each row represents a document, and each column represents a word. The document-term matrix then allows for LDA to assign a probabilistic distribution of the underlying themes within each review.

Table 2.1 illustrates the fundamental structure of a document-term matrix as applied to pharmaceutical review analysis. This matrix representation transforms unstructured patient narratives into quantifiable data by recording term frequencies across reviews. The first row demonstrates a review primarily concerned with drug effectiveness, mentioning "effective" twice while referencing side effects minimally. In contrast, the second row represents a review heavily focused on adverse reactions, with "side" and "effects" each appearing twice alongside mentions of specific symptoms like "pain" and "nausea". The third row exemplifies a predominantly positive review, with high frequencies for "effective"

(3 occurrences) and "helped" (2 occurrences), though notably including three mentions of "nausea", suggesting this side effect may be significant even in otherwise positive experiences. The fourth row presents a mixed review pattern, with moderate effectiveness mentions but higher frequencies for "pain" (3 occurrences), indicating this review likely discusses pain management experiences. This frequency-based representation enables LDA analysis of patient sentiment patterns and facilitates the identification of common themes across large collections of drugs reviews.

Table 2.1: Example of a Document-Term Matrix for Drug Reviews using Bag of Words (BoW) representation. Each cell represents the frequency of occurrence for specific terms within individual reviews, demonstrating how LDA processes textual data by quantifying word co-occurrence patterns across documents.

Review	**Effective**	**Side**	**Effects**	**Helped**	**Pain**	**Nausea**
Review 1	2	1	1	1	0	0
Review 2	0	2	2	0	1	1
Review 3	3	0	0	2	1	3
Review 4	1	1	1	0	3	2

Then, the algorithm uses a generative process based on Gibbs Sampling [23] to infer the topic distribution for each document, and the word distribution for each topic. Gibbs sampling works by iteratively assigning words in a document to a topic and evaluating, at each step, the likelihood of the word belonging to that topic. These assignments are refined over multiple iterations until the topic distributions stabilize.

The technical details and mathematical concepts of LDA are out of scope of this research, but are thoroughly explained in [11], where the authors detail the underlying formulas.

This algorithm involves different hyper parameters that affect the performance of the model such as:

1. **Number of Topics (K)**: Determines how many distinct topics will be uncovered in the corpus;
2. **Topic density (α)**: Controls how many topics appear in a document;
3. **Word density (β)**: Controls the number of words in a topic.

Once the model converges, LDA produces the word distribution for each topic, and the topic distribution for each document. For instance, Figure 2.3 represents two topics, with one consisting of words such as "nausea", "headache", "bleeding", "side" representing side effects, while the other reflects efficacy of the drug, with words like "effective", "relief", "improved" and "helped".

The document-topic distribution shows how much each topic contributes to the document, allowing the profiling of documents. Considering the following two example documents:

Document 1: *"This drug has a lot of side effects, I started bleeding and now I have a huge headache, I don't think this will improve."*

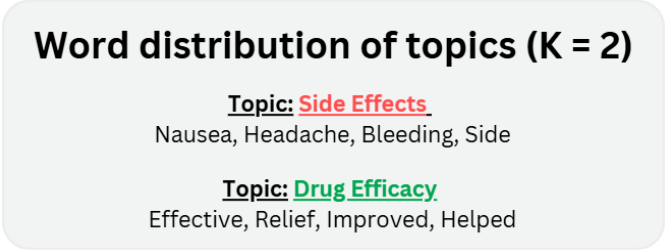


Figure 2.3: *Example of word distribution for two topics (K=2)*

Document 2: *"I started off feeling nauseous, however, after a while it seems to have improved and I feel relieved."*

Table 2.2: Example topic distribution for topics provided in Figure 2.3

	Side Effects	Drug Efficacy
Document 1	0.9	0.1
Document 2	0.1	0.9

Based on their textual content, the topic distributions might appear as shown in Table 2.2, with Document 1 being predominantly associated to "Side Effects" and Document 2 reflecting "Drug Efficacy".

Non-Negative Matrix Factorization

Non-negative Matrix Factorization (NMF) employs linear algebra principles to perform matrix factorization, decomposing a non-negative matrix $\mathbf{V}$ into two lower-dimensional non-negative matrices:

$$\mathbf{V} \approx \mathbf{W} \times \mathbf{H}$$

where $\mathbf{W}$, the **term-topic matrix**, represents the association between words and topics, and $\mathbf{H}$, the **topic-document matrix**, describes the contribution of each topic to a given document, as illustrated in Figure 2.4. NMF emerges as a versatile computational tool across multiple domains, demonstrating particular efficacy in dimensionality reduction [91] and source separation applications [97]. Within the context of TM, NMF facilitates the discovery of latent semantic structures through the decomposition of document-term matrices into interpretable constituent parts.

These features are derived from the contents of the documents, and the *feature-document* matrix describes clusters of related documents [99].

BERTopic

Bidirectional Encoder Representations from Transformers [45], is a Deep Learning (DL) model designed to understand the meaning of ambiguous language in text. Unlike traditional language models that process text sequentially, either from left-to-right or right-to-left, BERT reads text bidirectionally. This bidirectional approach allows the model to capture the full context of a word based on its surrounding words, addressing the limitations of unidirectional models and improving performance in various NLP tasks [87].

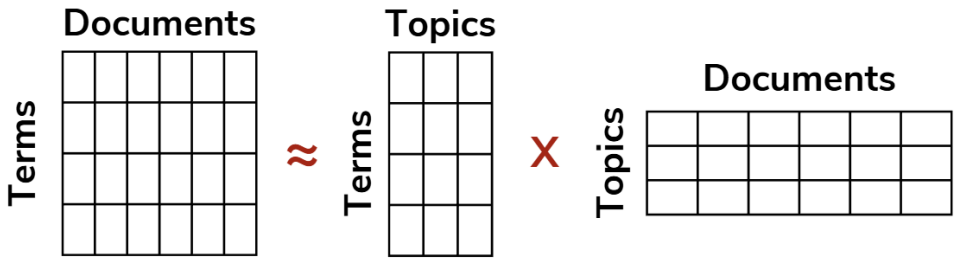


Figure 2.4: *Example of the Non-negative Matrix Factorization showing the decomposition into two key matrices: the term-topic matrix (***W***) that captures word-topic associations, and the topic-document matrix (***H***) that quantifies topic contributions across documents.*

Building on BERT's capabilities, **BERTopic** [28] is a TM technique that leverages **BERT word embeddings** alongside **class-based** Term Frequency Inverse Document Frequency (TF-IDF) to create coherent word clusters. This combination enables BERTopic to extract interpretable topics while preserving key terms essential for understanding the underlying themes within a corpus [29, 30].

In addition to word embeddings, BERTopic relies on text representation techniques like TF-IDF. Unlike the traditional BoW model, which only considers word frequency, TF-IDF [14] assigns weights to words based on their frequency within a document (Term Frequency) and their rarity across the corpus (Inverse Document Frequency). This approach provides a more nuanced representation of text by emphasizing both local and global word importance.

The strength of BERTopic lies in its ability to combine these techniques effectively. The **BERT word embeddings** capture the semantic meaning of sentences, generating high-dimensional vectors that reflect the context of the text. Each document is transformed into an embedding that encapsulates its meaning, enabling BERTopic to handle synonyms, contextual nuances, and topic overlap more effectively than traditional topic models.

The TM process in BERTopic is illustrated in Figure 2.5 and consists of the following four key steps:

1. **Document Embedding**: Generates document embeddings using a pre-trained BERT language model, capturing the semantic meaning of the text.

2. **Dimensionality Reduction** Uniform Manifold Approximation and Projection (UMAP) [54]: Reduces the high-dimensional embeddings into a lower-dimensional space.

3. **Clustering** Hierarchichal Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [55]: Groups similar document embeddings into clusters.

4. **Topic Extraction**: Extracts meaningful keywords from each cluster using TF-IDF, highlighting the most representative terms for each topic.

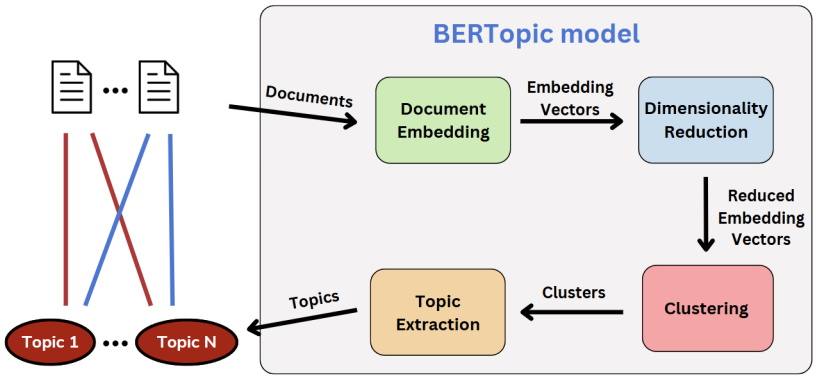


Figure 2.5: *Topic Modelling with BERTopic. Each document is first transformed into a high-dimensional embedding using pre-trained BERT models. The embeddings are then reduced using UMAP and clustered with HDBSCAN to group semantically similar documents. Finally TF-IDF extracts the underlying topics.*

2.3 Sentiment Analysis

Sentiment Analysis (SA) is a field of research related to computational linguistics, Natural Language Processing, and text mining. It can also be called *subjectivity analysis*, *opinion mining* and *appraisal extraction* [59]. The primary task of SA is to extract and analyse people's opinions, sentiments, attitudes or perceptions toward different entities such as topics, products, or services. Figure 2.6 illustrates examples of positive, neutral, and negative sentiments across different documents.

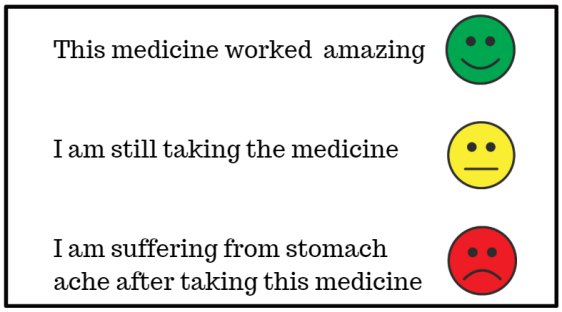


Figure 2.6: *Example of three sentiments in different sentences*

It poses as tool for businesses, governments and researchers to extract and analyse public mood and views, gaining business insights and allowing for better decisions [10]. SA can be performed at three levels: **document level**, **sentence level** or **aspect level**.

- **Aspect-level SA** focuses on detecting sentiments related to specific aspects of an entity, rather than identifying sentiment for the entire paragraph or sentence
- **Sentence-level SA** determines whether a given sentence expresses positive, neutral or negative opinion
- **Document-level SA**, which is the focus of this research, classifies an entire document (such as a review) as positive, neutral, or negative. This approach is particularly

effective when the document is written by a single individual and evaluates a single entity. However, it presents as unsuitable for document that address multiple entities.

As illustrated in Figure 2.7, Sentiment Analysis follows a typical NLP pipeline. It begins with text preprocessing such as tokenization, stop-word removal, part-of-speech tagging and stemming. Then, Feature Extraction and Selection techniques are applied to follow up with Sentiment Classification algorithms, further presenting the results. This workflow allows raw textual input to be transformed into meaningful insights.

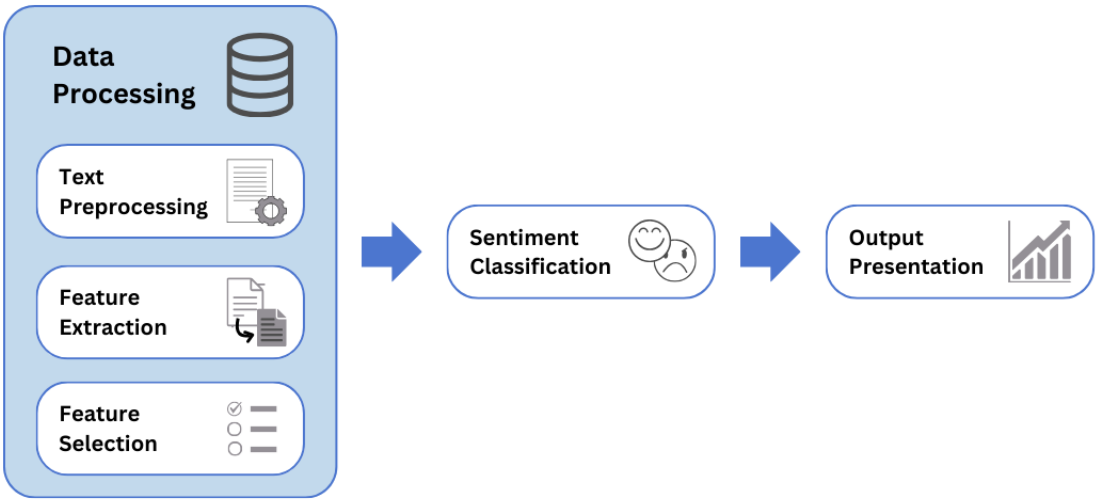


Figure 2.7: *Typical SA-NLP pipeline. It begins by data preprocessing, applying text processing techniques (e.g. tokenization, stemming); followed by feature extraction and selection to feed sentiment classification algorithms. Finally the outcomes are visually presented to the users.*

Being an active research field, applied to various domains, researchers propose, evaluate and compare different approaches constantly, aiming to increase performance and find novel solutions to the known challenges. Most literature separates SA approaches into three categories [89]:

- **Machine Learning**: In ML we can distinguish between **supervised learning**, which employs labelled datasets that combine text data with corresponding sentiment annotations to develop models capable of accurately determining the sentiment of new, previously unseen text; **unsupervised learning** approaches, which are utilized when labelled data is not available, allowing researchers to group documents into clusters based on similar sentiment characteristics; and **reinforcement learning**, which represents an adaptive approach where the model modifies its reward mechanisms through continuous sentiment feedback, allowing it to improve performance through ongoing interactions.

- **Lexicon-based**: These approaches rely on pre-defined sentiment lexicons, assigning polarity scores to words or phrases to calculate the overall sentiment of a document. The document is first tokenized, and sentiment values are assigned to individual

tokens. The final sentiment is computed using an algorithm that aggregates these values. These approaches do not require training or testing datasets, but fail to capture context-specific meanings. For instance, the word "small" in "The food portion is small" implies a negative sentiment, whereas "The camera is small" conveys a positive sentiment.

- **Hybrid**: These combine ML with lexicon-based methods, leveraging the robustness of lexical analysis alongside the adaptability of ML to handle ambiguity and integrate contextual information. Hybrid approaches are particularly effective in addressing challenges such as polysemy (words with multiple meanings) and domain-specific variations.

3 Related Work

This chapter aims to explore the current state-of-the-art on Recommender System (RS), covering specific methodologies and already known conclusions. Additionally, it explores Natural Language Processing (NLP) and Sentiment Analysis (SA) approaches and recent research to identify user sentiment within the healthcare sector.

It begins by covering the usage of RS in industries like e-commerce and video streaming, where they have been instrumental in enhancing user experience. It continues with an exploration of how NLP and SA techniques are used to extract insights from unstructured textual data tuned into medical data. Following this, key methodologies in RS are discussed, detailing algorithms, models, and approaches. Finally, the application of RS in healthcare is discussed, examining related work that has explored how these systems can be implemented to support clinical decisions.

3.1 Recommender Systems

The concept of **RS** first emerged in the late 1970s with early systems like **Grundy**, a computer-based librarian that suggested books to users based on personality stereotypes. In [78], the authors introduced the early foundations of RS by emphasizing the importance of treating users as individuals with distinct personalities, preferences, and goals. They proposed an initial approach of **modelling users based on stereotypes** to recommend novels tailored to their interests.

It was not until the 1990s that RS gained traction with projects such as **Tapestry** [24], and **GroupLens** [77]. Tapestry introduced **Content-based Filtering (CB)**, where users could search for documents based on their content and reactions recorded by other users. GroupLens, on the other hand, pioneered **Collaborative Filtering (CF)** by predicting ratings based on the heuristic that users who agreed in the past would likely agree in the future. Today, RS play a crucial role in **driving user engagement and boosting revenues** across various industries. Companies like **Amazon** leverage CF to recommend products based on user behaviour, implicit preferences, and item-to-item similarity [49, 80]. Similarly, **YouTube** [25] and **Netflix** rely on RS to personalize content by analysing user data such as reviews, watch history, watch time, and expressed interests. Netflix's

recommendation engine combines user ratings, genre preferences, and even contextual features such as time of day to deliver precise and personalized suggestions [65].

Within the field of movie RS, [96] **address the challenge of data sparsity** in large-scale datasets, where traditional CF struggles to identify meaningful user similarities due to the high dimensionality of user profile vectors. To overcome this limitation, the authors propose a **hybrid model-based RS** that integrates **Genetic Algorithm-optimized K-means clustering (GA-KM)** to transform the user space and reduce computational complexity while maintaining recommendation accuracy. The proposed method operates in **two distinct phases: offline and online**. The offline phase begins by applying **Principal Component Analysis (PCA)** to project the user-item rating matrix into a lower-dimensional space, mitigating the effects of data sparsity. The K-means clustering algorithm, further optimized by a **Genetic Algorithm (GA)**, is then applied to group users based on their inferred interests. In the online phase, the system assigns new users to the most relevant clusters and generates top#N movie recommendations based on their cluster neighbours. By limiting the similarity search to a specific cluster rather than the entire dataset, the method significantly improves efficiency while ensuring personalized recommendations. To evaluate the effectiveness of their approach, the authors conduct experiments using the MovieLens dataset [36], where each user has rated at least 20 movies. **The dataset is split into 80% training and 20% testing**, with the offline phase utilizing the training data to build the clustering model, while the testing data is used to evaluate recommendation performance. The proposed PCA GA-KM clustering approach is compared against other clustering-based CF models, demonstrating superior performance in recommendation accuracy and recall. However, it is noted that while their model outperforms baseline methods in terms of precision, it displays a higher Mean Absolute Error (MAE) for 20 recommendations. As part of their future work, the authors suggest further improvements to data reduction algorithms and a more detailed analysis of the impact of cluster size on recommendation quality and scalability.

Beyond numerical ratings and interval-based reviews, **unstructured textual feedback** offers valuable information for improving recommendation quality. For example, [53] employ **Latent Dirichlet Allocation (LDA)**, a **Topic Modelling (TM)** technique to uncover **hidden topics** within book descriptions and build a CB book RS.

The preprocessing pipeline includes common NLP practices such as **tokenization, lemmatization** and **stop-word removal**, followed by representing the data in a **BoW** format. To determine the optimal number of topics, the authors measure performance using a topic similarity matrix calculated using the **Jaccard Similarity**.

The similarity matrix generated from LDA allows for the system to identify and rank relevant books based on underlying themes, providing users with a list of the **top 10 most similar books**. This approach demonstrates the versatility of NLP techniques in enhancing RS performance by integrating contextual features into the recommendation process. To evaluate the performance, the author used **Intra and Inter Model Similarity**, achieving **0.865**, and **0.700** scores, respectively. For improvements, it is proposed to assign greater weight to more prevalent tokens, clear out noise, and normalise document length

by removing longer texts.

3.2 Natural Language Processing and Sentiment Analysis

In the scientific research conducted by [50], the author evaluates and compares the performance of **Baseline BERT**, **fine-tuned BERT** (on the last 4 layers), **Bio+Clinical BERT**, and a **Convolutional Neural Network (CNN)** in classifying the sentiment behind textual drug reviews in the **UCI ML Drug Review Dataset**, enabling inference of overall satisfaction from similar narratives such as patient surveys.

To simplify the analysis, the original 0-10 sentiment ratings were binned into three categories: Negative, Neutral and Positive. This approach ensures consistency and reduces ambiguity, particularly in addressing misclassified instances where highly positive reviews are wrongfully classified as negative, and vice-versa.

The results demonstrate that **Bio+Clinical BERT** outperforms the other models, achieving **Precision, Recall and F1-Score** values of 0.81, 0.80, and 0.81, respectively. The superior performance is attributed to its pre-training on medical-domain data, which enables a better interpretation of technical and domain-specific terms. Notably, **fine-tuned Baseline BERT** achieved comparable results, trailing the best model by only 1% in F1-Score.

By further analysing the **misclassifications**, the author identified three primary challenges:

- **Contradictory Reviews**: Instances where the numerical rating assigned to a review does not align with its textual sentiment.
- **Contradictory Language**: Reviews containing both positive and negative sentiments, making it difficult for the models to discern the overall polarity.
- **Non-Domain Sentiment Statements**: Reviews lacking medical terminology but expressing generic sentiments such as "Overall I'm happy" or "Well worth the payoff", where CNN models performed better at capturing the prominent phrases.

The author then concludes that while **Bio+Clinical BERT** demonstrates the best overall performance, there remains room for improvement. **CNN models** show a distinct advantage in processing **tangential information** such as capturing sentiment from key phrases among ambiguous reviews, by effectively filtering noise and focusing on the most relevant parts, while also being computationally more efficient. As future direction, the author suggests exploring **ensemble methods** to combine the strengths of multiple models. This approach would enhance the prediction of sentiment scores, enabling healthcare providers to better identify patients requiring further attention.

Extending the focus to **explicit recommendations**, [33] explore **lexicon-based SA** in online reviews. Their work highlights how **linguistic features** and **sentiment indicators** influence the detection of recommendations, combining text pre-processing techniques such as **stemming, lemmatization** and **Part-of-Speech (PoS) tagging** with classification

algorithms. In contrast to [50] model-driven approach (using BERT), the authors employ a **rule-based lexicon method**.

The reviews were first pre-processed using the IBM SPSS Modeller Text Analytics tool, where non-linguistic entities are removed, such as phone numbers, social security numbers, percentages, HTTP addresses and punctuation errors. Next, stemming and lemmatization were applied along with Part-of-Speech (PoS) tagging to create a dictionary, further enhanced by employing synonyms of words. Then, the set of lexicons was used to create seven categories with positive connotation, other seven for negative connotation and three for classifying wait-time and customer support. Each review is then classified as positive or negative, and fed into five ML classification algorithms: Probit, Binomial logistic, Chi-Squared Automatic Interaction Detection (CHAID), Classification and Regression Trees, and Random Forest. **CHAID** achieved the best performance, with **66.05% accuracy**.

As a result of their study, the authors found that feelings and attitudes were the most important predictors of explicit recommendations in a review.

While this study and [50] both utilize text cleaning and sentiment analysis techniques, they diverge in their modelling approaches: the referenced work applies transformers for domain-specific insight extraction, while [33] emphasize interpretable lexicon-based models, illustrating the complexity-interpretability trade-off in sentiment analysis.

Similarly, [74] explore sentiment patterns in the **UCI ML Drug Review Dataset** [48], focusing on lexicon-based Sentiment Analysis. Their methodology begins by addressing data quality, removing corrupted entries based on the *condition* feature, which affected the dataset's reliability. The textual preprocessing pipeline includes converting all text to lowercase, stripping HTML tags and punctuation, removing contextually uninformative words (custom stop words), and applying **lemmatisation** to ensure an accurate text analysis.

For sentiment classification, the authors employ two lexicon-based tools, **TextBlob** [88] and **VADER** [40]. Their analysis revealed that VADER produces more nuanced and distinct sentiment classifications than TextBlob. While TextBlob often labels a large portion of reviews as neutral, potentially losing meaningful insights, VADER provides clearer sentiment differentiation and is particularly effective in identifying negative reviews. This is crucial in the healthcare domain where negative sentiment often signals adverse drug reactions.

As future work [74] propose refining the preprocessing pipeline and investigating alternative Sentiment Analysis (SA) tools to either complement or validate their conclusions, aiming to support the development of a Drug RS.

3.3 Health Recommender Systems

Following the success of RS in commercial domains, researchers have transitioned the potential of RS in the healthcare industry, where their application could enhance patient care by facilitating access to personalized recommendations. In [72], the authors present

a survey of RS in the biomedical domain, motivated by the **exponential growth of biomedical data** and the associated challenges in information retrieval and analysis. The study reviews 60 original works published between 2015 and November 2021 across five scientific databases, categorizing them by RS technique and highlighting the predominance of model-based CF approaches. The results indicate that most datasets deviate from the standard ⟨user, item, rating⟩ format, are often only partially accessible, and frequently lack the documentation needed for reproducibility and extensibility. Furthermore, evaluations rely largely on classification metrics (precision, recall, AUROC). Despite these limitations, the survey emphasizes the growing relevance of **Knowledge Graph (KG)** as a means of integrating heterogeneous biomedical data, improving interpretability, and alleviating the cold-start problem. The authors also outline key challenges for KG-based biomedical recommendation—namely data integration, KG quality, scalability, domain expertise, interpretability, and evaluation—while stressing that addressing these issues is crucial to advancing reproducible research and enabling more effective applications in personalized medicine and precision healthcare.

[90] provides a review of **Health Recommender System (HRS)**, highlighting the need for these systems to manage complex and large volumes of health-related information. They emphasize the use of HRS in suggesting personalized diets, physical activities, medications, and treatment plans, allowing for patient-center decision-making, and enhancing the overall well-being of the users. The authors then elaborate on foundational techniques employed in HRS such as **CF** and **CB**, while mentioning the importance of **KB** recommendation in leveraging domain-specific knowledge, meeting the user's pre-defined requirements.

In the work done by [46], the authors address how the clinical decision-making is supported by the practitioners' experience and limited medical databases, and very prone to undesirable biases and human errors, affecting the quality of the aid provided to the patients. Specifically in heart disease, they expose the problem as being one of the highest death rate non-infectious diseases, associated with high cost in prevention and treatment. For this, they propose an **intelligent RS using time series prediction** to provide recommendations to heart disease patients in a tele-health environment, further supporting them in the necessity of medical tests. This RS analyses medical data of heart failure patients, assesses the short-term risk of heart disease and then provides recommendations based on the outcome of the predictions.

First, the authors preprocess the data, imputing missing data with a global constant and removing incorrect readings, categorised as noise. Then, for the time series recommendation the authors consider the number of heart rate tests performed in the past "k" days and if those readings are considered "normal". If both prerequisites are met, then a "no test needed" recommendation is provided, else, a "test is needed" is recommended. To use the system, human interaction with the system is needed to receive the inputs of the studied parameters.

To assess the performance of the model, a dataset with six patients, and over seven thousand records during a six-month period was used. It contained measurements as numerical features with medical attributes such as heart rate, blood pressure, oxygen saturation.

This dataset was used as a ground truth to the performance of the recommendation system, and the recommendations were compared with actual readings of the measurements in two metrics, precision and workload saving. The precision was calculated using the number of days with correct and incorrect informations, with workload saving using the prior mentioned, along with the total number of days in the dataset. For a sliding window of 5 days, the authors obtained precisions ranging from 75% to 100% across different patients, reducing on average 10% of the workload for patients from their daily medical tests.

While no future work is presented, further tests would provide better insights on the performance of the system, for example testing different time sliding windows and incorporating more features, records, and patients. [35] partnered with a leading European healthcare provider to address the challenges patients face when selecting family doctors, aiming to build strong, long-lasting and trusting patient-doctor relationships, and recognising that patients frequently struggle in identifying suitable doctors due to limited information and high search costs, ultimately relying on word of mouth recommendations. The authors propose a Hybrid RS to generate personalised doctor recommendations based on a hybrid matrix factorization model, representing patients and doctors as linear combinations of embeddings derived from demographic and behavioural data. This method combines CF by including past patient-doctor interactions and CB, relying on similarities in patient and doctor characteristics. To capture the dynamics of relationships, the authors introduce a weighted parameter to incorporate recency of interactions, assuming an exponential decay in relationship strength as time goes by.

The dataset consists of anonymised transactions between 2012-2017 in the hospital network. Each transaction details clinical consultations, describing the set of services provided to treat a clinical condition or procedure, with identifiers assigned to both patients and hospitals within the network. To evaluate their approach, the authors consider a temporal cross-validation, where the data is split into training and testing chronologically to preserve temporal consistency. Model performance is assessed using Hit Rate@K and Precision@K, comparing results against baseline heuristics such as "Recommend the most visited doctor" and traditional CF approaches. The results outperform baseline heuristics, and determine that using a hybrid combination of the interaction matrix and the predicted relationship scores presents the best results. For future work, [35] plan to deploy the RS in a real-world healthcare environment, aiming to collect explicit patient preferences and further refine the model. This approach addresses limitations in offline evaluations and enhance personalised patient-doctor matchmaking quality.

3.4 Drug Recommender Systems

Drug RS play a critical role in supporting clinicians by suggesting suitable medications tailored to a patient's medical profile, including their diseases, allergies, and known drug interactions. By leveraging advanced algorithms and patient data, these systems aim to optimize therapeutic decisions, enhance treatment outcomes, and mitigate potential risks associated with generalized prescribing practices.

The authors in [26] propose a drug RS for diabetes based on **CF** and **clustering**, aiming to assist health professionals in making decisions about treatments or determining appropriate prescriptions. The dataset used in their study comprises over 100,000 patient records from United States hospitals, collected between 1999 and 2008. Since each record contains more than 50 features, the authors employ **PCA** to reduce dimensionality. Following an initial exploratory data analysis, non-informative features, such as patient identification numbers and drugs with minimal administration rates were discarded. For clustering, the authors leverage both the **K-means** algorithm and the **Density-Based Spatial Clustering of Applications with Noise (DBSCAN)**, using the **Silhouette Score** to determine the optimal number of clusters. In the subsequent recommendation workflow, a new user is first assigned to one of the identified clusters. The system then locates similar users within that cluster using **CF**, further predicting and recommending medications. The authors evaluate the system in terms of both prediction accuracy and recommendation effectiveness. Their reported **Mean Square Error (MSE)** for prediction is 0.51, and the recommendation component achieves a precision of 61%. A recognized limitation is the **cold-start** problem, which arises when new medications appear in the system but have not yet been prescribed to any patients, thus lacking sufficient data to be recommended. As future work, the authors propose enhancing the system by leveraging metadata about these new items, combining both user and item information in a **hybrid recommendation** approach to mitigate the cold-start issue.

Similarly, [6] propose a universal medicine RS framework designed to reduce medication errors and enhance healthcare decision-making. The framework is composed of five key modules: **database system**, **data preparation**, **recommendation model**, **model evaluation**, and **visualization**. The dataset utilized in their study consists of 1,200 patient records, including features such as age, sex, blood pressure, cholesterol levels, sodium, potassium, and prescribed drugs. During the **Exploratory Data Analysis (EDA)** phase, the authors employ **correlation analysis** to identify and eliminate non-informative features. For instance, attributes like sex showed no significant correlation with drug prescriptions and were subsequently removed. To ensure uniform scaling and minimize noise during the training phase of the algorithms, the authors apply **min-max normalization**, standardizing the data across all features. For the **recommendation model**, [6] explore three machine learning approaches: **Support Vector Machine (SVM)**, **Backpropagation Neural Network**, and **ID3 decision trees**. The dataset was partitioned into 70% for training and 30% for testing, with ten experimental runs conducted for each model to mitigate random errors and ensure result reliability. Model performance was evaluated based on a balance between **accuracy** and **execution time**. Although the backpropagation neural network achieved slightly higher accuracy, Support Vector Machine (SVM) was ultimately selected as the optimal model due to its **93% accuracy** and **shorter running time**, offering a more favourable trade-off between performance and efficiency. The authors emphasize the critical importance of achieving high accuracy in predicting and recommending medications, as it directly influences the **quality of healthcare services** and ensures **patient safety**. As part of their **future work**, [6] propose the implementation

of a **mistake-check mechanism**. This mechanism requires **expert validation** whenever the model's recommendations diverge from established medical practices. Such a feedback loop not only enhances the reliability of prescriptions but also contributes to the continuous improvement of the RS by updating the underlying expert knowledge database.

Additionally, [32] identify key challenges in selecting optimal drug therapies tailored to individual patients, highlighting the limitations of traditional methods that rely on generalized clinical guidelines and often overlook complex medical data. To address these issues, the authors propose a drug RS based on **neighbourhood-based CF**, designed to leverage data from **Electronic Health Records (EHRs)** and **clinical registries** to support personalized medical decision-making. The proposed system integrates EHR and clinical registries to capture rich, real-world medical data. The RS identifies similar patient profiles to generate personalised drug recommendations. This approach enables the system to incorporate not only past treatment outcomes but also patient-specific characteristics, enhancing the relevance of the recommendations. To evaluate the system's performance, the authors utilize two primary evaluation metrics. First, the **prediction accuracy** of treatment outcomes is assessed using the **Root Mean Square Error (RMSE)**. Second, the **quality of the ranked list of recommendations** is evaluated through the **Mean Average Precision (MAP) at position 3, MAP@3**, which quantifies the correctness between the system's top-3 recommendations and the treatments actually administered in practice. The evaluation demonstrates that the RS effectively improves both the accuracy of outcome predictions and the quality of therapy recommendations. However, the authors acknowledge certain limitations, such as potential biases from the underlying data and the system's dependence on the availability of comprehensive EHR data. As future work, they propose refining the system by incorporating additional patient feedback and enhancing the explainability of recommendations to foster greater trust among healthcare professionals and patients.

Finally, [75] introduces DRecSys-SUSA, a drug RS built under Large Language Model (LLM) specifically fine-tuned for the medical context and user generated content, leveraging the UCI ML Drug Review dataset [48] as its foundational source.

The proposed methodology follows four principal steps. It begins with an initial data exploration (Exploratory Data Analysis (EDA)), examining the structure and distribution of the dataset to extract deeper insights into its composition. This is followed by a cleaning and pre-processing of the textual using the [9] toolkit to transform raw, noisy text into a cleaner format by removing HTML tags, stop words, and punctuation, alongside applying lowercasing, tokenisation and part-of-speech tagging.

In the third stage SA is performed to capture and quantify user sentiment expressed within the cleaned textual data, preparing it to be integrated into the final stage. This fourth phase involves text generation and fine-tuning of a LLaMA 2 model, which is used to produce relevant drug recommendations based on user inputs.

To personalise the recommendations, DRecSys-SUSA collects the user's personal and medical details and employs Semantic Key Retrieval to build a tailored prompt for a second

LLM. This model analyses the reported symptoms, maps them to the conditions found in the original dataset and generated a ranked list of recommended drugs.

For evaluation, [75] uses yet another LLM, ChatGPT 4.0 [68] to generate a structured test set derived from the original dataset. The performance is then assessed by comparing the quality of generated drug lists with and without incorporating user sentiment and semantic retrieval, using both accuracy and ranking-based metrics. The results indicate that the best-performing configuration combines both user sentiment and semantic key retrieval, achieving an F1-Score of 0.2853, with a Precision of 0.5269 and a Recall of 0.1948. Regarding ranking metrics, it attains a Hit Rate (HR) of 0.0408, a Mean Reciprocal Rank (MRR) of 0.0669, and a normalized Discounted Cumulative Gain (nDCG) of 0.0849. As future work, two major improvements are identified, as the author proposes leveraging larger language models and improving the RS by fine-tuning on larger and more domain-specific healthcare datasets, improving the ability to better understand clinical language and drug interactions, leading to improved recommendations.

3.5 Related Work Overview and Research Opportunities

This section aims to provide an overview of the reviewed literature, highlighting notable contributions and innovative approaches in improving recommendation algorithms. It also identifies opportunities for further exploration, particularly within the healthcare sector and drug prescription, as proposed in TopicDrugRec.

In the commercial domain, large scale retailers and entertainment platforms were among the first to adopt RS to personalise content and increase user engagement, ultimately boosting their revenue, and laying the groundwork for the adoption of RS within other industries. However, a challenge in RS is the issue of data sparsity, where high dimensionality and sparsity of user-item interactions impacts the system's capability to generate accurate recommendations. Wang et al. [96] address this issue by proposing dimensionality reduction techniques, followed by clustering to identify similar user groups. Their work enhances and presents an innovative approach in working with sparse vectors, and not only improving the efficiency, but also the performance of generating recommendations.

Beyond structured data, text emerges as an additional opportunity to generate recommendations. TM algorithms for extraction of underlying themes were adopted by [53]. The authors propose a pipeline of NLP techniques to convert qualitative data from books into a quantitative representation, and further integrate recommendation algorithms. Their work presents an opportunity in domains where user-generated content such as reviews, feedbacks, or health records contain implicit information. Despite the improvement, the authors lack the exploration of the impact of different text preprocessing strategies, for example, varying n-gram sizes or evaluating between stemming and lemmatization.

Additionally, incorporating user feedback through SA could further improvement their work by not only matching books based on their content but by also incorporating generalised user feedback.

[74] positively demonstrate how sentiment insight can be extracted using lexicon based

models to classify sentiment, presenting as an opportunity to complement the topic-based approach in [53], yielding more user-aware recommendations.

In drug recommendations, the explored literature showcased the application of RS fit into clinical context, relying on patient data, predictive modelling and clustering algorithms that can be used to suggest patient treatments. [26] and [6] demonstrated how EHR and recommendation algorithms can be integrated to personalise drug prescriptions, emphasising the importance of achieving high accuracy to ensure patient safety.

More recent approaches, such as the one seen in DRecSys-SUSA [75], build upon previous work by incorporating NLP pipelines, user sentiment, and large language models, effectively integrating state-of-the-art AI techniques into patient-centred drug recommendation. While the results are promising, the black-box nature of LLM presents significant limitations, as these models do not discriminate the decision-making process, making it difficult for clinicians to justify or understand specific recommendations which is critical in healthcare.

Additionally, LLMs do not provide control over lower level tasks, such as the tokenization strategies, stopword filtering or the use of n-grams which limits the ability to assess the impact of these NLP techniques.

Building on these gaps, this dissertation proposes a novel drug RS approach that integrates user reviews in textual data as well as sentiment patterns to provide drug recommendations. The modular approach of TopicDrugRec allows the study of how different components of the RS impact final drug recommendations, allowing for an analysis on possible improvements on a lower grain. Additionally, it studies the effect of different text processing techniques in effectively capturing underlying topics within unstructured data which further enhances the downstream task of providing relevant drug recommendations.

4 The TopicDrugRec System Methodologies

This chapter provides a comprehensive overview of the methodologies, tools, and frameworks employed in the development of **TopicDrugRec**, while simultaneously addressing the three research questions highlighted in the Contributions section. The research was conducted entirely in Python, with *Pandas* [70] serving as a fundamental library for data manipulation throughout the multiple processing stages [56].

Furthermore, the following sections elaborate the reasoning and choices taken in each step of the proposed solution in Figure 4.1:

1. **Exploratory Data Analysis (EDA):** This step focuses on understanding the dataset structure, interpreting features, and utilizing graphical visualizations to identify patterns, missing values, and potential biases in the data.

2. **Data Cleaning and Preprocessing:** To ensure data consistency, this step involves cleaning corrupted or incomplete data and imputing missing values when applicable. For Sentiment Analysis (SA), textual features are preprocessed through stop-word removal and the application of stemming and lemmatization techniques.

3. **Sentiment Analysis:** To address extreme response bias and validate the ratings associated to each drug, this step focuses in determining the sentiment of already preprocessed textual reviews. Furthermore, corrections are applied to align the sentiment expressed in the text with the numerical rating.

4. **Topic Modelling (TM):** With the sentiment-corrected data, this step is the main foundation of the RS algorithm, as it is where the latent topics are extracted from the user review (textual data) using various TM techniques, including probabilistic, deterministic and transformer-based algorithms.

5. **External Knowledge:** The fifth stage focuses on patient safety by integrating knowledge into the recommendations. It considers known negative drug interactions and highlights contraindications, side effects, and adverse reactions, supporting safer and more informed decisions.

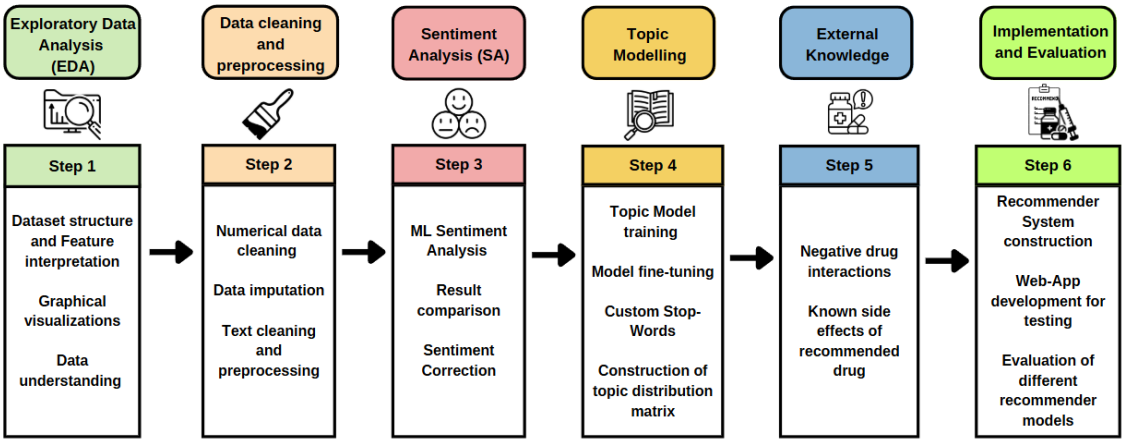


Figure 4.1: *Six-step methodology framework for implementing TopicDrugRec: The process flows from initial data exploration (Step 1) through data cleaning (Step 2), sentiment analysis (Step 3), topic modelling (Step 4), and external knowledge integration (Step 5), concluding with system implementation and evaluation (Step 6). Each step builds upon the outputs of preceding steps to create a comprehensive recommendation system.*

6. **Implementation and Evaluation:** Finally, the RS is implemented within a web application environment, facilitating practical testing. Evaluation metrics are then applied to compare the performance of the trained topic models and to assess overall effectiveness of the system in generating informed recommendations.

4.1 Exploratory Data Analysis

The main dataset used in this research is the *UCI ML Drug Review Dataset* [48], which was initially utilized for the Winter 2018 Kaggle University Club Hackathon [43]. Originally, this dataset was published on the UCI Machine Learning Repository as part of a scientific study [31] where the authors explored the possibility of applying sentiment analysis on drug reviews to identify the effectiveness of a drug, as well as the type of side effects caused.

The dataset was compiled by extracting user reviews and ratings on drug experiences from two independent sources: Drugs.com [20], and Druglib.com (no longer available online). Drugs.com provides user reviews for a specific drug, along with related conditions and a 10-star user rating reflecting overall satisfaction, while Druglib.com complemented it with additional reviews.

A notable difference exists between the two sources regarding the *condition* field. In Druglib.com, this field is of free-text, allowing users to describe conditions in their own words, which introduces potential variability through typographical errors, abbreviations, or inconsistent terminology. On the contrary, Drugs.com implements a different approach, where users select conditions from a predefined standardized list. This difference must be considered during the EDA phase, as condition field inconsistencies may introduce significant noise in the data, potentially affecting downstream analysis and topic extraction.

The dataset was originally compiled through web scraping, extracting data from raw HTML pages using the Beautiful Soup library [79] as part of its initial creation by the dataset

Figure 4.2: *Python libraries employed in the EDA phase: Pandas provides the foundational data manipulation framework, NumPy enables efficient numerical operations, while Matplotlib and Seaborn deliver complementary visualization capabilities for identifying data patterns, outliers, and quality issues within the drug review dataset.*

authors. This process of scraping user reviews from medical websites resulted in a dataset containing seven features, which are summarized in Table 4.1.

Table 4.1: Description of fields in the drug review dataset.

Field	**Description**
uniqueID	A unique identifier for each review entry.
drugName	The pharmaceutical name of the reviewed drug.
condition	The medical condition for which the drug was prescribed or taken.
review	Detailed patient feedback on the drug-condition pair, presented in textual format.
rating	A numerical 10-star patient rating indicating the degree of satisfaction with the drug's effectiveness.
date	Date when the review was submitted.
usefulCount	Number of users who found the review informative or helpful.

The dataset is publicly available at [48], structured into separate test and training files collectively containing **215,063 individual entries**, each with the **seven features** described above.

In addition to the fundamental library **Pandas** [70], this step also uses **NumPy** [67], **Matplotlib** [52], and **Seaborn** [81] to support data visualization, which helps identify patterns and reduce noise during cleaning and preprocessing, as illustrated by the libraries shown in Figure 4.2.

Through EDA, key insights into the structure and quality of the dataset are uncovered, laying foundation for detecting and addressing potential noise. Understanding the data sources and their characteristics allows for targeted approaches in the subsequent **cleaning and preprocessing** phase. Specifically, the variability identified in the *condition* attribute, where there are typographical errors or inconsistencies that must be addressed in order to optimize the performance of both SA and TM models.

These data refinement steps are designed to enhance overall data quality and consistency, aiming to contribute to more accurate drug recommendations and improved RS performance.

The following section details the steps that must be considered in the cleaning and preprocessing of the dataset to mitigate noise and ensure its suitability for the subsequent steps.

4.2 Data cleaning and preprocessing

With the key insights into the structure of the original dataset uncovered in the previous step, the next step is the **Data cleaning and preprocessing** of the data to ensure optimal quality for subsequent analysis.

In this step, techniques such as **Feature Importance** [42] can be used to determine the degree to which different features impact the machine learning model's predictions, however, due to the nature of the dataset, containing only 7 features, this step is not necessary.

Special attention is given to the *condition*, as it is one of the most critical features for generating recommendations. To ensure its consistency, this feature is standardised through the correction of typographical errors and the imputation of missing values. When a review lacks a specified condition, an additional check is made, if the associated drug is known to be prescribed for only one specific condition, the same condition is imputed accordingly. This decision preserves valuable data that would otherwise be discarded, while maintaining the integrity of drug-condition pairs.

This standardisation of the *condition* feature is made with two complementary strategies:

1. **Manual Corrections**: Typographical errors and inconsistencies are first identified manually, and corrected programmatically in Python, using a supplementary dictionary file.

2. **ICD11 API**: The ICD11 [101] is leveraged to retrieve approximate matches for conditions mentioned in the dataset, clustering them into 27 standardized medical categories, as outlined in [100]

Similar correction approaches are intentionally not applied to the *review* feature due to the capabilities of Natural Language Processing (NLP) libraries, particularly the Natural Language Toolkit (NLTK) [9], which process textual data with built-in error tolerance. These libraries offer techniques such as **lemmatization**, which reduces words to their base forms, and **stemming**, which removes common suffixes. An example of both of these techniques is depicted in Figure 4.3, where stemming reduces words to a common root form like "improv", while lemmatization returns the proper base form "improve".

In addition to these preprocessing steps, **stop word removal** is applied to the *review* feature to remove uninformative words that may introduce noise in the analysis. This process consists of filtering out commonly used words that do not contribute to meaningful information to the context, such as "is", "the" or "and". This is handled by NLTK, with extra domain-specific words added to filter out common but uninformative terms like "medicine", "drug", "medic", which are overrepresented and do not contribute to the analysis.

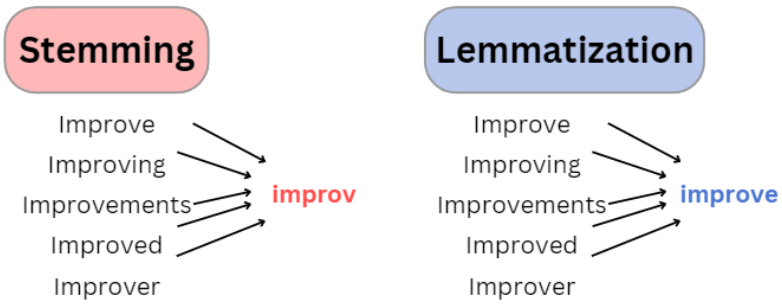


Figure 4.3: *Comparison of stemming and lemmatization: Stemming removes word suffixes, while lemmatization uses morphological analysis to return words into their base forms (e.g., "improve").*

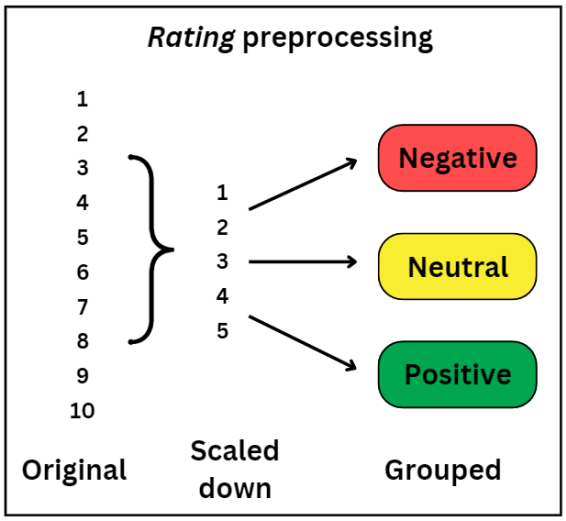


Figure 4.4: *Preprocessing methodology applied to the rating feature: The original 10-star scale is compressed to a 5-point scale and then categorized into three sentiment groups—negative (1-2), neutral (3), and positive (4-5).*

Stop word removal is only applied to the TM models, since the SA models require the full context to detect sentiment and polarity accurately. To ensure clean input for both TM and SA, entries with missing or duplicated reviews are removed from the dataset.

Alongside the review text, the *rating* feature is also adjusted to support SA. Originally on a 10-star scale, it was mapped to a simplified 1–5 range and grouped into three sentiment categories: ratings 1 and 2 as negative, 3 as neutral, and 4 and 5 as positive, as shown in Figure 4.4. This grouping reflects the nature of the dataset, where neutral reviews lack informational value and provide limited insight into the efficacy of the treatment or sentiment of the user, and therefore, their weight is reduced for the analysis.

The *usefulCount* feature is left untreated at this stage but is normalized during implementation to account for variability in review counts across all drugs from the same conditions, ensuring balanced impact in the recommendations.

The *date* feature was removed during preprocessing, as the focus of *TopicDrugRec* is on review content and sentiment rather than temporal patterns. Although the date could be

relevant for seasonal conditions like allergies or flu, drug efficacy and side effects tend to remain stable over time, making this information less useful for the analysis.

These data cleaning and preprocessing steps are implemented using a combination of Python libraries, including **NLTK** for text processing, **Pandas** for data manipulation and transformation, and **NumPy** for numerical computation, as depicted in Figure 4.5

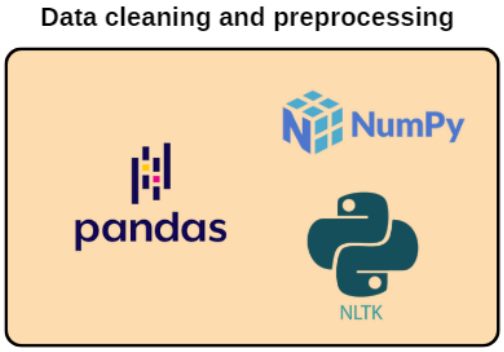


Figure 4.5: *Libraries employed in the Data Cleaning and Preprocessing phase.*

4.3 Sentiment Analysis

With the dataset cleaned, the next step is **Sentiment Analysis (SA)**, a critical process to address the inherent bias in the numerical ratings. This step directly addresses the first contribution question outlined in Section 1.3, where the objective is to employ SA tools to mitigate extreme response bias by aligning the sentiment expressed in textual reviews with their corresponding numerical ratings. This is done because textual reviews often provide richer context and more nuanced feedback compared to the numerical ratings, making them a valuable resource for refining the recommendations.

However, it is noteworthy that ratings and reviews often require careful interpretation, as they are widely influenced by a range of other personal and contextual factors that are not explicitly stated in the text. For example, two users may write similar reviews but assign different ratings depending on their expectations, treatment history or emotional state. One patient who has been dealing with a condition for longer may be more tolerant to side effects and therefore be more inclined to rate a drug positively, while another who has just started treatment may be more intolerant, rating it lower.

This research builds on the approach by [50], where fine-tuned transformer models like Bio+Clinical BERT and Baseline BERT were used for SA on the same dataset. That study highlighted key challenges in analysing drug reviews, such as contradictory language, ambiguous content, and general sentiment expressions that are not specific to the medical domain.

In this research, domain-specific BERT models were not employed. This decision is based on the nature of the textual reviews, which consist of casual language and general sentiments, rather than structured clinical or domain-specific expressions, as seen in Table 4.2.

Table 4.2: Examples of drug reviews from users. The samples show casual language and mixed sentiment, with review #1 combining negative side effects and positive outcomes, highlighting the challenge of sentiment classification.

#	Review
1	*I took this for a month and a half to two months and I had severe dry mouth and was constantly tired or sleeping. It helped my depression a lot though.*
2	*This medication, in combination with other medicines, gave me my life back. I had unsuccessful or allergic reactions to other meds that I had tried. I thank my doctor and this medicine every day for giving me my life back.*
3	*This is the only medication that has relieved the pain due to osteoarthritis in my knees. A side effect I had was sleepiness and drowsiness.*
4	*I used Hyzaar for the last 6 years. The best thing that happened to me, no side effects.*

Additionally, as demonstrated by [50], fine-tuned domain-specific models such as Bio+Clinical BERT did not yield substantially improved results when compared to fine-tuned baseline transformer architectures, further supporting the decision to utilize models pre-trained for informal and general text.

Instead of domain-specific models, this research comprehensively explores multiple complementary SA approaches, leveraging both transformer-based architectures such as **Twitter-roBERTa** [38] and **bert-base-multilingual-uncased-sentiment** [39], alongside rule-based tools such as **Valence Aware Dictionary and sEntiment Reasoner (VADER)** [40] and **TextBlob** [88].

Both BERT-based models were pre-trained on large text corpora but were not fine-tuned for drug review analysis. While the original dataset includes user ratings, these are not used as ground truth due to frequent inconsistencies between the text and the score, and a strong bias toward extreme values. For this reason, the SA models were applied directly to classify sentiment without supervised fine-tuning.

The **Twitter-roBERTa** model predicts sentiment as *Negative*, *Neutral*, or *Positive*. The **bert-base-multilingual-uncased-sentiment** model classifies reviews on a 1–5 star scale, similar to standard rating systems. Both models are combined to correct the original numeric ratings, which were previously rescaled to a 1–5 scale and grouped into *Negative*, *Neutral*, and *Positive* intervals, as explained in Section 4.2.

Additionally, the rule-based **VADER** and **TextBlob** tools are utilized as confirmatory tools to validate and cross-compare predictions generated by the primary BERT models, providing complementary perspectives on both sentiment polarity and language subjectivity. The complete pipeline for SA and subsequent rating correction is comprehensively illustrated in Figure 4.6.

VADER functions as a lexicon and rule-based SA tool designed to process general domain text, attuned to sentiments expressed in social media contexts. It employs a predefined lexicon and applies linguistic rules to the text based on intensity indicators, punctuation patterns, and capitalisation. Its primary output is the *compound score*, which represents

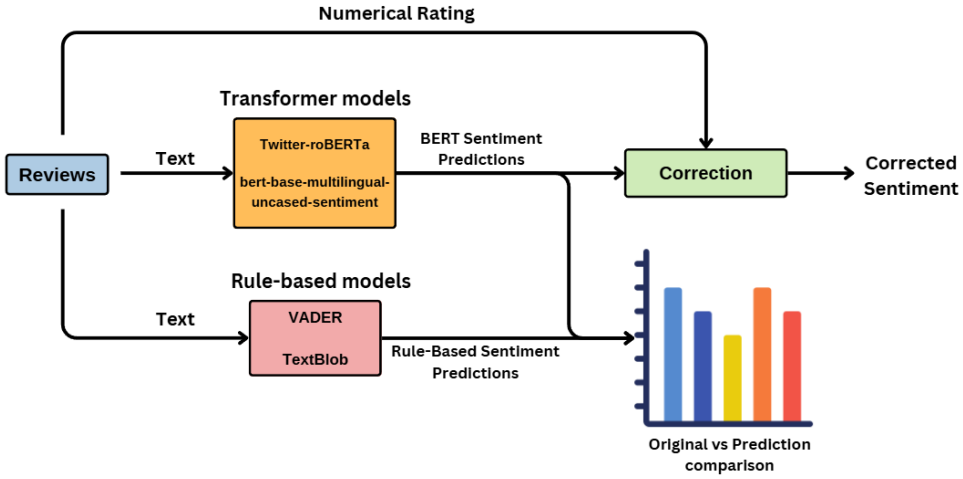


Figure 4.6: *Comprehensive pipeline for SA and rating correction: The process begins with the preprocessed text reviews that undergo parallel analysis by both transformer-based models (`Twitter-roBERTa` and BERT) and rule-based tools (VADER and TextBlob). The sentiment predictions from these models are then compared and integrated to produce a corrected rating value that reflects the actual sentiment expressed in the review text, addressing inconsistencies between numeric ratings and textual content.*

the normalised, weighted sentiment of a given text sentence on a scale ranging from -1 (extremely negative) to +1 (extremely positive).

Similarly, **TextBlob** is also rule-based, and provides two key metrics: **Polarity** and **Subjectivity**. Identical to *compound* in VADER, *polarity* is a floating point and ranges from -1 and +1, with intermediate values representing varying degrees of sentiment. Subjectivity measures how much of the text is opinion-based versus fact-based, ranging from 0 (objective) to 1 (subjective). In this research, the *compound* score from VADER and the *polarity* metric from TextBlob serve as analytical metrics, as both align with the measuring of sentiment in a drug review.

The choice between transformer-based or rule-based SA models depends on specific implementation constraints and requirements, as transformer-based architectures demonstrate superior performance in capturing the context and nuance of sentences by leveraging pretrained embeddings and contextual understanding, at the cost of more computational resources. On the contrary, rule based models fail to capture complex contextual patterns and subtle sentiment expressions, however, these require fewer computational resources and are faster.

The implemented approach allows sentiment-based rating correction and supports direct comparison between different SA model classes, helping assess their effectiveness in analysing drug reviews.

The technologies used in this step include the **Hugging Face Transformers** library for accessing the **BERT** models, **VADER** from the NLTK library [9], and the **TextBlob** library [88]. Additionally, **Matplotlib** [52] was used to compare predictions between transformer-based and rule-based models, as illustrated in Figure 4.7.

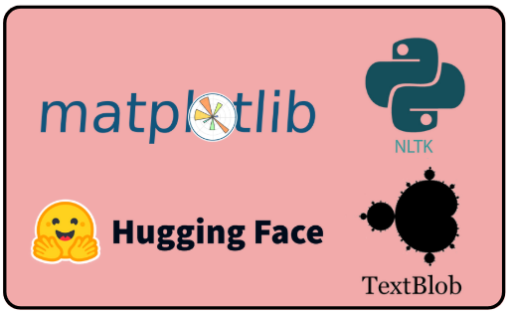


Figure 4.7: *Libraries used in the SA phase: Transformers (Hugging Face) for BERT-based models, NLTK for VADER, TextBlob for rule-based sentiment, and Matplotlib for performance visualization.*

4.4 Topic Modelling

With sentiment analysis aligned, the next step in the development of **TopicDrugRec** is addressing the second key contribution of this research (as outlined in Section 1.3): "*How can unstructured PRO be leveraged to enhance drug recommendations and improve medical treatment?*".

This is achieved through **TM**, which uncovers latent topics from the reviews, complementing the RS by reducing reliance on numerical ratings, and by considering more expressive patient feedback.

In this research, three distinct TM approaches were trained and evaluated to extract meaningful topics from the collection of user reviews: **Latent Dirichlet Allocation (LDA)**, **Non-negative Matrix Factorization (NMF)**, and **BERTopic**.

The evaluation methodology focuses on assessing the performance of different parametric configurations within each model type, particularly varying the number of extracted topics (represented as parameter K for LDA and $n\_components$ for NMF). Each model undergoes multiple training iterations with different number of topics, ranging from relatively low to high granularity, and is evaluated based on four quality metrics: **UMass Coherence**, **Diversity**, **Perplexity**, and **Similarity** (measured as Stability).

- **UMass Coherence** [62]: This metric measures the co-occurrence of topic words within documents. Unlike other topic coherence implementations which require an external corpus for calculations, $Umass$ coherence measure can be directly computed from the corpus making it more suitable for this research.
- **Diversity** [94]: Quantifies the proportion of unique words across topic representations, providing insight into how varied the topic is. Higher diversity scores indicate that topics capture distinct words, while lower scores indicate overlap between words. In this implementation, diversity is calculated for the top 10 most representative words of each topic.
- **Perplexity** [58]: Measures the quality of topic distribution in generative models like LDA and BERTopic. It evaluates how a trained model can predict patterns in unseen

data, with lower values indicating superior predictive performance. This metric is not applicable to non-probabilistic models, specifically NMF in this implementation.

- **Similarity** [5]: Assessed using **Pairwise Jaccard Similarity**, this metric reflects the proportion of shared vocabulary between topic representations. Lower scores indicate more dissimilar topics, suggesting broader thematic coverage, while higher scores point to possible redundancy or overlap.

The goal of this evaluation is to determine the most effective configuration for each model based on a balance of performance metrics, interpretability, and computational cost.

While the primary goal of this evaluation phase is not to rank TM algorithms, a comparative analysis is conducted to better understand how LDA, NMF, and BERTopic differ in coherence, diversity, similarity, and clustering behaviour. These insights help identify which model best aligns with the goal of **TopicDrugRec**.

The three TM approaches were implemented using different Python libraries suited to each method. LDA was trained with Gensim [104], which is efficient for large text corpora. NMF used Scikit-learn's matrix factorization modules [18], and BERTopic was implemented with its dedicated library [28], which combines BERT embeddings with clustering. The full set of libraries used is shown in Figure 4.8.

Figure 4.8: *Specialized libraries employed in the TM phase: Gensim provides the implementation framework for LDA, Scikit-learn supports the NMF approach, and the BERTopic library enables transformer-based TM with contextual embeddings.*

4.5 External Knowledge

Addressing the third contribution question outlined in Section 1.3 "*In the context of medical recommendations based on a RS, how can external knowledge be integrated to ensure the safety of the recommendations?*", this research proposes an **external knowledge layer** into the **TopicDrugRec** pipeline. This integration marks a difference between drug RS and those in domains like e-commerce or entertainment, where incorrect suggestions pose minimal risk to users. On the contrary, drug recommendations must be clinically validated to ensure patient safety and avoid harmful outcomes. For this reason, considering external medical knowledge is essential to ensure that generated recommendations are enhanced with clinical constraints and drug information, keeping patient safety the main priority.

Figure 4.9: *Libraries and tools employed for External Knowledge integration: Pandas for data manipulation, DrugStandard for drug name standardization, DailyMed for contraindications and side effects, and Apache Airflow for automated data updates.*

This extra layer is constructed by combining information from three external sources: *side effect*, obtained from a dataset built using structured information retrieved from Drugs.com [3, 20]; *contraindications* and *adverse reactions*, both extracted from DailyMed's clinical database [92]; and documented Drug-Drug Interaction (DDI), sourced from the specialized DDInter database [17, 102]. These datasets provide clinically relevant contextual information to support the post-processing stage of TopicDrugRec, enabling the user to identify and filter out unsafe drugs, contraindicated for specific conditions, associated with severe adverse reactions, or involved in moderate to major DDI.

The *contraindications* dataset integrated into this research was originally developed in a separate collaborative project [73]. In this foundational project, detailed pharmaceutical information was retrieved from DailyMed, Orange Book and Purple Book from FDA [92], specifically targeting clinically critical sections related to indications, contraindications, warnings and precautions for each medication. The data collection process stores the extracted clinical information in individual XML files, as well as in CSV format or in TXT, which are automatically updated on a monthly basis through a scheduled data pipeline managed by **Apache Airflow** [4].

For this research, our contribution involved transforming the XML files into a structured tabular format using the **Pandas** library. This transformation process included the parsing of the XML content, standardization of drug names using the DrugStandard library, and integration of the processed information into the recommendation pipeline, as illustrated in Figure 4.9.

4.6 Implementation and Evaluation

The implementation of TopicDrugRec combines a mixture of two RS algorithms to generate the drug recommendations: Content-based Filtering (CB) on extracted topic distributions, and Knowledge-based (KB) components that integrate medical knowledge. The system also incorporates an element of Collaborative Filtering (CF), a **sentiment-based weighted mechanism**, incorporating generalised user feedback into the recommendations. Due to the characteristics of UCI ML Drug Review dataset [85], which lacks explicit user profile information this approach does not rely on user-item interaction matrices. Instead,

sentiment is aggregated at a drug-condition level, capturing overall trends rather than individual preferences. As such, **TopicDrugRec** can be accurately characterized as a **Hybrid Recommender System**.

By considering the trained TM mentioned earlier (LDA, NMF, BERTopic), a recommendation matrix is built by considering the topic distributions, *usefulCount* and the corrected *sentiment* features. As seen in the work of [26], a similar approach to optimize recommendation inference speed is used, where **K-means clustering** is applied to group reviews with similar topic distributions.

The resulting recommendation matrix is characterized by the following features:

1. **Condition**: The original medical condition.
2. **Drug**: Drug prescribed for the specified condition.
3. **Topics (0 to N)**: The probabilistic distribution or weight assigned for each extracted topic in the trained topic model.
4. **UsefulCount**: Normalized value of the original UsefulCount.
5. **Sentiment**: Corrected sentiment score using the SA tools described in Section 4.3.

With the recommendation matrix established, **TopicDrugRec** uses the trained topic models and k-means clustering model to assign new, unseen user symptom descriptions to an existing cluster. The recommendation is then built upon the **topic distribution similarity** between the user symptom descriptions and documents from the assigned cluster, as well as the **usefulness score**, which gives a greater importance to reviews that were marked as informative, and the **sentiment score**, which prioritises drugs that received more positive reviews.

These three highlighted variables, **topic similarity, useful count, and sentiment**, function as hyperparameters within the recommendation process. The final ranking is determined using a weighted average, allowing adjustment of the importance assigned to each factor. This flexibility enables fine-tuning of the recommendation system to emphasize either content similarity (topic similarity), patient-reported helpfulness (useful count), or overall sentiment.

The final recommendation score for a drug is computed as follows:

$$Score = \frac{(\text{Topic Similarity} \times W_{ts}) + (\text{Sentiment} \times W_s) + (\text{UsefulCount} \times W_u)}{W_{ts} + W_s + W_u}$$

where:

- W_{ts} is the weight assigned to **Topic Similarity**.
- W_s is the weight assigned to **Sentiment**.
- W_u is the weight assigned to **UsefulCount**.

Patient safety is addressed through the integration of the previously described **external knowledge layer**, which provides structured, clinically relevant information on **DDI**,

Figure 4.10: *Flask provides the web application framework for the user interface, Docker enables containerized deployment for consistent performance across environments, Scikit-learn supports the implementation of K-means clustering and evaluation metrics.*

contraindications, **side effects**, and **adverse reactions**. After the recommendations are generated, users can access detailed pharmaceutical information for each suggested drug directly within the TopicDrugRec web application interface. This application is built using the **Flask** framework [27] and containerized within a **Docker** [60] environment to ensure consistent deployment, scalability, and portability across different computing environments, as illustrated in Figure 4.10.

The web interface presents comprehensive information regarding potential interactions with other already prescribed drugs, adverse reactions, and known contraindications, ensuring that users can make informed decisions based on both the recommendations and the clinical knowledge. The complete recommendation pipeline, from initial symptom input processing with TM to final drug selection and safety verification, is shown in Figure 4.11.

Following the implementation of TopicDrugRec's recommendation logic, the next step is to evaluate how well it performs in providing useful drug recommendations based on the user-reported symptoms. The RS is tested using an **offline evaluation method**, as real-world clinical validation would require extensive resources, extended timelines, and multiple iterations of testing and refinement that extend beyond the scope of the research.

The evaluation dataset is first split into **training** and **testing** subsets using a stratified sampling approach, ensuring balanced distribution of **condition-drug** pairs across both partitions to minimize potential evaluation bias.

The evaluation process then proceeds by selecting the optimal configurations for each TM approach, BERTopic, LDA, and NMF, based on their previously assessed intra-model performance across different topic ranges. **TopicDrugRec** employs these models to generate recommendations based on symptom descriptions from the testing set. For each test entry, the corresponding *condition* is used as the input query, and all drugs associated to that condition in the dataset form the ground truth. The system's **top-K recommendations** are then compared against this reference set to evaluate both accuracy and efficiency of the RS.

To ensure comprehensive performance assessment, **TopicDrugRec** integrates complementary **ranking-based** and **predictive metrics**. The ranking-based metrics, **Mean Average Precision (MAP)**, **Mean Average Recall (MAR)**, and **Mean Reciprocal**

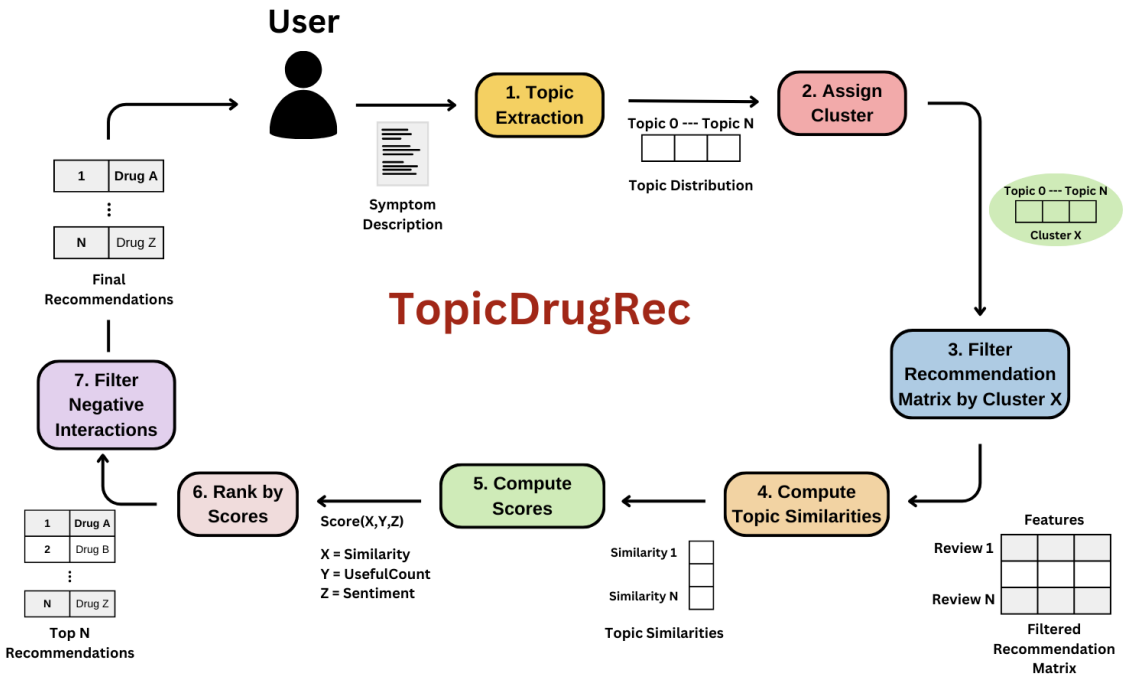


Figure 4.11: *TopicDrugRec recommendation pipeline: The process starts with user-provided symptom descriptions, which are preprocessed and passed through TM to identify underlying themes. These are then matched to similar reviews using K-means clustering, and based on this similarity, the system RS recommends drugs, considering user sentiment and review useful scores. Finally, the recommendations may be filtered through the external knowledge layer to remove drugs with known interactions, contraindications, or harmful side effects.*

Rank (MRR), measure how effectively the system retrieves and correctly ranks relevant recommendations. The predictive metrics, **Precision**, **Recall**, and **F1-Score**, evaluate the overall correctness of recommendations without considering specific ranking order.

These evaluations are conducted across multiple values of **K** recommendations to analyse performance at different levels of granularity. Within the healthcare context, selecting an appropriate K value is particularly critical to ensure clinically safe decision support, as medical practitioners rely on evidence-based suggestions with high confidence levels. Superior performance on a smaller number of top recommendations ensures that only the most relevant and appropriate treatments are suggested, reducing the risk of overwhelming clinicians with excessive information and mitigating potential risks associated with recommending too many or less relevant drugs.

5 Implementation and Evaluation of Core Components

This chapter presents and details the key stages involved in the implementation of **TopicDrugRec**, focusing on the preparation and analysis of the dataset, as well as the construction of the recommendation matrix prior to generating and evaluating the final recommendations. The aim is to provide an overview of the methodologies applied, the results obtained at each stage, and preliminary conclusions drawn from them.

The chapter begins with an Exploratory Data Analysis (EDA), which offers valuable insights into the dataset's structure and its features. This exploratory step is fundamental for identifying underlying patterns, feature imbalances, and possibly biases, such as the over representation of drugs or conditions, which can impact the performance of the generated recommendations. Additionally, these results inform future strategies to be considered in optimising the performance of the proposed Recommender System (RS). Following the EDA, the data cleaning and preprocessing steps are described, emphasising on how missing data was addressed and the techniques applied to clean and prepare the textual content.

Next, the results of the Sentiment Analysis (SA) are presented, comparing lexicon-based and embedding-based models in identifying inconsistencies between textual sentiment and their corresponding numerical ratings. This analysis directly addresses challenges that impact the quality of data collected in Patient-Reported Outcomes (PRO), such as **extreme response bias**, stated by [13]. The chapter then covers the Topic Modelling (TM) process, where different models and configurations are tested and evaluated to determine the best setups in uncovering underlying themes within the textual reviews, their distributions, and laying foundation to build the recommendation matrix.

This is followed by the integration of external knowledge, describing the datasets used and how they were prepared to incorporate clinical information, such as known drug interactions, contraindications, and side effects.

This chapter outlines an approach to drug recommendation that combines Natural Language Processing (NLP) techniques with medical knowledge to generate relevant and safe suggestions. Each step addresses specific challenges in drug recommendation, focusing on sentiment accuracy, topic extraction, and patient safety.

5.1 Exploratory Data Analysis

While the EDA done by Pinto et al. [74] on this dataset is both comprehensive and well-detailed, this step aims to cover two main objectives: validate their findings, and introduce another exploratory perspective by integrating International Classification of Diseases (ICD11) [101]. ICD11 enables the mapping of each condition in the dataset to its corresponding chapter (or disease group), providing a broader view of the diseases covered in the data. Parallel to supervised learning, the *condition* feature and its mapped ICD11 chapter can be interpreted as target labels, where, in the **TopicDrugRec** context, the goal extends further from predicting the most appropriate condition or disease group based on the user inputs. It also aims to generate meaningful drug recommendations suited for both specific conditions, as well as their broader ICD11 categories. This grouping allows for an evaluation of **TopicDrugRec** on two granularity levels: the individual condition level and the broader disease group level.

As detailed previously in Section 4.1, the UCI ML Drug Review [48] consists of 215,063 entries, with 7 features, describing user experiences about drug usage and their personal ratings, both in text format as well as numeric, as well as the number of people who find their review useful. The original dataset contains **2,388 unique drugs** and **889 distinct conditions**, as shown in Table 5.1.

Upon further analysis, it was observed that **some reviews were duplicate**. Despite this being part of the preprocessing step, it is mandatory to remove these entries, as their entries may skew data and introduce noise into the EDA. The step **reduced the dataset from 215,663 entries to 128,478**.

The fact that the number of unique drugs exceeds the number of conditions suggests that many drugs are prescribed for the same conditions. As a result, during the recommendation process, a single drug may appear multiple times in the recommendation list if **TopicDrugRec** identifies several conditions based on the user's input description.

Table 5.1: Number of Unique Drugs and Conditions in the Dataset

Feature	Review	Drugname	Conditions
Unique Values	128,478	2,388	889

Analysing the distribution of ratings, shown in Figure 5.1, a clear pattern of **extreme response bias** is identified, with pronounced peaks at ratings 1, 8, 9 and 10. The mean rating is 7, while the median is 8, indicating a **left-skewed distribution**, where the bulk observations (ratings) lie towards the lower end of the axis [21]. Although these original ratings might suggest positive patient experiences, the corresponding textual reviews do not always reflect the same level of happiness with the treatment, often mentioning adverse effects or some reservations regarding the treatment. This behaviour is seen in Table 5.2, where reviews assigned with perfect scores of 10 include expressions that may appear to contrast with this rating. Rather than categorising this as an inconsistency, these scenarios may reflect underlying personal or contextual factors that are not stated in the text but influence the rating. For instance, in the first review of a birth control method,

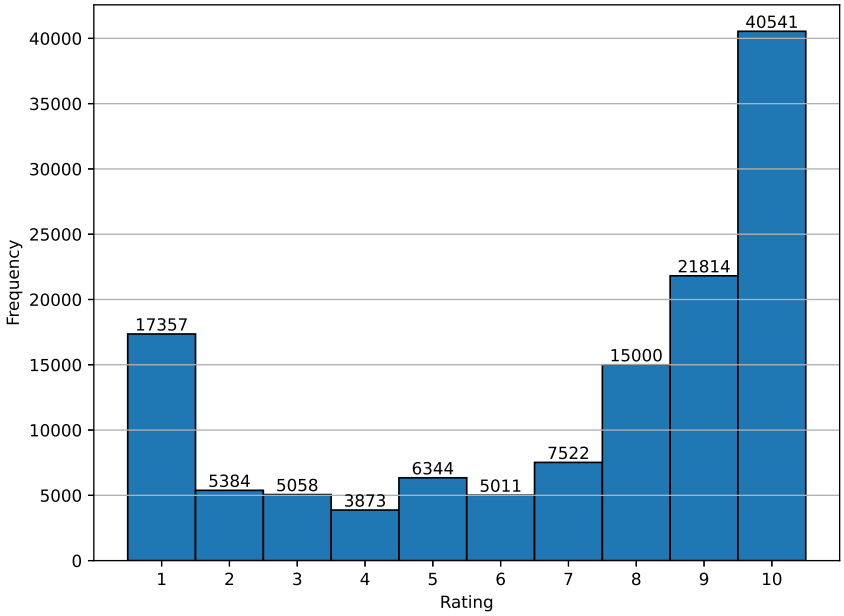


Figure 5.1: *Distribution of Drug Ratings. The histogram shows pronounced peaks at ratings 1, 8, 9, and 10, demonstrating extreme response bias. Despite a mean rating of 7 and median of 8, the distribution is left-skewed, with a significant number of extreme ratings at both ends of the scale. This pattern suggests users tend to rate medications either very negatively or very positively, with fewer moderate ratings.*

the user expressed strong dissatisfaction with side effects yet might still assign a high score if effectiveness at preventing pregnancy is valued more than overall satisfaction. Similarly, in the case of weight-loss medication, one reviewer reported experiencing *nausea* and *dizziness* during the first month of treatment, but these effects subsided and the user emphasized being "tickled to death" with the results after losing more than 40 pounds. In this situation, the rating would likely be strongly positive, as the long-term benefits of effective weight reduction outweighed the temporary adverse effects. Such factors highlight the importance of considering both structured and unstructured patient feedback, as relying solely on numerical ratings can distort the influence of the user-sentiment hyperparameter in **TopicDrugRec**.

As illustrated in Figure 5.2, which presents the top 10 most frequently reviewed conditions, **Birth Control** dominates the dataset, representing approximately **15.3%** of all reviews. It is followed by **Depression**, on a sharp decline to **5.6%** of reviews, while other significant conditions including **Pain, Anxiety, and Acne** each account for less than **4%** of the total reviews. The combined share of these top 10 conditions constitute **44.5%** of the entire dataset, **with the remaining majority distributed across a wide range of less common conditions**.

This reveals that the dataset is both **highly diverse** and **unbalanced**, meaning that the user experiences are not concentrated, but rather distributed across multiple medical condition groups. For a **Topic Modelling-based Recommender System** like **TopicDrugRec**, trained on user reviews, this unbalanced distribution presents a substantial challenge, as the over representation of certain conditions risks biasing the model towards

Table 5.2: Examples of Rating-Review Misalignment. This table illustrates situations where users provide high numerical ratings (9–10) despite writing negative textual feedback. In these cases, the ratings reflect the drug's effectiveness (e.g., preventing pregnancy or supporting weight loss), while the reviews focus on unpleasant side effects or dissatisfaction. This contrast highlights the need to consider both perspectives, as relying only on numerical ratings may obscure the actual user experience in **TopicDrugRec**.

Drug	**Condition**	**Review (Rated 10)**
Depo-Provera	Birth Control	*I've been on the shot for about a year now and I **absolutely hate** it!!!! I've **gained** so much weight and we're talking like 40 lbs. It's like I **can't stop eating**. My **cravings are absolutely horrible**. I'm **always hungry**. I work out here and there but I **can't seem to lose the weight**. I'm 17 years old. This birth control method is not something I would **suggest** to anyone **except** for the fact that it is **very effective against pregnancy**.*
Contrave	Weight Loss	*I started Contrave in April 2016 — at first I had the **nausea** and was **dizzy** but it finally stopped after about 30 days. I started out at 221 lb and now at 180. **Tickled to death with the results**.*

recommending drugs which are unsuitable for the exposed user symptoms.

To mitigate this imbalance, grouping conditions into broader ICD11 disease categories introduces an abstraction layer, which aggregates clinically related conditions, enabling the RS to provide recommendations in a more generalised manner, potentially reducing recommendation errors. Additionally, this approach also considers the clinical reality, where similar diseases often share similar treatment.

As shown in Figure 5.3, the most prevalent disease group is **symptoms, signs, or clinical findings, not elsewhere classified**, representing **17.8%** of all reviews. According to the ICD11 classification system, this category encompasses *"symptoms, signs, abnormal results of clinical or other investigative procedures (. . .) and ill-defined conditions regarding which no diagnosis (. . .) is recorded"* [86]. This indicates that a significant portion of reviews describe conditions that may span multiple diagnosis, becoming ambiguous and further complicating the recommendation pipeline.

Following the most prevalent category, **"Mental, behavioural or neurodevelopmental disorders"** represents **15.8%** of reviews, while **"Certain conditions originating in the perinatal period"** accounts for **15.6%**. These second and third most common disease groups align with the most prominent individual conditions identified earlier, reflecting the significant representation of mental health concerns and reproductive health issues throughout the dataset. The remaining disease groups in the top 10 each constitute between **6.0%** and **3.6%** of the total reviews, covering a wide range of physical and neurological conditions, from skin disorders to sleep related diseases, further demonstrating the clinical diversity captured within the dataset. Together, the top 10 disease groups represent approximately **83%** of all reviews, nearly doubling the coverage achieved by the top 10 individual conditions (44.5%). This increase in representation demonstrates how hierarchical disease classification enables better data generalisation and provides an approach to addressing class imbalance challenges in clinical scenarios.

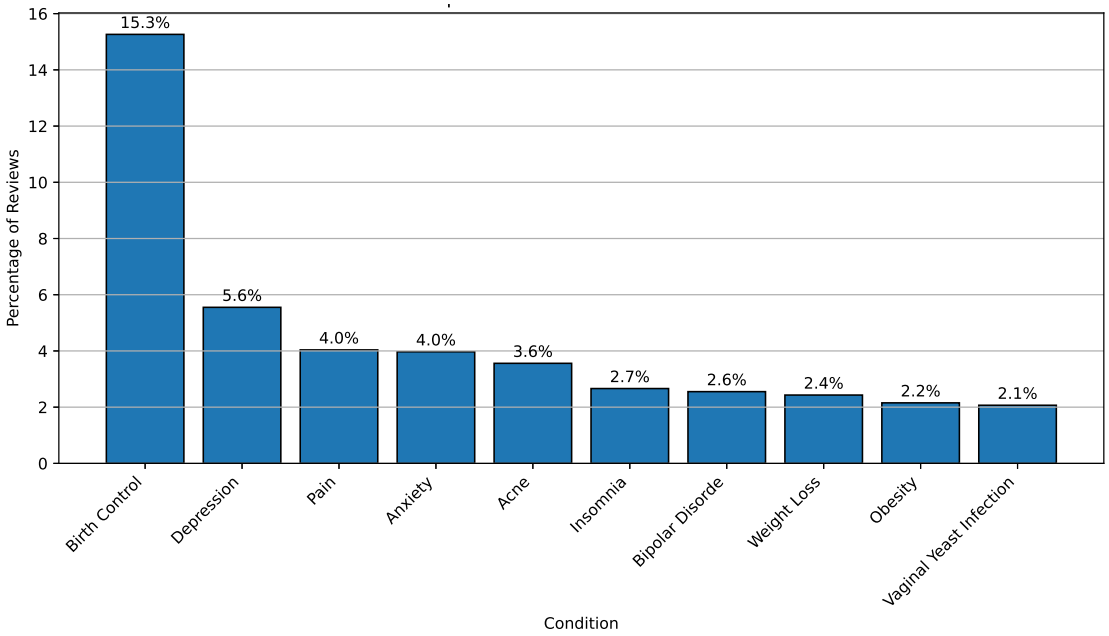


Figure 5.2: *Distribution of the Top 10 Most Common Conditions in the Dataset. The distribution is unbalanced, with Birth Control accounting for 15.3% of all reviews, followed by Depression at 5.6%, and Pain, Anxiety and Acne, each contributing to under 4%. These top 10 conditions represent 44.5% of the total reviews, while the remaining 55.5% are distributing across a wider range of less frequent conditions. These findings reinforce the diversity of the medical conditions in the original dataset, and the over representation of reviews related to reproductive health.*

To improve our understanding of the dataset and support the evaluation of the RS, an analysis was made on the number of unique drugs associated with the top 10 most common conditions and disease groups. This analysis helps revealing insights into the diversity of treatment options available for each medical category, which is important when deriving conclusions from performance metrics like Recall@K, which measures the system's ability to include relevant drugs among the top K recommendations.

As depicted in Figure 5.4, which presents the number of unique drugs for the most frequently occurring conditions, **Pain** demonstrates the highest treatment diversity with **141** unique drugs. This aligns with the generalised nature of pain as a symptom, often occurring as a sub-symptom of various underlying medical conditions. Following this, **Depression** (94 unique drugs) and **Birth Control** (87 unique drugs) also show a broad list of treatments, which can result in more diverse recommendations. Similarly, Figure 5.5 reveals that **Symptoms**, **signs**, or **clinical findings**, **not elsewhere classified** has the highest pharmacological diversity with **230** unique drugs, highlighting the broad spectrum of this disease group. Other categories, including **Diseases of the nervous system** (123 unique drugs) and **Diseases of the skin** (115 unique drugs), also demonstrate significant treatment variability. However, a challenge arises with conditions or disease groups that have **low unique drug counts**, as these imbalances may affect TopicDrugRec's performance for these specific medical scenarios. Despite this limitation, these conditions and groups were retained, as their exclusion would have substantially reduced the dataset to fewer than

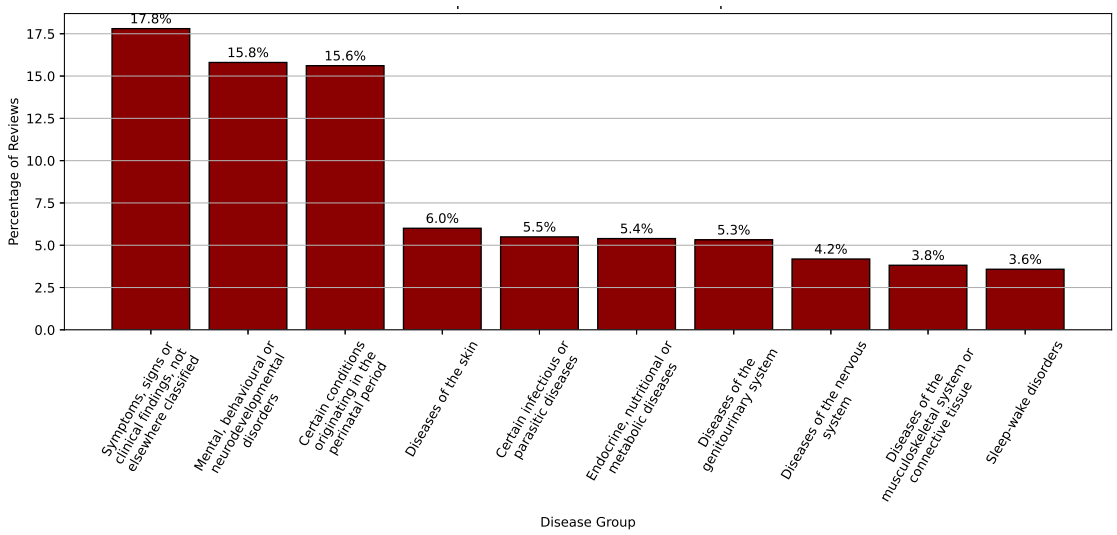


Figure 5.3: *Distribution of the Top 10 Most Common ICD-11 Disease Groups. When conditions are grouped by ICD-11 disease categories, it is clear that the top 3 most common disease groups become more balanced compared to the condition-level analysis. The category " Symptoms, signs, or clinical" findings accounts for 17.8% of all reviews, demonstrating a substantial portion of cases associated with ambiguous or unclear clinical representations. By considering a higher-level categorisation, introducing an abstraction layer, TopicDrugRec is able to account for a larger number of reviews across its 10 most frequent disease groups and potentially generate more accurate recommendations.*

100,000 entries, potentially impacting the training of the topic models used to generate topic distributions and subsequently construct the recommender matrix.

To better understand the diversity of drugs, and their coverage on different treatments, the EDA was extended to examine the top 10 most frequently prescribed drugs and quantify the number of unique conditions for which they were administered. As presented in Table 5.3, each of the top 10 most common drugs is utilised for treating at least 15 distinct medical conditions, indicating a broad range of applications across diverse medical conditions. Another important conclusion from this analysis is the **diversity in the conditions treated** by certain drugs. For example **Prednisone**, a corticosteroid primarily used to reduce inflammation [16, 19], is used to treat conditions as distinct as **Inflammatory Conditions, Asthma, and Gouty Arthritis**, exemplifying its therapeutic diversity. While the first two are typically associated with allergic inflammation, the latter is caused by extra uric acid that forms crystals in the joints [15].

The final step of EDA, the relationship between review length and perceived usefulness is analysed, aiming to assess whether longer reviews are seen as more useful than shorter, which is fundamental for the development of TopicDrugRec, given that the dataset comprises reviews that are often short. This characteristic poses challenges for TM algorithms, which depend on the co-occurrence of words that is often scarce in shorter content [64].

The scatter plot in Figure 5.7 depicts the relationship between review length and usefulness metrics. As illustrated in Figure 5.6, 99.4% of reviews contain fewer than 250 words. While longer reviews have a slightly higher median usefulness, no correlation between text length

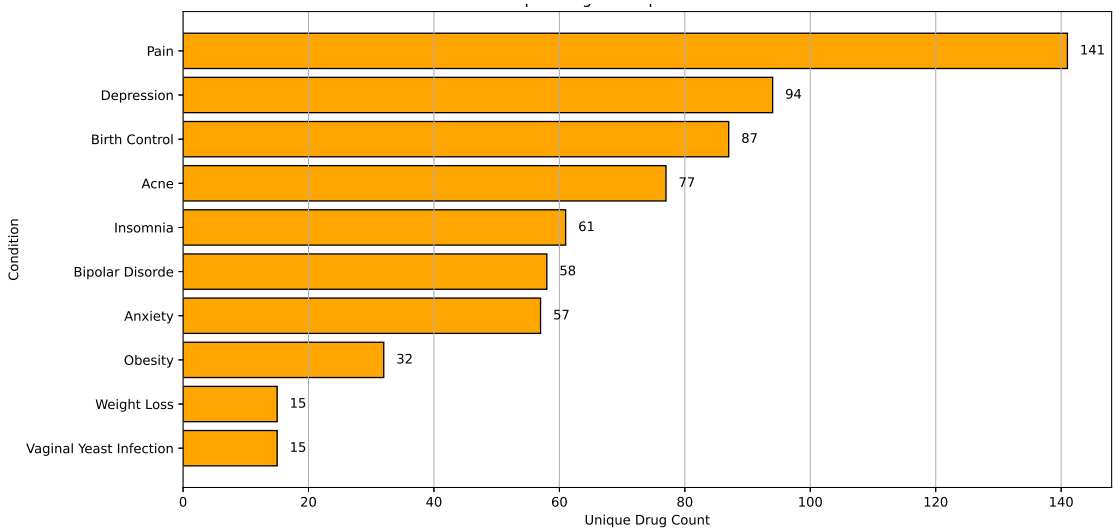


Figure 5.4: *Number of Unique Drugs Associated with the Top 10 Most Common Conditions. Pain shows the highest diversity with 141 unique drugs, followed by Depression (94) and Birth Control (87). This distribution reflects the wide variety of treatments across conditions, with some having extensive drug options while others present more limited treatments.*

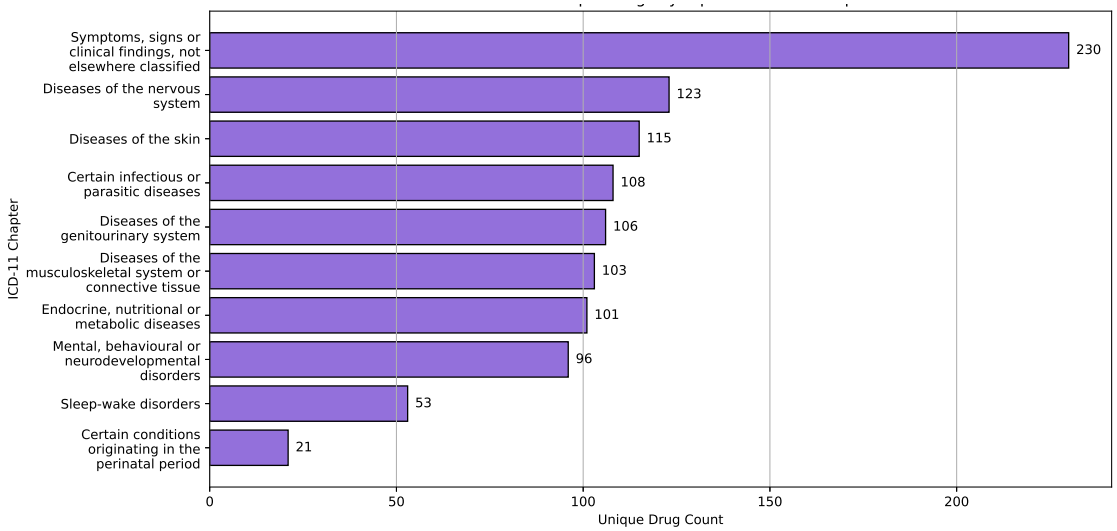


Figure 5.5: *Number of Unique Drugs Associated with the Top 10 Most Common Disease Groups. The category "Symptoms, signs, or clinical findings, not elsewhere classified" shows the largest diversity with 230 unique drugs, reflecting its diverse nature of the conditions. It is followed by "Diseases of the nervous system" (123) and "Diseases of the skin" (115), which also show substantial treatment variability.*

and perceived usefulness was found.

The most upvoted review in the entire dataset, with 1,291 votes, only had 187 words, which is significantly below the maximum number of words, but above the **average review length of 84.6 words**. This finding shows that conciseness may contribute more to a review's perceived usefulness rather than lengthy user experiences.

Despite the prevalence of relatively short texts in the dataset, under 50 words, the dataset contains enough content to apply both probabilistic and deterministic TM algorithms

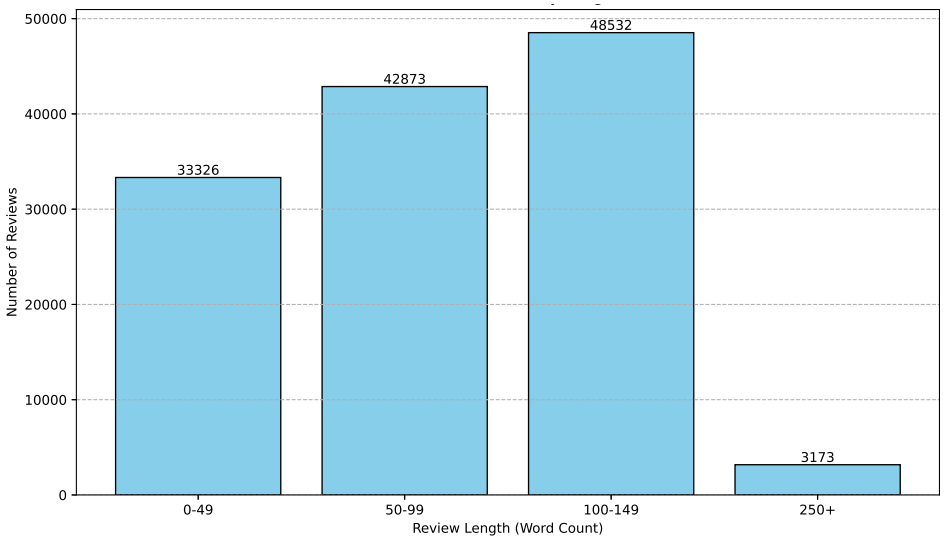


Figure 5.6: *Distribution of Review Lengths. The histogram shows that 99.4% of reviews have fewer than 250 words, with an average length of 84.6 words. This tendency toward shorter reviews reflects how users often focus on specific parts of their experience rather than writing long descriptions.*

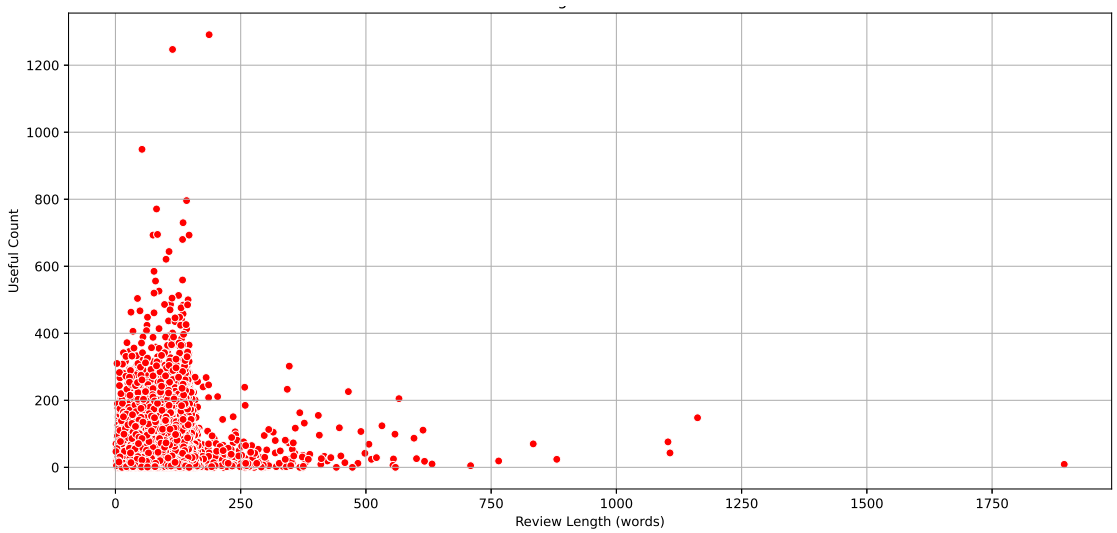


Figure 5.7: *Relationship Between Review Length and Usefulness Votes. This scatter plot illustrates the correlation (or lack thereof) between review length (word count) and perceived usefulness (vote count). While longer reviews show a slightly higher median usefulness score (26 vs. 16), no correlation exists between length and helpfulness. The most highly upvoted review (1,291 votes) contained only 187 words, demonstrating that information quality may matter more than length in medication reviews.*

effectively. Furthermore, embedding-based models like BERTopic are particularly useful, as they excel in capturing contextual nuances and semantic relationships in shorter texts, often achieving better results compared to Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF) [57].

Table 5.3: Top 10 most common drugs and number of unique conditions treated. This table shows the most frequently used drugs in the dataset, along with the number of unique medical conditions they are prescribed for and the top three conditions for each drug. The first column lists the drug name, the second column shows the number of conditions associated with that drug, and the third column presents the three most common conditions treated. Each drug is linked to at least 15 conditions, illustrating their broad use. *Prednisone* stands out for its flexibility, being prescribed for a wide range of conditions with different underlying causes.

Drug Name	#	**Top 3 Conditions**
Prednisone	33	Inflammatory Conditions, Asthma, Gouty Arthritis
Gabapentin	29	Anxiety, Pain, Fibromyalgia
Doxycycline	20	Acne, Bacterial Infection, Chlamydia Infection
Venlafaxine	20	Depression, Anxiety, Generalized Anxiety Disorder
Metronidazole	19	Bacterial Vaginitis, Bacterial Infection, Dental Abscess
Amitriptyline	19	Migraine Prevention, Insomnia, Pain
Triamcinolone	17	Dermatitis, Psoriasis, Alopecia
Clonazepam	17	Anxiety, Panic Disorder, Insomnia
Duloxetine	17	Depression, Fibromyalgia, Anxiety
Azithromycin	17	Chlamydia Infection, Bronchitis, Sinusitis

5.2 Data cleaning and preprocessing

Data cleaning and preprocessing are critical phases in any data-driven RS, especially in healthcare applications where data quality directly impacts recommendation accuracy and patient outcomes. As noted by García et al. [22], unprocessed data often contains noise, redundancies, and inconsistencies that can distort analytical results and lead to misleading conclusions in Machine Learning (ML) models.

Given the structure of the dataset, where the `condition` feature is collected as a free-text field (as noted in Section 4.1), addressing eventual typographical errors was a priority. Rather than relying on spell-correction libraries, which often struggle with medical terminology, we implemented a manual correction approach, where first we identified, and catalogued all typos within a JSON file, mapping original incorrect values to their verified medical conditions. This mapping was then programmatically applied to standardise the condition feature across the dataset.

Another approach to ensure the correctness of conditions involved incorporating their ICD11 approximate matches and retrieving their respective disease groups. This additional classification proved highly beneficial, as demonstrated in the preliminary analysis where the top 10 most common disease groups account for 83% of the reviews, compared to only 17.8% when using the top 10 individual conditions alone. This disease group mapping significantly improved the granularity of our analysis and improved the system's ability to identify meaningful patterns across related conditions.

During data validation, we identified **608 entries** with missing condition attributes, an expected challenge given the free-text nature of this field. To address this, we implemented

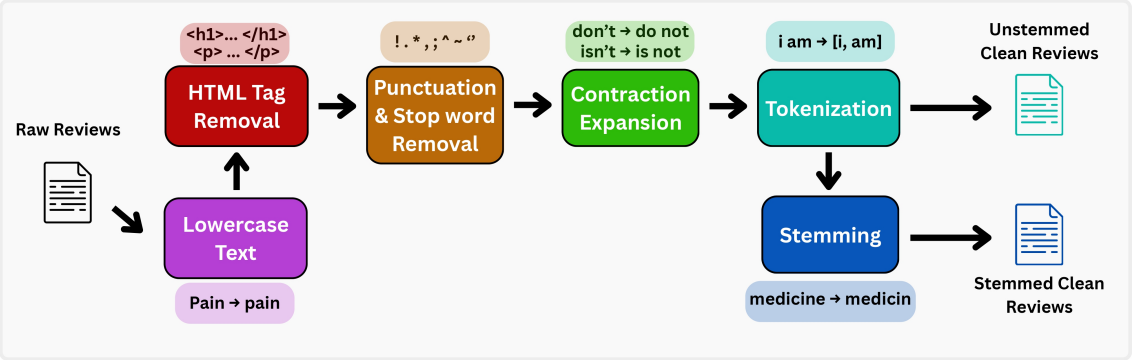


Figure 5.8: *Text Cleaning and Preparation Pipeline. The pipeline processes raw drug reviews through five sequential steps: (1) conversion to lowercase text, (2) HTML tag removal, (3) punctuation and stop word removal, (4) contraction expansion and tokenization, and (5) creation of parallel outputs—both stemmed clean reviews and unstemmed clean reviews, with examples described in the figure. This comprehensive preprocessing approach ensures consistent and normalized text input for subsequent TM algorithms.*

an imputation strategy for entries where the prescribed drug was associated with only one condition throughout the dataset, successfully **updating 40 entries**. The remaining entries with missing conditions were excluded from the analysis, as determining the the condition to which the user took the drug was not possible.

With the *condition* feature addressed, we implemented a five-step preprocessing pipeline for the *review* text, as illustrated in Figure 5.8. Initially, all text was converted to lowercase, ensuring consistency across the dataset and preventing the SA and TM algorithms from treating identical words differently based on capitalisation.

Next, we removed **HTML tags, punctuation, stop words, and unnecessary white spaces using regular expressions**. These elements, often resulting from the web scraping introduce noise that can impact the extraction of meaningful patterns. Stop words (such as “and”, “the”, “is”) were eliminated as they occur with high frequency but contribute minimal contextual value. This removal is imperative for TM, where the focus on word co-occurrence patterns can be impacted by the occurrence of these words.

Following text normalisation, we **expanded all contractions** (such as “donť”) into their **complete forms** (“do not”), ensuring textual consistency and reducing potential ambiguities. This expansion was implemented using an established open-source dictionary based on standard English contraction rules [98].

In the final preprocessing stages, we tokenized the text and created two distinct datasets: one containing stemmed reviews and another with unstemmed reviews. This processing approach enables comparative evaluation of whether stemming positively impacts the results of TM, or whether preserving words in their complete form provides richer contextual information for topic extraction. This experimental design is important for comparing algorithm performance, as LDA and NMF algorithms rely on word pattern recognition, while BERTopic leverages contextual embeddings that may benefit from the full semantic richness of unstemmed text [29].

Table 5.4 illustrates examples of raw and cleaned versions of a review. The table presents three versions of the same review: the original raw text with HTML tags and informal language patterns, the clean stemmed version with words reduced to their root forms, and the clean unstemmed version that maintains complete word forms while removing noisy elements. As demonstrated in the example, stemming significantly transforms words like "absolutely" to "absolut" and "lightly" to "lightli", with the trade-off of losing contextual information. This comparison provides insight into how different preprocessing approaches alter the textual data that serves as input for our TM algorithms.

Table 5.4: Examples of Raw and Clean Review (Stemmed & Unstemmed). This table demonstrates the transformation of a single drug review through our preprocessing pipeline, showing the original text with HTML entities and informal language patterns, alongside both the stemmed version with words reduced to their root forms and the unstemmed version that maintains complete word forms while removing noise elements.

Review Type	**Review**
Original Review	*"First month was awesome. Absolutely wonderful. Then I started lightly bleeding. I just thought it was my period. But it has lasted up until now which has been about 2 almost 3 months. I will not take the shot again. Having your period for months is hard on a girl. And on top of that my emotions have been insane. Like I am not a cryer. I never cry and ever since I've had the shot watching finding nemo makes me ball. And I get pissed at people for nothing!!! Its ridiculous. Like I dislike who I have become with the shot."*
Clean Stemmed Review	*first month awesom absolut wonder start lightli bleed thought period last almost month take shot period month hard girl top emot insan like cryer never cri ever sinc ive shot watch find nemo make ball get piss peopl noth ridicul like dislik becom shot*
Clean Unstemmed Review	*first month awesome absolutely wonderful started lightly bleeding thought period lasted almost months take shot period months hard girl top emotions insane like cryer never cry ever since ive shot watching finding nemo makes ball get pissed people nothing ridiculous like dislike become shot*

With the textual reviews cleaned and prepared for the downstream stages of TopicDrugRec, we next addressed the bias evident in the original numeric drug ratings (previously illustrated in Figure 5.1). We compressed the scale into a 1-5 range, where ratings of 1 and 2 were mapped to the *Negative* class, 3 to *Neutral*, and 4 and 5 to *Positive*. This transformation was made because the original scale had too many options, which may have contributed to indecisiveness, leading to more extreme answers observed in the dataset. For example, in the original scale, a user might assign a rating of 8 when their sentiment more closely aligned with an 6, just because of decision paralysis of too many choices. By reducing granularity, this new scale removes ambiguous ratings while maintaining a representation of user sentiment.

This scaling and grouping also aligns with the SA models adopted in this dissertation. The `twitter-roBERTa-base` model classifies text into *Negative*, *Neutral*, or *Positive* categories, while the `bert-base-multilingual-uncased-sentiment` classifies it in a 1-5 scale. By mapping the original numeric ratings into standardised sentiment categories, we ensured these could be easily used to compare with the models' outputs, facilitating their use in

validating and correcting the original ratings.

The resulting distributions, shown in Figures 5.9 and 5.10 show the adjustments made to the rating scale, while preserving the inherent asymmetry towards *Negative* and *Positive* reviews, however, at a more interpretable scale. The *Neutral* category now helps capturing uncertain or unclear reviews, creating a clearer distinction between positive and negative sentiments.

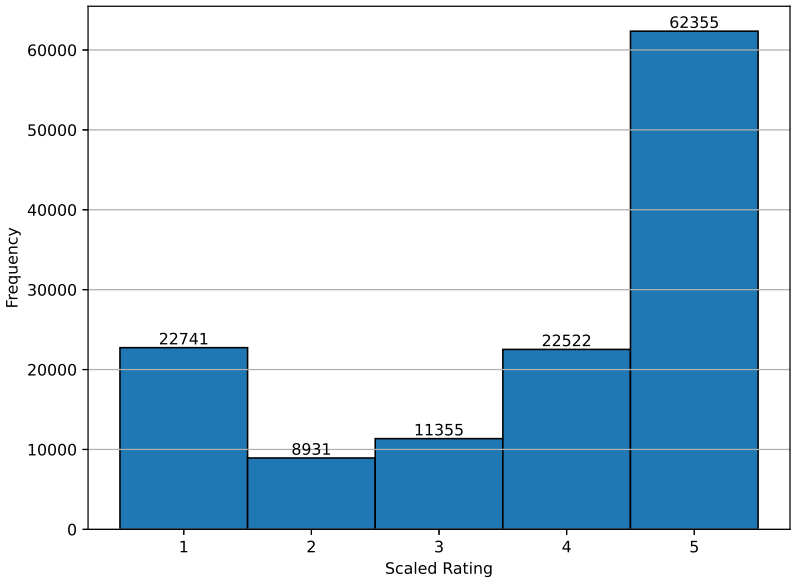


Figure 5.9: *Distribution of Drug Ratings After Scale Compression. The original 1-10 rating scale has been compressed to a 1-5 scale to reduce ambiguity and align with SA model outputs, while preserving the overall sentiment distribution pattern.*

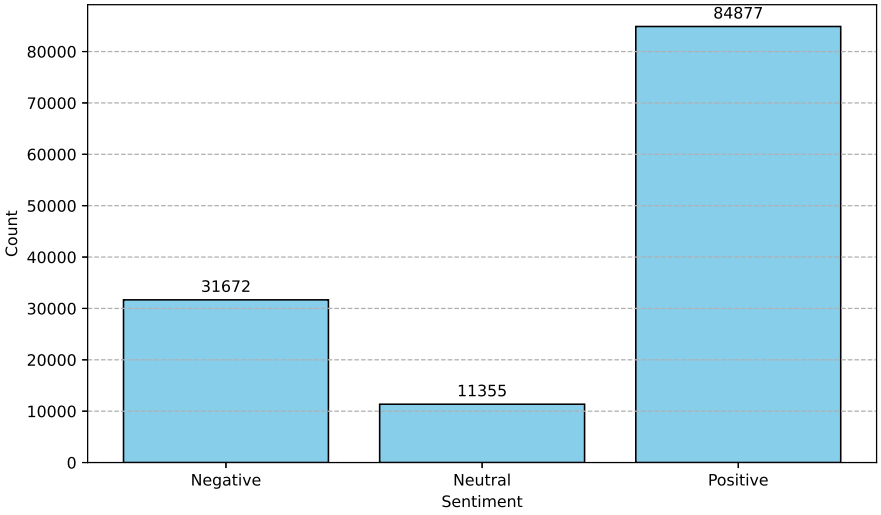


Figure 5.10: *Drug Ratings Categorized by Sentiment Class. Ratings have been grouped into three distinct sentiment classes: Negative (ratings 1-2), Neutral (rating 3), and Positive (ratings 4-5), illustrating the asymmetric distribution of sentiment while providing clearer boundaries between sentiment categories.*

5.3 Sentiment Analysis

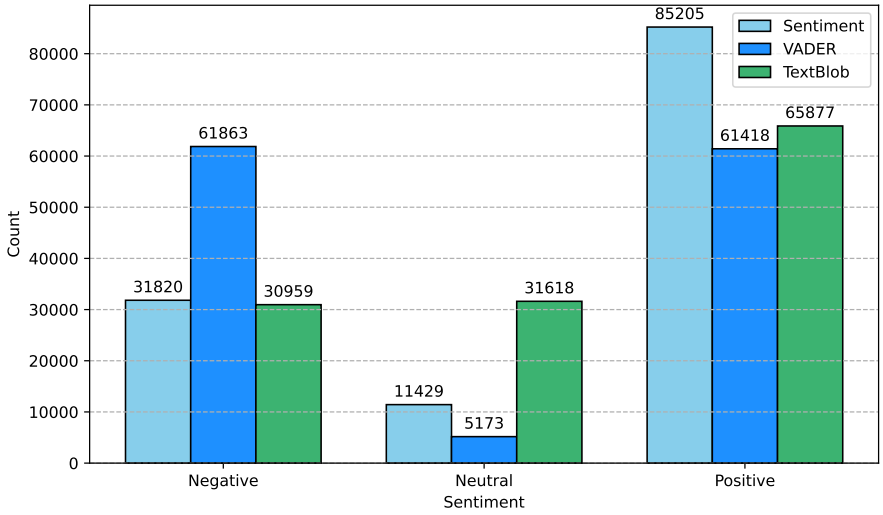
Following data preparation, the SA phase addresses our first research question by implementing both lexicon-based: Valence Aware Dictionary and sEntiment Reasoner (VADER) and TextBlob, and embedding-based: `bert-base-multilingual-uncased-sentiment` and `twitter-roBERTa` models to mitigate extreme response bias and correct misaligned reviews. Figure 5.11 presents a comparative analysis between clean (stemmed and unstemmed) reviews and the classifications produced by the lexicon-based models. For sentiment extraction using VADER and TextBlob, we applied the thresholds recommended by the original authors [40]:

- **Positive**: compound score ≥ 0.05
- **Neutral**: $-0.05 <$ compound score < 0.05
- **Negative**: compound score $\leq$ -0.05

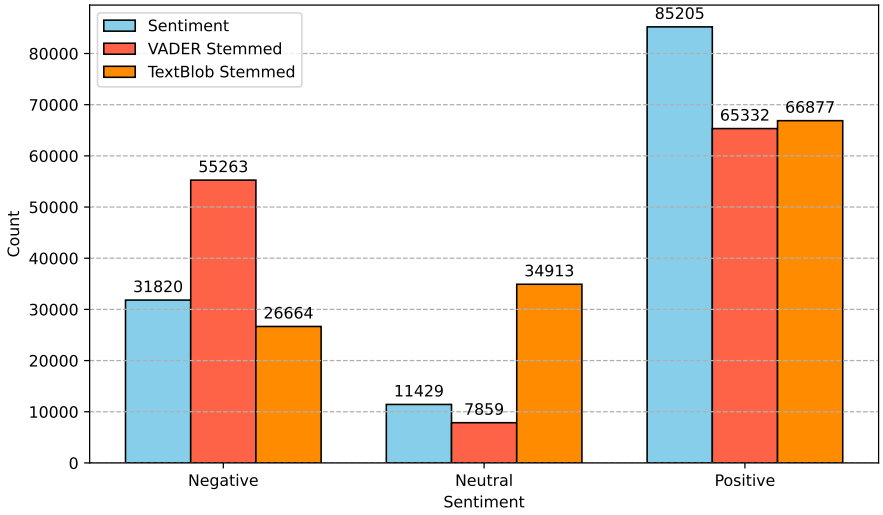
The results reveal that VADER tends to produce predominantly binary classifications, with minimal reviews falling into the neutral category, only 5,173 in unstemmed texts and 7,859 in stemmed versions. This suggests that VADER emphasises polarised sentiment, often pushing ambiguous reviews toward either extreme of the spectrum. Additionally, the increase in neutral classifications when stemming is applied indicates that this technique may reduce contextual information necessary for detecting sentiment intensity, resulting in more conservative classifications.

In contrast, TextBlob demonstrates a more balanced sentiment distribution. In the unstemmed scenario, it classifies similar numbers of reviews as Negative and Neutral, showing less polarisation than VADER. However, this balanced distribution suggests potential difficulty in distinguishing between negative and neutral sentiments. When stemming is applied to TextBlob, we observe a notable increase in neutral classifications, further indicating that reducing words to their root forms may remove emotional or contextual information critical for precise sentiment detection. To investigate why Negative and Neutral reviews present classification challenges for TextBlob, we computed the average subjectivity score for each sentiment category. As shown in Table 5.5, negative (0.378), neutral (0.395), and positive (0.403) reviews all display relatively low subjectivity scores. However, subjectivity gradually increases as reviews shift from negative to positive, with positive reviews demonstrating the highest subjectivity. This suggests that positive reviews tend to contain more personalised language, as positive sentiment often accommodates broader individual interpretation. Conversely, negative sentiment typically manifests in more direct and objective language patterns, making it more readily identifiable. The minimal difference in subjectivity between neutral and negative reviews (0.017) helps explain why models struggle to clearly distinguish between these sentiment categories.

Despite the changes introduced by stemming, both VADER and TextBlob consistently reflect the strong positive bias found in the original dataset. However, they apply more



(a) *Original unstemmed text analysis.*



(b) *Stemmed text analysis.*

Figure 5.11: *Comparison between Original and Stemmed Sentiment Distributions using VADER and TextBlob. This figure illustrates how preprocessing techniques affect lexicon-based sentiment analysis results highlighting differences in sentiment distribution patterns between the two preprocessing approaches.*

conservative sentiment classification, redistributing over 25,000 positive reviews into neutral and negative categories, reducing the positive count to approximately 60,000.

To address VADER's binary classification tendency and TextBlob's ambiguity in distinguishing between Negative and Neutral sentiments, we extended our analysis to include embedding-based models. These models not only provide sentiment predictions but also output confidence scores associated with each classification. `Twitter-roBERTa` predicts sentiment directly as Negative, Neutral, or Positive, while `bert-base-multilingual-uncased--sentiment` outputs a score on a 1-5 scale, which we subsequently mapped to the same three sentiment categories for consistency. These confidence scores are crucial in our sentiment correction methodology, so that when the original sentiment differs from model predictions, we select the corrected sentiment from the model demonstrating highest confidence.

Table 5.5: Average Sentiment Subjectivity Coefficient ($\bar{S}_{TB}$) by Sentiment. This table quantifies the degree of subjectivity in each sentiment category, demonstrating the gradual increase in subjective language as reviews shift from negative to positive, with the minimal difference between negative and neutral subjectivity scores (0.017) helping explain classification difficulties.

Sentiment	Negative	Neutral	Positive
$\bar{\mathbf{S}}_{\mathbf{TB}}$	0.378	0.395	0.403

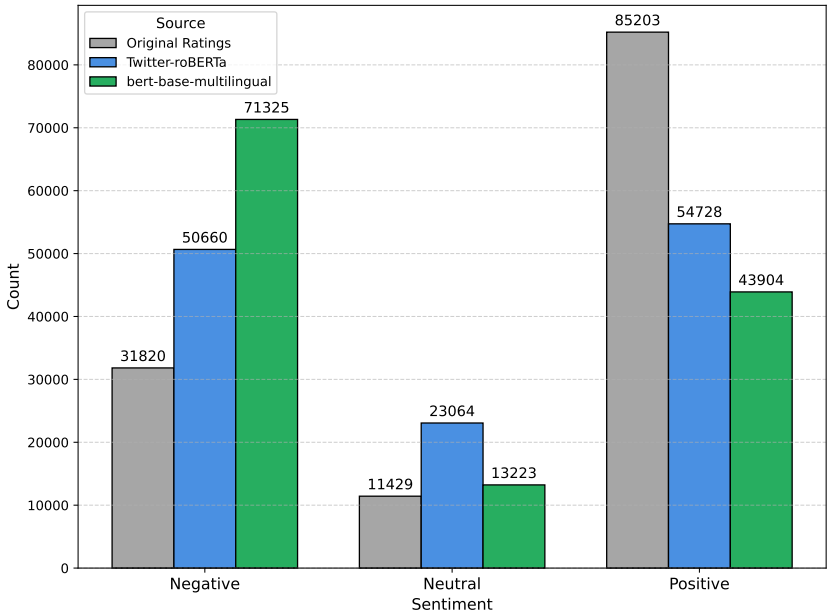


Figure 5.12: *Sentiment Distributions Comparison Between Original and Embedding-based Models. This visualization contrasts the original sentiment distribution with those produced by* `twitter-roBERsTa` *and multilingual BERT models, highlighting the more conservative interpretation of positive sentiment and the increased detection of negative sentiment by embedding-based approaches.*

As illustrated in Figure 5.12, both embedding-based models significantly reduce the number of positive reviews compared to the original distribution, reflecting a more conservative interpretation of the data. While the original dataset contains over 85,000 reviews labelled as Positive, `twitter-roBERTa` and multilingual BERT reduce this count to approximately 54,728 and 43,904, respectively. On the contrary, reviews classified as Negative increase substantially, with multilingual BERT assigning over 71,000 reviews, more than double the original count, while `twitter-roBERTa` assigns 50,660. These changes show that the embedding-based models can pick up signs of dissatisfaction in reviews that the original ratings did not clearly show.

Neutral sentiment classifications also increase with both models, with `twitter-roBERTa` producing more neutral classifications than multilingual BERT. This suggests that `twitter-roBERTa` adopts a more conservative approach to ambiguity, favouring neutrality, while multilingual BERT demonstrates a more assertive interpretation, pushing sentiment toward polarity.

To evaluate prediction confidence, we calculated the mean (μ) and standard deviation (σ) of confidence scores for each sentiment label. Table 5.6 presents these results, where BERT refers to `bert-base-multilingual-uncased-sentiment` and roBERTa to `twitter-roBERTa`. The analysis reveals that `twitter-roBERTa` displays higher confidence across all predicted sentiments compared to multilingual BERT. Specifically, roBERTa exhibits higher confidence scores for Negative (μ=0.770) and Positive (μ=0.756) predictions, demonstrating more assertiveness in classifying polarised opinions. Neutral predictions by roBERTa show lower average confidence (μ=0.501) alongside the smallest standard deviation (σ=0.085), indicating more cautious behaviour when processing ambiguous reviews. This statistical interpretation aligns with the observed distribution of sentiment predictions in Figure 5.12, where Positive and Negative classifications are almost equally distributed, accounting for 42.6% and 39.4% of reviews, respectively.

Table 5.6: Mean (μ) and Standard Deviation (σ) of Confidence Scores by Predicted Sentiment. This analysis quantifies the certainty levels of each model's predictions, revealing `twitter-roBERTa`'s higher overall confidence, particularly for polarized sentiments, while demonstrating more cautious behaviour with neutral classifications compared to multilingual BERT.

Predicted Sentiment	μ_{BERT}	σ_{BERT}	μ_{roBERTa}	σ_{roBERTa}
Negative	0.485	0.157	**0.770**	0.173
Neutral	0.408	0.109	**0.501**	0.085
Positive	0.505	0.164	**0.756**	0.165

Our final sentiment correction results, addressing extreme response bias, are depicted in Figure 5.13, which compares original and corrected sentiment distributions. The number of positive reviews decreased by approximately 11.2%, from 85,203 to 75,699, while negative reviews increased by approximately 43.2%, from 31,820 to 45,552. Neutral reviews declined significantly by 37%, from 11,429 to 7,201. This reduction in neutral classifications is particularly beneficial for TopicDrugRec, as these reviews typically contain ambiguous content that could potentially hinder performance.

To further understand each model's contribution to the correction process, Figure 5.14 illustrates the proportion of labels assigned by `twitter-roBERTa` versus multilingual BERT across sentiment categories. As expected, `twitter-roBERTa` was responsible for the majority of final sentiment predictions, aligning with its higher average confidence scores. In the Positive class, it accounted for 94.1% of all corrected reviews, while in the Neutral and Negative classes, it contributed 78.5% and 84.7%, respectively.

5.4 Topic Modelling

The previous steps combined data understanding and preparation, establishing the foundation for one of the core phases of TopicDrugRec: Topic Modelling (TM). This section organizes results by each of the studied models. For each, we evaluated multiple configurations, including variations in n-grams and stemming for Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF). We also tested different numbers of topics

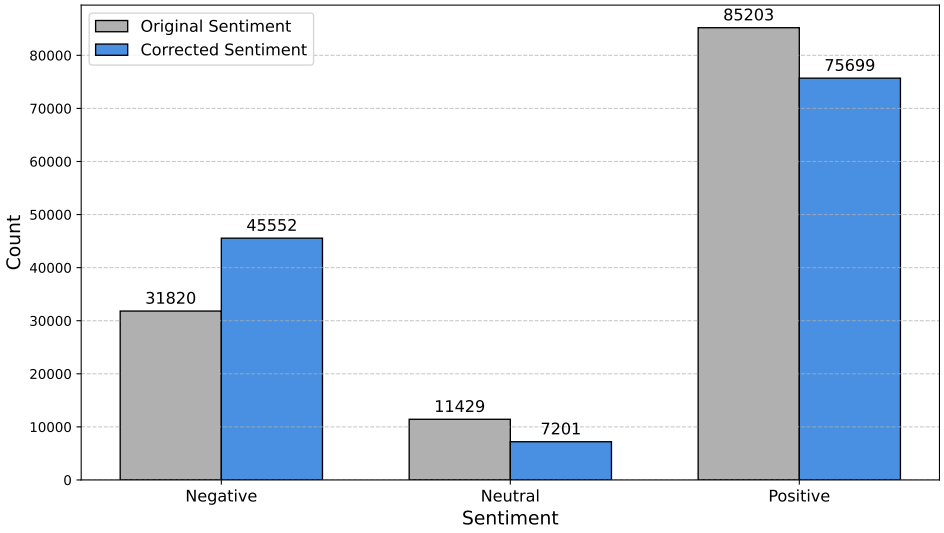


Figure 5.13: *Sentiment Distribution: Corrected vs Original. This figure illustrates the results of our sentiment correction methodology, showing a significant redistribution of reviews across sentiment categories—with an 11.2% decrease in positive reviews, a 43.2% increase in negative reviews, and a 37% reduction in neutral classifications.*

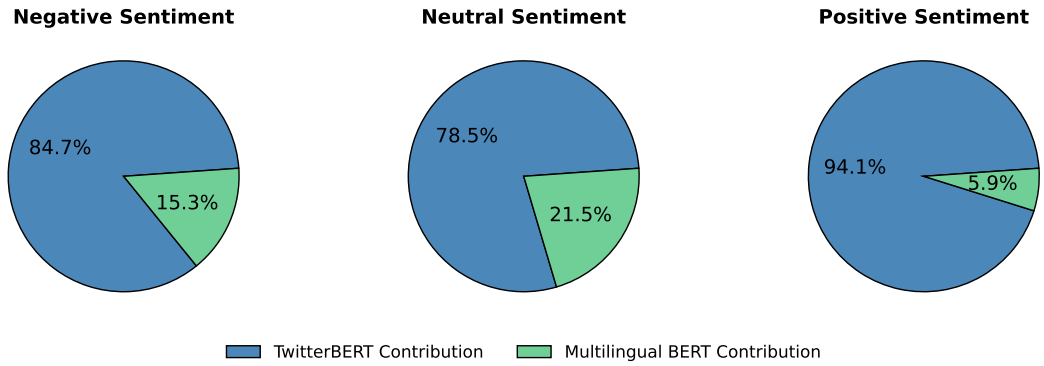


Figure 5.14: *Contributions of* `twitter-roBERTa` *and Multilingual BERT to Each Corrected Sentiment Label. This visualization breaks down how each embedding-based model contributed to the final corrected sentiment labels, highlighting* `twitter-roBERTa`*'s dominant role in the correction process across all sentiment categories, particularly for positive sentiment (94.1%).*

across all models, ranging from 2 to 50 in increments of 5. Additionally, we assessed the K-Means clustering algorithm, which was implemented to optimize drug recommendation inference speed, to determine the optimal number of clusters

5.4.1 Latent Dirichlet Allocation Results

Figure 5.15 presents the evolution of the four evaluation metrics for LDA across different configurations. As seen, at lower topic numbers, models trained with unigrams achieved higher diversity compared to bigram configurations. In particular, the unstemmed version maintained consistently high diversity across all topic numbers, while the stemmed reached similar values from 10 topics onwards.

In contrast, bigram models demonstrated lower initial diversity, requiring a higher number

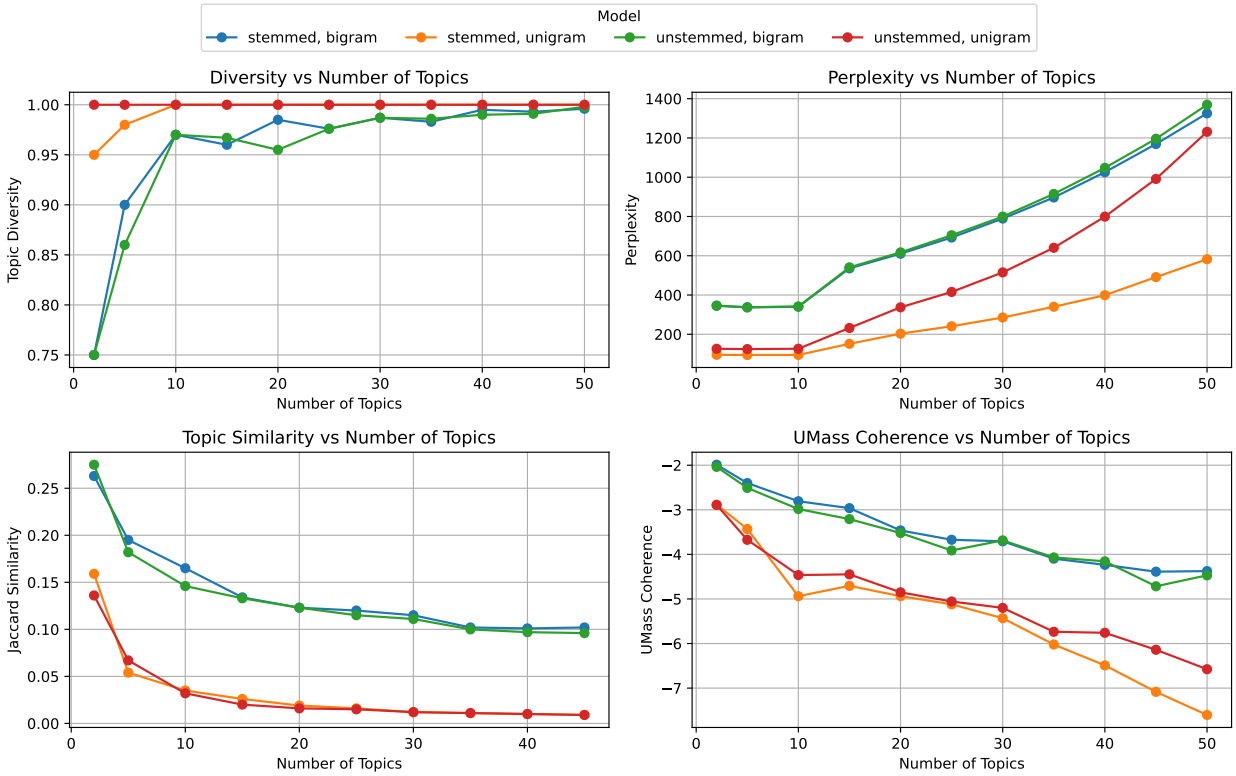


Figure 5.15: *Comparative analysis of LDA topic model performance across four evaluation metrics, Topic Diversity, Perplexity, Topic Similarity, and UMass Coherence, under different preprocessing configurations. Models were evaluated using stemming (stemmed vs. unstemmed) and n-gram settings (unigram vs. bigram), across a range of topic numbers (2 to 50). Each line represents a different text preprocessing configuration:* ***stemmed bigram*** *(blue line),* ***stemmed unigram*** *(orange line),* ***unstemmed bigram*** *(green line), and* ***unstemmed unigram*** *(red line).*

of topics to achieve diversity values comparable to those reached earlier by unigram models. This behaviour can be attributed to the fundamental nature of tokenization techniques. Unigrams, by capturing individual terms, produce a broader vocabulary and more varied topic representations. Conversely, bigrams form composite tokens that reduce the number of distinct terms, which at low topic numbers frequently results in overlapping content across topics.

For perplexity, all configurations show an increase as the number of topics grows, indicating higher model uncertainty with more complex topic structures. However, stemmed models consistently achieved lower perplexity values than their unstemmed counterparts, with this advantage being most notorious in the unigram version. The **stemmed unigram** LDA model demonstrated notably lower perplexity across all topic counts, with the difference becoming particularly significant at 50 topics. An important observation is that the gap between stemmed and unstemmed unigram models widened as the number of topics increased. This consistent advantage of stemming in reducing perplexity can be attributed to vocabulary reduction, which decreases the dimensionality of the feature space. Additionally, as the number of topics rose, the sparsity in topics also increased, making the benefits of reducing vocabulary to root forms more pronounced.

Table 5.7: Model evaluation metrics for different LDA configurations.

Model	#Topics	UMass Coherence	Perplexity	Diversity	Similarity
Unstemmed Unigram	**10**	**-4.464**	**126.380**	**1.000**	**0.032**
Unstemmed Unigram	20	-4.851	337.381	1.000	0.016
Stemmed Unigram	25	-5.120	240.863	1.000	0.016

Regarding topic similarity, the unstemmed unigram configuration consistently achieved lower Jaccard similarity scores across the entire range of topic counts. This behaviour reflects the richer and less constrained vocabulary of these models, which facilitates the creation of more distinct topics with reduced word overlap. Contrary to perplexity, where stemming improves word prediction performance, in similarity measurements, stemming reduces word uniqueness, leading to higher overlap between topics.

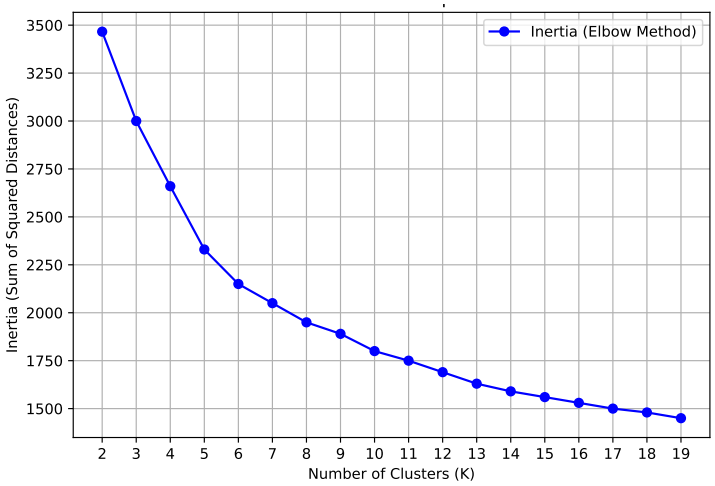
Finally, the unigram configurations outperformed the bigram versions, with the stemmed unigram achieving the best values. Similar to the behaviour in perplexity, the gap between the stemmed and unstemmed unigram models widens as K increases. This trend can also be explained by the model generating increasingly sparse topic distributions at higher topic ranges, an effect which is mitigated by reducing words to their root forms through stemming.

Based on our analysis of the evaluation metrics and additional topic interpretability assessment, we selected the **Unstemmed Unigram – 10 topics** configuration as the optimal LDA model. This configuration provides the best balance across all evaluation metrics while utilizing fewer topics, which is crucial since a lower topic count significantly accelerates the inference process by reducing the number of probability calculations required when determining the distribution for each user symptom in **TopicDrugRec**.

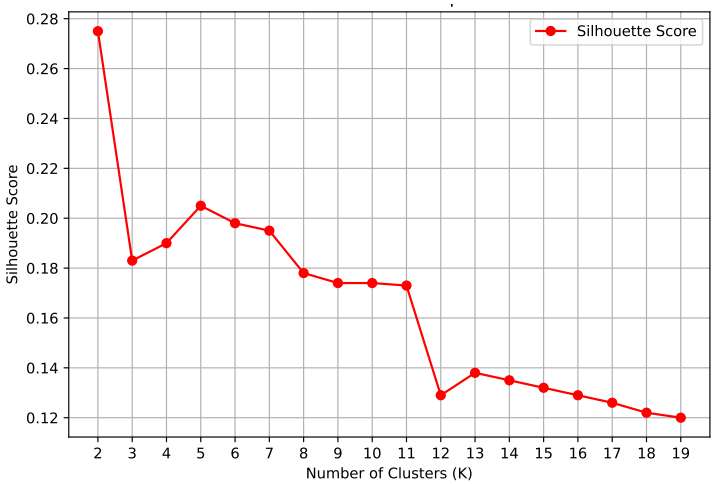
The three best-performing model configurations are presented in Table 5.7, with our selected model highlighted in bold.

To further enhance topic inference efficiency, we implemented the K-means clustering algorithm. For determining the optimal number of clusters, we applied two complementary evaluation methods: the Elbow Method [41] and the Silhouette Score [34]. The Elbow Method identifies the point where adding additional clusters no longer substantially reduces the sum of squared distances within clusters (inertia), while the Silhouette Score evaluates cluster quality by measuring both within-cluster cohesion and between-cluster separation, with higher scores indicating superior clustering.

As illustrated in Figure 5.16, the inertia curve exhibits diminishing returns as the number of clusters increases. Notably, from $K = 8$ onward, the reduction in inertia becomes marginal, decreasing from approximately 2,000 to just below 1,500 across 11 cluster increments. While the Silhouette Score reaches its maximum value at $K = 2$, such a low cluster count would result in excessively generic grouping of the data. The optimal trade-off between minimizing inertia and maximizing the silhouette score happens at $K = 7$, which offers a more balanced clustering with sufficient granularity, this was the number of clusters chosen for the LDA model.



(a) *Elbow Method: Inertia across different numbers of clusters.*



(b) *Silhouette Score: Cluster separation quality across different K.*

Figure 5.16: *Evaluation of the optimal number of clusters using (a) the Elbow Method and (b) the Silhouette Score for the LDA model.*

5.4.2 Non-negative Matrix Factorization Results

Unlike LDA, NMF does not rely on probabilistic generation, and as such, this model is not evaluated using the perplexity metric [69]. Figure 5.17 displays the performance of different NMF configurations across the suggested range of topics.

In terms of topic coherence, a clear separation between the four NMF models configurations is observed, with stemming and n-gram choices leading to distinctly different coherence scores. This behaviour contrasts with the one seen in the LDA model, where the differences in coherence were primarily justified by the number of n-grams, with unigram models consistently outperforming bigrams.

Given the nature of the NMF algorithm, which relies on linear decomposition, this model may not be as sensitive to dictionary preprocessing transformations as LDA is, being more dependent on the term frequency structure rather than probabilistic occurrences, for example, if the input includes "blood" or "blood pressure", NMF focuses on how frequently these terms appear across documents, whereas LDA aims to relate these terms as a semantic

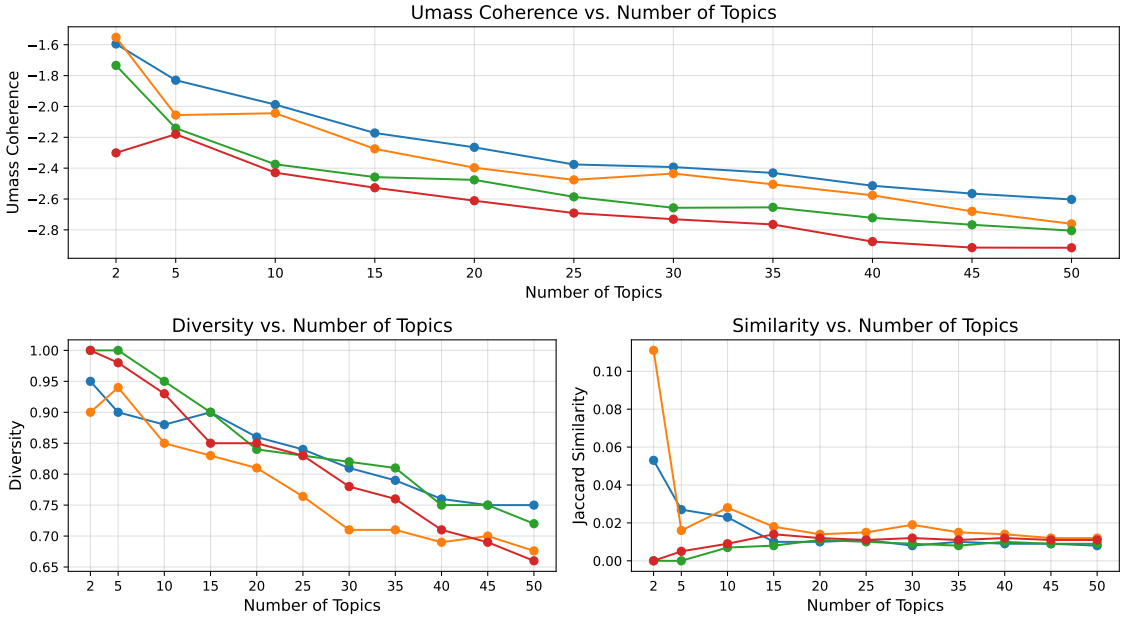


Figure 5.17: *Comparative analysis of topic modelling performance metrics for NMF across different configurations. The figure presents three evaluation metrics, UMass Coherence, Topic Diversity, and Jaccard Similarity, as a function of the number of topics (2-50). Each line represents a different text preprocessing configuration:* ***stemmed bigram*** *(blue line),* ***stemmed unigram*** *(orange line),* ***unstemmed bigram*** *(green line), and* ***unstemmed unigram*** *(red line). Higher values are better for UMass Coherence (less negative), higher values are better for Diversity, and lower values are preferred for Similarity.*

relationship.

As the number of topics increases, we observe improved performance in terms of coherence, with the unstemmed unigram configuration consistently achieving the best scores overall. However, this improvement is marginal; for instance, the unstemmed bigram model with 15 topics reaches a coherence value very close to the best-performing configuration, and at higher topic counts, the difference between unstemmed unigram, unstemmed bigram, and stemmed bigram becomes less pronounced. This convergence in coherence reinforces the idea that NMF is less reliant on specific word representations than LDA.

Moving to diversity, NMF displays a contrasting trend compared to the previously analysed LDA models. As the number of topics increases, diversity progressively declines for all configurations. This suggests that the top terms in each topic begin to overlap, reducing the uniqueness of topic vocabularies. More specifically, the unigram configurations, both stemmed and unstemmed consistently perform worse in this metric, declining as the number of topic count grows. This reduction in diversity, however, does not necessarily translate to an increase in topic similarity, as this metric measures the uniqueness of the top 10 terms, whereas similarity captures the overlap between all topic terms across the model.

Among the evaluated configurations, two stand out at 15 topics: stemmed bigram and unstemmed bigram, with both achieving a diversity score of 0.90. Additionally, an interesting pattern is seen around the 20 to 25 topic mark, where all configurations converge to similar diversity values, indicating a temporary balance in performance, regardless of the

Table 5.8: Model evaluation metrics for different NMF configurations.

Model	#Topics	Umass Coherence	Diversity	Similarity
Unstemmed Bigram	**15**	**-2.458**	**0.900**	**0.008**
Unstemmed Bigram	50	-2.805	0.720	0.009
Stemmed Bigram	50	-2.603	0.750	0.008

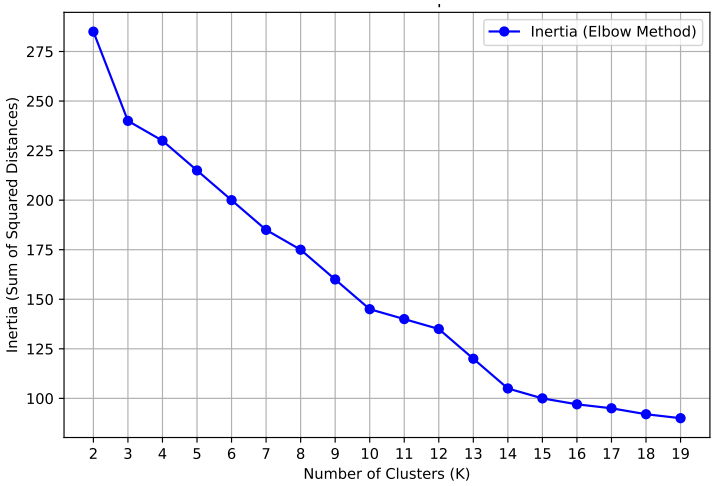
text preprocessing techniques. However, from 30 topics onwards, a clear separation reappears, with bigram-based models maintaining higher diversity compared to their unigram counterparts. This may indicate the key-point at which the effect of creating more complex token configurations (2-gram, 3-gram, etc.) presents as a more performant alternative in terms of topic diversity.

Finally, the similarity metric suggests that from 15 topic onwards, all models converge relatively the same to an approximate value of 0.01 similarity, indicating that the generated topics share very few or no common words between them. Since this metric is calculated using Pairwise Jaccard Similarity, a value of 0.01 means only 1% of the words are shared between two topics. These shared words may include stop words or highly frequent domain-specific terms that persisted across the topics, despite the preprocessing stage.

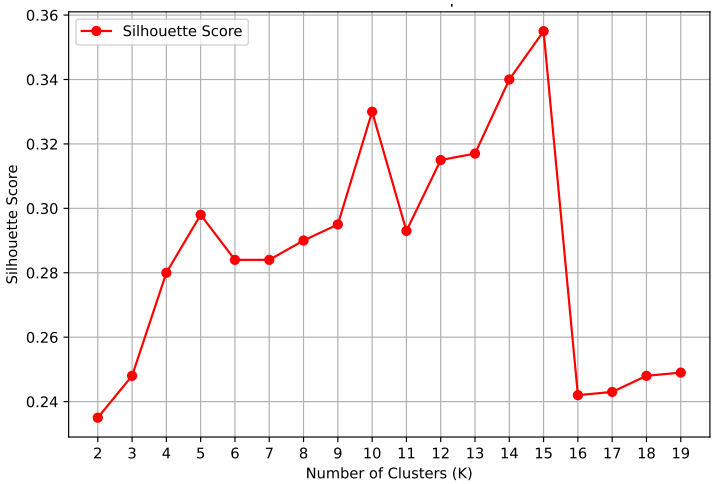
This convergence also supports that beyond a given threshold (10 topics in this scenario), NMF is capable of extracting distinct topics, regardless the tokenization or stemming strategies. Furthermore, it reinforces the differences between diversity and similarity, where, in this case, the configurations with lower diversity managed to achieve minimal similarity, indicating that even though top keywords may be reused, the underlying topic themes remain unique.

Considering this trade-off between performance and computational efficiency, the **unstemmed bigram configuration with 15 topics**, whose performance metrics can be seen in Table 5.8 highlighted in bold, presents as the most balanced choice compared to the second and third best models, even though it does not achieve the highest diversity score nor the best topic coherence, it maintains respectable performance in both metrics.

Figure 5.18 illustrates the results of our cluster optimization analysis for the selected NMF model using two complementary techniques. The Elbow Method, shown in the top plot, reveals a sharp decrease in inertia up to $K = 14$, after which the curve flattens significantly. This inflection point indicates that additional clusters beyond this threshold produce diminishing returns, as they no longer substantially improve intra-cluster performance. Simultaneously, the Silhouette Score analysis, presented in the bottom plot, peaks at $K = 15$, with $K = 14$ achieving the second highest score with only a marginal difference. Given this minimal performance gap, we selected $K = 14$ as the optimal number of clusters, as the negligible improvement offered by an additional cluster does not justify the increased computational complexity it would introduce.



(a) *Elbow Method: Inertia across different numbers of clusters.*



(b) *Silhouette Score: Cluster separation quality across different K.*

Figure 5.18: *Evaluation of the optimal number of clusters using (a) the Elbow Method and (b) the Silhouette Score for the NMF model.*

5.4.3 BERTopic Results

Unlike the previously evaluated TM approaches, BERTopic leverages transformer-based embeddings that are capable of understanding full semantic context, including the grammatical structure and relationships between words. Therefore, even stop words contribute meaningfully to the embedding space, allowing for more cohesive topic extraction. As such, stemming or stop-word removal are unnecessary in this scenario, and may even potentially hinder the performance of the task by removing contextual information from the drug reviews. This is explicitly stated in the documentation, which advises "*(. . .) removing stop words as a preprocessing step is not advised as the transformer-based embedding models that we use need the full context to create accurate embeddings.*" [29].

Figure 5.19 presents the evaluation metrics applied to the cleaned original reviews. The figure includes a dashed grey line at 38 topics, representing BERTopic's 'auto' topic estimation feature output, which determines the optimal number of topics based on the embedded data structure.

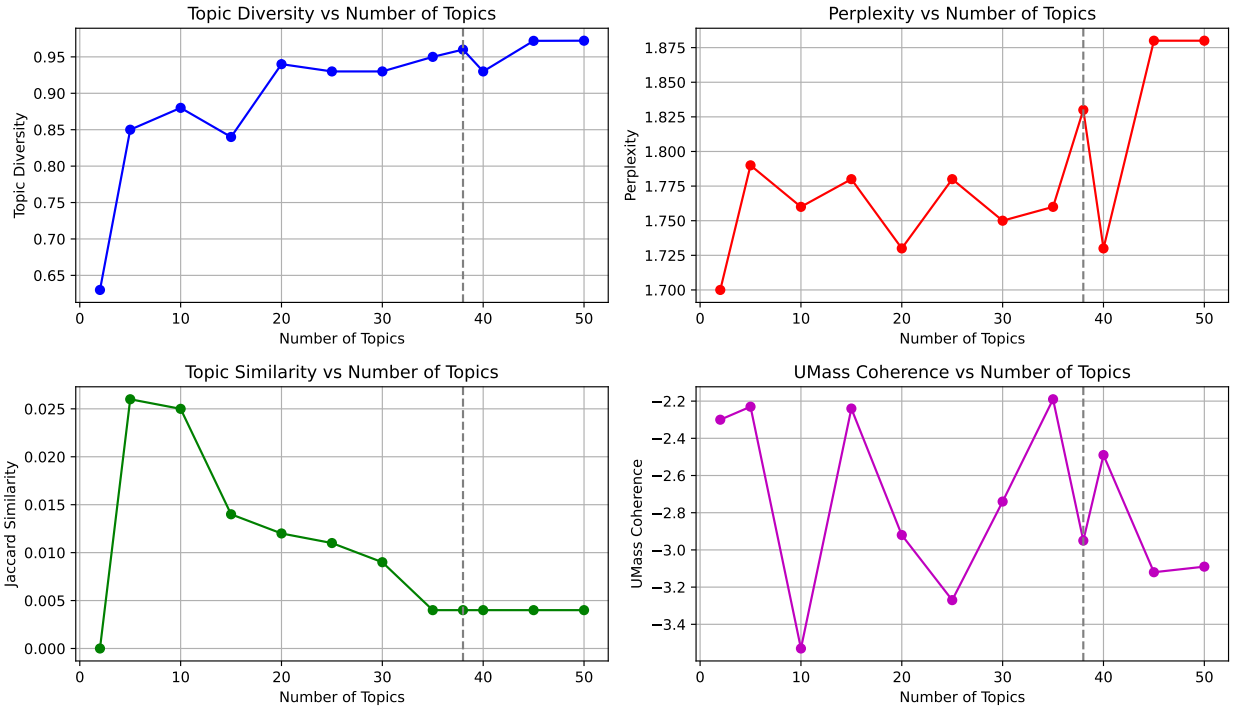


Figure 5.19: *BERTopic evaluation metrics across topic configurations (2-50). The panels display Diversity (top-left), Perplexity (top-right), Similarity (bottom-left), and UMass Coherence (bottom-right). The dashed vertical line at 38 topics represents BERTopic's automatic topic estimation output. Higher diversity values indicate more distinct topic vocabularies, lower perplexity values suggest better predictive performance, lower similarity scores reflect reduced topic overlap, and lower UMass Coherence values indicate better topic coherence.*

As opposed to the results from the probabilistic, and matrix factorization models , diversity starts at its lowest value when the number of topics is smallest. In this scenario, the model generalises broad themes across the many drug reviews, which is leading to the reuse of top terms across all topics. However, as the number of topics increases, diversity starts to improve, reaching its maximum value of 1 at 45 and 50 topics, meaning that the top terms of each topic are completely distinct from each other. As opposed to the other studied models, this pattern suggests that BERTopic benefits from the contextual embeddings, becoming better at isolating underlying subjects. Notably, the automatically estimated number of topics (38), presents a relatively high diversity of approximately 0.93, displaying a good trade-off of diversity and number of topics.

Perplexity ranges from approximately 1.70 to 1.88, representing an overall increase of around 10.6% as the number of topics grows from 2 to 50. While this increase is not major, perplexity values remain unstable across most topic counts, with fluctuations observed until around 45 topics, after which they stabilise. At lower topic counts, BERTopic produces broad clusters that capture wide contexts, resulting in lower perplexity. As the number of topics increases, perplexity begins to rise and fluctuate, which may suggest topic overlap or data noise. The 38-topic configuration sits between the lower end of perplexity and the stabilisation point seen from 45 topics onwards.

Table 5.9: Top-performing BERTopic configurations with evaluation metrics. The optimal 38-topic model (automatically estimated) is highlighted in bold. Lower UMass Coherence indicates better topic coherence, lower Perplexity suggests better model fit, higher Diversity shows more distinct topic vocabularies, and lower Similarity reflects reduced topic overlap.

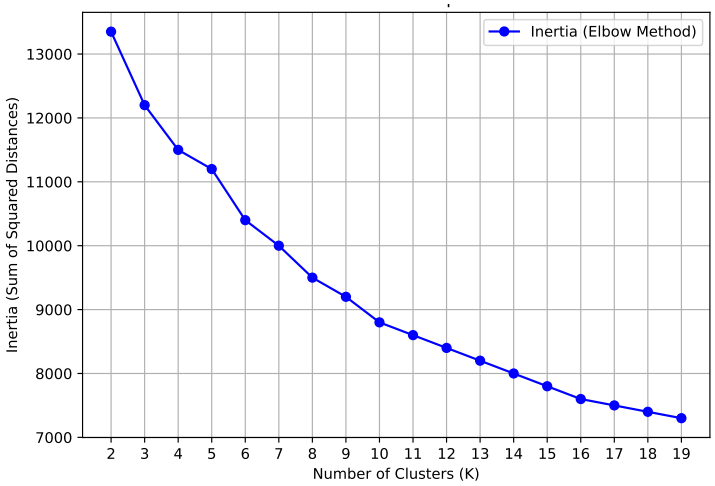
#Topics	Umass Coherence	Perplexity	Diversity	Similarity
38	**-2.953**	**1.832**	**0.960**	**0.004**
45	-3.121	1.889	0.970	0.004
25	-3.275	1.786	0.930	0.011

Despite the overall low similarity values, this metric shows a decline as the number of topics increases. At lower topic counts, term overlap is more pronounced, leading to higher similarity. As more topics are introduced, each topic's vocabulary becomes more distinct, reducing overlap and lowering similarity scores. A key distinction between BERTopic and other analysed models is its consistently low similarity values, attributable to transformer-based embeddings that create distinct clusters even with fewer topics, reducing term redundancy. Supporting the automatic topic estimation feature, the 38-topic configuration falls within the range of topics (35-50) with the lowest similarity values.

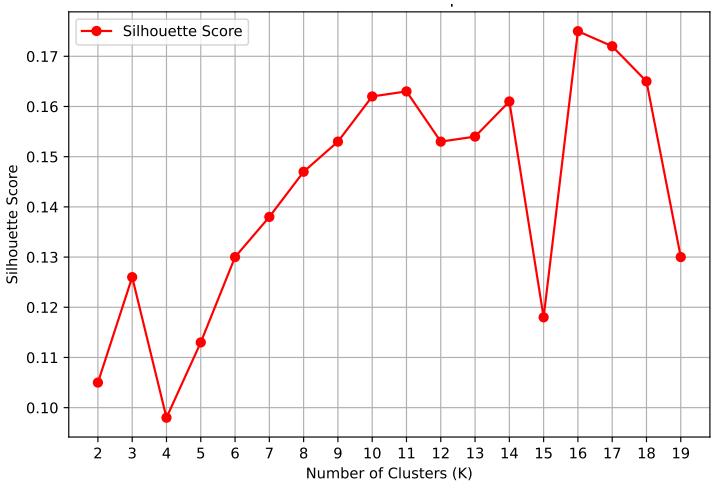
The UMass coherence exhibits an irregular fluctuating trend rather than a consistent improvement trajectory. In general, it is possible to see that at higher topic counts, the coherence declines, indicating better topic coherence, however, there are multiple sharp oscillations throughout the topic range. Unlike LDA or NMF, BERTopic does not demonstrate a clear trend of improved coherence with more topics. Instead, optimal configurations appear at 25, 38, and 45 topics. The strong performance at 38 topics, which also aligns with very good metric performance-topic count in the other evaluation metrics reinforces the quality of BERTopic's automatic topic estimation. While LDA and NMF require manual tuning and rely on very extensive analysis to find the optimal number of topics, BERTopic's approach is more straightforward and produces good metric performances, as seen in this use case, simplifying the TM pipeline.

Based on comprehensive analysis of all metrics, the 38-topic configuration was selected as the optimal BERTopic model, highlighted in bold in Table 5.9 alongside the two other best models.

Figure 5.20 presents the results of the Elbow Method and Silhouette Score analysis for determining the optimal number of clusters (K) for the selected 38-topic BERTopic model. The Elbow Method plot shows that inertia decreases sharply until $K = 7$, indicating significant improvement in intra-cluster distance as data points become closer to their respective cluster centroids. Beyond this point, the diminishing rate of decline suggests that additional clusters provide marginal performance gains. The Silhouette Score, which measures cluster separation and cohesion, reaches its maximum value at $K = 15$ (0.175). However, the score at $K = 7$ (approximately 0.14) remains competitive while providing a more parsimonious solution. Based on this analysis, $K = 7$ was selected as the optimal number of clusters, balancing good cluster quality with model simplicity and computational efficiency. This selection provides a more interpretable clustering structure while maintaining robust



(a) *Elbow Method: Inertia across different numbers of clusters.*



(b) *Silhouette Score: Cluster separation quality across different K.*

Figure 5.20: *Cluster optimization analysis for the 38-topic BERTopic model using: (a) the Elbow Method to identify the point of diminishing returns in cluster compactness, and (b) the Silhouette Score to evaluate separation between clusters and cohesion within clusters.*

separation between topic groups.

5.5 Comparative Analysis of Topic Modelling Approaches

5.5.1 Comparative Analysis of LDA and NMF Models

After evaluating both LDA and NMF models across multiple configurations, it is instructive to compare their performance on shared evaluation metrics. In terms of topic coherence, NMF consistently outperformed LDA across comparable configurations, with the selected unstemmed bigram NMF model (15 topics) achieving a coherence score of -2.458 compared to -4.464 for the optimal unstemmed unigram LDA model (10 topics). This substantial difference (approximately 45% improvement) suggests that NMF's non-probabilistic approach may be better suited for extracting coherent topics from drug review data. Despite the algorithm's dependence on preprocessing techniques, this finding aligns with research by [69], who demonstrated that NMF often yields more interpretable topics in specialized

Table 5.10: Comparative evaluation of optimal LDA and NMF models for the TopicDrugRec system.

Model	Configuration	#Topics	Coherence	Diversity	Similarity	K
LDA	Unstemmed Unigram	10	-4.464	1.000	0.032	7
NMF	Unstemmed Bigram	15	-2.458	0.900	0.008	14

domains characterized by distinct terminology patterns.

Regarding topic diversity, both the optimal models performed well, with the LDA model achieving perfect diversity (1.000) and the NMF model showing strong performance (0.900). The difference between the two models highlights fundamental algorithmic distinctions. LDA, being a probabilistic model, distributes words across topics to maximize distinctiveness, which often results in higher diversity, especially as the number of topics increases. On the other hand, NMF is based on matrix factorization, constructing topics as non-negative combinations of terms. As more topics are added, NMF tends to reuse high-weight words across topics due to the nature of its decomposition, leading to increased overlap and, consequently, slightly lower diversity.

Both models performed well in minimizing topic similarity, with Jaccard similarity scores of 0.032 for LDA and 0.008 for NMF. The lower similarity score for NMF indicates less overlap between topics, suggesting that, despite its slightly lower diversity score, the NMF model may produce more clearly separated topics overall. This apparent contradiction between the diversity and similarity metrics underscores the need to evaluate TM using multiple criteria, as each metric captures a distinct aspect of the model's performance.

The clustering optimization process revealed notable differences, with the optimal LDA model requiring fewer clusters ($K = 7$) compared to the NMF model ($K = 14$). This disparity suggests that NMF topics may capture more granular patterns that benefit from a greater number of clusters, while LDA topics tend to represent broader thematic categories that can be grouped effectively into fewer clusters. This distinction in the number of clusters has significant implications for the computational efficiency of TopicDrugRec, as the LDA approach involves fewer probability calculations during the inference process.

In terms of computational efficiency, the LDA model offers advantages in both topic count (10 vs. 15) and cluster count (7 vs. 14), suggesting potentially faster inference times. However, the NMF model outperforms LDA in terms of coherence and topic separation, which could lead to more precise recommendations. This performance-efficiency trade-off must be carefully weighed in the final implementation of TopicDrugRec, with further benchmarking recommended to identify which approach best balances the system's needs for both accuracy and responsiveness.

5.5.2 Extending the LDA and NMF Comparison to BERTopic

Extending the previous analysis of LDA and NMF models (Table 5.10), this section extends the analysis to include the embedding-based model, BERTopic, providing a more detailed evaluation of all three TM approaches. Table 5.11 summarizes the optimal configurations

Table 5.11: Comparison of optimal topic model configurations across LDA, NMF, and BERTopic. Closer to 0 UMass Coherence indicates better topic coherence, higher Diversity represents more distinct topic vocabularies, lower Similarity reflects reduced topic overlap, and K represents the optimal number of clusters.

Model	Configuration	Topics	Coherence	Diversity	Similarity	K
LDA	Unstemmed Unigram	10	-4.464	1.000	0.032	7
NMF	Unstemmed Bigram	15	-2.458	0.900	0.008	14
BERTopic	Original Text	38	-2.953	0.960	0.004	7

and performance metrics for each of the studied models.

While the previous analysis showed that NMF achieved the best coherence score (-2.458), BERTopic followed closely with a coherence of -2.953, still outperforming LDA (-4.464). In terms of topic diversity, BERTopic scored 0.960, which is higher than NMF (0.900) and close to LDA's perfect score of 1.000. This suggests that BERTopic's transformer-based embeddings are effective at distinguishing topics while still capturing coherent themes, similarly to what was observed with LDA. In addition, BERTopic achieved the lowest similarity score (0.004) across all models, improving on NMF (0.008) and significantly outperforming LDA (0.032). The clustering results further support this trend: despite generating more topics (38), BERTopic's optimal cluster count was $K = 7$, the same as LDA. This indicates that, while BERTopic captures more detailed topics, these can still be organised into broader clusters, consistent with LDA's structure. Beyond metric-based evaluation, BERTopic also introduces several practical advantages. It eliminates the need for extensive preprocessing such as stop-word removal and stemming, as it performs best with raw or minimally processed text. Moreover, its built-in topic estimation feature automatically selected the 38-topic configuration used in this analysis, reducing the need for manual tuning and accelerating the implementation process.

5.6 External Knowledge

Addressing the final research question, TopicDrugRec integrates structured external knowledge on contraindications and Drug-Drug Interaction (DDI) to enhance the safety and relevance of its suggestions.

Three external sources were incorporated to enrich the recommender system with biomedical context: Side effects [3] curated from Drugs.com [20], contraindications extracted from DailyMed [92] and drug–drug interactions obtained from DDInter [17]. To ensure consistency in drug naming across all datasets and the recommender matrices from each TM model, the open-source Drug Standards library [7] was used to map brand names and variations to their corresponding chemical entities (for example, "Benadryl" to "Diphenhydramine").

The initial Drug Reviews dataset contained 2388 unique drug names, while the Side Effects dataset provided information for 820 unique drugs. In this dataset, as the side effects were curated from Drugs.com in free-text form, the *side_effect* feature contained long and unstructured descriptions rather than the isolated side effects. To address this, Named

Table 5.12: Top 5 most frequently extracted side effects after NER processing.

Side Effect	Count
hives	1317
rash	465
swelling	344
allergic reaction	292
pain	267

Entity Recognition (NER) was applied using the `en_ner_bc5cdr_md` model from SciSpacy [66], which was specifically trained for identifying entities in biomedical text, particularly, diseases.

Following this extraction, and subsequent alignment of side effects with the drugs present in the recommender matrices, 612 distinct side effects were identified, **"hives"**, **"rash"**, **"swelling"**, **"allergic reaction"**, and **"pain"** are the three most common, further represented in Table 5.12.

The contraindications data from DailyMed was originally extracted as a collection of individual JSON files, each containing structured information such as the drug name, active ingredients, contraindications, and, in some cases, additional sections including warnings, precautions, and adverse reactions [73]. The resulting files were later consolidated into a tabular format for downstream processing and integration into this work.

The compiled dataset included a total of **301 unique drug entries**, with many drugs appearing across multiple JSON files under different brand names. This variation required the use of the **DrugStandard** library to harmonise drug names according to a standardised naming convention consistent with those used in the recommender matrices. After standardisation, entries referring to the same active substance were **grouped**, and their associated contraindications, warnings, and adverse reactions were **aggregated into a single record per drug**.

Out of the initial 301 drugs, **252 were successfully matched** to corresponding entries in the recommender matrix. Among these, **231 drugs contained valid contraindication information**, while **adverse reaction data was available for all 252** matched drugs.

In addition to contraindications and adverse reactions, this stage also incorporated **DDI**, sourced from the publicly available **DDInter** database [17]. This dataset consists of pairwise interaction records between drugs, each labelled with a severity level classified as *Minor*, *Moderate*, *Major*, or *Unknown*. Every entry includes both drugs involved in the interaction, identified by their DDInter IDs and corresponding names, along with the severity of the interaction.

As with the previous external knowledge sources, drug names were first standardised using the **DrugStandard** library to ensure consistency with the drugs represented in the TopicDrugRec recommender matrices, filtering out interactions involving drugs that could not be matched. Initially, DDInter included **1,939 unique drugs** and **160,235**

interaction entries, and after alignment with TopicDrugRec's vocabulary, the dataset was reduced to **816 drugs**, covering a refined total of **71,054 interactions**.
To facilitate integration, the dataset was reshaped from a sparse pairwise format into a structured representation with the following fields: `drug_name`, `Minor_Interactions`, `Moderate_Interactions`, `Major_Interactions`, and `Unknown_Interactions`. Each row corresponds to a single drug, with each interaction column containing a list of drugs that interact with it at the specified severity level.

To complete the external knowledge integration, the three external knowledge datasets, *side effects*, *contraindications*, and *DDI* were merged using a **left join** [83], with the `side_effects` dataset serving as the base. The join was performed on standardised drug names across all sources, resulting in a final dataset containing **2,462 entries** and **8 features**. Table 5.13 presents the coverage of each feature, expressed as the percentage of non-null entries. The DDI features show high coverage, particularly `Moderate_Interactions` at 96.75%, indicating that most drugs are associated to at least one moderate interaction. Furthermore, `contraindications` and `adverseReactions` exhibit the lowest coverages, 34.85% and 38.91%, respectively, however, this is explained due to the smaller scope of the dataset within the TopicDrugRec recommender matrices.

Table 5.13: Feature coverage in the final external knowledge dataset, expressed as the percentage of non-null entries per column. While `drug_name` and `side_effects` are fully populated, DDI fields display high coverage, with `Moderate_Interactions` being the most represented. In contrast, `contraindications` and `adverseReactions` exhibit the lowest coverage, as they originate from the smallest external knowledge source, covering only 252 drugs.

Feature	Coverage (%)
`drug_name`	100.00
`side_effects`	100.00
`Minor_Interactions`	80.91
`Moderate_Interactions`	96.75
`Major_Interactions`	87.53
`Unknown_Interactions`	72.54
`contraindications`	34.85
`adverseReactions`	38.91

5.7 Web Application

To enable the use of **TopicDrugRec** in a real-world setting, a web-based application was developed and tested locally, running inside a Docker container. The source code is available at `https://github.com/matpato/TopicDrugRec`, and includes all necessary files to build and run the Docker image locally, allowing users to interact with the recommender system directly.

Figure 5.21 shows the main interface of the application, where users can input symptoms in free-text form and optionally select a condition they believe is relevant. This application is a proof-of-concept, designed to be tested by clinical or technical staff, where the professional

listens to the patient's complaints and describes the symptoms in clinical terms. Users can also choose the topic model to be used (LDA, NMF, or BERTopic), select the number of desired recommendations, and adjust the hyperparameters that control the importance of topic similarity, sentiment, and usefulness.

A practical use case is summarised below:

1. **Input Description:** Appears withdrawn and emotionally flat. Describes feeling like everything is pointless and has stopped caring about things that once mattered. Eats very little, barely sleeps, and avoids people. Mentions thinking about death often, not wanting to wake up, and feeling like others would be better off without them—clear signs of deep depression and suicidal thoughts.

2. **Model:** BERTopic

3. **Hyperparameters:** $Similarity = 1; UsefulCount = Sentiment = 0.7$

4. **Number of Recommendations:** 5

The results of this use case are shown in Figure 5.21. In this scenario, the clinician might reasonably expect the condition to be "Depression". On the left-hand side, the system displays the top 5 recommended drugs, ranked by their final score. On the right-hand side, the table shows the drugs associated with the expected condition. This table is paginated and allows the user to explore all relevant drugs in the dataset enabling manual inspection of the results.

After reviewing the recommendations, the clinician can navigate to the Drug Info tab (visible in the top right of Figure 5.21) to access additional information about any recommended drug. For instance, searching for `FLUOXETINE` opens a view that presents its known side effects, contraindications, and drug–drug interactions, as shown in Figure 5.22.

5.8 TopicDrugRec Evaluation

This final evaluation section presents an assessment of the TopicDrugRec RS, evaluating its ability to generate meaningful and clinically relevant drug recommendations based on user-described symptoms. Building on the core components developed throughout this dissertation, the objective is to determine how effectively unstructured textual data can be leveraged to train TM algorithms and transform unseen patient input into treatment suggestions, mitigating the information overload on clinicians. In doing so, this section contributes to addressing the second research question, outlined in section 1.3. For this evaluation step, the dataset was split into a training (70%) and test (30%), where the first is used to train the topic models and create the recommendation matrix, while the unseen test set was used to simulate patient inputs, apply the trained model then generate and evaluate the quality of the recommendations.

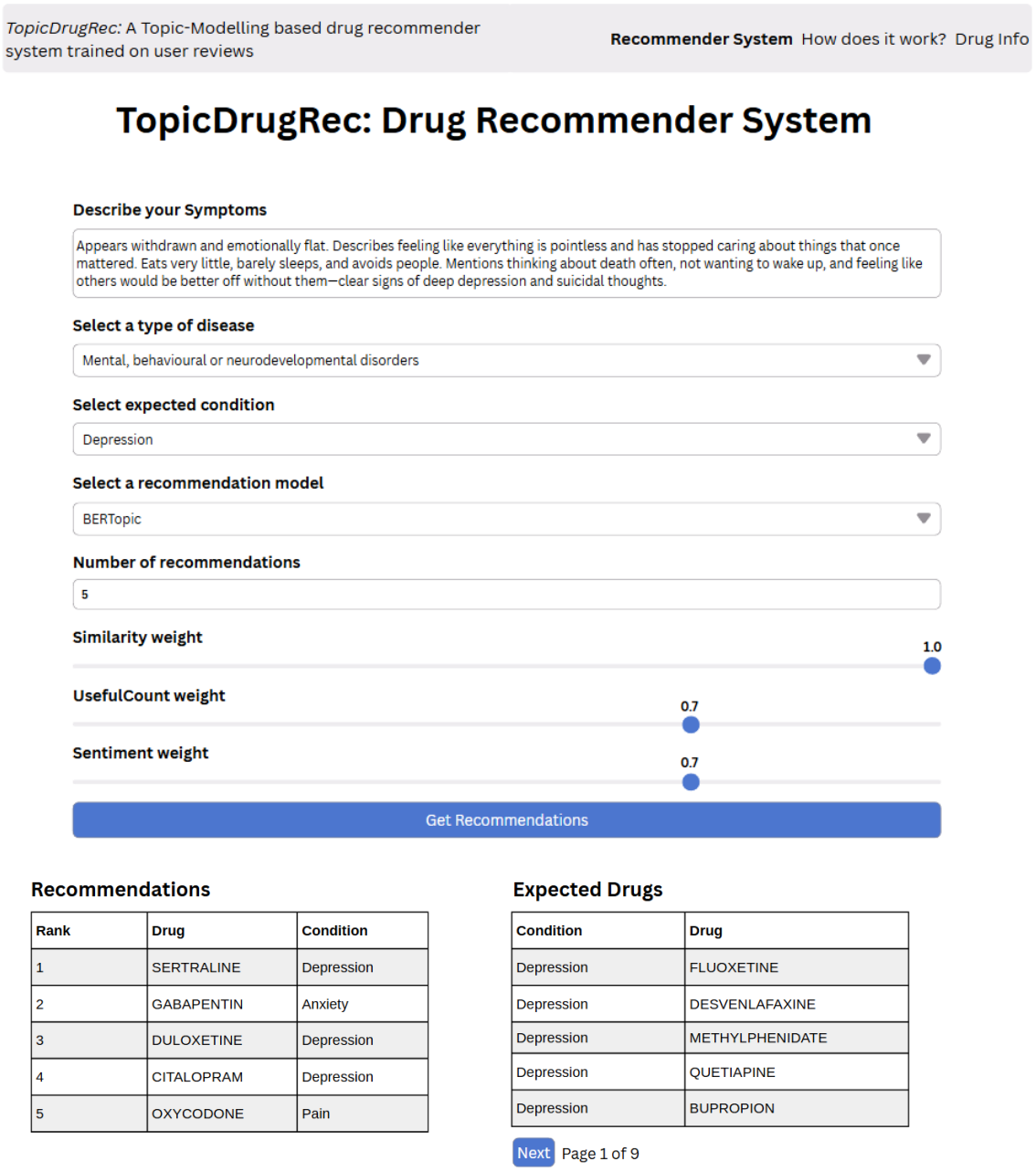


Figure 5.21: *Example of use of the TopicDrugRec web application showing a possible use case. The first text box displays how clinicians can enter free-text symptoms, select the expected conditions according to their knowledge, adjust the recommender model configurations and choose the list size. The second panel shows the resulting top-5 drug recommendations alongside the list of drugs known to treat the expected condition, allowing for direct comparison.*

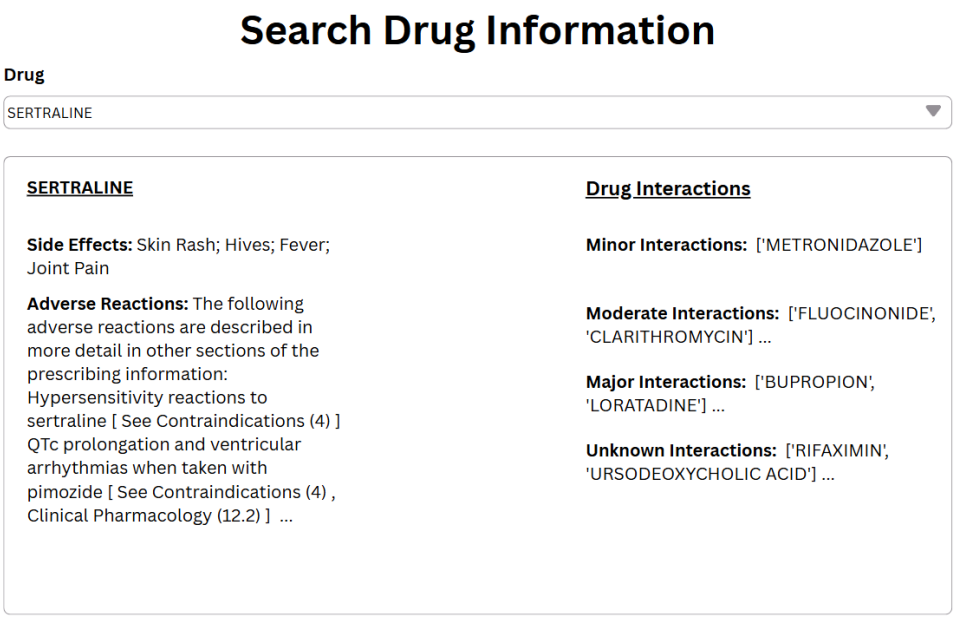


Figure 5.22: *The drug information section in TopicDrugRec is used for clinicians to look for known side effects, contraindications and known DDIs for the recommended treatments.*

The evaluation is structured around four steps. First, an ablation test investigates the contribution of the three key components that compose the scoring function to the performance of recommendations. **This test focuses on the highest performing topic model configurations identified previously**. To further explore the interaction between components, a complementary random parameter search was conducted using a fixed set of 20 configurations. The random search approach was chosen as grid-search is computationally costly, and each iteration of tests required approximately two hours.

Next, TopicDrugRec is tested at two levels of granularity, ICD11 chapter groupings and individual condition labels, to assess which structure yields better performance. Furthermore, the third analysis explores whether drug recommendation performance can be improved by employing an ensemble approach, in which the outputs of the different recommendation models are aggregated into a final recommendation list, aiming to capture the diversity in recommendations from all models to produce a more relevant list of suggested drugs. Finally, the impact of varying the number of recommendations is examined, highlighting the trade-offs between offering a wider range of options and maintaining the precision in the drug recommendations, while also drawing a conclusion for the best performing model and number of recommendations suited for mitigating information overload on clinicians.

5.8.1 Impact of scoring weights: Ablation and Random Parameter Search

In Machine Learning (ML), an ablation test is the removal of an individual component of a system, aiming to assess the relative contribution of individual components to its performance [82]. In the context of TopicDrugRec, this was achieved by setting all weights to zero except one, isolating the effect of `similarity`, `usefulCount`, or user `sentiment`.

In parallel, a random parameter search was conducted to explore a broader range of weight combinations. Each of the three parameters was sampled from a continuous uniform

Table 5.14: Recommendation performance of the LDA model when activating only one scoring hyperparameter at a time. The results show that considering just topic similarity leads to better outcomes across all metrics compared to using only `usefulCount` or sentiment.

W_{ts}	W_s	W_u	Prec@10	Rec@10	F1-Score	MAP@10	MAR@10	MRR@10
1.0	0.0	0.0	**0.455**	**0.102**	**0.142**	**0.345**	**0.176**	**0.622**
0.0	1.0	0.0	0.376	0.062	0.091	0.256	0.129	0.533
0.0	0.0	1.0	0.202	0.022	0.038	0.113	0.064	0.246

Table 5.15: Performance metrics for the NMF model in the ablation test, showing that performance is maximised when relying just on topic similarity. By considering user sentiment and `usefulCount` the performance of recommendations drops significantly across all metrics.

W_{ts}	W_s	W_u	Prec@10	Rec@10	F1-Score	MAP@10	MAR@10	MRR@10
1.0	0.0	0.0	**0.513**	**0.115**	**0.161**	**0.409**	**0.210**	**0.676**
0.0	1.0	0.0	0.425	0.081	0.116	0.282	0.154	0.526
0.0	0.0	1.0	0.331	0.041	0.068	0.21	0.127	0.309

distribution over the interval $[0, 1]$. More than twenty distinct combinations were evaluated per topic model (LDA, NMF, and BERTopic), using $k = 10$ recommendations per evaluation on an ICD11 granularity level.

The results from the ablation tests (Tables 5.14, 5.15, and 5.16) show the individual contribution of each scoring component in isolation. Across all three models the configuration with full weight on semantic similarity ($W_{ts} = 1.0$) consistently outperformed the alternatives across all evaluation metrics.

This confirms that the similarity component is the most contributive step of the TopicDrugRec RS, achieving the highest values in terms of Precision@10 (0.455, 0.513, 0.507), Recall@10 (0.102, 0.115, 0.116), F1-Score (0.142, 0.161, 0.161), MAP@10 (0.345, 0.409, 0.398), MAR@10 (0.176, 0.210, 0.203), and MRR@10 (0.622, 0.676, 0.676) for LDA, NMF, and BERTopic respectively.

In contrast, when only `usefulCount` ($W_u = 1.0$) or sentiment ($W_s = 1.0$) was active, performance dropped significantly across all metrics. By emphasizing perceived usefulness over topic similarity, TopicDrugRec is emphasizing how relatable the review was to other patients, rewarding popularity rather than clinical relevance. Similarly, the sentiment towards a drug may indicate satisfaction for one condition, and dissatisfaction for another, which inherently introduces noise, and by increasing W_s, the same noise is also amplified, removing relevant treatments from the list of suggestions.

For example, in LDA, the F1-Score decreased from 0.142 to 0.091 (`usefulCount` only) and 0.038 (`sentiment` only), while MRR@10 fell from 0.622 to 0.533 and 0.246 respectively. These trends are also reflected in NMF where Precision@10 fell from 0.513 to 0.331 when using only sentiment, and MAP@10 dropped from 0.409 to 0.210. Likewise, for BERTopic, the use of only `sentiment` yielded an F1-Score of 0.043 and an MRR@10 of 0.315.

Table 5.16: Recommendation performance of BERTopic recommender model during the ablation analysis. This analysis shows that like the other models, topic similarity is the dominant hyperparameter in improving recommendation performance.

W_{ts}	W_s	W_u	Prec@10	Rec@10	F1-Score	MAP@10	MAR@10	MRR@10
1.0	0.0	0.0	**0.507**	**0.116**	**0.161**	**0.398**	**0.203**	**0.676**
0.0	1.0	0.0	0.414	0.06	0.092	0.306	0.157	0.551
0.0	0.0	1.0	0.215	0.025	0.043	0.135	0.061	0.315

Table 5.17: Top 3 LDA weightings without ablations. Considering high usefulness ($W_u = 0.9$) and moderate topic similarity ($W_{ts} = 0.5$) improves precision and most rank based metrics, matching the other weighting schemes recall and only falling short in MAP.

W_{ts}	W_s	W_u	Prec@10	Rec@10	F1-Score	MAP@10	MAR@10	MRR@10
0.5	0.1	0.9	**0.397**	0.065	**0.096**	0.279	**0.145**	**0.552**
0.2	0.9	0.8	0.395	0.065	0.092	**0.280**	0.143	0.543
0.2	0.9	0.7	0.393	0.065	0.095	0.273	0.143	0.535

Although the ablation tests highlight the dominant role of semantic similarity when used in isolation, they do not consider potential interactions between the scoring components when combined. As such, the random search results (Tables 5.17, 5.18, and 5.19) offer further insight into how these components influence performance when used together.

While configurations with higher weights on similarity generally performed well, small contributions from `sentiment` or `usefulCount` occasionally led to marginal gains. The same configuration of hyperparameters ($W_{ts} = 0.5$, $W_s = 0.1$, $W_u = 0.9$) emerged as the best combination among the random search results across all three models, yet all still falling below the best ablation result.

In the case of NMF, this configuration achieved the highest F1-Score (0.123), along with the best results in Precision@10 (0.473), MAP@10 (0.357), MAR@10 (0.182), and MRR@10 (0.648). Although it did not surpass the ablation performance obtained through a purely CB approach, the relatively small average drop of approximately 3% across all metrics suggests that NMF may be receptive to a hybrid weighting strategy, where incorporating perceived patient usefulness can still possibly deliver comparable results. Similarly, for BERTopic, the same configuration also performed well, achieving an F1-Score of 0.116, Precision@10 of 0.450, MAP@10 of 0.340, Mean Average Recall (MAR)@10 of 0.171, and Mean Reciprocal Rank (MRR)@10 of 0.649. On the other hand, LDA was the most affected recommender model, as even the best hybrid weighting configuration significantly dropped its performance compared to similarity-only, indicating that pushing `usefulCount` or user sentiment does not compensate the performance of the Topic Modelling (TM) component.

These hybrid weighing observations further reinforce that topic similarity is the main decisive factor in providing accurate treatments. This conclusion is shown in Table 5.20 which shows that all three models yielded their best performance under the same weight configuration ($W_{ts} = 1.0$, $W_s = 0.0$, $W_u = 0.0$), which will be adopted for the following evaluation steps.

Table 5.18: Top 3 random search configurations for the NMF model, excluding ablation cases. The same pattern from LDA emerges, with high usefulness and moderate topic similarity achieving best results. Notably, increasing the weight on user sentiment to 0.3 or 0.5 degraded recommendation performance across all metrics.

W_{ts}	W_s	W_u	Prec@10	Rec@10	F1-Score	MAP@10	MAR@10	MRR@10
0.5	0.1	0.9	**0.473**	**0.084**	**0.123**	**0.357**	**0.182**	**0.648**
0.7	0.3	0.8	0.468	0.081	0.120	0.347	0.178	0.630
0.9	0.5	0.7	0.457	0.081	0.120	0.334	0.170	0.624

Table 5.19: Top 3 random search configurations for the BERTopic model, excluding ablation cases. The best results are achieved with the same weight configuration as the other models. As user sentiment is given more emphasis, the performance degrades, aligning with previous results.

W_{ts}	W_s	W_u	Prec@10	Rec@10	F1-Score	MAP@10	MAR@10	MRR@10
0.5	0.1	0.9	**0.450**	**0.080**	**0.116**	**0.340**	**0.171**	**0.649**
0.7	0.3	0.8	0.434	0.076	0.111	0.315	0.160	0.638
0.8	0.5	1.0	0.431	0.076	0.110	0.307	0.156	0.627

Table 5.20: Best scoring configuration per model using ICD-11 chapter-level recommendation tasks. All three models achieved highest performance under a pure content-based configuration by disconnecting user sentiment and perceived usefulness from the weighting scheme ($W_{ts} = 1.0, W_s = 0.0, W_u = 0.0$). Among them, NMF yielded best performance, falling short in Recall to BERTopic by a minimal margin.

Model	W_{ts}	W_s	W_u	Prec@10	Rec@10	F1-Score	MAP@10	MAR@10	MRR@10
LDA	1.0	0.0	0.0	0.455	0.102	0.142	0.345	0.176	0.622
NMF	1.0	0.0	0.0	**0.513**	0.115	**0.161**	**0.409**	**0.210**	**0.676**
BERTopic	1.0	0.0	0.0	0.507	**0.116**	**0.161**	0.398	0.203	**0.676**

A particularly notable aspect is that both NMF and BERTopic achieved the MRR@10 score of 0.676. This indicates that, on average, the first relevant drug appears between the first and second position in the list of top-10 recommendations.

Focusing on the best results of NMF, while the first relevant item typically appears near the top of the list, the Precision@10 score of 0.513 indicates that, on average, about five out of the top ten recommendations are relevant. The relatively low Recall@10 of 0.115 should not be interpreted as poor performance, since the total number of relevant drugs per test case often exceeds the evaluation window of $k = 10$. In the drug recommendation context, limiting the number of recommendations is intentional, as it is preferable to provide fewer but highly accurate suggestions rather than a longer list that may include less relevant drugs. Furthermore, the Mean Average Precision@10 of 0.409 suggests that while NMF retrieves relevant drugs effectively, it does not always succeed in consistently ranking all of them near the top when multiple relevant items are present.

5.8.2 ICD11 vs Singular Condition

Considering the best performing weight configuration identified previously, this evaluation step explores the impact of disease granularity on recommendation performances. More specifically, it compares the results when the input to TopicDrugRec is expressed on a broader ICD11 level versus a narrower defined singular condition, for example, "Mental, behavioural or neurodevelopmental disorders" versus "Major depressive disorder".

The underlying hypothesis is that broader ICD11 groups may yield better recommendation performance due to higher drug coverage compared to singular conditions, whereas individual conditions promote more clinical precision. As such, the goal is to assess the trade-off between generalisation and specificity.

Table 5.21 presents the performance of each model when using condition-level input, offering a more fine-grained evaluation compared to the broader ICD11 groupings discussed earlier in Table 5.20. As expected, using singular conditions improved recall across all models, indicating that the smaller drug subset allows TopicDrugRec to match a greater number of relevant drugs to the input user symptoms.

However, this increase in recall is accompanied by several trade-offs as all models exhibit lower MRR@10 compared to the ICD11 counterparts, meaning that while more relevant items are retrieved, these are ranked lower in the list. Additionally, the precision metrics (Precision@10 and MAP@10) also decline, suggesting less effectiveness in identifying and ranking relevant drugs within the list of recommendations.

While NMF outperformed the other models at the broader ICD11 level, BERTopic emerged as the strongest performer in the condition-level evaluation. It achieved the highest scores across all metrics except Mean Average Recall@10, where it was narrowly outperformed by NMF. This performance shift can be attributed to BERTopic's use of contextualised embeddings, which are better suited to capturing fine-grained semantic nuances present in condition-specific inputs. These embeddings allow the model to more accurately interpret user-described symptoms and map them to relevant treatments, ultimately improving recommendation quality in scenarios requiring greater specificity.

All in all, the shift to condition-level inputs results in a general decline across most performance metrics, revealing a reduced ability not only to identify relevant drugs but also to rank them effectively. For instance, BERTopic's Precision@10 dropped from 0.507 (ICD11 level) to 0.343 (condition level). At the same time, Recall@10 increased significantly from 0.116 to 0.329, meaning that the model is able to retrieve a greater proportion of relevant drugs on a higher level of granularity, however, this is explained by the reduced number of total candidate drugs associated with a specific condition, contrary to the broader set found in ICD11 groupings. This increase in recall led to the corresponding improvement in F1-Score, however, the aim of a drug RS is to accurately suggest drugs for a specific condition, rather to identify as many relevant ones as possible.

These results demonstrate that despite the improved recall at the condition level, **the ICD11 based approach yields better performance**, particularly in both the precision

Table 5.21: Performance metrics using condition-level inputs across all models.

Model	Precision@10	Recall@10	F1-Score	MAP@10	MAR@10	MRR@10
LDA	0.274	0.258	0.227	0.234	0.170	0.427
NMF	0.331	0.321	0.277	0.309	**0.238**	0.495
BERTopic	**0.343**	**0.329**	**0.286**	**0.324**	0.233	**0.515**

metrics and ranking quality, which are essential in providing clinically relevant drug recommendations. As such, for the remainder of the evaluation steps, the model comparison and impact of the number of recommendations will be conducted using ICD11 inputs.

5.8.3 Ensemble Model: Intersection of Recommendations

Based on the results seen in Table 5.20, which established NMF as the best performing model on ICD11-level inputs, this test explores the potential of an ensemble approach in further improving recommendation quality by considering the agreement of each individual recommendation models.

As such, the proposed ensemble model follows a two step approach as depicted in Figure 5.23. First, it considers the intersection of the top-K recommendations generated by each independent model for a given symptom description in the test set, where recommendations that are present in all three lists are considered of higher relevance and kept. The ordering of the recommendations is kept in accordance to the output of NMF due to its performance on the ranking metrics, especially in MRR.

Then, in the second step, it ensures the list reaches the target size of K recommendations, by backfilling the remaining slots sequentially, keeping the same hierarchy of performance, starting by the recommendations from NMF, then BERTopic, and finally LDA if necessary.

Table 5.22 shows the metrics for the two step ensemble model. To judge its effectiveness, these values are compared with the strongest individual model that was previously identified (NMF) in Table 5.20.

At $K = 10$ the list is almost made up of drugs on which all three models agreed, with precision falling few precision points compared to the individual NMF model (from 0.513 to 0.468). However, on the contrary, the recall decreased drastically, indicating that most relevant drugs appear only in either one or two recommender models, and end up being excluded from the list of recommendations. This decline in recall can be translated into a real life medical scenario, where a physician will be presented with a set of very conservative options (drugs which all three models recommended), at the cost of missing a lot of relevant alternatives.

When increasing the list to $K = 20$ and $K = 30$, the effects of the backfilling from NMF become noticeable, as recall reaches 0.098 for $K = 30$, while precision suffers a marginal decline, going from 0.468@10 to 0.454@30. This increase in recall means that the ensemble model is now recapturing the list of previously excluded drugs on which the models disagreed. At the same time, the ranking metrics never catch up to the results of NMF.

As such, while the metrics suggest that the ensemble model **is not the best choice** when

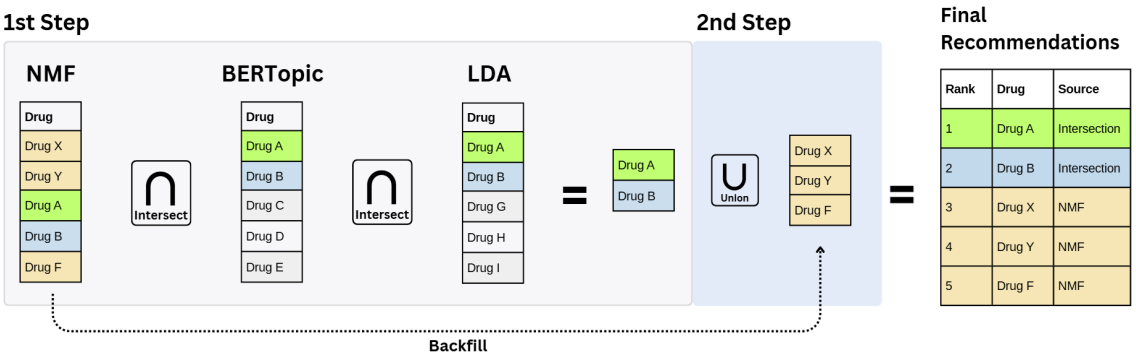


Figure 5.23: *Workflow of the two-step TopicDrugRec Ensemble Model. In the first step an intersection of the top-K recommendations is generated. In the second step, remaining slots are filled sequentially using NMF recommendations, followed by BERTopic's and LDA's.*

Table 5.22: Performance metrics of the Ensemble model for K = 10, 20 and 30 recommendations. The results highlight the impact of increasing K in Recall. Precision decays marginally as K increases, with Recall increasing due to the backfilling.

K	Precision	Recall	F1-Score	MAP	MAR	MRR
10	**0.468**	0.035	0.062	**0.382**	**0.211**	0.629
20	0.463	0.067	0.110	0.356	0.177	0.665
30	0.454	**0.098**	**0.148**	0.342	0.161	**0.69**

the objective is to maximise precision and recall, it comes with an underlying aspect which is the level of agreement. By ensuring that the top of the suggested drugs list is made up of unanimous agreement across all recommender models, the ensemble model can be seen as a cross-validation of drug suggestions. Translating to a real world scenario, this can be seen as getting validation from different clinicians before applying a treatment to a patient.

5.8.4 Impact of the Number of Recommendations

The final recommendation evaluation aims to assess how varying the number of recommendations influences the performance of TopicDrugRec particularly in terms of ranking and precision based metrics. More specifically, the underlying hypothesis that is addressed is that by expanding the recommendation list in three levels ($K = 10$, $K = 20$, and $K = 30$), an increase in recall is expected due to potentially covering more relevant drugs, however, it is expected that this comes at the expense of correctly identifying relevant treatments, as well as ranking them.

To investigate this trade-off, this analysis is built upon previous results, focusing on ICD11-level inputs, with the previously discussed optimal scoring configuration ($W_{ts} = 1.0$, $W_s = 0.0$, and $W_u = 0.0$), along with the best performing topic model configurations in terms of drug recommendation and is presented in separate tables, one for each model, with the highest performing metrics highlighted in bold.

Table 5.23 represents the results from varying K in the LDA model and how they reflect in the metrics. Starting from the shortest list of recommendations ($K = 10$), this list delivers the best precision at 0.455, meaning that almost half the drugs in the list are truly relevant, however, with a recall of 0.102 this indicates that only a tenth of clinically relevant drugs

Table 5.23: LDA performance metrics for $K = 10, 20$ and 30 recommendations. The results show that lower K performs best for Precision. For Recall and Ranking based metrics the results are best for $K = 30$, with the biggest over performance in Recall and F1-Score.

K	Precision	Recall	F1-Score	MAP	MAR	MRR
10	**0.455**	0.102	0.142	0.345	0.176	0.622
20	0.437	0.189	0.212	0.335	0.179	**0.626**
30	0.422	**0.263**	**0.255**	**0.347**	**0.217**	**0.626**

Table 5.24: NMF performance metrics for $K = 10, 20$ and 30 recommendations. The results validate that lower K achieves best Precision and MAP. Recall increases substantially from $K = 10$ to $K = 30$, and despite lower performance, it achieves the best harmonized F1-Score.

K	Precision	Recall	F1-Score	MAP	MAR	MRR
10	**0.513**	0.115	0.161	**0.409**	0.210	0.676
20	0.486	0.210	0.238	0.394	0.209	0.679
30	0.464	**0.289**	**0.284**	0.401	**0.245**	**0.68**

are captured. With a MAP of 0.345, this indicates that a third of the relevant items are positioned in higher rankings of the list, while an MRR of 0.622 implies that on average the first clinically relevant drugs appears around the 1st and 2nd position in the list.

By doubling the number of recommendations ($K = 20$), the Recall improves by 85%, achieving 0.189, because, as expected, the recommender is now considering lower-scoring recommendations that were pruned from the previous list of drugs. However, Precision decays to 0.437, indicating that **despite covering more relevant drugs, a bigger number of irrelevant ones was included within the list of recommendations**. This is particularly important, because despite F1-Score showing an increase of approximately 49% by shifting from 0.142 to 0.212, in patient healthcare, it is more important to provide clinically relevant drugs rather than covering more relevant ones. Similarly, in the ranking metrics, MAP lowers to 0.335, showing that despite more relevant items, these are now positioned lower in the list. Furthermore, MRR increases to 0.626, meaning that despite introducing more drugs to the list, the first relevant hit remains within the first 1 to 2 positions, which also uncovers that the new relevant drugs are positioned farther down the list rather than moving the first relevant drug downward.

Extending to $K = 30$, Recall increases again to 0.263, an increase of almost 250% compared to the baseline $K = 10$. However, the same trend in Precision is registered, with the value decreasing to 0.422, approximately -3% compared to $K = 20$. Interestingly, MAP bounces up to 0.347, marginally surpassing the $K = 10$ list, suggesting that some of the relevant drugs may sit on reviews to which topic similarity is lower, further reinforcing the broad nature of the text in the original dataset. Furthermore, MAR reaches 0.217, demonstrating that this configuration is now focusing on covering a wider list of relevant drugs at the expense of introducing more irrelevant ones, while MRR remains the same, further reinforcing that extending the recommendations just incorporates lower scoring hits at the tail of the list.

Table 5.25: BERTopic performance metrics for $K = 10, 20$ and 30 recommendations. For Precision based metrics, $K = 10$ achieves the best performance, with a marginal decay as K increases. For Recall based metrics, the contrary is seen, with the best performance being for higher values of K.

K	Precision	Recall	F1-Score	MAP	MAR	MRR
10	**0.507**	0.116	0.161	**0.398**	0.203	0.676
20	0.480	0.211	0.237	0.381	0.201	0.679
30	0.459	**0.292**	**0.283**	0.389	**0.238**	**0.680**

The NMF results, displayed in Table 5.24 display a similar trend, with Precision decaying and Recall increasing as the number of recommendations. Contrary to LDA, the MAP peaks at $K = 10$ with a value of 0.409, compared to the second best value of 0.401 for $K = 30$. One notable aspect is that the ten recommendations from NMF hits the highest precision in the entire study (0.513) and the best MAP (0.409).

By expanding the list to 20 recommendations, the recall almost doubles, with precision decaying from 0.513 to 0.486, and by 30 drug recommendations the recall reaches its peak value of 0.289 without degrading early-ranking quality by presenting a MAP of 0.401 compared to the 0.409 of the baseline ($K = 10$). Similar to LDA MAR increases with the expansion of the list of recommended drugs, while MRR plateaus at approximately 0.680.

Similar to the previous models, BERTopic displays the same patterns, represented in Table 5.25. Precision at $K = 10$ (0.507) falls shortly to the best performing in the study from NMF, however, at $K = 30$, recall comes up ahead, achieving a value of 0.292 while the MRR remains unchanged at 0.68.

In conclusion, all three topic model based recommender systems confirm the previously stated precision-recall hypothesis, as the number of recommendations increases the recall also increases, with precision taking a hit. LDA falls short on all metrics, whereas NMF and BERTopic present very close metrics on all extent of K recommendations. However, **the ten drug list of NMF delivers the highest precision and the best ranking performance**. Considering the premise of information overload on clinicians, and the aim of this dissertation of proposing a drug RS to enhance patient treatment by compacting a list of drugs given a symptom description, **the 10 item NMF-based recommender system emerges as the best choice.**

6 Final Considerations

6.1 Conclusions

Pursuing the rise in data collection and the increase in medical decision-making oriented to patient profiles, alongside long and highly demanding working shifts and intense working environments, healthcare professionals struggle to make data-driven decisions due to its immense volume.

TopicDrugRec was proposed in this dissertation as a **drug RS** built on Patient-Reported Outcomes (PRO) of previously administered drugs in the form of **textual reviews** from the UCI ML Drug Review dataset. The system aims to reduce information overload on clinicians and promote more direct and personalised patient interaction.

In addition to textual information, the RS incorporates quantitative features from the same dataset, such as **user sentiment** and **perceived usefulness**, to explore how different ML techniques can address the problem of **extreme response bias**.

Given the importance of patient safety within the medical context, a safety-oriented approach was also adopted by integrating external biomedical knowledge, including **side effects**, **drug-drug interactions**, and **contraindications**.

The proposed solution follows a six-step methodology. It begins with an Exploratory Data Analysis (EDA) to understand the dataset, its features and what they represent, followed by data cleaning to prepare it for downstream tasks. These include Sentiment Analysis (SA) to correct sentiment bias and Topic Modelling (TM) to extract latent themes from user experiences, forming the basis of the recommendation algorithm. The fifth step enhances patient safety by integrating external knowledge, while the sixth consists of a web application for using, testing, and evaluating the system's performance.

The first four steps (EDA, Data Cleaning, SA, and TM) provided a structured approach to implementing TopicDrugRec, enabling the isolation of influential factors in generating meaningful drug recommendations, which can be tuned by the end-user.

Throughout the different stages of the proposed methodology, insights beyond the primary objective of recommending suitable drugs were uncovered.

For instance, during the EDA, we found that the dataset is highly diverse in terms of both drugs and conditions, while also exhibiting high extreme response bias, where a large scale of reviews were clustered at the extremes of the numerical scale. Additionally, from the 889 conditions represented, the top 10 alone accounted for 44.5% of the entries, highlighting the need for aggregating them in order to reduce the lack of representation and imbalance. As such, the medical conditions were grouped into categories using ICD11 codes, leading to a high increase in coverage, where the top 10 ICD11 groups represented 83% of the dataset.

In the sentiment analysis stage, two different approaches were evaluated, comparing lexicon and embedding-based models. The lexicon-based models, VADER and TextBlob revealed disagreements between the sentiment predictions and original ratings, both showing a tendency to positive bias, which was more pronounced in TextBlob. In contrast, the embedding-based models like `twitter-roBERTa` offered a more negative view of the reviews, frequently categorising negative sentiment with higher confidence. Applying these models to the original sentiment labels led to a shift in the distribution of reviews, where positive classified decreased by approximately 11.2%, neutrals by 37%, while negative reviews increased 43.2%. This outcome demonstrated how pre-trained embedding-based models can be leveraged as unsupervised SA tools to offer a different perspective in the interpretation of user generated content beyond unknown personal context.
These findings answer the first research question, demonstrating that ML libraries not only are able to mitigate extreme response bias, but also enable different perspectives into uncategorised sentiment in user reviews.

The Topic Modelling stage evaluated three algorithms that lay foundation for the TopicDrugRec Recommender System, LDA, NMF and BERTopic. The impact of several Natural Language Processing techniques, such as different combinations of n-grams (1-gram and 2-gram) and stemming were studied in relation to topic coherence and interpretability. LDA performed best on a lower number of topics (10) and simple unigrams, without any stemming. Similarly, NMF yielded better results in a configuration with less topics and using bigrams (2-gram), reinforcing that more topics does not necessarily lead to clearer topic separation. In contrast, BERTopic's highlights appeared on the ease of configurations, where its automatic parameter tuning enabled for faster experimentation, making it suitable for situations where quick iteration and testing are necessary.
This TM stage targetted the second research question, providing a foundational baseline that enabled the recommendation algorithm from TopicDrugRec, demonstrating a way of leveraging PROs to enhance/accelerate drug recommendations and medical treatment.

In the final stages of the implementation, the performance of TopicDrugRec was evaluated in its ability to generate relevant drug recommendations. The experiments were conducted to understand how different configurations impacted the recommendation quality, using metrics such as Precision, Recall, MAP, MAR and MRR.

The first assessment focused on an **ablation test** using random search of parameters to determine the optimal contribution of each feature (similarity, sentiment and usefulness) to the quality of the generated recommendations. The results indicated that relying solely on

semantic similarity ($W_{ts} = 1.0$, $W_s = 0.0$, $W_u = 0.0$) yielded the most impact in all models, with NMF standing out with a Precision@10 of 0.513 and MRR@10 of 0.676. These results suggest that, on average, at least half of the top 10 recommended drugs were relevant, and the first correct recommendations appears within the first or second position in the list.

The second experiment analysed how the ICD11 grouped conditions versus singular input conditions impacted the performance of the recommendations, aiming to understand whether increasing the dataset coverage led also translated to better results. The results showed that ICD11 categories significantly improved recommendation performance, particularly within the NMF model, where it reached Precision@10 = 0.513, Recall@10 = 0.115, MAP@10 = 0.409, MAR@10 = 0.210 and MRR@10 = 0.676. In contrast, relying on the ungrouped conditions as input for the same model configuration led to worse performance, for instance, Precision@10 dropped down to 0.331, and MRR@10 to 0.495.

The results demonstrate that generalising singular conditions into broader categories not only improves dataset coverage and mitigates sparsity, but also enhances the overall quality of recommendations. In a real-world clinical scenario, this approach is similar to the diagnostic process, where an initial possible diagnostic is framed into a category of conditions. Using the ICD11 approach, TopicDrugRec generates a list of supported treatment options for that group of conditions, and, from there, the clinician can refine the selection by narrowing them down based on nuanced symptoms descriptions or contextual factors not explicitly captured by the Recommender System.

Thirdly, the **ensemble model** was explored by intersecting the top recommendations from multiple configurations, aiming to increase consistency and reliability across the models. While this approach enables for a better agreement between recommendations, as it integrates only drugs which were agreed by the three models, as well as the remainder of the top-performing one, it came at the cost of reduced coverage, achieving a much lower Recall@10, of 0.035, and a lower Precision@10 of 0.468. This means that the strict agreement between all three models may lead to excluding potentially relevant suggestions, limiting its overall effectiveness.

However, this approach can be linked to real-world scenarios in which multiple professionals independently assess the same patient and provide their treatment suggestions. When a consensus is reached, it often leads to a perceived increase in confidence and reliability of the treatment. In this sense, the ensemble model may be suitable in scenarios where agreement are priority, and as such, the trade-off between recall and the confidence may be acceptable in some clinical contexts.

Finally, the last test assessed the **impact of the number of recommendations** by analysing different values of K (10, 20 and 30) and how these affect the quality of the generated recommendations. The results demonstrated that Precision was highest for lower levels of K, indicating a more focused list of relevant treatments, however with limited coverage as seen by the lower Recall values. In contrast, increasing K resulted in a substantial improvement in Recall-based metrics and a tangible gain in MRR, suggesting that more relevant drugs were captured overall, at the cost of introducing potentially

irrelevant recommendations.

These findings highlight the trade-off between precision and recall that should be considered when using TopicDrugRec in real world situations. Lower levels of K are better suited for scenarios where a concise and highly accurate list of suggestions is preferred. On the other hand, higher values of K may be more appropriate in exploratory use cases, where the goal is to present a broader range of possibly relevant treatments.

Although these values may appear modest, it is important to emphasise that they reflect the potential of Recommender System to support the medication prescription process. Clinical validation by qualified professionals remains essential, however, these algorithms can reduce their cognitive load. These results also highlight the potential of recommendation algorithms in the healthcare sector, offering insights into how these tools can be used in data-driven medical decisions.

Ultimately, this dissertation demonstrates the possibility of implementing **TopicDrugRec** as a baseline tool for drug recommendation in healthcare. By leveraging user generated content in the form of Patient-Reported Outcomes, addressing sentiment biases and integrating external biomedical knowledge, it lays groundwork for a RS that could reduce information overload for clinicians. In particular, **this integration of external knowledge contributes to the third research question**, by showing how the biomedical knowledge can be integrated into the **TopicDrugRec** Web Application. Beyond offering drug recommendations, it provides a functionality where clinicians can consult detailed information on side effects, adverse reactions and potential drug-drug interactions.

While the results highlight its potential usefulness, TopicDrugRec should be considered as a proof-of-concept that requires refinement and user validation. This assessment can only be done through test sessions with healthcare professionals, ensuring that it is tailored to their needs and possible to integrate in day-to-day work.

6.2 Limitations

While the experimental results of the different stages of TopicDrugRec demonstrated its potential as a support tool in drug prescription, there are known limitations that must be considered when interpreting these findings. These limitations have influenced TopicDrugRec's capabilities but also provide direction for future work.

The UCI ML Drug Review dataset, although publicly available and rich both textual content and numerical features, it lacks many attributes that could have enhanced the personalisation of recommendations. Notably, it does not include demographic information such as age, sex, or ethnicity, nor does it provide details on past treatments, dosages or drug experiences. In real world scenarios these characteristics are fundamental in determining the suitability of a treatment. For example, the same drug may lead to different side effects or efficacy depending on the patient's demographics. Without these, TopicDrugRec cannot capture these patterns that influence both the sentiment expressed in the reviews and the relevance of recommendations.

Moreover, the dataset also exhibits a significant class imbalance, with the top 10 most represented conditions accounting for less than half of all entries. This imbalance was particularly highlighted in the ICD11 grouping experiment, where generalising conditions into broader disease categories substantially improved coverage and resulted in more relevant recommendations. This improvement underscored how sparsity in the dataset hinders the performance of the generated recommendations.

Additionally, the response bias in numerical ratings complicates the alignment between sentiment and rating. This is due to the inherent difficulty of classifying text as Positive, Negative or Neutral without additional contextual information. Factors outside of the dataset's scope such as user expectation, treatment history or tolerance to side effects can lead to reviews with similar wording receiving vastly different ratings.

Furthermore, TopicDrugRec currently remains as a proof-of-concept and has not been validated in a real world clinical workflow. All evaluations were performed offline, using standard train-test splits, without feedback or validation from clinicians or patients.

The integration of external biomedical knowledge, despite being a topic of major importance in this dissertation, was constrained by partial coverage. Since the external datasets did not map directly to the UCI ML Drug Review vocabulary, contraindications and side effects only covered approximately 35% of the drugs present in the dataset. This incomplete mapping limits the system's ability to flag all potentially harmful recommendations. Moreover, these biomedical knowledges sources are static and do not automatically incorporate the latest research findings, which means they may become outdated if not periodically updated.

6.3 Future Work

Building on the findings and limitations identified in this dissertation, several avenues for future work can be explored to improve both the accuracy and applicability of TopicDrugRec, as well as to expand its scope within the healthcare domain.

- **Development of a Curated Dataset**: One significant limitation found in this dissertation is the restricted nature of the UCI ML Drug Review dataset. Considering a curated dataset that combines the textual reviews alongside structured attributes, such as patient demographics, treatment timelines, dosage information and past drug experiences would enable a deeper understanding on how each factor influences the treatment perception and outcomes. By considering an historical dataset, with longitudinal follow-up data on the treatment of the patients TopicDrugRec would be able to track these changes in sentiment or the reporting of side effects over time.
- **Dynamic Integration of Biomedical Knowledge**: The integration of external knowledge in TopicDrugRec aims to address patient safety when recommending new drug prescriptions. However, due to its static nature and partial coverage it may affect long-term effectiveness. Future work could incorporate automated mechanisms to fetch and integrate updates on the biomedical knowledge, ensuring drug-drug interactions, contraindications and side effects remain up to date.

- **Addressing the Impact of Side Effects on Drug Ratings**: Identifying which side effects tend to cause significant drops in ratings, and which are frequently disregarded may help differentiate between clinically tolerable and intolerable adverse reactions. This would also allow TopicDrugRec to adjust the ranking and relevance of recommendations by adapting the scoring mechanism to better reflect patient tolerance.

- **ICD11 Focused Recommendation Models**: The ICD11 grouping demonstrated benefits in mitigating sparsity and improving coverage and recommendation quality. Building on these improvements, future work could explore a dedicated Recommender System trained for each ICD11 condition group, rather than a single global model as demonstrated in this dissertation.

- **Real-World Clinical Validation**: To ensure that TopicDrugRec is relevant in real-world scenarios, this Recommender System should be tested in real clinical environments. This would involve deploying the web-application and collaborating with healthcare professionals, integrating it into their decision process and collecting structured feedback in usability, trust, and impact on the drug prescription quality.

- **Improved Explainability**: The ability to justify a recommendation is essential for building trust between the healthcare professional and the patient. TopicDrugRec does not currently provide explainability, however, future work could focus on implementing explanations for why a specific set of drugs was recommended and how their ranking was determined. This could be achieved through methods such as highlighting the most relevant review excerpts, summarising sentiment patterns, and detailing any safety considerations applied.

Bibliography

[1] A. Abdelrazek, Y. Eid, E. Gawish, W. Medhat, and A. Hassan. "Topic modeling algorithms and applications: A survey". In: *Information Systems* 112 (Feb. 2023), p. 102131. ISSN: 0306-4379. DOI: 10.1016/j.is.2022.102131 (cit. on p. 8).

[2] C. C. Aggarwal. *Recommender Systems*. Springer International Publishing, 2016. ISBN: 9783319296593. DOI: 10.1007/978-3-319-29659-3 (cit. on p. 7).

[3] J. Anievarghese. *Drugs, Side Effects and Medical Condition*. https://www.kaggle.com/datasets/jithinanievarghese/drugs-side-effects-and-medical-condition. Accessed: 2025-05-06. 2023 (cit. on pp. 37, 68).

[4] *Apache Airflow*. https://airflow.apache.org/. Accessed: 2025-05-13. Apache Software Foundation, 2025 (cit. on p. 37).

[5] S. Bag, S. K. Kumar, and M. K. Tiwari. "An efficient recommendation generation using relevant Jaccard similarity". In: *Information Sciences* 483 (2019), pp. 53–64. ISSN: 0020-0255. DOI: https://doi.org/10.1016/j.ins.2019.01.023. URL: https://www.sciencedirect.com/science/article/pii/S0020025519300325 (cit. on p. 36).

[6] Y. Bao and X. Jiang. "An intelligent medicine recommender system framework". In: *2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA)*. IEEE, June 2016, pp. 1383–1388. DOI: 10.1109/iciea.2016.7603801 (cit. on pp. 23, 26).

[7] M. L. Bernauer. *mlbernauer/drugstandards: Python library for standardizing drug names*. 2017. DOI: 10.5281/ZENODO.571248 (cit. on p. 68).

[8] D. Bertram. "Likert scales". In: *Retrieved November* 2.10 (2007), pp. 1–10 (cit. on p. 7).

[9] Bird, Steven and Klein, Ewan and Loper, Edward. *Natural Language Toolkit*. Accessed: 2025-01-15. 2025. URL: https://www.nltk.org/ (cit. on pp. 24, 30, 34).

[10] M. Birjali, M. Kasri, and A. Beni-Hssane. "A comprehensive survey on sentiment analysis: Approaches, challenges and trends". In: *Knowledge-Based Systems* 226 (Aug. 2021), p. 107134. ISSN: 0950-7051. DOI: 10.1016/j.knosys.2021.107134 (cit. on p. 13).

[11] D. M. Blei, A. Y. Ng, and M. I. Jordan. "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022 (cit. on pp. 9, 10).

[12] R. Carvalho, M. Pato, and N. Datia. *A topic modelling-based recommender system for drugs using user experience reviews [TopicDrugRec]*. 2025 (cit. on p. 5).

[13] E. M. Chang, E. F. Gillespie, and N. Shaverdian. "Truthfulness in patient-reported outcomes: factors affecting patients' responses and impact on data quality". In: *Patient Related Outcome Measures* Volume 10 (June 2019), pp. 171–186. ISSN: 1179-271X. DOI: 10.2147/prom.s178344 (cit. on pp. 2, 41).

[14] H. Christian, M. P. Agus, and D. Suhartono. "Single Document Automatic Text Summarization using Term Frequency-Inverse Document Frequency (TF-IDF)". In: *ComTech: Computer, Mathematics and Engineering Applications* 7.4 (Dec. 2016), p. 285. ISSN: 2087-1244. DOI: 10.21512/comtech.v7i4.3746 (cit. on p. 12).

[15] C. Clinic. *Gout: Symptoms, Treatment & Prevention*. Accessed: 2025-04-05. 2023. URL: https://my.clevelandclinic.org/health/diseases/4755-gout (cit. on p. 46).

[16] C. Clinic. *Prednisone Tablets: Uses & Side Effects*. Accessed: 2025-04-05. 2023. URL: https://my.clevelandclinic.org/health/drugs/20469-prednisone-tablets (cit. on p. 46).

[17] *DDInter: Drug–Drug Interaction Database*. https://ddinter.scbdd.com/. Accessed: 2025-05-06. 2024 (cit. on pp. 37, 68, 69).

[18] S. learn Developers. *Scikit-learn: Machine Learning in Python*. Accessed: 2025-01-15. 2025. URL: https://scikit-learn.org/stable/ (cit. on p. 36).

[19] Drugs.com. *Prednisone: Uses, Dosage, Side Effects, Warnings*. Accessed: 2025-04-05. 2024. URL: https://www.drugs.com/prednisone.html (cit. on p. 46).

[20] *Drugs.com - Prescription Drug Information, Interactions & Side Effects*. https://www.drugs.com. Accessed: 2025-01-12 (cit. on pp. 28, 37, 68).

[21] L. D. Erik B. Erhardt Alan T. Arnholt et al. *8.3 Skewed Left Distributions | Passion Driven Statistics — statacumen.com*. https://statacumen.com/teach/S4R/PDS_book/skewed-left-distributions.html. [Accessed 03-04-2025] (cit. on p. 42).

[22] S. García, J. Luengo, F. Herrera, et al. *Data preprocessing in data mining*. Vol. 72. Springer, 2015. ISBN: 978-3-319-10246-7 (cit. on p. 49).

[23] A. E. Gelfand. "Gibbs Sampling". In: *Journal of the American Statistical Association* 95.452 (2000), pp. 1300–1304. DOI: 10.1080/01621459.2000.10474335 (cit. on p. 10).

[24] 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 (Dec. 1992), pp. 61–70. ISSN: 1557-7317. DOI: 10.1145/138859.138867 (cit. on p. 17).

[25] C. Goodrow. *On YouTube's Recommendation System*. Accessed: November 5, 2024. 2021. URL: https://blog.youtube/inside-youtube/on-youtubes-recommendation-system/ (cit. on p. 17).

[26] L. F. Granda Morales, P. Valdiviezo-Diaz, R. Reátegui, and L. Barba-Guaman. "Drug Recommendation System for Diabetes Using a Collaborative Filtering and Clustering Approach: Development and Performance Evaluation". In: *Journal of Medical Internet Research* 24.7 (July 2022), e37233. ISSN: 1438-8871. DOI: 10.2196/37233 (cit. on pp. 23, 26, 38).

[27] M. Grinberg. *Flask web development: developing web applications with python*. "O'Reilly Media, Inc.", 2018 (cit. on p. 39).

[28] M. Grootendorst. "BERTopic: Neural topic modeling with a class-based TF-IDF procedure". In: *arXiv preprint arXiv:2203.05794* (2022). arXiv: 2203.05794 [cs.CL]. URL: https://arxiv.org/abs/2203.05794 (cit. on pp. 12, 36).

[29] M. Grootendorst. *BERTopic API Documentation*. Accessed: 2025-02-05. n.d. URL: https://maartengr.github.io/BERTopic/api/bertopic.html (cit. on pp. 9, 12, 50, 63).

[30] M. Grootendorst. *Getting Started with BERTopic - c-TF-IDF*. Accessed: 2025-02-05. n.d. URL: https://maartengr.github.io/BERTopic/getting_started/ctfidf/ctfidf.html (cit. on p. 12).

[31] F. Gräßer, S. Kallumadi, H. Malberg, and S. Zaunseder. "Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning". In: *Proceedings of the 2018 International Conference on Digital Health (DH '18)*. New York, NY, USA: ACM, 2018, pp. 121–125. DOI: 10.1145/3194658.3194677. URL: https://doi.org/10.1145/3194658.3194677 (cit. on p. 28).

[32] F. Gräßer, F. Tesch, J. Schmitt, S. Abraham, H. Malberg, and S. Zaunseder. "A pharmaceutical therapy recommender system enabling shared decision-making". In: *User Modeling and User-Adapted Interaction* 32.5 (Aug. 2021), pp. 1019–1062. ISSN: 1573-1391. DOI: 10.1007/s11257-021-09298-4 (cit. on p. 24).

[33] J. Guerreiro and P. Rita. "How to predict explicit recommendations in online reviews using text mining and sentiment analysis". In: *Journal of Hospitality and Tourism Management* 43 (June 2020), pp. 269–272. ISSN: 1447-6770. DOI: 10.1016/j.jhtm.2019.07.001 (cit. on pp. 19, 20).

[34] H. Gültekin. *What is Silhouette Score?* Accessed: 2025-04-27. 2020. URL: https://medium.com/@hazallgultekin/what-is-silhouette-score-f428fb39bf9a (cit. on p. 59).

[35] Q. Han, M. Ji, I. M. d. R. de Troya, M. Gaur, and L. Zejnilovic. *A Hybrid Recommender System for Patient-Doctor Matchmaking in Primary Care*. 2018. DOI: 10.48550/ARXIV.1808.03265 (cit. on p. 22).

[36] F. M. Harper and J. A. Konstan. *MovieLens 1M Dataset*. Accessed: 2025-02-05. 2003. URL: https://grouplens.org/datasets/movielens/1m/ (cit. on p. 18).

[37] D. Howell, S. Molloy, K. Wilkinson, E. Green, K. Orchard, K. Wang, and J. Liberty. "Patient-reported outcomes in routine cancer clinical practice: a scoping review of use, impact on health outcomes, and implementation factors". In: *Annals of Oncology* 26.9 (Sept. 2015), pp. 1846–1858. ISSN: 0923-7534. DOI: 10.1093/annonc/mdv181 (cit. on p. 1).

[38] Hugging Face. *cardiffnlp/twitter-roberta-base-sentiment*. Accessed: 2025-01-15. 2025. URL: https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment (cit. on p. 33).

[39] Hugging Face. *nlptown/bert-base-multilingual-uncased-sentiment*. Accessed: 2025-01-15. 2025. URL: https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment (cit. on p. 33).

[40] C. J. Hutto and E. Gilbert. *VADER (Valence Aware Dictionary and sEntiment Reasoner)*. Accessed: 2025-01-15. 2025. URL: https://pypi.org/project/vaderSentiment/ (cit. on pp. 20, 33, 53).

[41] B. In. *How to Use the Elbow Method to Determine the Optimal Number of Clusters*. Accessed: 2025-04-27. 2023. URL: https://builtin.com/data-science/elbow-method (cit. on p. 59).

[42] B. In. *Feature Importance in Machine Learning: What It Is and How to Use It*. Accessed: 2025-01-15. 2025. URL: https://builtin.com/data-science/feature-importance (cit. on p. 30).

[43] *Kaggle*. https://www.kaggle.com. Accessed: 2025-01-12 (cit. on p. 28).

[44] M. Kankaraš and S. Capecchi. "Neither agree nor disagree: use and misuse of the neutral response category in Likert-type scales". In: *METRON* 83 (2025), 111–140. DOI: 10.1007/s40300-024-00276-5 (cit. on p. 2).

[45] M. V. Koroteev. "BERT: a review of applications in natural language processing and understanding". In: *arXiv preprint arXiv:2103.11943* (2021) (cit. on p. 11).

[46] R. Lafta, J. Zhang, X. Tao, Y. Li, and V. S. Tseng. "An Intelligent Recommender System Based on Short-Term Risk Prediction for Heart Disease Patients". In: *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*. IEEE, Dec. 2015, pp. 102–105. DOI: 10.1109/wi-iat.2015.47 (cit. on p. 21).

[47] I. Leviatan, B. Oberman, E. Zimlichman, and G. Y. Stein. "Associations of physicians' prescribing experience, work hours, and workload with prescription errors". In: *Journal of the American Medical Informatics Association* 28.6 (Oct. 2020), pp. 1074–1080. ISSN: 1527-974X. DOI: 10.1093/jamia/ocaa219 (cit. on p. 2).

[48] J. Li. *KUC Hackathon - Winter 2018*. https://www.kaggle.com/datasets/jessicali9530/kuc-hackathon-winter-2018. Accessed: 2025-01-12. 2018 (cit. on pp. 20, 24, 28, 29, 42).

[49] G. Linden, B. Smith, and J. York. "Amazon.com recommendations: item-to-item collaborative filtering". In: *IEEE Internet Computing* 7.1 (Jan. 2003), pp. 76–80. ISSN: 1089-7801. DOI: 10.1109/mic.2003.1167344 (cit. on p. 17).

[50] Y. Ling. *Bio+Clinical BERT, BERT Base, and CNN Performance Comparison for Predicting Drug-Review Satisfaction*. 2023. DOI: 10.48550/ARXIV.2308.03782 (cit. on pp. 19, 20, 32, 33).

[51] K. MacMillan and J. D. Wilson. "Topic supervised non-negative matrix factorization". In: *ArXiv* abs/1706.05084 (2017). URL: https://api.semanticscholar.org/CorpusID:32288359 (cit. on p. 9).

[52] Matplotlib Developers. *Matplotlib: Visualization with Python*. Accessed: 2025-01-15. 2025. URL: https://matplotlib.org/ (cit. on pp. 29, 34).

[53] A. McAllister, I. Naydenova, and Q. Nguyen Duc. *Building a LDA-based Book Recommender System*. Accessed: 2024-11-07. 2019. URL: https://humboldt-wi.github.io/blog/research/information_systems_1819/is_lda_final/ (cit. on pp. 18, 25, 26).

[54] L. McInnes and J. Healy. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction". In: *ArXiv* abs/1802.03426 (2018). DOI: 10.48550/arXiv.1802.03426 (cit. on p. 12).

[55] L. McInnes, J. Healy, S. Astels, et al. "HDBSCAN: Hierarchical density based clustering." In: *J. Open Source Softw.* 2.11 (2017), p. 205. DOI: 10.21105/joss.00205 (cit. on p. 12).

[56] W. McKinney et al. "Data structures for statistical computing in Python." In: *SciPy* 445.1 (2010), pp. 51–56 (cit. on p. 27).

[57] D. Medvecki, B. Bašaragin, A. Ljajić, and N. Milošević. "Multilingual transformer and BERTopic for short text topic modeling: The case of Serbian". In: (2024). DOI: 10.48550/ARXIV.2402.03067 (cit. on p. 48).

[58] C. Meister and R. Cotterell. "Language Model Evaluation Beyond Perplexity". In: (2021). DOI: 10.48550/arXiv.2106.00085. arXiv: 2106.00085 [cs.CL] (cit. on p. 35).

[59] Y. Mejova. "Sentiment analysis: An overview". In: *University of Iowa, Computer Science Department* (2009), p. 5 (cit. on p. 13).

[60] D. Merkel. "Docker: lightweight linux containers for consistent development and deployment". In: *Linux journal* 2014.239 (2014), p. 2 (cit. on p. 39).

[61] M. Michael. "What are recommender systems? Use cases, types, and techniques". In: *Aporia* (2023). URL: https://www.aporia.com/learn/recommender-systems/what-are-recommender-systems-use-cases-types-and-techniques/ (cit. on p. 7).

[62] D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum. "Optimizing semantic coherence in topic models". In: *Proceedings of the 2011 conference on empirical methods in natural language processing*. 2011, pp. 262–272 (cit. on p. 35).

[63] S. Mulukuntla and M. Gaddam. "Data-Driven Healthcare: Trends in Machine Learning and AI for Disease Prediction and Prevention". In: *ESP Journal of Engineering & Technology Advancements* 1.1 (2021), pp. 25–33. ISSN: 2583-2646. DOI: 10.56472/25832646/JETA-V1I1P106. URL: https://www.espjeta.org/ (cit. on p. 2).

[64] B. A. H. Murshed, S. Mallappa, J. Abawajy, M. A. N. Saif, H. D. E. Al-ariki, and H. M. Abdulwahab. "Short text topic modelling approaches in the context of big data: taxonomy, survey, and analysis". In: *Artificial Intelligence Review* 56.6 (Oct. 2022), pp. 5133–5260. ISSN: 1573-7462. DOI: 10.1007/s10462-022-10254-w (cit. on p. 46).

[65] Netflix. *How Netflix's Recommendations System Works*. 2024. URL: https://help.netflix.com/en/node/100639 (cit. on p. 18).

[66] M. Neumann, D. King, I. Beltagy, and W. Ammar. "ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing". In: *Proceedings of the 18th BioNLP Workshop and Shared Task*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 319–327. DOI: 10.18653/v1/W19-5034. eprint: arXiv:1902.07669. URL: https://www.aclweb.org/anthology/W19-5034 (cit. on p. 69).

[67] NumPy Developers. *NumPy: The fundamental package for scientific computing with Python*. Accessed: 2025-01-15. 2025. URL: https://numpy.org/ (cit. on p. 29).

[68] OpenAI. "GPT-4 Technical Report". In: *CoRR* abs/2303.08774 (2023). DOI: 10.48550/ARXIV.2303.08774. arXiv: 2303.08774. URL: https://doi.org/10.48550/arXiv.2303.08774 (cit. on p. 25).

[69] D. O'callaghan, D. Greene, J. Carthy, and P. Cunningham. "An analysis of the coherence of descriptors in topic modeling". In: *Expert Systems with Applications* 42.13 (2015), pp. 5645–5657 (cit. on pp. 60, 66).

[70] Pandas Developers. *Pandas: Python Data Analysis Library*. Accessed: 2025-01-15. 2025. URL: https://pandas.pydata.org/ (cit. on pp. 27, 29).

[71] 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).

[72] M. Pato, M. Barros, and F. M. Couto. "Survey on Recommender Systems for Biomedical Items in Life and Health Sciences". In: *ACM Comput. Surv.* 56.6 (Feb. 2024). ISSN: 0360-0300. DOI: 10.1145/3639047. URL: https://doi.org/10.1145/3639047 (cit. on p. 20).

[73] A. C. Pereira, M. Pato, and N. Datia. "Mapping Drug Interactions and Therapeutic Clusters through Knowledge Graph Visualization". In: *2025 29th International Conference Information Visualisation (IV)*. 2025. DOI: 10.1109/IV68685.2025.00074 (cit. on pp. 37, 69).

[74] A. S. Pinto, M. Pato, and N. Datia. "Enhancing Drug Reviews Insights through Exploratory Data Analysis and Sentiment Analysis". In: *2024 28th International Conference Information Visualisation (IV)*. 2024, pp. 190–195. DOI: 10.1109/IV64223.2024.00042 (cit. on pp. 20, 25, 42).

[75] A. S. S. Pinto. *Drug recommendation system based on symptoms and user sentiment analysis (DRecSys-SUSA)*. 2025 (cit. on pp. 24–26).

[76] S. Ravoire, M. Lang, E. Perrin, A. Audry, P. Bilbault, M. Chekroun, L. Demerville, T. Escudier, L. Guéroult-Accolas, C. Guillot, M. Malbezin, P. Maugendre, J. Micallef, M. Molimard, F. Montastruc, E. Pierron, L. Reichardt, and F. Thiessard. "Advantages and limitations of online communities of patients for research on health products". In: *Therapies* 72.1 (Feb. 2017), pp. 135–143. ISSN: 0040-5957. DOI: 10.1016/j.therap.2016.11.058 (cit. on p. 1).

[77] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. "GroupLens: an open architecture for collaborative filtering of netnews". In: *Proceedings of the 1994 ACM conference on Computer supported cooperative work - CSCW '94*. CSCW '94. ACM Press, 1994, pp. 175–186. DOI: 10.1145/192844.192905 (cit. on p. 17).

[78] E Rich. "User modeling via stereotypes". In: *Cognitive Science* 3.4 (Oct. 1979), pp. 329–354. ISSN: 0364-0213. DOI: 10.1016/s0364-0213(79)80012-9 (cit. on p. 17).

[79] L. Richardson. *Beautiful Soup: A Library for Screen-Scraping HTML and XML*. https://www.crummy.com/software/BeautifulSoup/. Accessed: 2025-01-12 (cit. on p. 28).

[80] M. S. *Recommendation System: A Complete Guide*. https://medium.com/@mitalis2905/recommendation-system-a-complete-guide-d147906fb452. Accessed: November 5, 2024. 2024 (cit. on p. 17).

[81] Seaborn Contributors. *Seaborn: Statistical Data Visualization*. Accessed: 2025-01-15. 2025. URL: https://seaborn.pydata.org/ (cit. on p. 29).

[82] S. Sheikholeslami. "Ablation Programming for Machine Learning". MA thesis. KTH, School of Electrical Engineering and Computer Science (EECS), 2019, p. 52 (cit. on p. 73).

[83] *SQL LEFT JOIN Keyword - W3Schools*. https://www.w3schools.com/sql/sql_join_left.asp. Accessed: 2025-05-12. 2025 (cit. on p. 70).

[84] J. Su, Y. Guan, Y. Li, W. Chen, H. Lv, and Y. Yan. *Do recommender systems function in the health domain: a system review*. 2020. DOI: 10.48550/ARXIV.2007.13058 (cit. on p. 8).

[85] F. G. Surya Kallumadi. *Drug Reviews (Drugs.com)*. 2018. DOI: 10.24432/C5SK5S (cit. on pp. 3, 37).

[86] *Symptoms, signs and abnormal clinical and laboratory findings*. [Accessed 04-04-2025]. 2019. URL: https://icd.who.int/browse10/2019/en#/XVIII (cit. on p. 44).

[87] TechTarget Contributor. *BERT Language Model*. Accessed: 2025-02-05. n.d. URL: https://www.techtarget.com/searchenterpriseai/definition/BERT-language-model (cit. on p. 11).

[88] TextBlob Contributors. *TextBlob: Quickstart - Sentiment Analysis*. Accessed: 2025-01-15. 2025. URL: https://textblob.readthedocs.io/en/dev/quickstart.html#sentiment-analysis (cit. on pp. 20, 33, 34).

[89] H. Thakkar and D. R. Patel. "Approaches for Sentiment Analysis on Twitter: A State-of-Art study". In: *CoRR* abs/1512.01043 (2015). arXiv: 1512.01043. URL: http://arxiv.org/abs/1512.01043 (cit. on p. 14).

[90] 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 (Dec. 2020), pp. 171–201. ISSN: 1573-7675. DOI: 10.1007/s10844-020-00633-6 (cit. on p. 21).

[91] S. Tsuge, M. Shishibori, S. Kuroiwa, and K. Kita. "Dimensionality reduction using non-negative matrix factorization for information retrieval". In: *2001 IEEE International Conference on Systems, Man and Cybernetics. e-Systems and e-Man for Cybernetics in Cyberspace (Cat. No. 01CH37236)*. Vol. 2. IEEE. 2001, pp. 960–965 (cit. on p. 11).

[92] U.S. National Library of Medicine. *DailyMed*. Accessed: 2025-05-06. 2024. URL: https://www.dailymed.nlm.nih.gov/dailymed/ (cit. on pp. 37, 68).

[93] M. Uta, A. Felfernig, V.-M. Le, T. N. T. Tran, D. Garber, S. Lubos, and T. Burgstaller. "Knowledge-based recommender systems: overview and research directions". In: *Frontiers in Big Data* 7 (Feb. 2024). ISSN: 2624-909X. DOI: 10.3389/fdata.2024.1304439 (cit. on p. 8).

[94] S. Vargas and P. Castells. "Rank and relevance in novelty and diversity metrics for recommender systems". In: *Proceedings of the Fifth ACM Conference on Recommender Systems*. RecSys '11. Chicago, Illinois, USA: Association for Computing Machinery, 2011, 109–116. ISBN: 9781450306836. DOI: 10.1145/2043932.2043955. URL: https://doi.org/10.1145/2043932.2043955 (cit. on p. 35).

[95] J. Walsh, C. Dwumfour, J. Cave, and F. Griffiths. "Spontaneously generated online patient experience data - how and why is it being used in health research: an umbrella scoping review". In: *BMC Medical Research Methodology* 22.1 (May 2022). ISSN: 1471-2288. DOI: 10.1186/s12874-022-01610-z (cit. on p. 1).

[96] Z. Wang, X. Yu, N. Feng, and Z. Wang. "An improved collaborative movie recommendation system using computational intelligence". In: *Journal of Visual Languages & Computing* 25.6 (Dec. 2014), pp. 667–675. ISSN: 1045-926X. DOI: 10.1016/j.jvlc.2014.09.011 (cit. on pp. 18, 25).

[97] F. Weninger, J. Le Roux, J. R. Hershey, and S. Watanabe. "Discriminative NMF and its application to single-channel source separation." In: *Interspeech*. 2014, pp. 865–869 (cit. on p. 11).

[98] Wikipedia. *Contraction (grammar) — Wikipedia, The Free Encyclopedia*. http://en.wikipedia.org/w/index.php?title=Contraction%20(grammar)&oldid=1282548593. [Online; accessed 10-April-2025]. 2025 (cit. on p. 50).

[99] Wikipedia contributors. *Non-negative Matrix Factorization*. Accessed: 2025-02-04. 2024. URL: https://en.wikipedia.org/wiki/Non-negative_matrix_factorization (cit. on p. 11).

[100] World Health Organization. *ICD-11 Newsletter, November 2015*. Accessed: 2025-01-15. 2015. URL: https://www.who.int/docs/default-source/classification/icd/icd11/icd11-newsletter-nov2015.pdf?sfvrsn=2a181fbc_2 (cit. on p. 30).

[101] World Health Organization. *International Classification of Diseases, 11th Revision (ICD-11)*. Accessed: 2025-04-02. 2019. URL: https://icd.who.int/ (cit. on pp. 30, 42).

[102] G. Xiong, Z. Yang, J. Yi, N. Wang, L. Wang, H. Zhu, C. Wu, A. Lu, X. Chen, S. Liu, T. Hou, and D. Cao. "DDInter: an online drug–drug interaction database towards improving clinical decision-making and patient safety". In: *Nucleic Acids Research* 50.D1 (Oct. 2021), D1200–D1207. ISSN: 1362-4962. DOI: 10.1093/nar/gkab880. URL: http://dx.doi.org/10.1093/nar/gkab880 (cit. on pp. 3, 37).

[103] Y. Zhang, R. Jin, and Z.-H. Zhou. "Understanding bag-of-words model: a statistical framework". In: *International journal of machine learning and cybernetics* 1.1 (2010), pp. 43–52 (cit. on p. 9).

[104] R. Řehůřek and P. Sojka. *Gensim: Topic Modelling for Humans*. Accessed: 2025-01-15. 2025. URL: https://radimrehurek.com/gensim/ (cit. on p. 36).