National College of Ireland

A Content Based Recommender System for Medicine using Machine Learning Algorithm

MSc Research Project
Data Analytics

Utkarsh Mathur
Student ID: x19232977

School of Computing
National College of Ireland

Supervisor: Aaloka Anant

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Utkarsh Mathur
Student ID:	x19232977
Programme:	Data Analytics
Year:	2021
Module:	MSc Research Project
Supervisor:	Aaloka Anant
Submission Due Date:	31/01/2022
Project Title:	A Content Based Recommender System for Medicine using Machine Learning Algorithm
Word Count:	6596
Page Count:	21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Utkarsh Mathur
Date:	31st January 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A Content Based Recommender System for Medicine using Machine Learning Algorithm

Utkarsh Mathur
x19232977

Abstract

Background: With the introduction of covid-19 virus and diseases it spreads in such a minimal period, Doctors around the world has to be more prepared with technology and the ability to take an effective decision in terms of proposing medicines or treatment of illnesses should be empowered. As e-commerce grows in the medical industry, more and more healthcare products are being sold online. Involved stakeholders in the study can include healthcare facilities such as hospitals and clinics, as well as online retailers of OTC (Over-the-Counter) medications and other non-profit organizations with extensive patient records. The above-mentioned stakeholders may benefit from making treatment decisions based on prior patients with comparable symptoms who used a Machine Learning recommender model based on patient history.

Objective: When it comes to creating a recommendation system for prescribing drugs, this research will focus on Machine learning models, as well as examining various metrics to determine whether Content Based Recommender system approaches may be utilized to build an Drug Recommendation Model. The CSV file is obtained from the UCI Machine Learning repository, which provides the reviews and ratings of different users over the same medications. This dataset doesn't involve any personal information and its available on open source platform so no breach of personal information has happened and consecutively ethics form submission not required. This investigation will thus empower patients / stakeholders to take better informed decision and initiate proper medication themselves without the intervention of physician. It may be determined which recommender system outperforms others using the evaluation scores.

Results: After using various metrices for similarity calculation, the model was build upon cosine similarity with an average score of 0.098. The drug recommendation model deployed on cloud also shows that 4 out 5 recommendations are correct. Hence, 90percent accuracy is achieved for this recommender model.

1 Introduction

1.1 Overview

Most hospitals have information on the health conditions of its patients. As a result, medical professionals face a data glut that must be simplified in order to provide useful insights. To achieve that, the combination of recommender technology and a machine learning algorithm is the most cutting-edge model currently available. In order to provide

the greatest possible user experience, a healthcare recommender system can assist stakeholders in making well-informed decisions. Many of the world's largest digital companies are already employing recommender systems to tailor the user experience on their websites, which in turn generates money for the businesses. For movie recommendations, there's Netflix. For e-commerce recommendations and movie recommendations there's Amazon. For friend recommendations there's Facebook. With the emergence of new disease and costlier treatment, it becomes immensely necessary to implement some machine learning techniques to reduce cost and effecient treatment, this idea is also discussed in the paper, where author of (Han et al.; 2018) describes the cost of health care is expected to climb in many emerging countries due to an aging population. Investing in systems that strengthen patient-doctor relationships and increase primary care doctors' gatekeeping responsibilities is therefore critical. As a result of this collaboration, the authors were able to design a system for connecting patients with primary care providers in order to sustain continuity of care. As a result, they employ a hybrid technique that aims to provide every patients with a customized list of doctors. Patients, doctors, and the relationships between them were gathered from various sources to build a dataset that was used by researchers. Patients who have seen the same doctor's patients can be used to infer patient preferences for primary care doctors. This information can be utilized to identify similar patients or clinicians in different social circumstances, for example. These traits were discovered through the use of scientists in the field and exploratory data analysis. As a result, the hybrid recommender system is able to use a collection of relevant features to represent the records of consultations. Both the heuristic baseline and a traditional recommendation system for CF exhibit more accuracy in our results than the heuristic baseline.

In this research proposal the author will develop a recommender model based on patient disease history and medicine used over time in the medical recommendation system. Using this model doctors can prescribe medication to follow with greater precision and a better recommendation system to patients with almost identical comorbidities. By using content-based and user-item based collaborative filtering, these can be implemented. Hybrid collaborative filtering, according to the author, is the best way to deal with the cold start and scant data issues that is bottleneck for all recommendation systems. As per (Dai et al.; 2018) Collaborative Filtering (CF) based on Principal Component Analysis (PCA) is most commonly used to reduce the dimensionality of user/patient data using the Matrix Factorization approach. Figure 1 shows the diagram of basic recommender system and how it works.

1.2 Motivation

Because of its potential impact on the medical industry, this was chosen as a research topic. It is increasingly important for people to have a better recommendation on the products they use for their health and lifestyle, and this recommender system will also benefit stakeholders such as healthcare professionals, doctors, and volunteers working in the medical field to have better and more sustainable recommendations for their patient's health and well-being. The goal of this study is to develop a recommender system that may be used to make medication recommendations to patients and help them make better medical decisions. On the UCI Machine Learning repository, there is freely available data of user evaluations and ratings of various medicines for specific diseases and symptoms. To see whether collaborative filtering recommendation system is appropriate for

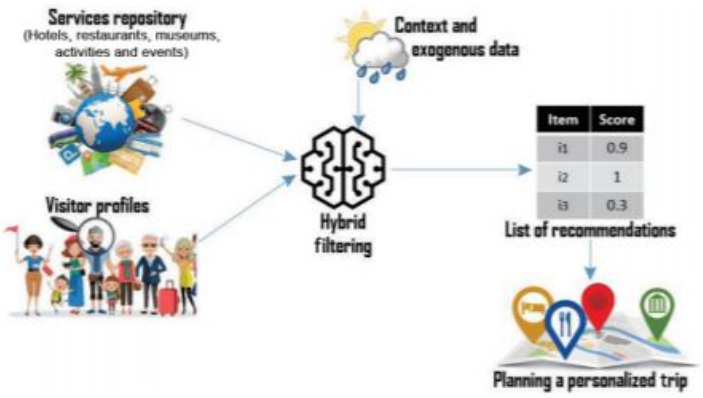

Figure 1: Proposed Architecture for Recommender System. (Fararni et al.; 2021)

recommending medicines in light of cold start and high sparse data problems, this study also seeks to check that.

1.3 Objective

The goal of this study is to develop a recommendation system that recommends Medicine to the user based not only on predicted ratings but also on the closeness of the user's interests. The recommender engine will match products and ratings offered by the user's closest neighbors to determine the most relevant stuff for the user.In this research, NLP technique which is deep learning algorithm will be used to determine the users review by their comments on a specific medicine. Word vectorization was done to make each word in the form of vector and then the similarity of each medicine among each other according to disease or symptoms using Cosine similarity and Pearson baseline was calculated. The similarity and distance metrices used in Recommender system is also evaluated. These are Pearson Correlation, Spearman's Correlation, Jaccard's Similarity, Cosine Similarity, Kendalls tau correlation. Apart from this Euclidean distance and Manhattan distance was calculated.

1.4 Research Question

To what extent may content based recommender system tackles the issue of cold start and high data sparsity while recommending medications to patients?

1.5 Structure Of Report

The further structure of this report will have following sections: section 2 describes the previous related researchers in the field of recommender system models used by different technique, section 3 discusses the methodology followed in the completion of research objective / Question, section 4 discusses about designing of environment for performing experiments using different technique, section 5 shows the implementation of codes and various Data mining techniques to build a recommendation model, section 6 evaluates the various parameters and trade-off being considered in projects and finally section 7

discusses the result and conclusion of the report. It finally describes the future work of scope.

2 Related Work

Using systems that propose recommendations based on user behavior and patterns, it may alleviate the millenium problem of Data overload and get insight for improved analytics. This is also based on the reviews of those products or items. Building a model for proposing medical items or treatments to patients with comparable comorbidities is the subject of this study.Note that here there are five subsection 2.1, subsection 2.2, subsection 2.3, subsection 2.4 and subsection 2.5

2.1 Research into the use of recommender systems in various industries

Recommender systems have proven to be an efficient method of reducing information overload in the age of ever-increasing online data. Since recommender technologies are largely used in many web applications, their potential to alleviate many of the difficulties associated with over-choice cannot be understated. Many fields of research, including computer vision and NLP, have recently seen a surge in interest in deep learning because of its exceptional performance as well as its ability to learn feature representations from scratch.Research on information retrieval and recommender systems has lately demonstrated the usefulness of deep learning. The field of deep learning in recommender systems appears to be flourishing. In (Zhang et al.; 2019) A taxonomy of deep learning-based recommendation models, as well as a complete description of the state-of-the-art, are provided and devised. Finally, they present new perspectives on this exciting new development in the field by expanding on existing trends. Here, the author provides a comprehensive evaluation of the most notable papers on deep learning-based recommender systems that have been published to date. The authors also discuss some of the most important issues in the field, as well as anticipated future developments. Deep learning and recommender systems have both been hot research subjects for decades.

(Stark et al.; 2019) describes that a drug recommendation system can aid doctors and nurses in prescribing the proper medication. Thanks to modern technology, it is feasible to create recommendation systems that lead to shorter decisions. Several contemporary pharmaceutical recommendation systems use unique algorithms. As a result, it's vital to understand how these systems work now, their benefits and drawbacks, and where more research is needed. This study examines and compares existing methods for medicine recommendation systems and provides future research targets. This study provided a systematic literature evaluation of medicine recommendation engines. They searched six databases for 13 research that met our strict criteria. Ontology-based and rule-based techniques dominate machine learning and data mining research. The research evaluated parameters such as data storage, interface, data collection, data preparation, platform/technology/algorithm. Non-disease research lacked data storage, interface, data collection, pre-processing, and custom algorithms.Music recommendation system using CNN module is also discussed in (Elbir et al.; 2018) where authors discusses that after extracting acoustic data from music, machine learning techniques were utilized to classify music genres and create music suggestions. Convolutional neural networks were used to

classify and recommend music genres, as well as compare the findings. This project uses convolutional neural networks and digital signal processing to categorize and recommend songs. The study created a service that offers music based on user requests after analyzing how features are obtained. Initially, features were extracted using digital signal processing, and then a CNN was taught to do so. The acoustic qualities of songs are then used to determine the best suggestion algorithm. SVM outperforms other approaches in classification accuracy. As a result, changing the window size or window type had relatively negligible effects on performance.

2.2 Content-Based Filtering in Recommender Systems Research

There are different kind of Filtering used in Recommender systems and the most used ones are content-based filtering and collaborative Filtering. Based on the user's preferences and choices, the Content-Based (CB) technique suggests things or products to the user. In order to suggest new things with similar attributes, content-based filtering operates on the matching level of the attributes of content. In (Pal et al.; 2017) and (Zhao; 2019) the authors concludes that as an alternative to traditional content-based filtering, this work presents a simple method for finding correlations between two features using set intersection and predicting the similarities between two items for recommendations using this method. Naive Bayes and other text classifiers have been utilized in content-based algorithms in the past. In addition, the algorithm is tested and compared to the PureCF and SVD algorithms. After evaluation, the MAE values that are created allow for accurate comparisons. A larger dataset may yield different results, even if Hybrid content recommendation had superior MAE values and increased the dataset's sparsity between 1 percent and 2 percent.Figure 2 shows that at sparsity levels of 98.5percent, there is a little difference between the outcomes of Pure CF and the authors'technique, but when the sparsity level is increased to 99percent, the algorithm out-performs the Pure CF.

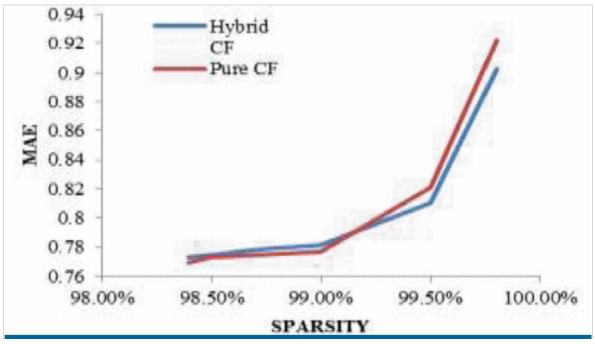

Figure 2: Comparison of MAE values with increasing sparsity for hybrid CF and Pure CF. (Pal et al.; 2017)

2.3 Study of Recommender System Collaborative Filtering Approach [1]

Collaborative filtering is the most common way to survive more advanced recommendation engines. Netflix, Amazon, Facebook, and other major digital companies employ this type of recommendation system in a wide range of industries. In the paper (Gurbanov and Ricci; 2017), the authors discusses about the use of technique for recommendation system and limitation faced while using Collaborative filtering. Furthermore, the paper combines Sequence Mining (SM) and Collaborative Filtering (CF) to forecast user actions. When a target user performs an unobserved action on an object, the proposed model predicts based on the users' completed actions at that time. The model does not use any user or item information. A large real-world dataset shows that hybrid models outperform standalone SM and CF. Any SM and CF model can be subcomponents in the suggested hybrid model. Our technique considers time delays between actions and their sequential order. Another issue is that there is no method to harness the interaction effect or the impact of one action on another. As a result, the SM component's probability computations can be wrong for odd sequences. Finally, the proposed hybrid system has the same cold-start issue as any other CF system. The authors want to compare the proposed hybrid model to FPMC and CBCF and test it on more datasets in the future. They also want to include information about user and item attributes in the model to help with the cold-start issue. Similalry, there are researchers who concludes that having a large dataset can be sometime difficult for collaborative filtering approach for a novel solution but hybrid model with matrix factorization can resolve the problem and they get better results.

In (Dong et al.; 2017) the authors illustrates that in recommender systems, collaborative filtering (CF) is a commonly employed method for resolving a wide range of real-world issues.Users' preferences for products are encoded in a matrix known as a user-item matrix in traditional CF-based methods for learning to provide recommendations. Because rating matrices in real-world applications are often sparse, CF-based algorithms' recommendation performance degrades dramatically. Data sparsity and cold start issues can be addressed with enhanced CF methods that make use of more side information. However, the sparseness of the user-item matrix and the side information may mean that the learnt latent components are ineffective. A hybrid model that uses side information and collaborative filtering from the rating matrix to jointly perform deep user and item latent factor learning is proposed to address this issue by the authors, who draw on developments in learning effective representations in deep learning. On three different datasets, extensive testing has shown that the hybrid model is superior than other methods at leveraging side information, and this leads to better outcomes overall. (Fararni et al.; 2021) employs Hadoop-based infrastructure and the MapReduce algorithm to solve the problem of dealing with a large dataset in CF, which is frequently encountered. The author in (Shaikh; 2020) develops a TOP-N Nearest Neighbor Based Movie Recommender System. The research includes evaluating machine learning models. Conclusion: Cross-validation of machine learning models using K-fold and LOO (Leave One Out). RMSE and MAE are used to assess accuracy (Mean Absolute Error). For RMSE and MAE, KNNBaseline SVD, KNNWithMean(Ib) and KNNWithMean(Ib) came second (Ub). Using SVDpp in Matrix factorization and KNNBaseline Content-based filtering, K-fold CV folds are 2 percent more accurate than LOOCV folds. Sprawl is a term used to describe data scarcity.

[1] `https://towardsdatascience.com/tagged/collaborative-filtering`

There will be further iterations on many recommendation machine learning models (such as restricted Boltzmann machines and auto encoders). SageMake can use big data from AWS cloud services.

2.4 Recommender System Research Using Deep Learning

Deep neural networks and their implementation over recommender systems have expanded rapidly in recent years. This field is exploding with new ideas and methods. We cannot overestimate the value of recommender systems given their extensive use in online applications and capacity to solve various issues connected to over-choice. Deep learning has recently gained popularity in numerous academic disciplines, including computer vision and natural language processing, because to its superior performance and the ability to build feature representations from scratch. Deep learning has lately shown its efficiency in information retrieval and recommender systems research. Deep learning in recommender systems is booming. The author in (Wang et al.; 2014) suggests people regularly use CF when recommending. This method is widely used in recommender systems. Users' ratings are the only source of CF information. Because the evaluations are often low, CF-based approaches perform poorly. Auxiliary data, such as item content data, can aid. CTR combines two components that learn from numerous sources of data. However, scarce auxiliary data may render CTR's latent representation ineffective. They propose a collaborative deep learning hierarchical Bayesian model that extends recent deep learning successes from single identifier inputs to CF-based inputs (CDL). CDL can increase understanding on three real-world datasets. They demonstrated cutting-edge content information performance by combining deep representation learning and collaborative filtering. First hierarchical Bayesian model that combines deep learning with RS. The researchers also developed a sampling-based Bayesian back-propagation method for CDL. More powerful alternatives may replace the bag-of-words model. CDL also supports deep learning models besides SDAE. Convolutional neural networks, for example, can consider word order. Further improving performance with deep learning models.

In (Tran et al.; 2021) authors suggests nowadays, a large amount of clinical data is scattered across various websites, making it difficult for users to find useful information. The abundance of medical information (e.g., on drugs, medical tests, and treatment suggestions) has made it difficult for doctors to make patient-centered decisions. These issues highlight the need for recommender systems in healthcare to help both end-users and medical professionals make better health decisions. They review existing research on healthcare recommender systems in this article. Unlike other related overview papers, ours delves into recommendation scenarios and approaches. Food, drug, health status prediction, healthcare service, and healthcare professional recommendations are examples. The authors also develop working examples to better understand recommendation algorithms. Finally, the authors discuss future challenges in developing healthcare recommender systems.

2.5 Recommender System Matrix Factorization Research

(Bhavana et al.; 2019) explains most recommender systems employ Matrix Factorization to reduce the number of dimensions in the underlying data set.When it comes to unsupervised machine learning, it uses Principle Component Analysis (PCA). Cold start (meaning a new user has no preferences or reviews to compare or propose the things)

and very sparse data are two of the key challenges and limits of any recommender system (which means items have no reviews or ratings from the user to build a correlation matrix between user-item). Our research will be based on the above-mentioned issue. In the paper (Guan et al.; 2017) the author proposes that the algorithms are gaining popularity due to their promising performance on recommender systems. Some algorithms suffer from data sparsity. Active learning algorithms work well in recommender systems because they ask users to rate items as they enter the system. This research proposes an enhanced SVD (ESVD) matrix factorization model that combines standard matrix factorization methods with active learning-inspired ratings completion. A link between prediction accuracy and matrix density is also constructed to further investigate its potentials. In order to increase forecast accuracy, the authors suggest the Multi-layer ESVD (MESVD). The Item-wise ESVD (IESVD) and User-wise ESVD (UESVD) are provided to manage imbalanced datasets with considerably more users than items. The approaches are tested on the Netflix and Movielens datasets. Comparing them to classic matrix factorization and active learning approaches, the results show that they are more accurate and efficient. Most recommender systems struggle with a shortage of data. Additionally, they suggest using classic matrix factorization methods to best estimate a matrix with missing data. In particular, the overall EVSD model proposes high density through popular goods and active users, inspired by active learning. However, as all ratings are added simultaneously (ESVD, IESVD, and UESVD) or iteratively (a preset number of repetitions), the suggested methods considerably minimize the computational cost (MESVD).

In (Bodhankar et al.; 2019), the author explains the methods used to overcome the difficulties of developing a recommendation system based on a social network with user interest. In addition, this overview study presented many methods for constructing the recommendation system. Based on user location, user interest, and interpersonal interest in the social network, the described method is able to identify users. The method described here utilizes social matrix factorization and base matrix improved recommendation results to arrive at its conclusions about what to recommend.However, factor analysis presupposes that there is a linear association between factors and the variables that were calculating correlations. This method has its drawbacks. A Recommender system is extremely important for the healthcare business and will immediately assist the stakeholders in making more informed judgments in recommending drugs to the patient, as evidenced by this extensive research. For data preparation, content-boosted collaborative filtering might be employed. Combining these techniques with CNN and Matrix factorization can help overcome the drawbacks of cold start data and sparse datasets.

3 Methodology

The focus of the research project is to build a recommender system which suggests or recommends the OTC drugs based on the symptoms or disease of the patient. CRISP-DM approach is used for initial exploration of Dataset and basic techniques of Data Analysis are used.Figure 3 shows the CRISP-DM (Cross Industry Standard Process for Data Mining) approach followed as a Data Analytics methodology.

3.1 Stakeholder / Business Approach

Recommender system are evolving every other day and its significance in Healthcare industry is booming after the pandemic of 2020. Researchers show the usefulness of this

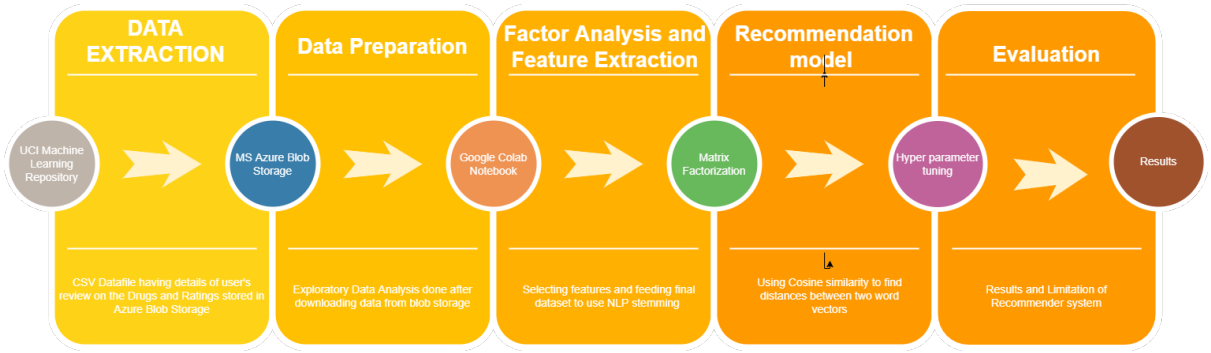

Figure 3: Methodology flow of CRISP-DM

technology in paper (Khoie et al.; 2017) where they discuss that healthcare organizations survey patients and staff about their experiences. Hospital administrators frequently use graphs and charts to provide patient satisfaction data. A deeper data analysis to find crucial patient satisfaction aspects is rarely followed by such visualization. Researchers present an unsupervised method for analyzing patient satisfaction survey data. Identification of similar patient communities and the primary factors contributing to their satisfaction. It will help hospitals identify patient groups or clusters most likely to be satisfied and take proactive steps to improve patient care. Finding links between patient demographics and satisfaction indicators using unsupervised exploratory data analysis.

3.2 Data Extraction

This dataset file was extracted from the UCI Machine Learning Repository[2] and contains little over 200,000 observations with distinctive properties of OTC (Over-the-Counter) drugs. It was created by UCI Machine Learning Repository. This is a multivariate data set with six attributes in total. The description is about the drug and users/patients who have provided reviews and ratings out of ten stars, which represent the overall satisfaction of the patient with the drug. In addition, by vectorizing the language, it is possible to perform sentiment analysis on the review attribute of a certain Drug. Individually identifiable information (PII) is not contained within this data, which is offered as opensource. First and foremost, cleaning up the data is essential. This includes removing null values and removing any columns that are no longer needed. Starting with the data analysis, the plotting of the top 10 highly rated medicines for each symptom is possible.

In Table 1 shows details about drugs dataset.

3.3 Data Pre-processing and Analysis

The CSV data file used for the project is kept at Microsoft Azure blob Storage and can be fetched from cloud whenever required. The Analysis of Dataset consists of Data Cleaning by removing na values and dropping duplicate rows. Doing Factor Analysis to determine which feature can be used for model building and accordingly modifying few features. Furthermore, using attributes Rating and Useful count, the most reviewed medicines are sorted out and graph is plotted. Using Python regex library some unwanted characters were removed from Reviews and fed to word vectorizer to calculate score of individual

[2]Dataset link : `https://bit.ly/3lZ3u6U`

Table 1: Drugs Dataset Descriptor

Attributes	Description	Values
UniqueID	Drugs's identification number	Uniqueness of each drug
Drug Name	OTC drug names	Around 200,000 glossary of drugs
Condition	Patient's Symptoms or disease	200,000+ health conditions
Review	Given by previous patients	Reviews are in sentences
Rating	Users' rating	Ratings are out of 10
Useful Count	Upvote given to specific Drugs	Used for finding most reviewed Drugs

word. For initial data visualition , Top10 drugs as per most reviewed by users, the graph got plotted.

3.4 Recommender system Methodology

3.4.1 Matrix Factorization

In this project , Matrix factorization played very crucial role for feature decomposition as per the rules of PCA (Principle Component Analysis) which is an unsupervised machine learing algorithm. After Feature extraction and factor analysis the dataset was fed for Matrix factorization and word vectorization.

3.4.2 Singular Vector Decomposition

The reviews were converted into vector and then using NLTK library were used for stemming the words to its root word in english language. The SVD model was applied using CountVectorizer class of sckitlearn library and further fed for building a vector matrix

3.5 Similarity Metrices used in Recommender System

3.5.1 Pearson Correlation

The Pearson correlation coefficient is quite sensitive to data values that are out of the ordinary. A single value that is significantly different from the other values in a data set might have a significant impact on the value of the correlation coefficient. A low Pearson correlation coefficient does not necessarily imply that there is no relationship between the variables in question. It is possible that the variables have a nonlinear connection.

3.5.2 Cosine Similarity

Using this similarity metrics, the similarity between vectorized words and most frequently used words were calculated and sorted in descending order. Since, jacard's distance is inversly propotional to similarity, which means more the distance between two vectors lesser the similarity between them. This is the basic algorithm used in finding the recommendations.

3.5.3 Spearman's Correlation

In the case of Pearson's correlation and Spearman's correlation, the Pearson correlation is equal to the Spearman correlation between the rank values of those two variables (whether linear or not). There are no repeated data values when the variables are perfect monotone functions of one another, hence the Spearman correlation is always between -1 and +1.The author of (Akoglu; 2018) has also discussed about Spearman's correlation coefficient and its significance.

3.5.4 Kendall Tau's Correlation

It's also called Kendall's tau coefficient. Kendall's Tau and Spearman's rank correlation coefficients employ data ranks. If your data violates one or more hypotheses, Kendall rank correlation (non-parametric) might be used instead of Pearson's correlation. For samples with many tied rankings, non-parametric Spearman correlation may be an option. Using Kendall rank correlation, two sets of data can be ordered similarly. Rather than looking at individual observations, Kendall's method looks at patterns of concordance or discordance between pairs of observations.

3.5.5 Jaccard's Similarity

The Jaccard Similarity algorithm can be used to determine how similar two objects are. The computed similarity might then be used into a recommendation query. As an example, you can use the Jaccard Similarity algorithm to display the products that were purchased by comparable customers, in terms of the previous products that they have purchased As also shown in (Fletcher and Islam; 2018)

3.6 Distance Metrices used in Recommender system

3.6.1 Euclidean Distance

Mathematics uses a line segment as a measure of how far two points are from each other in Euclidean space. This distance may be determined using the Pythagorean theorem, hence it is sometimes referred to as the Pythagorean distance. When two medications are compared, the distance between them is calculated, and the similarity between them is inversely proportional to this distance.

3.6.2 Manhattan Distance

Real-valued vectors can be measured using the Manhattan Distance (also known as the Taxicab Distance or City Block Distance). A chessboard or a city block is a good example of a vector that may be used to describe an object on a consistent grid. Intuition about what the metric calculates: the quickest path a taxicab would travel between city blocks is reflected in the name (coordinates on the grid). In an integer feature space, it may make sense to calculate Manhattan distance rather than Euclidean distance for two vectors.

4 Design Specification

Our recommendation engine's proposed architecture will be discussed in this section. Both content-based and user-based collaborative filtering are used in a Recommendation

system's foundation. The basic types of recommender system is shown in below Figure 4. The implementation of our Recommendation engine techniques was made possible with help of these designs' specifications. It will be discussed in the following section 5

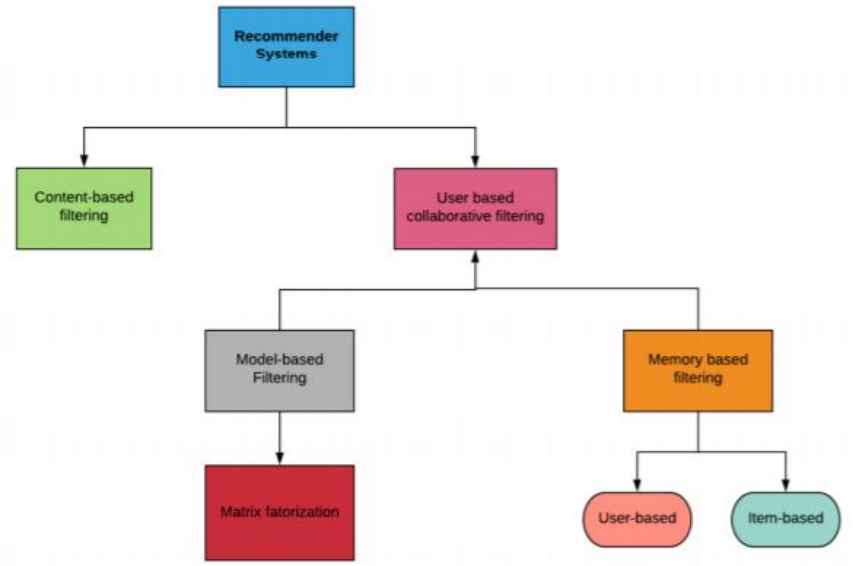

Figure 4: Types of Recommender System

4.1 Content Based Filtering Recommendation System

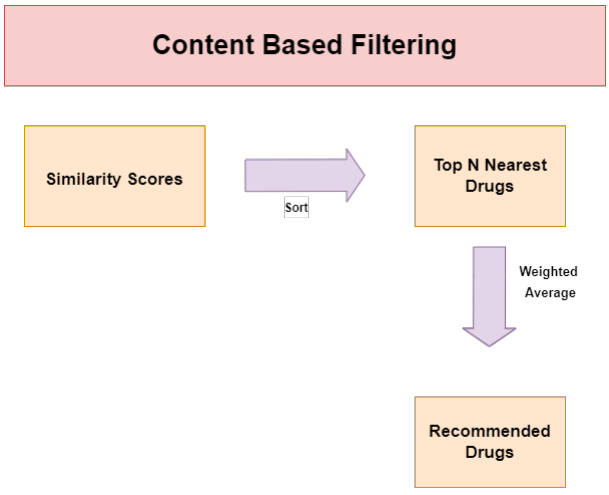

Figure 5: Process followed in content based filtering

In this project content based filtering recommender system was used by matching the attributes of a drug and review given by different patient to that drug as shown in Figure 5. The Drug also matches its corresponding disease or symptoms given in the

dataset and then further by using Cosine similarity score the relation is calculated. Word to Vector using CountVectorization is done in order to have vectors for each words.The Formula used in calculation of similarity which was directly import from python library is shown as in Figure 6

$$cos(U, I) = \frac{\sum_1^n U_i * I_i}{\sqrt{\sum_1^n U_i^2} * \sqrt{\sum_1^n I_i^2}}$$

Figure 6: Formula for cosine similarity calculation

4.2 Content Boosted Collaborative Filtering (CBCF)- hybrid method

High sparsity and cold start in data can be avoided using a hybrid method that combines Content Boosted with Collaborative Filtering. Figure 7 shows a flowchart showing how the symptoms of a patient and the reviews of other patients on a specific drug for that symptom can be combined and weighted sums calculated by CB and CF algorithms to predict recommendations. User-item rating matrix decomposition using the Singular Value Decomposer algorithm will be performed using a Convolution Neural Network. SVD is a Matrix Factorization technique based on the unsupervised learning PCA principle.

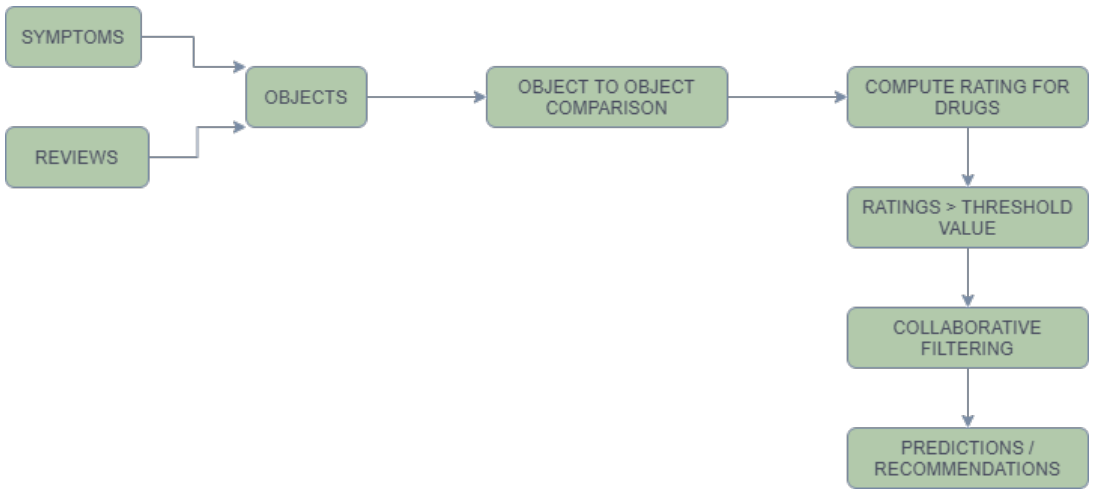

Figure 7: Diagram of Hybrid Method

5 Implementation

The entire project got implemented using Python as the core language and all its Data Science library was utilized to the fullest. The code were written in Jupyter Notebook as ipynb extension and is available as project's artefacts. In this project cloud technology is also used for keep the data file in Azure Blob storage and also the entire Recommender

engine model was deployed on Heroku Cloud (Salesforce application). These platforms are variables which can be replaced by different platform or vendors but the underlining technique remains constant. Data Extraction, Data gathering, Data pre-processing and Features of Dataset was already discussed in section 3. So, this section will comprise of all the methods implemented to build a recommendation engine.

5.1 Recommendation Engine Building

- Drugs with most reviews are sorted based on Ratings given by users and upvotes given by other patients to specific drugs. This has been graphed using Matplotlib Library.
- Next step implemented in the making of model is to split down the reviews into each words and making it in the form of List. Attributes drug name and condition was also separated by commas and converted into List in order to add with Reviews column. The metadata of each drug was created into tags cloumn.
- Singular Value Decomposition (SVD) was used for Matrix Fatorization technique to make the complex matrix even simpler and uses PCA at its core.
- Now, the Tag columns was joined and the words were stemmed using 'PorterStemmer' and 'WordNetLemmatizer' from nltk library in order to reduce all the words to their respective root word.
- Comparing output from both Lemmatizer was chosen for further analysis as it gave more meaningful output.
- Further to make these roots word in to vector in order to calculate distances / similarity between them, CountVectorizer was implemented from sckit learn feature extraction class.
- 'Bag of Words' technique was used in this part to fetch most frequent words in the entire dataset and then calculating score against each Drug name based on its metadata, an array of vectors was formed.
- Various similarity and distance metrices were tested and evaluated with random comparing two vectors. The results will be discussed in details in Section 6
- Finally, 'Cosine Similarity' function was imported from sckit learn metric pairwise class and similarity among different drugs was calculated and matrix was found
- Some hyper Parameter tuning was also implemented to check if the score changes and how will impact overall recommendations.
- A function was created to sort the drugs from the similarity matrix for any condition user select and recommended top 5 nearest drugs can be popped up with their unique id and name.

5.2 Building Website and Deployment on Cloud App

After the Recommender system model is build, it is now required to be deployed on cloud so that people can access it and can take recommendation of medicines as and when required. Figure 8 is the screenshot of the website hosted on heroku cloud app. Below are the process and techniques used to deploy the application.

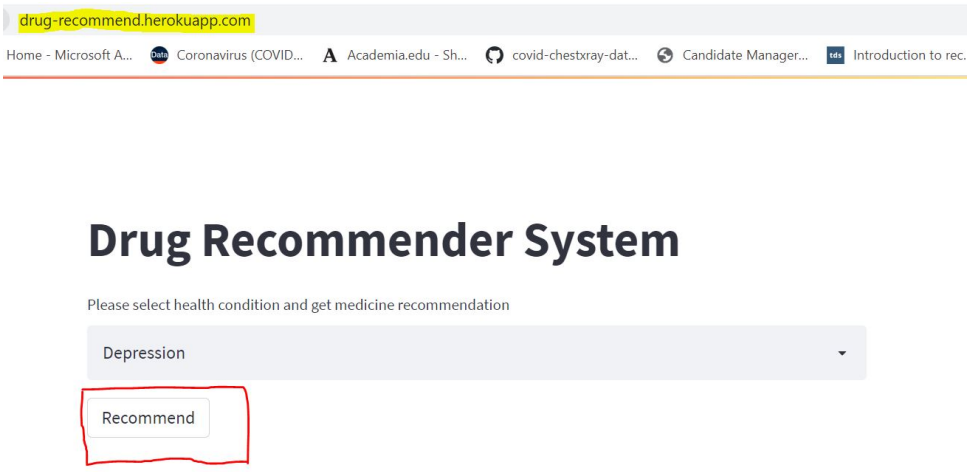

Figure 8: Drugs Recommendation Website

- To build a website which has a recommendation engine, Streamlit[3] has been used. 'Spyder' was used as the IDE for writing Python codes.
- Spyder IDE was used to build a website after importing a pickle file from a Jupyter notebook that contained a clean dataframe with the required columns. Also, the 'cosine similarity' function's similarity matrix vector was imported as a pickle (.pkl) file.
- In order to serialize and deserialize Python object structures, the pickle module implements binary protocols. Unpickling is the inverse of "pickling," in which a byte stream is reconstructed from a binary file or bytes-like object.[4]
- Streamlit function was defined to create basic website designing and 'Recommend' clickable button was also deployed. Drug Recommendation Website: `http://drug-recommend.herokuapp.com`. This may take 10-15 seconds to load a page as the application is hosted on free tier of cloud.
- To deploy the application file in Heroku Cloud, various other files were also created like Procfile, requirement.txt file and setup file which is being described in configuration manual and artefacts file.

[3]Source link : `https://streamlit.io/`
[4]Source link : `https://docs.python.org/3/library/pickle.html`

	drugName	most_reviewed
0	Sertraline	12910.0
1	Mirena	12470.0
2	Adipex-P	7960.0
3	Oxycodone	6255.0
4	Celexa	5544.0
5	BuSpar	5265.0
6	Clomid	4860.0
7	Topiramate	4670.0
8	Denosumab	4480.0
9	Amoxicillin	4080.0

Figure 9: Top 10 most reviewed drugs

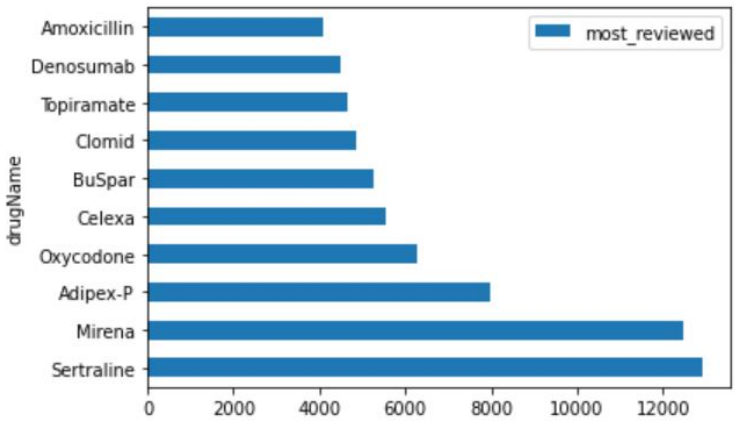

Figure 10: Barh graph plot

6 Evaluation

The Recommender system is based on the technique of Data Mining and Analysis. The suggestion made by these system are based on two widely used technique or method called Content-Based Recommender System and Collaborative Filtering based Recommender system. The product metadata or features which can be compared with other products metadata or features is called content based recommender system and the second method involves users profile on a particular product and matches with other users' profile for giving recommendation. Irrespective of these two methods, the evaluation metrices which is used in these cases are Similarity and Distance between two products and their score.

In this project, we have built a drug recommender system which is based on content-based recommendation technique. The metadata is created by concatenating different features of a drug. The evaluation metrices used in this project are: Pearson Correlation, Spearman's Correlation, Jaccard's Similarity, Cosine Similarity,Kendalls tau correlation, Euclidean Distance and Manhattan Distance.

Figure 9 and Figure 10 represents the Top 10 Drugs most reviewed and Barh graph plotted out of it.

6.1 Case Study 1 : Comparison between Text Normalization Technique

Since, in this project Patients reviews are given in the form of text (String Type) and metadata prepared is also text format. Hence, Natural Language Processing will be involved for Text normalization and further processing. Lemmatization and Stemming is implemented to check which can be better utilized for Count Vectorization. Below are the screenshot of results came when both implied one after the other. Remove or stem the last few characters of a word, which can lead to inaccurate meanings and misspellings, is known as stemming. In the context of the sentence, Lemmatization considers the word and transforms it to its meaningful base form, which is known as the Lemma. Multiple Lemmas can exist for the same term. Hence, in this project , Lemmatization was chosen

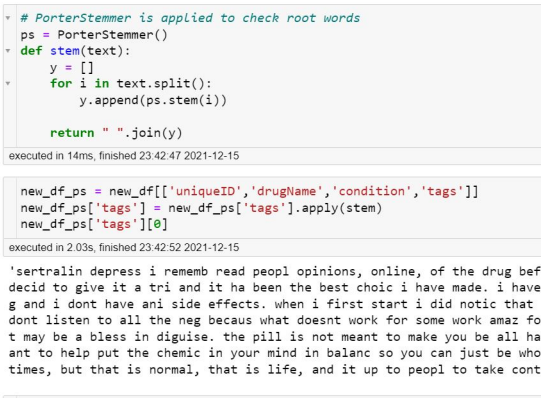

Figure 11: Output of PorterStemmer NLTK

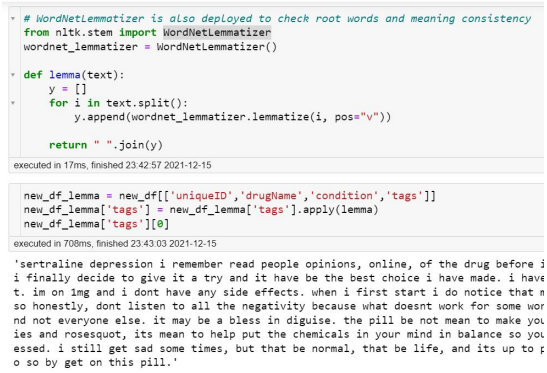

Figure 12: Output of WordNetLemmatizer NLTK

to move forward with text normalization of Tags/Metadata.

6.2 Case Study 2 : Word to Vector using Count Vectorization

Figure 13 shows the snippet of code written for Count Vectorization and Vector matrix. It can be seen that the array has high sparsity in it and this the trade-off point where content based recommendation system cannot overcome the problem of cold start and high sparsity in data. May be by using hybrid technique it can be achieved but this is not the objective of the project.

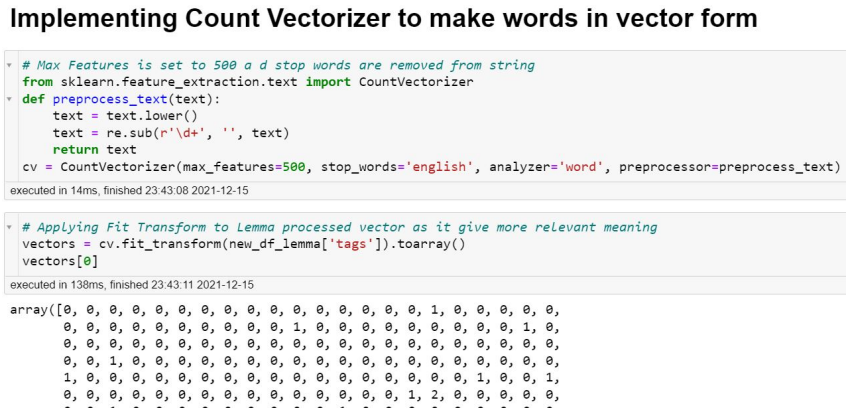

Figure 13: Count Vectorization done on Lemma Processed String

6.3 Case Study 3 : Comparison between all the Similarity Metrices

This section will show the various results obtained by calculating similarity scores between two drugs based on its content and suggest best similarity method to build recommender system. From the below tables it is evident that Jaccard's similarity have very good score as compared to any other method but it has bias towards vector similarity. Hence for this project Cosine Similarity was used for building recommender model as it has second

highest similarity score. Table 2 shows similarity score between two random drugs vector using different method

Table 2: Similarity Metrices Comparison Table

Methods	**Pearson**	**Spearman**	**Kendall Tau**	**Cosine**	**Jaccard**
Iteration 1	0.151	0.070	0.094	0.155	0.320
Iteration 2	0.062	0.027	0.053	0.198	0.327
Iteration 3	0.113	0.082	0.113	0.044	0.238
Iteration 4	0.049	0.120	0.155	0.162	0.226
Iteration 5	0.075	0.171	0.037	0.115	0.305
Iteration 6	0.079	0.017	0.064	0.048	0.323
Iteration 7	0.052	0.088	0.185	0.000	0.233
Iteration 8	-0.048	0.053	0.021	0.121	0.233
Iteration 9	-0.004	0.045	0.110	0.000	0.234
Iteration 10	0.029	0.056	0.151	0.136	0.302
Average	**0.056**	**0.073**	**0.097**	**0.098**	**0.274**

6.4 Discussion and Result

After Experiments and performing hyper-parameter tuning of various variables, it can be concluded that this recommender system is reliable and efficient in the field of Healthcare domain. The novelty of this research paper was to work on dataset of OTC drugs which has lots of users review. Using this dataset following stakeholders as mentioned in Introduction can get benefit and very less research paper is published on Healthcare recommender system. There are many research paper based on deep learning recommendation system which mostly deals with Movies recommendation or music recommendation. The problem of High Sparsity and Cold Start which is a millennium problem of recommender system is not resolved using content based recommendation. The Dataset of this project could have been more insightful if more features were available. Hybrid system inclusive of Content Boosted Collaborative Filtering probably can give much better recommendation which is future scope of this report.

The results obtained were quite nice score based on the Vectorization of words the drugs were recommended. The average cosine similarity score for drug recommendation came out to 0.098 which is comparatively decent than other similarity calculation methods available. Also, the recommender system was deployed in the cloud and website is hosted, hence apart from Data Analytics methodology this project utilizes the skill of cloud (Microsoft Azure) and deployment using Heroku (Salesforce App). It can be calculated from the website which is deployed that for any condition (disease/symptoms) selected by user the recommendation shown are correct upto 4 out of 5, which means accuracy of the model is 90percent.

7 Conclusion and Future Work

To answer the research question as upto what extent the content based recommender system tackles with the issue of cold start and high sparsity in data, then it can be considered that users review were quite persistent in this dataset, hence high sparsity

problem tried to be tackled on and gives a decent recommendation. But,if cold start problem is to be concluded then it is big challenge that any drug coming to the database has no review or any metadata for data analysis, hence it becomes the bottleneck of the system. This has also been discussed by (Wei et al.; 2021). The findings of this research paper was in accordance with the related literature review provided. Those researches were totally aligned with recommender model building but in the fields other than medical domain.

Using this model the stakeholders can get benefits in the time of crisis. Since, Doctors profession is said to nobel profession, hence working in this field will definitely be a nobel task. This experiment has contributed fully to the knowledge in the field of Data Analytics as well as Cloud Technology.

The limitation of the recommender model is still cold start and Adaption with changing users' behaviour. May be hybrid techniques could solve the problem which is future scope of work. May be utilizing more deep learning algorithm and ever changing IT environment can solve the limitations of this project. Big Data handling using Cloud Technology may result in faster processing and better accuracy can be achieved.

8 Acknowledgement

First and Foremost I wanted To thank almighty God for blessing me with the love and giving me ability to learn, perform better and implement my task successfully. Secondly, I would like to thank my parents for their never-ending believe in me and supporting me with everything throughout my journey. I would also, like to thanks professors and non-teaching staff of National College of Ireland for helping me out whenever needed. A special thanks to Dr. Catherine Mulwa and Noel Cosgrave and Dr. Anu Sahni for making it clear the details of the topic chosen and what is required in master research project. My personal gratitude to Aaloka Anant who was my supervisor for the Research project for his ever extending help and guidance whenver needed throughout the period. Thanks to all the authors of the research paper cited in the report.As a last note, I'd want to mention that the entire master's journey involved a lot of rigorous, challenging effort and attention to the study. The 13 weeks of my final thesis project were extremely tough and thrilling, but with the help of my mentors, family, and friends, I was able to complete it.

References

Akoglu, H. (2018). User's guide to correlation coefficients, *Turkish journal of emergency medicine* **18**(3): 91–93.

Bhavana, P., Kumar, V. and Padmanabhan, V. (2019). Block based singular value decomposition approach to matrix factorization for recommender systems, *CoRR* **abs/1907.07410**.
URL: *http://arxiv.org/abs/1907.07410*

Bodhankar, P. A., Nasare, R. K. and Yenurkar, G. (2019). Designing a sales prediction model in tourism industry and hotel recommendation based on hybrid recommendation, *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 1224–1228.

Dai, Q., Hong, X., Cai, J., Liu, Y., Zhao, H., Luo, J., Lin, Z. and Chen, S. (2018). Deep learning based recommendation algorithm in online medical platform, *in* J. Ren, A. Hussain, J. Zheng, C.-L. Liu, B. Luo, H. Zhao and X. Zhao (eds), *Advances in Brain Inspired Cognitive Systems*, Springer International Publishing, Cham, pp. 34–43.

Dong, X., Yu, L., Wu, Z., Sun, Y., Yuan, L. and Zhang, F. (2017). A hybrid collaborative filtering model with deep structure for recommender systems, *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, AAAI Press, p. 1309–1315.

Elbir, A., Bilal Çam, H., Emre Iyican, M., Öztürk, B. and Aydin, N. (2018). Music genre classification and recommendation by using machine learning techniques, *2018 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pp. 1–5.

Fararni, K. A., Nafis, F., Aghoutane, B., Yahyaouy, A., Riffi, J. and Sabri, A. (2021). Hybrid recommender system for tourism based on big data and ai: A conceptual framework, *Big Data Mining and Analytics* **4**(1): 47–55.

Fletcher, S. and Islam, M. Z. (2018). Comparing sets of patterns with the jaccard index, *Australasian Journal of Information Systems* **22**.
URL: *https://journal.acs.org.au/index.php/ajis/article/view/1538*

Guan, X., Li, C.-T. and Guan, Y. (2017). Matrix factorization with rating completion: An enhanced svd model for collaborative filtering recommender systems, *IEEE Access* **5**: 27668–27678.

Gurbanov, T. and Ricci, F. (2017). Action prediction models for recommender systems based on collaborative filtering and sequence mining hybridization, *Proceedings of the Symposium on Applied Computing*, SAC '17, Association for Computing Machinery, New York, NY, USA, p. 1655–1661.
URL: *https://doi.org/10.1145/3019612.3019759*

Han, Q., Ji, M., de Rituerto de Troya, I. M., Gaur, M. and Zejnilovic, L. (2018). A hybrid recommender system for patient-doctor matchmaking in primary care, *CoRR* **abs/1808.03265**.
URL: *http://arxiv.org/abs/1808.03265*

Khoie, M. R., Sattari Tabrizi, T., Khorasani, E. S., Rahimi, S. and Marhamati, N. (2017). A hospital recommendation system based on patient satisfaction survey, *Applied Sciences* **7**(10).
URL: *https://www.mdpi.com/2076-3417/7/10/966*

Pal, A., Parhi, P. and Aggarwal, M. (2017). An improved content based collaborative filtering algorithm for movie recommendations, *2017 Tenth International Conference on Contemporary Computing (IC3)*, pp. 1–3.

Shaikh, M. I. (2020). *Top-n nearest neighbourhood based movie recommendation system using different recommendation techniques*, Master's thesis, Dublin, National College of Ireland.
URL: *http://norma.ncirl.ie/4418/*

Stark, B., Knahl, C., Aydin, M. and Elish, K. (2019). A literature review on medicine recommender systems, *International Journal of Advanced Computer Science and Applications* **10**(8).
URL: *http://dx.doi.org/10.14569/IJACSA.2019.0100802*

Tran, T. N. T., Felfernig, A., Trattner, C. and Holzinger, A. (2021). Recommender systems in the healthcare domain: state-of-the-art and research issues, *Journal of Intelligent Information Systems* **57**: 171–201.

Wang, H., Wang, N. and Yeung, D. (2014). Collaborative deep learning for recommender systems, *CoRR* **abs/1409.2944**.
URL: *http://arxiv.org/abs/1409.2944*

Wei, Y., Wang, X., Li, Q., Nie, L., Li, Y., Li, X. and Chua, T. (2021). Contrastive learning for cold-start recommendation, *CoRR* **abs/2107.05315**.
URL: *https://arxiv.org/abs/2107.05315*

Zhang, S., Yao, L., Sun, A. and Tay, Y. (2019). Deep learning based recommender system: A survey and new perspectives, *ACM Comput. Surv.* **52**(1).
URL: *https://doi.org/10.1145/3285029*

Zhao, X. (2019). A study on e-commerce recommender system based on big data, *2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)* pp. 222–226.