Monolingual Sentiment Analysis on Pharmaceuticals Drug Reviews to Recommend Pain Medication to Patients

Sharon Wanjiru Ndichu

118461

ICS 4 C

Supervisor Name

Tiberius Tabulu

Submitted in Partial Fulfilment of the Requirements of the Bachelor of Science in Informatics and Computer Science at the School of Computing and Engineering Science

Strathmore University

Nairobi, Kenya

December 2022

Declaration and Approval

I declare that this is work produced from my own research and understanding on this project. It is not subject to theft of any publication or writing of other people. Where information from other sources is used, it is well referenced in the research document.

Student Name: Sharon Wanjiru Ndichu

Admission Number: 118461

Student Signature: ________________________ Date:________________________

The Document of **Sharon Wanjiru Ndichu** has been reviewed and approved by **Mr. Tiberius Tabulu**

Supervisor Signature: ________________________ Date:________________________

Acknowledgement

I express my gratitude to everyone who assisted me in one way or the other with the completion of my documentation. First and foremost, I express my appreciation to Strathmore University for offering resources such as library materials to assist me with my research. In addition, I appreciate Mr. Tiberius Tabulu, my project supervisor, for his guidance throughout the project documentation and development. I appreciate my friends who assisted me in the project and to my parents for providing support and resources needed for my project. Lastly, I am grateful to God for giving me good health and strength to complete this project.

Abstract

As the world's population grows, fewer health facilities and professionals are available to treat each individual. Due to a lack of resources, it is difficult for everyone to visit a health facility for medication. Also, there have been instances where individuals have died or experienced negative side effects because of the drugs they were given. Recommendation using reviews has been critical in most companies for individuals and also for improvements from companies. However, there is little research and projects using NLP in the medical field. This project recommends drugs based on sentiment analysis from users' reviews. A pre-existing dataset from Kaggle on patients' pharmaceutical drug was analyzed was used to train and test the model. The recommender system used TextBlob with TF-IDF vectorization for sentiment analysis and the classification model that was used was XGB and LGBM. The methodology used was Design Thinking. The solution was aimed to help those who self-medicate and also hiccups in accessing a specialist due to low doctor to patient ratio. The model gave a list of drugs and the final score calculated from the sentiment analysis combined with rating and useful count.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ANN- Artificial Neural Network

API – Application Programming Interface

EMRs- Electronic Medical Records

ERD- Entity Relationship Diagrams

ICD- International Classification of Disease

IFCR- Implicit Feedback and Crossing Recommendation

LGBM – Light Gradient Boosting Machine

NLP- Natural Language Processing

OTC – Over the Counter

RNN- Recurrent Neural Network

SDG- Standard Development Goal

SSAD- Structured systems analysis and design method

SVM- Support Vector Machine

UNII- Unique Ingredient identifier

TF-IDF- Term Frequency-Inverse Document Frequency

XGB -Extreme Gradient Boosting

Chapter 1: Introduction

1.1 Background Information

Technology has transformed the world into a global village in which individuals exchange their thoughts and knowledge, which is readily available when required. Health information is one of the most frequently searched topics on the internet According to research, 60% of people searched the internet for health-related topics, with 35% of them searching for diagnosis (Rao et al., 2020). To avoid negative effects including death, the health information available must be accurate. Accurate drug recommendation was done by processing the data available on the internet to more useful information using natural language processing language such as sentiment analysis on pharmaceutical drug review.

Currently, the doctor-to-patient ratio is limited. Since it takes 6 to 12 years to become a doctor, the number of doctors does not increase at the same rate as the population (Rao et al., 2020). As a result, there aren't enough doctors to treat each patient and the patients mostly self-treat for medication through methods such as over the counter medication and seeking information from the internet to self-medicate due to the low ratio. To provide information to patients who are unable to contact a specialist owing to this scarcity, an aid for drug recommendation was necessary.

There are also clinical mistakes which arise from misdiagnosis of a patient. About 40% medical practitioners make prescription mistakes, because they reference to only what they know which is limited (Wittich et al., 2014) . The number of medical facilities and resources available are not enough for the entire population, hence some of the diseases such as pain can be self-diagnosed. Having a medicine recommendation model from drug reviews helped to solve this gap. The drug reviews are feedbacks given by previous patients on drugs and the information was analyzed using sentiment analysis which is a natural language processing on textual data to get the attitude of a user.

Additionally, new discoveries and test on drugs are being made every day. There are more drugs that are added to the existing ones and some being banned for usage. It becomes progressively difficult for doctors to make a recommendation from only treatment or medication (Rao et al., 2020). Therefore, there was a need of a recommender model on drugs from the patients who have tried these medications to boost the already existing knowledge.

With increased use of the internet, there is a need to make use of the varied information that people have given. Some pharmaceuticals have included review sections to their products as one of the methods to measure performance and obtain feedback from customers. Many individuals rely on reviews before purchasing a product because they show the level of trust that other customers have in it. Monolingual sentiment analysis is identifying and categorizing opinions expressed in a single language text to determine the attitude towards a product (L. Zhang & Liu, 2017). Sentiment analysis was used in drug recommendation to know the attitude of users towards a medication and drugs recommended using other user satisfaction.

A drug review sentiment analysis model aids in providing individuals with additional medical recommendations. It can improve recommendation because other patients can use other patients' opinions to choose the drugs with the best feedback. Also, it can help specialist know the feedback of patients on drugs they administer and recommend the best medication to their users. SDG goal 3 of good health and well-being is attained because it reduces the hiccups experienced in medicine administering by a specialist or self.

1.2 Problem Statement

Currently, most of the health informatics available on the internet is not credible, just raw data from users. Most of the patients depend on recommendation from people, pharmacists or opinions on the internet. This prevents people from accessing good quality information (Goyal et al., 2020). The problem that was addressed in this model is lack of credible data from the internet for self-treatment.

The solution used the reviews, rating and useful count given on pharmaceutical drugs to know the feedback from users and the solution to be used for recommendation to other patients. The features were used as a measure of satisfaction. It used sentiment analysis library TextBlob and TF-IDF to do sentiment analysis and provide an accurate analysis for medicines.

1.3 Objectives

1.3.1 General Aims

The general objective is to do sentiment analysis on pharmaceuticals drug reviews to recommend medication for pain to other patients.

1.3.2 Specific Objectives

i. To research the parameters considered for drug recommendation to patients and ways in which they are recommended.

ii. To investigate algorithms used for NLP.
iii. To investigate the challenges associated with current drug recommendation systems.
iv. To review the solutions that recommend drugs to a patient.
v. To develop a drug recommendation model using sentiment analysis on pharmaceuticals drug reviews.
vi. To validate the model deployed using FastAPI.

1.4 Research Questions

i. What are the parameters considered and ways of drug recommendation to patients?
ii. What are the algorithms used in NLP?
iii. What are the challenges associated with current recommendation of systems?
iv. What are the solutions that recommend drugs to a patient?
v. How was the proposed solution developed?
vi. How was the solution deployed using FastAPI?

1.5 Justification

There are many challenges facing drug recommendation for patients. Some of the challenges are low ratio of a specialist per patient(Rao et al., 2020), the absence of enough health facilities and resources for each patient (Wittich et al., 2014) and clinical mistakes from specialists(Rao et al., 2020).With drug recommendation model , patients do not need to see a specialist unless symptoms persist or are critically ill hence reducing the hiccup of low ratio of specialist per patient and lack of enough facilities and resources for patients. Moreover, clinical mistakes can be reduced by pharmacist knowing which medications has the best feedback from patients and which ones do not have positive feedback.

The solution used drug review from patients, to able to quantify their sentiments with a score. With this analysis, the model gives a list of the drugs and their analysis for patients or specialist to boost the already existing recommendation system.

1.6 Scope and Delimitations

The project mostly focused on drugs to treat pain because it is a common ailment that patients experience and the sentiment analysis of the reviews of the medications. The proposed drug recommender model will give a list of medications associated with pain and the results from sentiment analysis on the reviews.

1.7 Limitations

This project is only limited to giving a list of medications and their analysis which is the score from sentiment analysis but the actual prescription needs the consultation of a pharmacist. This is because patients require recommendations based on other characteristics such as age, which the recommender model will not provide.

Chapter 2: Literature Review

2.1 Introduction

This chapter discusses the parameters used in drug recommendation, drug recommendation avenues, solutions for recommending drugs to patients, as well as the challenges associated with these drug recommendations avenues and solutions. It also explores related works in drug recommendation systems and gaps in related works. Conceptual framework is also illustrated in this chapter.

2.2 Parameters Used in Drug Recommendations

2.2.1 Composition of Drugs

The substances used in drugs is important to know the side effects or allergic reactions that may occur when recommending a drug (Rainsford, 2009). Additionally, knowing the constituents of drugs is important to prevent extreme cases such as fatalities which may arise from the adverse effects of a drug.

2.2.2 Age

Age is a crucial factor when recommending medicines. Different age groups take different medications because the age can affect how drugs are absorbed and broken down by the body(Mangoni & Jackson, 2004). Some medications may be either too strong or too weak for an individual based on age, hence age is a factor considered in medicine recommendation.

2.2.3 Side effects

Side effects of drugs is an important factor to consider before administering of drugs. The side effects may be contributed by factors such as other medication being taken and age of a patient. Medicine may cause adverse side effects if they are not considered such as dizziness, nausea failure of organs or even death (O'Donovan et al., 2019). Therefore, side effects a very important factor to consider in medicine recommendation.

2.2.4 Addictive components

Addictive components is a key factor to consider when administering drugs to patients. Addiction occurs when a drug causes pleasure caused by changes to the brain and liking to an effect of a drug (Jakovljevic et al., 2015). In addition, craving of drugs hence the addictive components of a drug are important to consider in recommendation of drugs.

2.3 Drug Recommendation Avenues

2.3.1 Over the Counter Recommendation

Over the counter medicines are common and mostly use factors such as age and regular medicines to treat an ailment, with the most common over the counter medication including: cough medications, painkillers, codeine-based medicines, sedatives, antihistamines and decongestants(Cooper, 2013). The danger of using this method include adverse side effects such as dizziness and nausea from medicine because of the composition of the drugs. Also, allergic reactions such as rashes or itching may occur due to the constituents of a medicine (Chautrakarn et al., 2021). With such kind of recommendation, effects are not clearly considered and if the drug is worth to buy.

2.3.2 Recommendation from Information Sought from the Internet

Most consumers utilize online health information to self-treat. Health information is widely available on the internet. According to Chautrakarn (Chautrakarn et al., 2021), most of the medication information is incorrect hence it can become a source of poisoning which can cause death. In addition, there might be side effects and allergic reactions because the composition of medicines is not considered.

2.4 Algorithms Used in NLP

2.4.1 Support Vector Machine

A support vector machine is a supervised machine learning model that can use both classification and regression algorithms. However, it is mostly used in classification problem. SVM plots each point in n-dimensional space where n is number of features with the value of the feature being the value of a particular co-ordinate data(Manning et al., 2009). The decision boundary is called Hyperplane that performs classification that differentiates two classes.

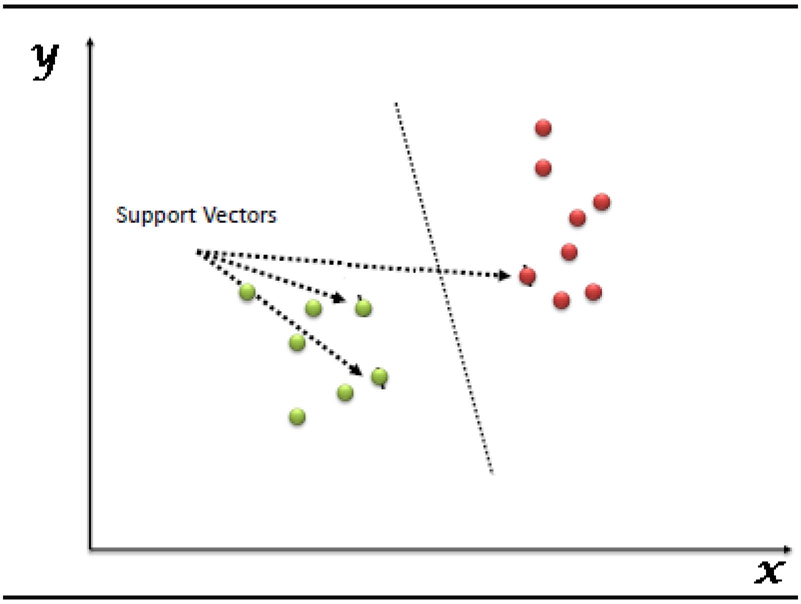


Figure 2.1 Support Vector Diagram

2.4.2 Recurrent Neural Network

A recurrent neural network uses sequential data or time series data for artificial neural network. Common applications of these deep learning techniques include language translation and natural language processing, which are ordinal or temporal problems. The information on neural network loops back on itself. It takes into account both the current input and the lessons it has learnt from prior inputs when making a decision. In this case, RNNs can have a certain amount of internal memory and are suitable for processing sequential data (Goodfellow et al., 2016).

2.5 Challenges Associated with the Current Drug Recommendation

The current methods of drug recommendation do not consider insights into those affected. There isn't any qualitative research on the medications' effects. Medicines compositions are not considered hence there is probability of side effects and allergic reactions. A study by Cooper (Cooper, 2013), has shown OTC use proxy, self-report and non-OTC specific data. Moreover, there is a challenge identifying credible data from information sought from the internet(Cline & Haynes, 2001), which makes the current recommendation method not suitable.

2.6 Related Works

2.6.1 Ontology-Based Drug Recommendation

A research by Doulaverakis (Doulaverakis et al., 2012),used a semantic-enabled online service that can provide drug-disease interaction information. Medical data is converted to ontological form and compared to medical knowledge using international standards such as ICD-10 and UNII. It makes a drug recommendation based on the infection, sensitivity, and drug interactions of the patient. However, this system utilizes a lot of system resources which is undesirable when implementing. It also requires a lot of medical expertise from a developer hence may not solve the problem of the evolving drugs which may not capture the effects on a patient.

2.6.2 Drug Recommendation System by Implicit Feed Back and Crossing Recommendation

According to chen (Chen et al., 2018) , this recommendation is done using the Electric Medical Records. The EMRs are designed to implement collecting, searching, statistical analysis and drug diagnosis. IFCR carries out deep analysis to achieve the most effective drug. IFCR involves three steps, the first is raw representation from EMRs, then model the representations by a non-negative matrix factorization for a set of robust features. Finally, the extracted features are used for the recommendation. The IFCR uses the cross recommendation where there are several symptoms. Figure 2.2 below is a sample pseudocode for cross recommendation:

Algorithm 1 Crossing Recommendation Model

Inputs:

```
SymptomList: the list Symptom for patients;
a:the number for one Symptom;
b: the totel number for the drugs
```

Outputs:

```
The list of drugs
function CROSSINGRECOMMENDATION(SymptomList, a, b)
    CRDict ← null
    for all symptom ∈ SymptomList do
        SingleRecommendationDict ← Matrix[symptom]
        SingleRecommendationDict ← TOPK(SingleRecommendationDict, a)
        for all (medicine, rate) ∈ SingleRecommendationList do
            rate ← (√rate + 1)/a
            if medicine ∈ CRDict then
                CRDict[medicine] ← CRDict[medicine] + rate
            elseCRDict[medicine] ← rate
            end if
        end for
    end for
    return TOPK(CRDict, b)
end function
```

Figure 2.2 Cross Recommendation Model

However, the IFCR does not incorporate the patient's emotions on the drugs, it is only based on the EMRs which do not include reviews. It therefore can recommend a drug but the recommendation is not based on the view of other patients.

2.6.3 Drug Recommendation based on Tensor Decomposition

A research by Zhang (Y. Zhang et al., 2014) , used an algorithm based on tensor decomposition. It uses the 'User-Item-Tag' three tuple to model a tensor. Personalized recommendation can be received by patients based on extracted important tensor according to a drug predicted rating. Evaluation index used in this model is accuracy and recall. Finally, Top-N drug recommendation list is gotten from each user using tensor decomposition. It is then checked if the drug is in the recommendation list with the patient and tags. Figure 2.3 shows 'User-Item-Tag' three tuple tensor:

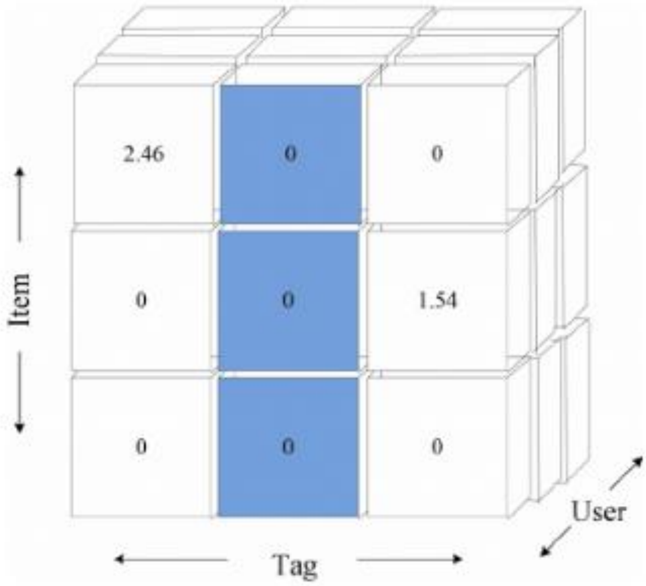


Figure 2.3 “User-Item-Tag” three tuple tensor

However, this analysis has shortage of collaborative filtering when dealing with big and sparse data. It makes use of information such as the name, description, and rating. This approach does not allow for the analysis of user sentiment about a medication.

2.7 Gaps in Related Works

Most of the works are focused on prescribing drugs based on patient records, ratings, and symptoms. They also use the expertise on drug such as the ontology-based recommendation system. However, these works do not include patients' sentiments, which can be utilized to determine a patient's emotional response to a drug, as well as whether it is successful and if they are satisfied with it.

2.8 Conceptual Framework

The conceptual diagram below shows how the Kaggle drug review dataset passed through data cleaning and preprocessing. In the data cleaning and preprocessing there was checking of null values, removal of unnecessary values such as punctuations and checking duplicate values. Text tokenization was also done in the data preprocessing. The cleaned data was split into 75% training dataset and 25% testing dataset. The training set was used in building of the model and it needs a lot of data to learn hence a higher ratio. The test dataset uses unseen data to test if the model is working as required. After the data cleaning and preprocessing, sentiment analysis was done on the data which involves giving a score of polarity and sentiment. The sentiment analysis used LGBM and XGB to classify the sentiments. The model also used other features such as rating and useful count for final score to cluster the medication for pain and deployed on FastAPI where a user can see the list of the recommended drugs to treat pain after choosing a condition.

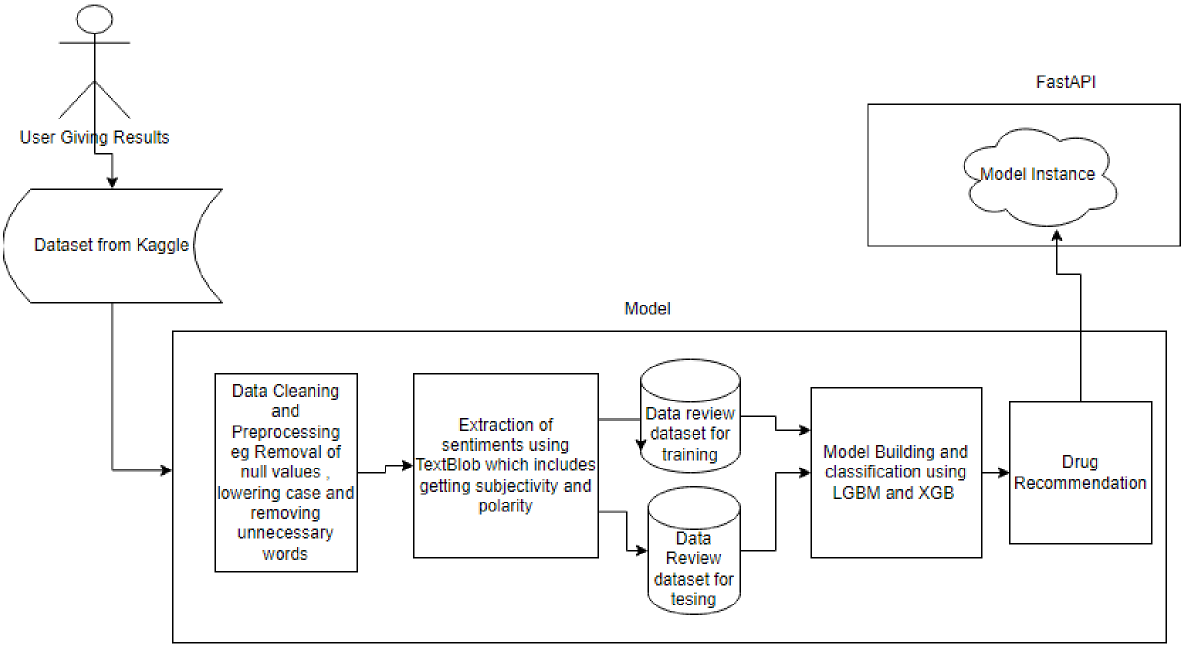


Figure 2.4 Conceptual Diagram

Chapter 3: Methodology

3.1 Introduction

This chapter covers the methodology approach that was used in this application development and the steps involved in the development approach. Moreover, the chapter covers analysis and design diagrams.

3.2 Methodology

The methodology that was used in the model development is design thinking. It is a method of problem-solving that puts the requirements of the users first. Solution-based approach is used to solve problems in design thinking, where it places more of an emphasis on finding solutions to problems unlike a problem-based approach that looks for limitations on why a problem exist (Raju, 2021). Figure 3.1 shows the design thinking methodology.

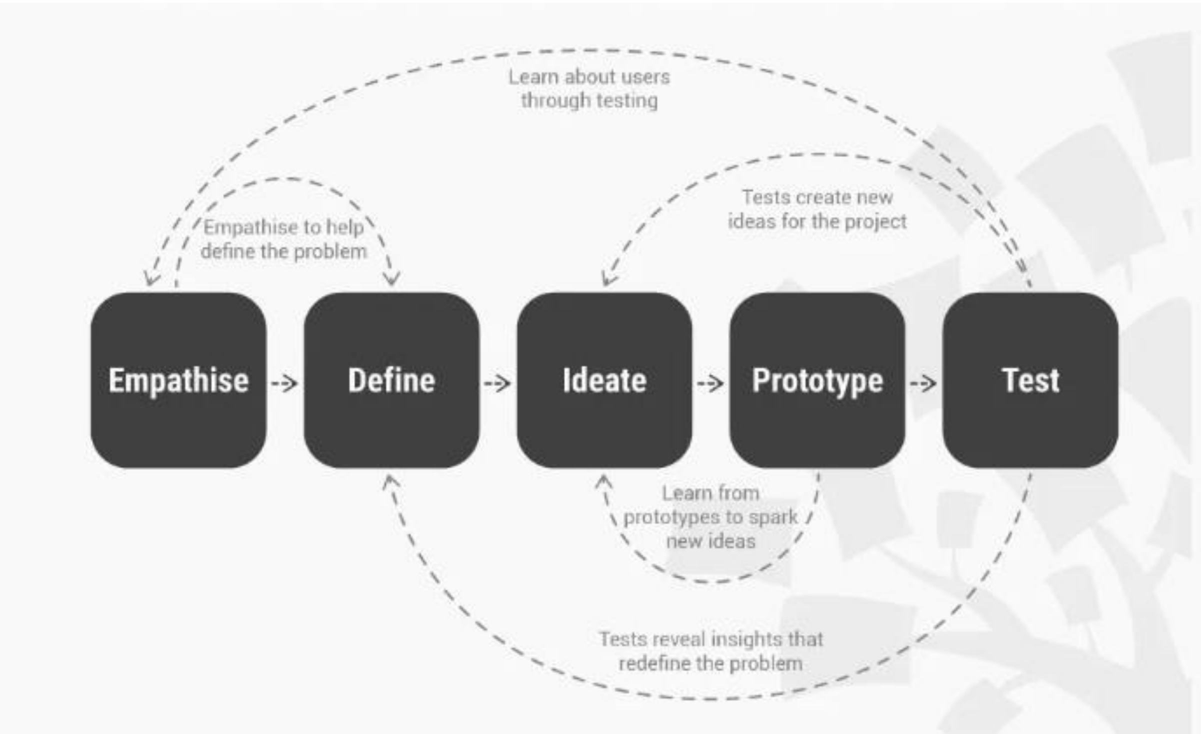


Figure 3.1 Design Thinking Methodology

3.2.1 Empathize

Empathy was the first step in the design thinking methodology because it helps understand the problem that is trying to be addressed. A designer observes or studies with empathy how users are affected by a situation. This stage is important for knowing the user's requirements besides the problem to provide a more personalized solution.

3.2.2 Define

In the second step, findings from the empathize stage are used to define the problem that is trying to be solved. Considered factors include the challenges users face, the issues they encounter frequently, and how a problem affects them. This step is important because problems can be defined once they have been synthesized.

3.2.3 Ideate

In this step, brainstorming of how to solve the problems defined takes place. This stage tries to create new ideas, whereby there may more be one than more idea. Possible problems from the user environment are considered when suggesting a proposed solution. Analysis diagrams using SSAD design paradigm in section 3.3 which include the use case diagrams, sequence diagram, system sequence diagram, ERD, context diagram and the dataflow diagram was used to ideate how a user and the model interacts. In addition, design diagrams in section 3.4 which includes the Dataset schema, wireframe and system architecture was drawn in this stage to show the visualization of how the project is supposed to be. This step is important because the best solution for a problem is sought.

3.2.4 Prototype

This step involves actual development of a model. Effectiveness of a solution is tested in this stage. All the possible solution may be tested in this step and the effectiveness checked. This stage is important because the less effective option can be dropped and the best solution adopted for use. Tools and techniques discussed in section 3.6 were used in the development of the prototype. The tools and techniques that was used include the FastAPI, Google Collaboratory, sentiment analysis, python language, Kaggle dataset and the GitHub repository. These tools and techniques are important for the development of the prototype because they provide the platform, language and storage for the project.

3.2.5 Test

This final stage tests the best solution from the prototype. The end results of this stage are refined because it is an iterative methodology. Blackbox testing , integration testing and white box testing was used to test the model. The black box testing is done against specification to discover faults because of not having completely fulfilled specification whereas white box testing is done against implementation and discovers faults in the implementation. Accuracy testing is done on the model and results given in percentage to know if a model is overfitting or underfitting, this is done by dividing the number of correctly classified samples with total number of samples in the drug review dataset. Test cases was also drawn, the first test case is where the results of drugs are expected when the ailment is pain and a test case where no results is shown when the ailment is not pain.

3.3 Analysis Diagrams

The analysis methods that were used were based on the SSAD methodology. The analysis diagrams that were used in this model are use case, sequence diagram, system sequence diagram, ERD, context diagram and the data flow diagrams level 0 and 1.

3.3.1 Use Case Diagram

A use case diagram describes a system and how the users of a model uses it but not the actual workings of the model. The requirements of the entire application or a specific portion are described. It displays the system's interactions with both internal and external actors(Waykar, 2015).It was applied in this application to display user-system interactions.

3.3.2 Sequence Diagram

A sequence diagram describes the events in a project and the specific order the project follows. It shows how these processes occur simultaneously. Horizontal lines with messages between them in the sequential sequence of their occurrence are placed between these activities (Al-Fedaghi, 2021). The sequence diagram was used in the model to show the activities in the model and how they follow each other.

3.3.3 System Sequence Diagram

A system sequence diagram shows the whole system sequence diagram. This includes the use case, external actors, and internal events of the system. It generally shows how the whole system works not only the internal processes of the application. It is used in the model to show how the application interacts with the external environment.

3.3.4 Entity Relationship Diagram

It is a relationship model that illustrates the entities to a system. It shows the relationship between the entities such as the people, objects, places and events within the system. The ERD in this model was used to show the relationship between various entities in the model (Li & Chen, 2009).

3.3.5 Context Diagram

A context diagram is used to show the entire system as one process. It illustrates the information flow between the system and external entities. A context diagram was used for a clear visualization of the whole application and how it works (Ibrahim & Yen, 2010).

3.3.6 Data Flow Diagrams, Level 1

In the data flow diagram level 1 the whole application is represented as a single process but sub-processes are added in the level 1 process.It is more descriptive and includes the processes of the context diagram ("Levels in Data Flow Diagrams (DFD)," 2019). It was used in the model to show a more detailed system process.

3.4 Design Diagrams

This is the process for defining the model, architecture, and their components. The system design satisfies specific requirements. The system design diagrams that was used are the database schema, wireframes, system architecture (Odhiambo, 2018).

3.4.1 Dataset Schema

It refers to the visual representation of the drug review dataset. The dataset schema shows the entities that was used in the drug review sentiment analysis.

3.4.2 Wireframe

A wireframe is a framework that shows the design and functionality of a user interface. The project's wireframes were used to show the interface elements that are present on the relevant pages.

3.4.3 Python

Python programming language is a high-level language used in development. It is suitable to use because it is open source, scalable and easy to learn. It works well for natural language processing because of the rich processing features and simple syntax.

3.4.4 System Architecture

System architecture defines the behavior, the structure, the interactions, and the views of a system. It addresses the properties, concepts, architectural principles and characteristics of the model. The system architecture describes the non-functional decisions and the functional decisions of a system. It acts as the blueprint and shows the coordination and communication in the system.

3.5 Deliverables

3.5.1 Proposal

The proposal for the project was delivered, which provided an outline of how the project was expected to be. It consisted of the abstract, chapter 1, chapter 2 and chapter 3. The abstract gave an overview of what is expected in the project. Chapter 1 consisted of the background information of the project, problem statement, objectives, research questions, justification, scope, delimitations and limitations. Moreover, it covered chapter 2 which discussed parameters used in drug recommendation, drug recommendation avenues, related works and the gaps in this related works. This chapter also illustrated the conceptual diagram. Lastly, chapter 3 covered the methodology which includes the methodology, design paradigm that consists of the analysis diagrams, design diagrams, deliverables, the tools and techniques used in the project building. The Gantt chart was also presented to show the timeline of the project activities.

3.5.2 Model

The recommender model was delivered. The model did sentiment analysis using TextBlob which polarity and sentiment score. This feature was used together with rating and useful count to give a medication recommendation list to users.

3.5.3 Application Programming Interface

The API allows the model to accessed by the user. The model was interfaced on the FastAPI, then a user can access the model and use it. The user keys in the disease which is Pain and a list of the medication and sentiment analysis displayed.

3.6 Tools and Techniques

The tools and techniques discussed are the FastAPI, Google Collaboratory, sentiment analysis technique, python, GitHub repository and Kaggle dataset.

3.6.1 Google Collaboratory

Google Collaboratory is an environment hosted on google drive to provide the tool for coding and building the project model. It is suitable for machine learning projects and data analysis. Google Collaboratory allows the project to be hosted on cloud making it suitable for use.

3.6.2 FastAPI

FastAPI is a web framework for building restful APIs with python. The FastAPI was used to deploy the model and allow the users to interact with the model.

3.6.3 Kaggle

Kaggle is a community hosted online for data scientists and machine learning engineers. Kaggle provides dataset to be used in model training and testing.

3.6.4 GitHub Repository

GitHub repository is used to store a project’s development and collaboration. The GitHub repository can also have README file to give an insight of a project.

3.6.5 Sentiment Analysis

Sentiment analysis is a technique that tries to get emotions and opinions from a text. The sentiment analysis can either be negative or positive.

Chapter 4: System Analysis and Design

4.1 Introduction

In this chapter, system requirements which include the functional and non-functional requirements are discussed. Moreover, the system analysis and design diagrams in Chapter 3 are drawn and discussed.

4.2 System Requirements

System requirements are the specifications that are required by the system to make it functional and satisfy the user needs and make the system work. These requirements are both functional and nonfunctional. Some of the system requirements reviewed in the project include.

4.2.1 Functional Requirements

The functional requirements covers what the system is supposed to do. These functional requirements include:

i. Authentication Module

The authentication module is used to verify a user. The login and the registration page were used in the authentication. The emails, username and passwords are collected. An email can only be used once. The password used password hashing function to hash the password.

ii. User Interface

The user interface is used for a user interaction with the system. The user searches for drugs using a user interface hosted on the web and receive the recommendation list.

iii. Pain Medication Recommendation Model

The model recommends the most suitable medication to treat pain from users' sentiments. The model uses dataset from Kaggle to make recommendations. The model goes through processes before it can finally give a recommendation list. The first process includes data cleaning and preprocessing Extraction of sentiments is then done to the data. Splitting of the data into training and testing dataset for training and testing is done and a model is then built using TextBlob to cluster the pain medications.

4.2.2 Non-Functional Requirements

Nonfunctional requirements is used to define how the application behaves and the limits of its functionality. The non-functional requirements of this model includes:

i. Security

The authentication was used to make the model secure and less prone to hacking or illegal access. Password hashing was one of the method used to convert the entered password to a hash and when logging in the password is compared to the hashed password using the password verify functionality. Also, for security purposes on the authentication module, an email can only be used once.

ii. Model Accuracy

Accuracy is used to measure the correctness of the prediction made by the model. The model needs to have minimum errors for it to have high accuracy levels. This ensures that the users are not mislead by the predictions.

iii. Data Integrity

Data integrity ensures that the data used is accurate, consistent and reliable. This was done at the data cleaning and preprocessing stage of the dataset which includes removal of duplicate data, elimination of unnecessary words and removal of missing values. This ensures that the data used in making the model is of high quality.

iv. Performance

Performance was used to assess if the model accurately achieves the task of recommending the highest rated pain medication. Moreover, it checked how the model performed when it was deployed to a web application through the FastAPI.

4.3 System Analysis Diagrams

4.3.1 Use Case Diagram

The use case in Figure 4.1 shows how actors of the system interacts with the model. The actor is any user of the system who interacts with the model to get the recommendation by the system.

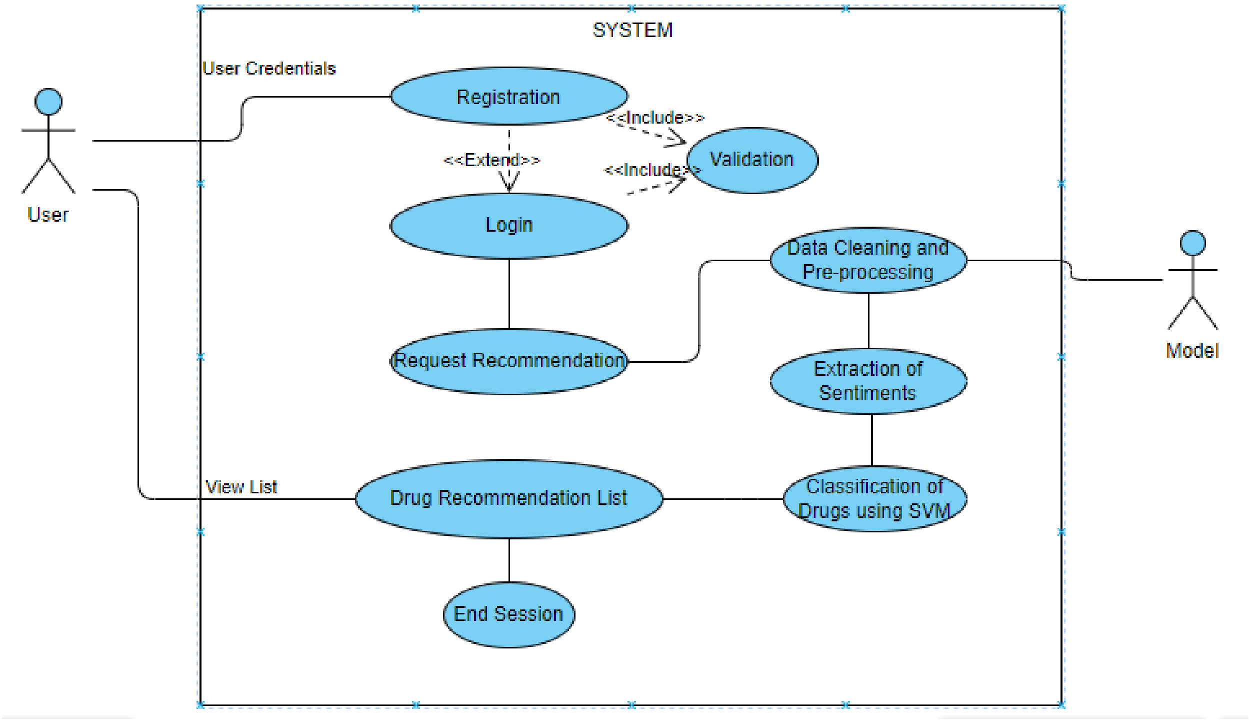


Figure 4.1 Use case Diagram

4.3.2 Sequence Diagram

A sequence diagram illustrates the sequence of messages between objects in an interaction. It describes how and in what order a group of objects works together. The sequence diagram is sshown in Figure 4.2:

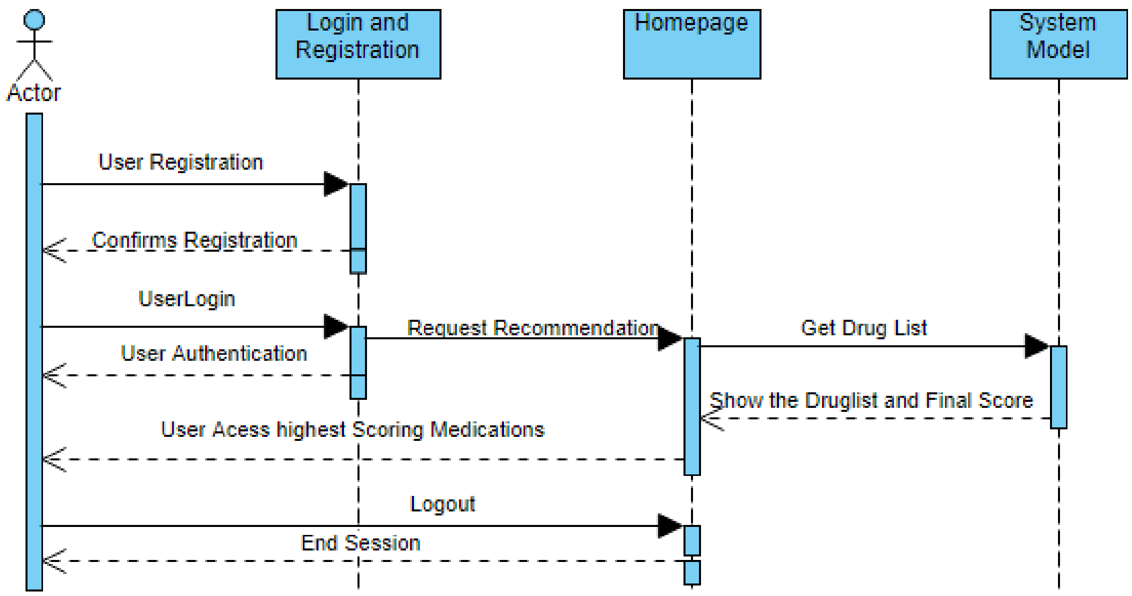


Figure 4.2 Sequence Diagram

4.3.3 System Sequence Diagram

System sequence diagram shows the interaction of the user with the system showing the input and output events. The Figure 2.2 shows the system sequence diagram:

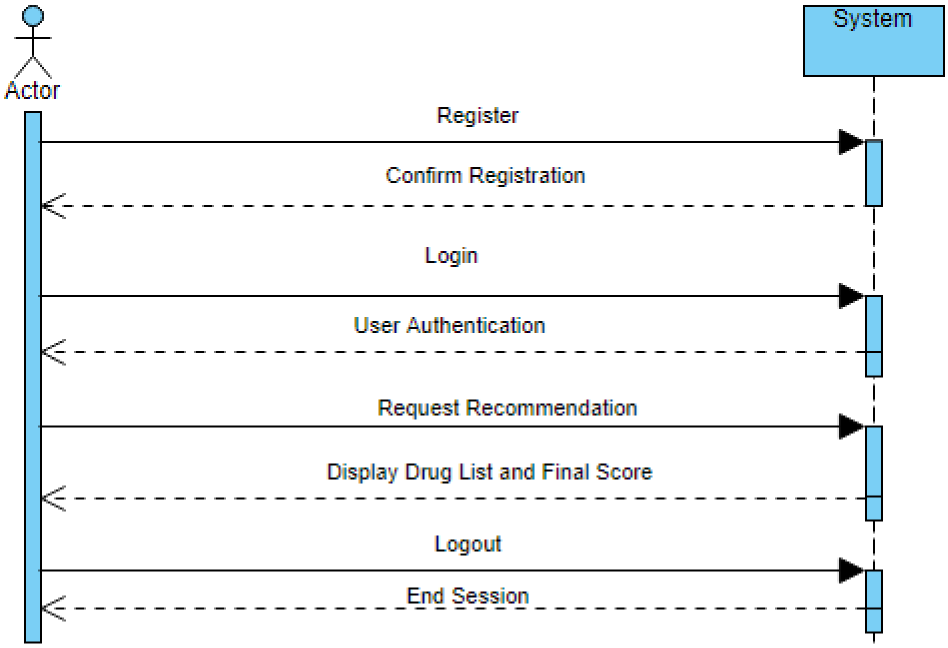


Figure 4.3 System Sequence Diagram

4.3.4 Entity Relationship Diagram

An Entity Relationship Diagram shows the relationship between each entity. It was used to show a model of the final system and attributes. The Figure 4.4 shows the ERD:

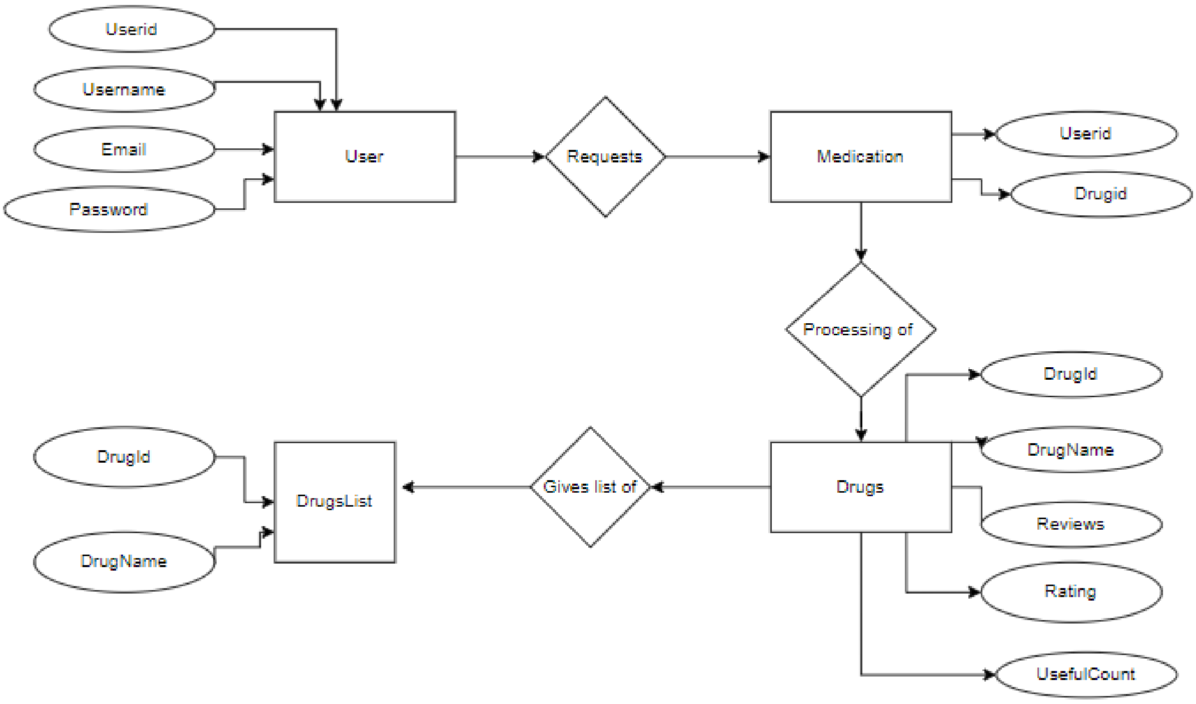


Figure 4.4 Entity Relationship Diagram

4.3.5 Context Diagram

A context diagram was used to display the system as a whole. It shows all the external entities and how they interact with the system. The application was put in the middle and the external entities that surround the system without going deep into the system.The Figure 4.5 shows the context diagram:

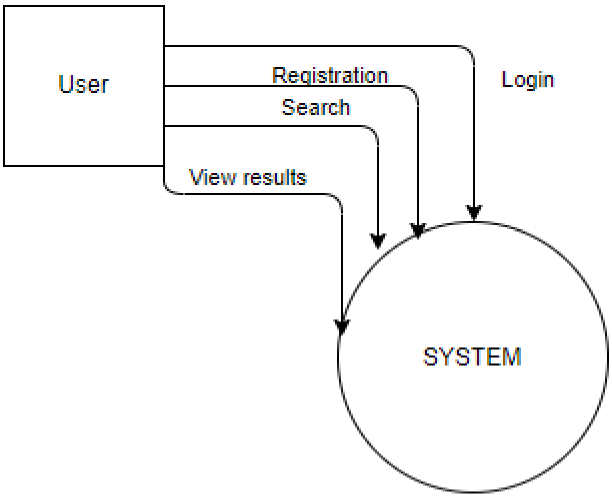


Figure 4.5 Context Diagram

4.3.6 Level 1 Diagram

The Level 1 one diagram shows how the entities of the application interact with each other and how they interact with the application.

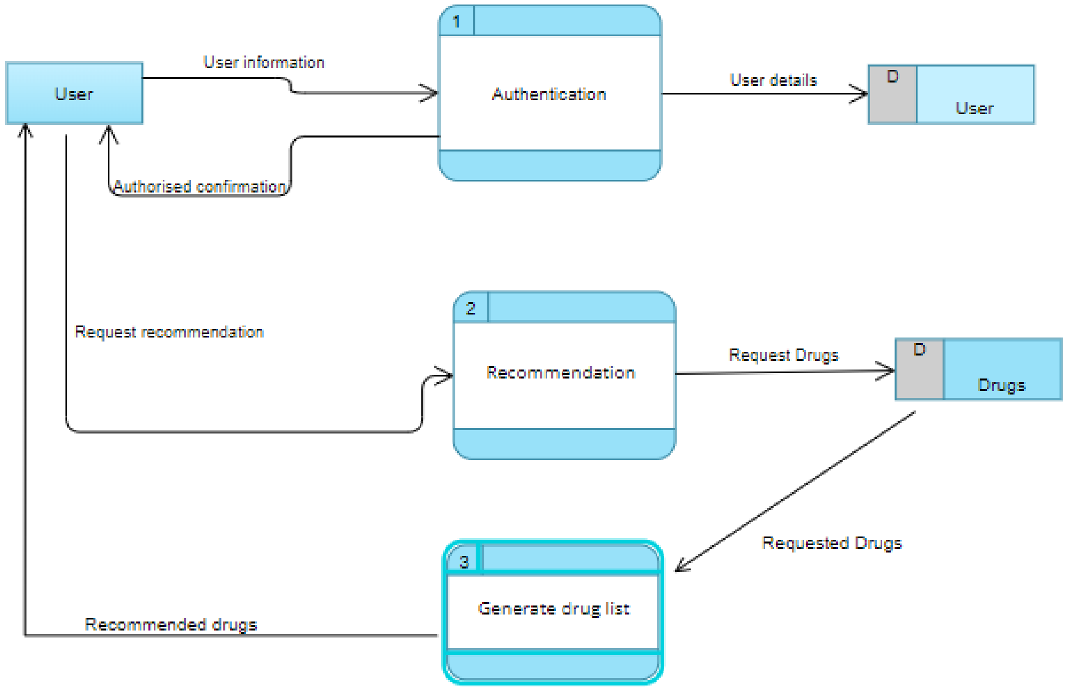


Figure 4.6 Level 1 Diagram

4.4 Design Diagrams

4.4.1 Dataset Schema

A database schema was used to show a blueprint of how the dataset was constructed. It was used to show the relationship between the datasets.

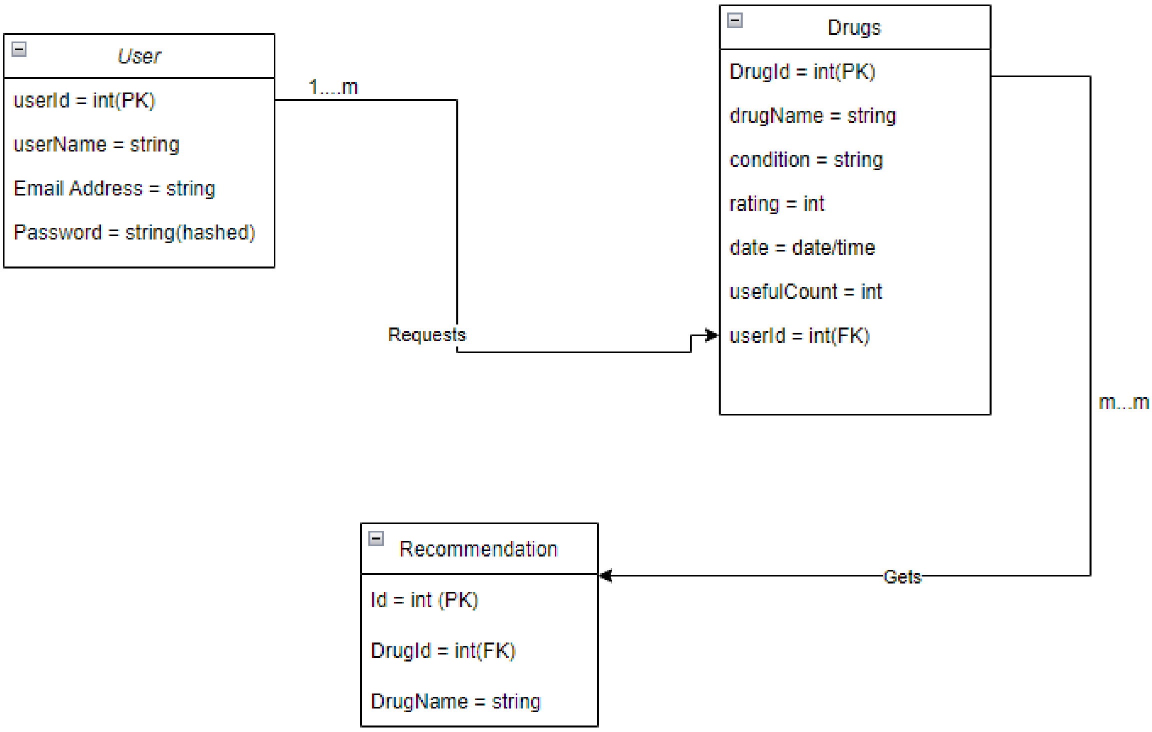


Figure 4.7 Database Schema

4.4.2 Wireframe

Wire frames shows different interfaces. It shows the models how different interfaces looks like. Below are some of the wireframes.

i. **Registration**

Figure 4.8 shows the registration wireframe.

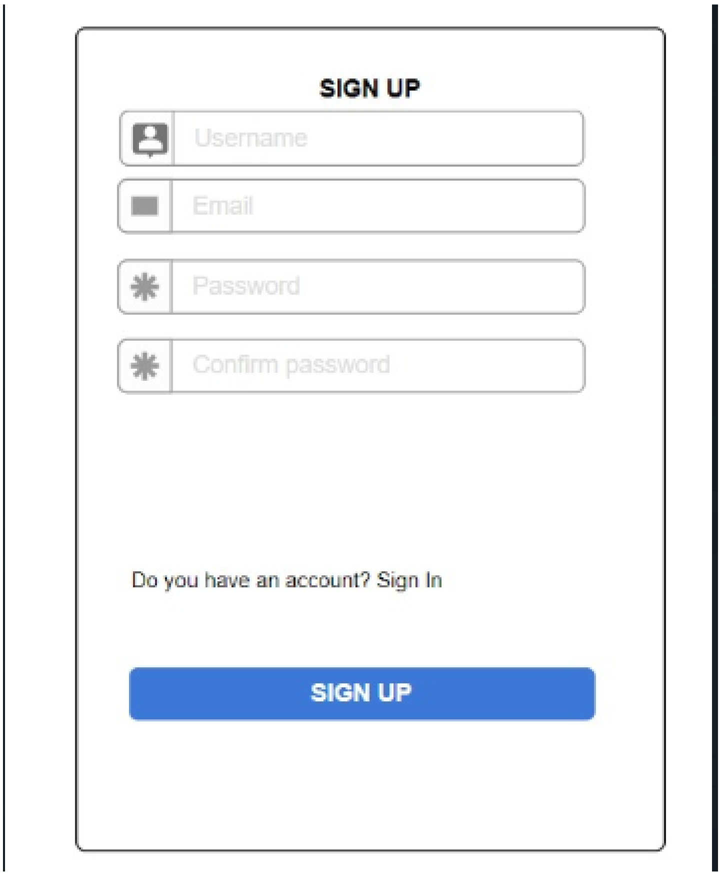


Figure 4.8 Registration Form

ii. Login

Figure 4.9 shows login framework.

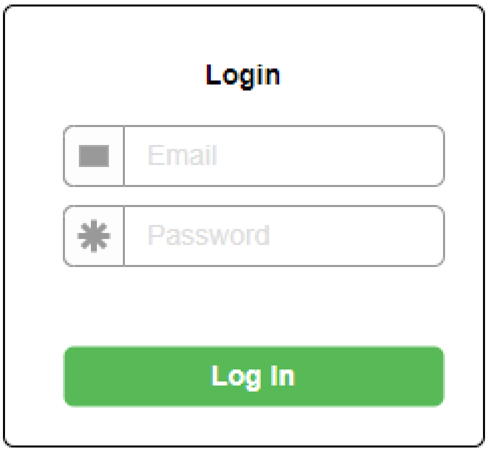


Figure 4.9 Login Page

iii. Search Input

Figure 4.10 shows the search input page wireframe.

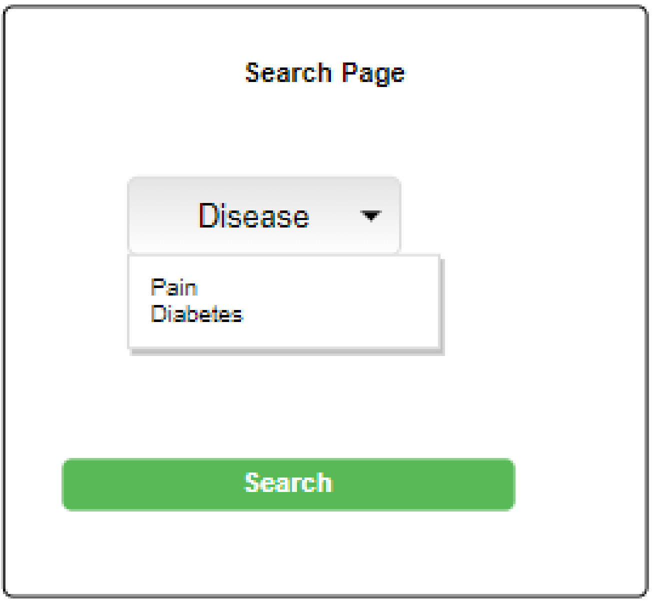


Figure 4.10 Search Input

iv. Recommendation List

Figure 4.11 shows the recommendation list framework

#	Drug Name	Score
1	Drug1	Score1
2	Drug2	Score 2
3	Drug3	Score 3
4	Drug4	Score 4
5	Drug5	Score 5

Figure 4.11 Recommendation List

4.4.3 System Architecture

The system architecture is used to show the blueprint of the application and the coordination of the application. The Figure 4.12 shows the system architecture.

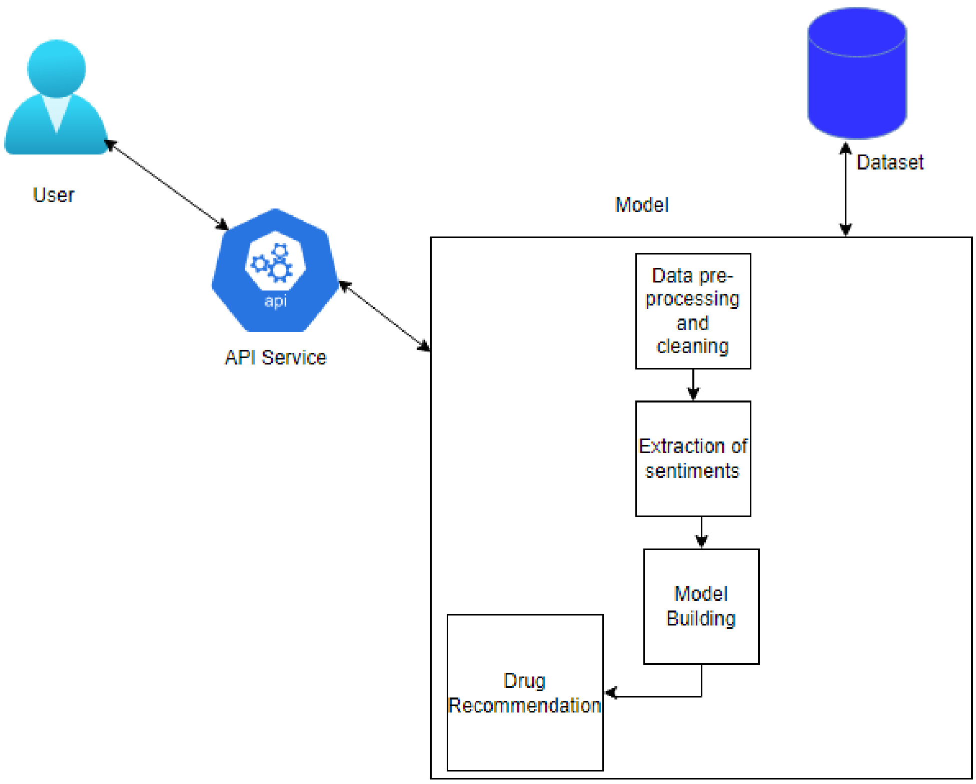


Figure 4.12 System Architecture

Chapter 5: System Implementation and Testing

5.1 Introduction

This chapter discusses the system implementation which includes description of the implementation environment, description of the dataset, testing paradigm and the testing results that were obtained.

5.2 Description of the Implementation Environment

The testing environment includes the hardware specifications and software specifications for computers and servers required to run the model.

5.2.1 Hardware Specification

Hardware specification shows the requirement that was required to run the model. Table 5.1 shows the hardware, its description and the justification.

Table 5.1 Hardware Requirements

Item	Description and Justification
RAM	4GB for higher processing speed
Processor	Intel Core i5 - -8250U CPU with 1.8 GHz for fast performance
Hard Storage	500GB of space to run the software for fast performance

5.2.2 Software Specification

Software specification shows the software requirement that was required to run the model. The Table 5.2 shows the software, its description and the justification.

Table 5.2 Software Requirements

Item	Description and Justification
Operating System	Windows 10 and 11 for compatibility with the API for deployment

IDE	Google Colab for writing and executing python code
ML software	Scikit Learn for classification
API	Fast API for deployment
Web Interface	Google Chrome, Mozilla Firefox
Libraries	NLTK , TextBlob to do natural language processing

5.3 Description of Dataset

The model's dataset was retrieved from Drug Review Dataset from UCI ML repository. The dataset can be found on Kaggle. The features of this dataset are unique ID, drug name, condition, review, rating, date and useful count shown in Appendix 7.

This dataset was chosen because it contains feedback from users which can be used in feature analysis and sentiment analysis. The most important features are the reviews, ratings and useful count because they are more of a user view on the medication. Some more features were added to assist in Natural Language Processing feature engineering which include the word count unique word count, punctuation count and stop word count. Word count was added for tokenization of words and identify unique words. Appendix 8 shows the added features.

There was also addition of sentiment and sentiment processed. The sentiment is from the uncleaned reviews and the sentiment processed feature which is from cleaned reviews done using TextBlob. Appendix 9 shows these columns.

The dataset has 215,063 entries of which was divided into 75% training dataset and 25% testing dataset to assess the performance of the model. There is a larger number of the training because the model need as much data as possible to for patterns which is meaningful. The patterns are used to make classification and prediction hence there is need of a higher ratio of training.

5.4 Description of Training

Supervised learning was used to training this dataset because each feature was associated with a label. The following stages were involved in training:

5.4.1 Data Preprocessing

The data preprocessing part involved removing null values because they were 0.55% making them almost insignificant and all were from the condition field. The reviews were also cleaned in text-preprocessing. The punctuations were removed to help treat text equally. Punctuations are used to add grammatical structure and not to show the meaning. The text was also lemmatized which involves converting words to their root form by using the nltk library. Lemmatization helps to group different inflected forms of words into root form, having the same meaning. Also, there was an extension of contractions which can be used to remove stop words. Finally, there was tokenization of the reviews. Tokenization is important because words are split into smaller units that can be easily assigned meaning.

5.4.2 Feature Extraction

Text can't be dealt with directly hence a good setup of the data is needed to build the classifier for sentiment analysis by converting into numerical form. There was use of TF-IDF vectorizer which gives low importance to the terms that appear often in the dataset. Term Frequency is how often a word appears. Inverse Document Frequency is an algorithm to determine how relevant words in a text are. TF- IDF is hence the multiplication of TF and IDF.

There was also the manual feature extraction which was done to increase the accuracy of the model. This includes useful count, condition encoded using a label encoder from Scikit library, rating and the sentiment and polarity extracted from the cleaned and uncleaned reviews.

5.4.3 Classification

LGBM and XGB classifiers were used for classification to predict sentiments. They were both trained on 2000 epochs because training accuracy with continued epochs. The models were saved with pickle to avoid training each time the model is loaded.

5.4.4 Drug Recommendation

The recommendation was done by combining the two models. The results were multiplied by the generalized useful count for the final score for the drugs. A higher score illustrates that the drug got better feedback from the user. The normalization of the drug was done to know the distribution in the useful count.

5.5 Description of Testing

The prediction of the test dataset was measured with precision, recall, f1score and Accuracy. The precision is number of true positives divided by the total positive predictions. Recall is the ratio between the number of positive sample and are correctly classified as positive and measures the model ability to detect positive samples. F1 score is the combination of prediction and recall using their harmonic mean. Accuracy was to measures the correctness of the model. Appendix 2 shows the LGBM performance metrics and Appendix 3 shows the XGB performance metrics.

5.5.1 Testing Paradigm

i. Integration testing

Integration testing was performed when checking if the two classifiers were working together and giving the right prediction. It was checked if the two classifiers when combined does not distort the performance. Integration testing was also done on deployment whether the model is giving right predictions on deployment.

ii. White Box Testing

White box testing was used to verify that the internal implementations were working as required. Some of the white box testing that was done is checking if the TextBlob library is giving the correct polarization and sentiment calculations as shown in Appendix 4. It was also done to verify that the reviews were cleaned by removing punctuations, lowering case and lemmatizing as shown in Appendix 10. Additionally, it was used to confirm that the features that were being added were included in the dataset.

iii. Black box testing

Black box testing was used to test the external structure. Black box testing is mostly for the end user experience, it was used in testing the user interface to select from a list of drugs and the user selects the condition and expects to get a list of the top five drugs with the best final prediction mean as shown in Appendix 6.

5.6 Testing Results

Testing was done to check if the model was meeting the functional and non- functional requirements.

5.6.1 User Interface

Test Case	Description	Test Case	Experimental Outcome	Result	Pass/Fail

TC001	A list of unique conditions from the dataset to be displayed	Input a condition: Pain	User will get the condition want	Pain condition is displayed	Pass Refer to Appendix 5
TC002	Top 5 drugs for a condition should be displayed	Input a condition: Pain	User will get 5 top medications for pain	Pain condition medication	Pass Refer to Appendix 5

5.6.2 Model Accuracy and Performance

Test Case	**Description**	**Test Case**	**Experimental Outcome**	**Result**	**Pass/Fail**
TC001	Model should give a good accuracy	Test LGBM and XGB model	Good accuracy of above 80%	The model gave good accuracy	Pass Refer to Appendix 2
TC002	Model should have good F1 score	Test LGBM and XGB model	Good F1 score of above 80%	The model gave a good F1 score	Pass Refer to Appendix 2
TC003	Model should have good precision	Test LGBM and XGB model	Good precision of above 80%	The model gave a good precision score	Pass Refer to Appendix 2

TC004	Model should have good recall	Test LGBM and XGB model	Good recall of above 80%	The model gave a good recall	Pass Refer to Appendix 2

5.6.3 Model Data Integrity

Test Case	Description	Test Case	Experimental Outcome	Result	Pass/Fail
TC001	The data used for the model should not have any null value	Query the dataset to return number of null values	No null values	The data did not have any null values	Pass Refer to Appendix 6

The Accuracy for LGBM model was 0.89 or 89%. This a good accuracy because the aim of the model was to assist people who self-medicate and the model means that 89% of the predictions were correct hence has assisted a large population of people. However, because drug prescription is critical the model accuracy can be increase the correct predictions.

The F1 score is 0.82 for LGBM model. It is important in the drug recommendation model because the classification models. It shows that the model ability to relay true performance is good.

Chapter 6: Conclusions, Recommendations and Future Works

6.1 Conclusion

Due to challenges in a low doctor ratio to patients, large populations that self-medicate and also mis prescription there is need for another way to recommend medication. The model gave the top 5 scoring drugs for a condition. It was done based on sentiment analysis by modeling it using LGBM and XGB. The models gave good accuracies as shown in Appendix 2 and 3.

The recommender system can act as a booster for the overwhelmed medical field or for people who want to self-medicate. This may be a better way because it incorporates the end user feedback to recommend to other people.

6.2 Recommendation

It is recommended that the model be used with a high-capacity RAM such as 8GB.The user should also be accessible to web browsers such as Google Chrome or Mozilla Firefox to be able to access the model. Also, for higher speed it should be accessed with 4G or 5G internet connectivity. Also, it is recommended that a dataset with more features can be used. More features in a dataset can help reduce biasness in the model.

6.3 Future Works

The model worked as it was intended to but there are certain limitations to it. This is because the accuracies are not very high so it might not be ready for real life application. The sentiment analysis classifiers may be explored so that the best classifiers can be determined to give better predictions. Also, a dataset with richer features such as the demographics can be used for rich analysis.

The future works may include integrating a system with a model that a user can input symptoms and conditions and the recommended drugs given. Also, different approaches can be used as deep learning and comparisons made to improve optimization for a better recommender system.

References

Al-Fedaghi, S. (2021). UML Sequence Diagram: An Alternative Model. *International Journal of Advanced Computer Science and Applications (IJACSA)*. https://hal.archives-ouvertes.fr/hal-03243870

Chautrakarn, S., Khumros, W., & Phutrakool, P. (2021). Self-Medication With Over-the-counter Medicines Among the Working Age Population in Metropolitan Areas of Thailand. *Frontiers in Pharmacology, 12*. https://www.frontiersin.org/articles/10.3389/fphar.2021.726643

Chen, C., Zhang, L., Fan, X., Wang, Y., Xu, C., & Liu, R. (2018). A Epilepsy Drug Recommendation System by Implicit Feedback and Crossing Recommendation. *2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, 1134–1139. https://doi.org/10.1109/SmartWorld.2018.00197

Cline, R. J. W., & Haynes, K. M. (2001). Consumer health information seeking on the Internet: The state of the art. *Health Education Research, 16*(6), 671–692. https://doi.org/10.1093/her/16.6.671

Cooper, R. J. (2013). Over-the-counter medicine abuse – a review of the literature. *Journal of Substance Use*, *18*(2), 82–107. https://doi.org/10.3109/14659891.2011.615002

Doulaverakis, C., Nikolaidis, G., Kleontas, A., & Kompatsiaris, I. (2012). GalenOWL: Ontology-based drug recommendations discovery. *Journal of Biomedical Semantics, 3*(1), 14. https://doi.org/10.1186/2041-1480-3-14

Goyal, V. A., Parmar, D. J., Joshi, N. I., & Champanerkar, K. (2020). *Medicine Recommendation System*. *07*(03), 5.

Ibrahim, R., & Yen, S. Y. (2010). Formalization of the Data Flow Diagram Rules for Consistency Check. *International Journal of Software Engineering & Applications*, *1*(4), 95–111. https://doi.org/10.5121/ijsea.2010.1406

Jakovljevic, M. (Michael), Lazarevic, M., Jurisevic, M., & Jovanovic, M. R. (2015). When cure becomes an illness—Abuse of addictive prescription medicines. *Frontiers in Pharmacology*, *6*. https://doi.org/10.3389/fphar.2015.00193

Levels in Data Flow Diagrams (DFD). (2019, March 18). *GeeksforGeeks*. https://www.geeksforgeeks.org/levels-in-data-flow-diagrams-dfd/

Li, Q., & Chen, Y.-L. (2009). Entity-Relationship Diagram. In Q. Li & Y.-L. Chen (Eds.), *Modeling and Analysis of Enterprise and Information Systems: From Requirements to Realization* (pp. 125–139). Springer. https://doi.org/10.1007/978-3-540-89556-5_6

Mangoni, A. A., & Jackson, S. H. D. (2004). Age-related changes in pharmacokinetics and pharmacodynamics: Basic principles and practical applications. *British Journal of Clinical Pharmacology*, *57*(1), 6–14. https://doi.org/10.1046/j.1365-2125.2003.02007.x

O'Donovan, B., Rodgers, R. M., Cox, A. R., & Krska, J. (2019). Use of information sources regarding medicine side effects among the general population: A cross-sectional survey. *Primary Health Care Research & Development*, *20*, e153. https://doi.org/10.1017/S1463423619000574

Rainsford, K. D. (2009). Importance of pharmaceutical composition and evidence from clinical trials and pharmacological studies in determining effectiveness of chondroitin sulphate and other glycosaminoglycans: A critique. *The Journal of Pharmacy and Pharmacology*, *61*(10), 1263–1270. https://doi.org/10.1211/jpp/61.10.0001

Raju, K. S. (2021, April 23). *What Is Design Thinking And Why Is It Important?* GreatLearning Blog: Free Resources What Matters to Shape Your Career! https://www.mygreatlearning.com/blog/importance-of-design-thinking/

Rao, T. V. N., Unnisa, A., & Sreni, K. (2020). *Medicine Recommendation System Based On Patient Reviews*. *9*(02), 5.

Waykar, Y. (2015). Role of use case diagram in software development. *International Journal of Management and Economics*.

Wittich, C. M., Burkle, C. M., & Lanier, W. L. (2014). Medication Errors: An Overview for Clinicians. *Mayo Clinic Proceedings*, *89*(8), 1116–1125. https://doi.org/10.1016/j.mayocp.2014.05.007

Zhang, L., & Liu, B. (2017). Sentiment Analysis and Opinion Mining. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning and Data Mining* (pp. 1152–1161). Springer US. https://doi.org/10.1007/978-1-4899-7687-1_907

Zhang, Y., Zhang, D., Hassan, M., Alamri, A., & Peng, L. (2014). CADRE: Cloud-Assisted Drug REcommendation Service for Online Pharmacies. *Mobile Networks and Applications*, *20*, 348–355. https://doi.org/10.1007/s11036-014-0537-4

Appendix

Appendix 1: Sentiment Analysis Using TextBlob

```
text_blob_object = TextBlob(df_data['review'][100])
print(text_blob_object.sentiment)

Sentiment(polarity=0.07743055555555556, subjectivity=0.48819444444444454)
```

Appendix 1: Sentiment Analysis TextBlob

Appendix 2: LGBM Metrics

```
    num_iterations=2000

)

model = clf.fit(X_train, y_train)

 #predictions
predictions_ml = model.predict(X_test)
print("The accuracy of the model is ", accuracy_score(y_test,predictions_ml), '\n')
print ("The confusion matrix is \n")
print( confusion_matrix(y_test, predictions_ml),  '\n ')

print (classification_report(y_test,predictions_ml))

The accuracy of the model is  0.8927456004189187

The confusion matrix is

[[12805  3314]
 [ 2421 34931]]

              precision    recall  f1-score   support

           0       0.84      0.79      0.82     16119
           1       0.91      0.94      0.92     37352

    accuracy                           0.89     53471
   macro avg       0.88      0.86      0.87     53471
weighted avg       0.89      0.89      0.89     53471
```

Appendix 2: LGBM metrics

Appendix 3: XGB Metrics

```
#predictions
predictions_xgb = model_xgb.predict(X_test)
print("The accuracy is : " ,  accuracy_score(y_test,predictions_xgb), '\n')
print ("The confusion matrix is \n")
print( confusion_matrix(y_test, predictions_xgb),  '\n ')

print (classification_report(y_test,predictions_xgb))
```

```
The accuracy is :  0.7605617998541265

The confusion matrix is

[[ 6430  9689]
 [ 3114 34238]]

              precision    recall  f1-score   support

           0       0.67      0.40      0.50     16119
           1       0.78      0.92      0.84     37352

    accuracy                           0.76     53471
   macro avg       0.73      0.66      0.67     53471
weighted avg       0.75      0.76      0.74     53471
```

Appendix 3: XGB Metrics

Appendix 4: TextBlob Code

```
#from nltk.corpus.reader.reviews import Review
def find_pol(te):
    return TextBlob(te).sentiment.polarity

df_data['sentiment'] = df_data['review'].apply(find_pol)
df_data['sentiment_processed'] = df_data['review_clean'].apply(find_pol)

#df_data.head()
```

Appendix 4: TextBlob Code

Appendix 5: Interface

```
condition  Pain

Top 5 Drugs
     drugname  final_pred
0  OxyContin       695.0
1  Oxycodone       695.0
2  Oxycodone       485.0
3     Elavil       308.0
4   Tramadol       285.0
```

Appendix 5: Interface

Appendix 6: Null Values Removal

```
from pandas.core import missing
percent_missing = df_data.isnull().sum()*100/ len(df_data)
missing_value_df = pd.DataFrame({'column_name': df_data.columns, 'percent_missing': percent_missing})
missing_value_df
```

	column_name	percent_missing
uniqueid	uniqueid	0.000000
drugname	drugname	0.000000
condition	condition	0.555186
review	review	0.000000
rating	rating	0.000000
date	date	0.000000
usefulcount	usefulcount	0.000000

[15]

```
print('Shape of df before dropping null values',df_data.shape)
df_data.dropna(inplace=True)
df_data.reset_index(drop=True)
print('Shape of df before dropping null values',df_data.shape)
```

Appendix 6: Null Value Removal

Appendix 7: Original Dataset

	uniqueID	drugName	condition	review	rating	date	usefulCount
0	206461	Valsartan	Left Ventricular Dysfunction	"It has no side effect, I take it in combinati...	9	20-May-12	27
1	95260	Guanfacine	ADHD	"My son is halfway through his fourth week of ...	8	27-Apr-10	192

Appendix 7: Original Dataset

Appendix 8: Feature added Dataset for sentiment analysis

word_count	unique_word_count	punctuation_count	upper_word_count	title_word_count	stopword_count	mean_word_len
11	11	3	1	6	6	4.272727
68	68	23	2	13	68	5.338235
78	78	30	6	15	57	5.166667

Appendix 8: Sentiment Analysis Dataset

Appendix 9: Dataset with sentiment and sentiment processed

sentiment	sentiment_processed	word_count	unique_word_count	punctuation_count	upper_word_count	title_word_count
0.000000	0.000000	11	11	3	1	6
0.168333	0.188021	68	68	23	2	13

Appendix 9: Sentiment and sentiment processed

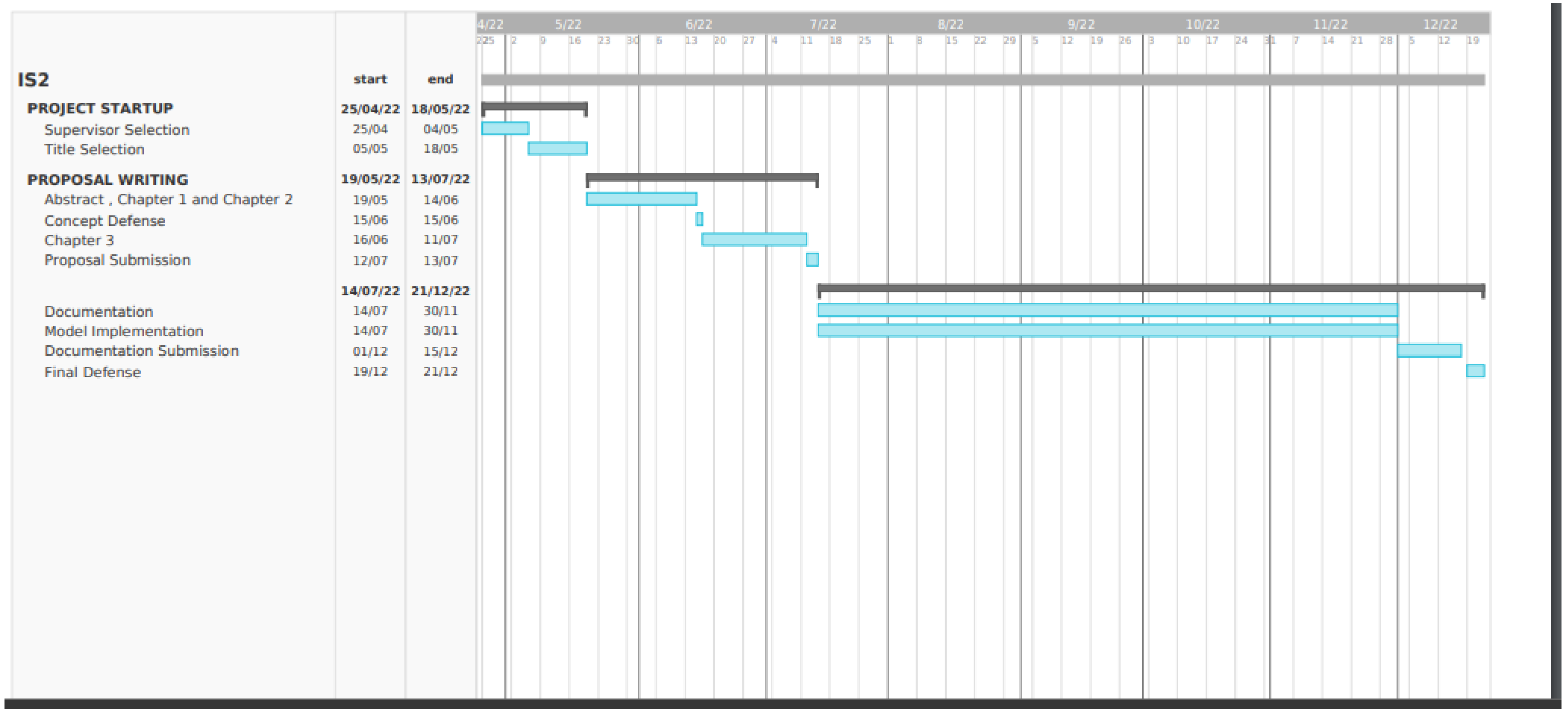
Appendix 10: Cleaned Review

```
df_data['review_clean'][100]
```

```
'ive latuda little 2 half year almost completely stopped psychotic symptom except still hear voice mainly try go sleep but no delusion paranoia drug take cogentin combination because cause shake lot main si
de effect experience include anhedonia shakiness jaw clenching inability sit still however im happy because actually work antipsychotic med tried not doesnt cause endless hunger experienced drug like saphri
s haldol zyprexa risperdal noted max daily dose 160mg'
```

Appendix 10: Cleaned Review

Appendix 11: Gantt Chart



Appendix 11 Gantt Char