Disease Prediction and Drug Recommendation System

CMPE 255, Group 3

Khushboo Ekhande (014370837)

Anurag Upadhyay (014371240)

Rakshitha Sathyakumar (014511705)

Sughosh Krishnamurthy (014370954)

Abstract

Data Mining is a method that requires analyzing and exploring large blocks of data to extract meaningful trends and patterns. Data mining techniques can be applied to various fields including medical databases. There are thousands of people all over the world facing health and medical diagnosis problems. Hospital Information System (HIS) generates massive data but gaining useful knowledge from the diagnosis case data is a big challenge. Using the methodologies used in this project, patients can easily get information about the disease they are suffering from and the drug helpful for dealing with that disease by just entering the symptoms he/she is showing.

Using predictive analysis on the diseases, we can recommend drugs to the users by considering the different features from the dataset. The experimental results from this paper can be further utilized for research purposes and for other various medical utilities.

Introduction

One of the most commonly found concerns among patients when confronted with any medical condition is which physician to trust. It is a known fact that the health of an individual significantly affects his/her quality of life. A survey in 2013 by the Pew Internet and American Life Project found that 59% of adults have looked online for health topics and with 35% of respondents focusing on diagnosing a medical condition online. There are more people every day caring about the health and medical diagnosis problem but still many who lose their lives due to medical errors. According to the administration's report, more than 200 thousand people in China and over 100 thousand in the USA, die each year due to medication errors. More than 42% medication errors are caused by doctors because they write prescriptions based on their experience which is quite limited. Hence, finding appropriate physicians to diagnose and treat medical conditions is one of the most important decisions a patient must make.

Advancement in Data mining and Recommender Technologies allow us to explore possibilities of potential knowledge from diagnosis history records and reviews and ratings on drugs to help doctors prescribe the correct medication and decrease the medication errors effectively.

The objective of this Data Mining paper is to design and implement a universal Disease Prediction and Drug Recommendation System that applies various Data Mining technologies to the recommendation system. By combining information from different sources we are using various prediction algorithms along with NLP for sentiment analysis and recommendation. Rest of the report talks about Data Gathering, Pre-processing, Methodology, Results and Conclusion of our project.

Related Work

- A considerable amount of research has been previously done in this field. But most of it focuses either on only Disease Prediction or Sentiment Analysis of the Drugs.
- In the proposed model, we are trying to combine datasets for prediction as well as a recommendation which makes it a complete recommendation system.
- Also, we are providing the patients with side effects (if any) of the drugs that we recommend which is an added benefit for them.
- Few techniques used in this project are our own implementations and give better results.

In the paper presented by K.Gomathi et al, they have demonstrated results of Decision Tree and Naive Bayes models in predicting three diseases, Heart Disease,Diabetes and Breast Cancer. [1]

Another paper presented by M.A.Nishara Banu et al, they have predicted heart diseases. By applying K-Mean on the medical dataset, they have clustered the relevant data, upon which MAFIA(Maximal Frequent Itemset Algorithm) is applied to generate rules and identification of frequent pattern which is fed to the C4.5 (Decision Tree) model to classify patterns. [2]

Datasets and Preprocessing

Healthcare information is protected by HIPPA. Sharing of medical records of patients without their knowledge is prohibited. Getting access to government health records and datasets required multiple permissions. Hence, for our project, we are using the datasets that were readily available on the internet and were open to downloads.

Dataset One

Data Gathering

The 10-year medical record dataset was obtained from Medical records 10 yrs - dataset by arvin6 | data.world.

It consists of four CSV files, namely

- encounter.csv
- encounter_dx.csv
- lab_results.csv
- medication_fulfillment.csv

The encounter.csv consists of 1176 rows and 17 columns, encounter_dx.csv consists of 3063 rows and 6 columns, lab_results.csv consists of 7509 rows and 21 columns, and the medical_fulfillment.csv consists of 5447 rows and 28 columns. In order to obtain useful insights and understand if the dataset has required information to deliver the problem statement, we preprocessed and merged the dataset to cater to our needs.

Data Preprocessing

After gathering the raw data and understanding the shape of all the four CSV files, we preprocessed each table and combined the tables to get a single merged dataset with the required columns. To preprocess, we executed certain commands to understand the data. Like, count of each row, the number of unique values, dropped a few columns which were irrelevant to achieve the end goal, dropped columns which had no data values, and combined a few columns to make it more usable.

Once the basic preprocessing was done, we then found out that the four tables were managed using the Star Schema Model and medical_fulfillment.csv is the fact table and ecounter.csv, encounter_dx.csv, and lab_results.csv are dimension tables. The star schema

follows the one-many relationship. We found out the Primary Key and Foreign Key and merged the four tables to form a single table. The medication_fulfillment table has 'Ecounter_ID' as the primary key. We then merged the Medication_fulfillment table with columns 'severity' and 'description' from the encounter_dx table using left join on 'Encounter_ID' by running a SQL query. The resulting table has 1176 rows.

Next, to check if 'order_ID' is the primary key of lab_results.csv, we ran a query to check the count of each row in 'order_ID' column and found out that it's not the primary key as the number of unique values didn't match the number of rows of the table. We then found out that 'Order_ID' and 'Result_LOINC' together make the composite primary key. Since none of the columns from lab_results.csv was useful, we didn't use any columns for the merged dataset.

With encounter.csv remaining, we found that 'Encounter_ID' is the primary key and extracted the CC column and merged it using left join on 'Encounter_ID'. Now we have a complete merged dataset of the required columns. It consists of 1176 rows. From the merged dataset, we extracted the Drug_Name, description, severity, and CC and grouped them to get the total number of each drug and their associated disease and description. The extracted columns consist of many null values as shown in the figure below.

	Drug_Name	description	severity	CC	cnt
0	OMS 50	Chronic Obstructive Pulmonary Disease	critical	critical shortness of breath	119
1	Ciprofloxacin	None	None	None	105
2	Isotonic Saline (0.9%)	None	None	None	101
3	Lisinopril	None	None	None	83
4	Potassium Chl	Type 1 Diabetes	severe	severe increased thirst	78
...	...	...	...	...	...
76	metoprolol	Hypertension	severe	moderate difficulty walking	1
77	oxycodone-acetaminophen 10-325	None	None	None	1
78	trimethoprim	None	None	None	1
79	trimethoprim	Pyelonephritis	severe	Pyelonephritis	1
80	valsartan	Chronic Congestive Heart Failure	severe	mild palpitations	1

Table 1: Drug name grouped by description, severity, and CC

Hence, we ran another query to get the total number of rows which consist of null values corresponding to a drug name. As a total of 764 rows had null values, we were only left with 416 rows of relevant data to train our classification model. Hence, we concluded that this dataset does not serve the purpose and looked for more datasets that would align with the delivery of our problem statement.

Dataset Two

Data Gathering

For the accurate recommendation of drugs, we first predict the diseases based on symptoms and then recommend the drugs based on the ratings. For this purpose, we have tried to gather information from two main datasets.

1. Symptoms dataset:

- This dataset is used to take symptoms as input and predict the disease as an output.
- Dataset is obtained from the Disease-Symptom Knowledge Database which is a knowledge database of disease-symptom associations generated by an automated method based on information in textual discharge summaries of patients at New York-Presbyterian Hospital admitted during 2004.
- This dataset contains 3 columns:
    1. Disease
    2. Count of Disease Occurrence
    3. Symptom

	Disease	Count of Disease Occurrence	Symptom
0	UMLS:C0020538_hypertensive disease	3363.0	UMLS:C0008031_pain chest
1	NaN	NaN	UMLS:C0392680_shortness of breath
2	NaN	NaN	UMLS:C0012833_dizziness
3	NaN	NaN	UMLS:C0004093_asthenia
4	NaN	NaN	UMLS:C0085639_fall

Table 2: Raw Symptoms dataset

- There are a total of 149 unique diseases in this dataset and 405 symptoms.
- Each disease contains 4-5 symptoms corresponding to it.
- This dataset is sent for pre-processing so that it can be used to train models to classify and predict the disease.

2. Drug Review Dataset:

- This dataset is used in order to take the predicted disease as input and recommend appropriate drugs based on reviews and ratings (Sentiment Analysis).
- The dataset is gathered from the UCI Machine Learning Repository for Drug Review which provides patient reviews on specific drugs along with related conditions and a 10-star patient rating reflecting overall patient satisfaction.
- The Repository had two datasets (Test and Train) which are combined for analysis and visualization purposes as they had the same number of columns.
- It contains 7 columns and 215063 rows:
    1. ID
    2. Drug name
    3. Condition
    4. Review
    5. Rating
    6. Date
    7. Useful count

	uniqueID	drugName	condition	review	rating	date	usefulCount
0	206461	Valsartan	Left Ventricular Dysfunction	"It has no side effect, I take it in combinati...	9.0	20-May-12	27
1	95260	Guanfacine	ADHD	"My son is halfway through his fourth week of ...	8.0	27-Apr-10	192
2	92703	Lybrel	Birth Control	"I used to take another oral contraceptive, wh...	5.0	14-Dec-09	17
3	138000	Ortho Evra	Birth Control	"This is my first time using any form of birth...	8.0	3-Nov-15	10
4	35696	Buprenorphine / naloxone	Opiate Dependence	"Suboxone has completely turned my life around...	9.0	27-Nov-16	37

Table 3: Raw Drug Review Dataset

- There are 3671 unique Drug names and 916 unique Conditions (Disease) in this dataset along with the rating and reviews corresponding with the drug names.
- This dataset is then pre-processed and visualized to gain more information for effective drug recommendation.

3. Side Effects Dataset

- We are also planning to include the dataset containing side effects of specific drugs in order to help patients identify the risks involved in the drug that is being recommended.
- This dataset is again gathered from UCI Machine Learning Repository for Side Effects of Drugs along with some raw data gathered from druglib.com.
- This Dataset contains a lot of columns similar to the drug review dataset, but it does not have a lot of rows. Hence only the "Side Effects" column from this dataset will be combined with the other two datasets along with the side effects found from druglib.com

urlDrugName	rating	effectiveness	sideEffects	condition	benefitsReview	sideEffectsReview
enalapril	4	Highly Effective	Mild Side Effects	management of congestive heart failure	slowed the progression of left ventricular dys...	cough, hypotension , proteinuria, impotence , ...
ortho-tri-cyclen	1	Highly Effective	Severe Side Effects	birth prevention	Although this type of birth control has more c...	Heavy Cycle, Cramps, Hot Flashes, Fatigue, Lon...

Table 4: Raw Dataset containing Side effects of drugs

Data Preprocessing

In this project, we have used and worked upon multiple datasets. All the datasets were obtained in raw format. To preprocess all the datasets, a few common steps and measures were carried out. Those were:

- All the datasets were first checked for the count of null and missing values.
- All such values were either handled or dropped from the dataset.
- After that, every column's unique values were found and their frequencies.
- Using standard libraries, the dataset was visualized and outliers, if any, were found out.
- Any irrelevant information was deleted from the dataset.

1. Symptoms Dataset

- This dataset had to be cleaned in order to gain information from it.
- First, the “Count” column was dropped as the information given in it is irrelevant for this project.
- Then, the null values in the Disease column were handled using the drop function.

```
data = df.fillna(method='ffill')
```

- Cleaning of Disease and Symptom columns was done to get only the name and removed unnecessary data.

	Disease	Symptom
0	hypertensive disease	[pain chest, shortness of breath, dizziness, a...
1	diabetes	[polyuria, polydypsia, shortness of breath, pa...
2	depression mental	[feeling suicidal, suicidal, hallucinations au...
3	depressive disorder	[feeling suicidal, suicidal, hallucinations au...
4	coronary arteriosclerosis	[pain chest, angina pectoris, shortness of bre...

Table 5: Symptoms dataset after preprocessing

- For classification of the symptoms based on diseases, we have converted it into a new CSV format file which now has symptoms as the columns and diseases as rows. Using one hot encoding, we have mapped every symptom with all the diseases and adding value 1 if it is present for disease and 0 otherwise. Below is a screenshot of the one-hot encoded dataset. This will help us predict the diseases when symptoms are given as input.

	Disease	Heberden's node	Murphy's sign	Stahli's line	abdomen acute	abdominal bloating	abdominal tenderness	abnormal sensation	abnorma hard consiste
0	Alzheimer's disease	0	0	0	0	0	0	0	0
1	HIV	0	0	0	0	0	0	0	0
2	PneumocystisÂ cariniiÂ pneumonia	0	0	0	0	0	0	0	0
3	accidentÂ cerebrovascular	0	0	0	0	0	0	0	0
4	acquiredÂ immuno-deficiency syndrome	0	0	0	0	0	0	0	0

5 rows × 405 columns

Table 6: Dataset after marking symptoms present for a disease as 1 else 0.

2. Drug Review Dataset

- This dataset contained two sets (Train and Test) which were combined to be able to visualise and analyze the data on a larger dataset. And also because they both had the same columns so could be combined easily.
- The dataset obtained was pretty clean and did not require a lot of pre-processing. Still, a few rows with null values were dropped and columns were renamed.
- The dataset contains a lot of information and visualizing was an interesting task.
- Many different graphs were plotted showing results of drugs with most reviews, most popular drugs, most common diseases, etc.

- One such visualization is shown below which has the names of a few most popular drugs:

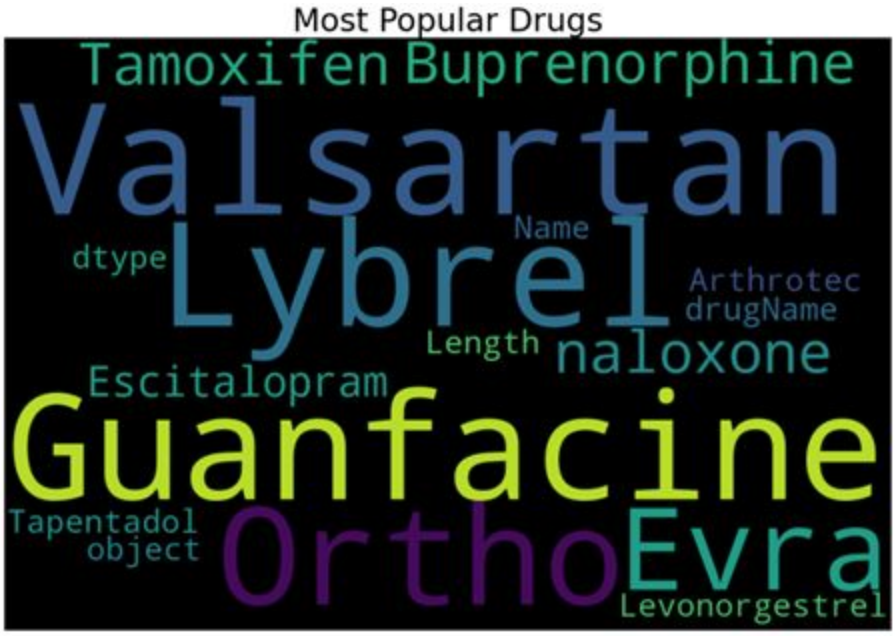

Figure 1: Visualizing the most popular drugs based on the ratings

3. Side Effects Dataset

- This dataset has a lot of information similar to the one in the Drug Review dataset and it was also clean.
- A few null values are handled and irrelevant columns are dropped.
- This dataset is to be merged with the drug review dataset to map only the side-effects for the specific drugs.

4. Merged Dataset

- The Dataset containing symptoms is merged with the one with reviews for final drug prediction.

	Drug	Disease	Review	Rating	UsefulCount	Symptoms
0	Aspirin	transient ischemic attack	"No side effects, easy to take, no more sympt...	10.0	10	['speech slurred', 'dysarthria', 'facial pares...
1	Clopidogrel	transient ischemic attack	"I've been taking this medicine for a lit...	10.0	8	['speech slurred', 'dysarthria', 'facial pares...
2	Clopidogrel	transient ischemic attack	"I took ibuprofen (2 caps at night for severe ...	6.0	13	['speech slurred', 'dysarthria', 'facial pares...
3	Clopidogrel	transient ischemic attack	"After my VAD Stroke I am on plavix. I have a...	5.0	9	['speech slurred', 'dysarthria', 'facial pares...
4	Bayer Children's Aspirin	transient ischemic attack	"No side effects, easy to take, no more sympt...	10.0	10	['speech slurred', 'dysarthria', 'facial pares...

Table 6: Merged Dataset

- Only the relevant columns are kept in the merged dataset and rest are dropped to reduce the dimensionality.

Methodology

The main goal of our project is to recommend a drug to a patient based on the symptoms she/he has. In accordance with our objective to implement a drug recommender system there are two main subcategories which are to be addressed i.e a disease prediction model and a recommendation model.

The disease prediction model is a probabilistic model which will give the predictions based on the symptoms. For this purpose, we are using the Disease-Symptom Knowledge Database which has over 149 unique diseases mapped with 405 symptoms. Upon preprocessing the data the data was transformed and mapped to a data frame which consists of columns derived from the list of unique symptoms and the list of unique target diseases as labels. The preprocessed symptoms dataset was used to map the newly created disease prediction data frame by setting the values of symptoms present for a particular disease as 1 and 0 if not. The data was trained using Multinomial Probabilistic Model and ExtraTree Classifier Model using different Symptoms as the training features and Diseases as the labels.
Note: Presently we are working on the models and we will be trying out other classifier models like Random Forest or SVM.

Upon obtaining the most probable diseases the next task is to map the list of drugs that can be prescribed for this particular disease using the Merged Dataset which has been created by preprocessing the symptoms and the drug review dataset. Once we obtain the list of possible drugs the next task is to be able to recommend the best drug for the patient. For this purpose, we will be adopting sentiment analysis using the Natural Language Processing approach on the drug review data set to understand the trend in the positive and negative reviews given by the patients. For this purpose, we will be building an NLP model based on word2vec and doc2vec methods to build a neural network classifier to classify the reviews. Based on the predicted reviews from the model we will be obtaining the weighted value based on a total number of positive reviews and useful count for a particular drug and recommend a drug based on the highest weight.
Note: We have completed till building the sentiment analysis neural network model using word2vec technique, due to low accuracy we will be looking into doc2vec. Mapping of the three, the prediction model, the sentiment analysis model and the recommender model is still pending. We are also looking into side effects dataset to build a better recommender system by weighting in the side effects factor for the recommended drug via review analysis and disease prediction.

Results

- We have completed data pre-processing, sentiment analysis and disease prediction as of now. The pre-processing results are mentioned earlier in this report.
- For Disease prediction we have used two models, one is Multinomial Naive Bayes and Extra Tree Classifier both of which have the performance of around 90% accuracy.

```
Actual Disease: malignant neoplasm of lung
Predicted Disease using Model: carcinoma of lung


Actual Disease: malignant neoplasm of prostate
Predicted Disease using Model: carcinoma prostate


Actual Disease: malignant tumor of colon
Predicted Disease using Model: carcinoma colon


Actual Disease: oralcandidiasis
Predicted Disease using Model: candidiasis


Actual Disease: pericardial effusion body substance
Predicted Disease using Model: effusion pericardial


Actual Disease: primary malignant neoplasm
Predicted Disease using Model: malignant neoplasms


Actual Disease: septicemia
Predicted Disease using Model: sepsis (invertebrate)


Actual Disease: systemic infection
Predicted Disease using Model: sepsis (invertebrate)


Actual Disease: tonic-clonic seizures
Predicted Disease using Model: tonic-clonic epilepsy


Total number of incoorect predictions: 15
Accuracy of the MNB model in Predicting Disease:  89.93288590604027 %
```

- For sentiment analysis, we have used the NLP model on reviews and ratings and classified the reviews as positive or negative sentiments.
- Given a review, the model predicts whether it is a positive or negative sentiment.
- This will further be used to map drug names to the reviews and recommend drugs which have only positive reviews.

- Code snippet of the accuracy of Sentiment Analysis Model:

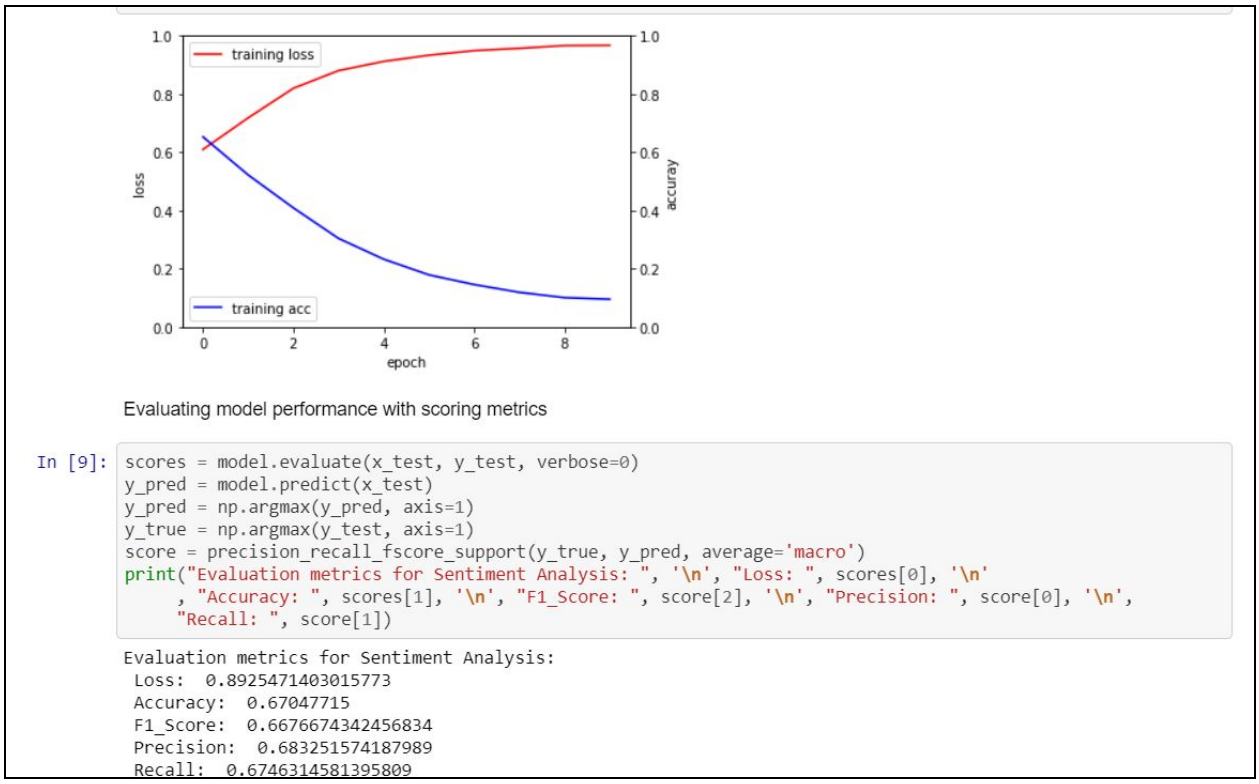

Figure 2: Model Evaluation

- The accuracy is 67.04% for now, but we are working on improving the same. We have used tfid-vectorizer which is not providing good accuracy. One way to improve that is trying doc2vec using the gensim library.
- Sample of a few predicted sentiments:

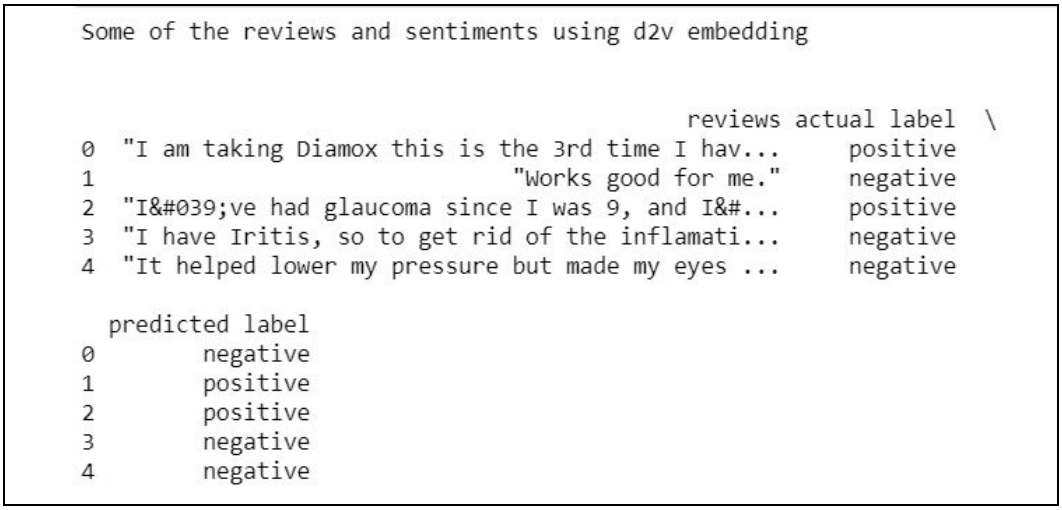

Figure 3: Sentiment analysis reviews

- Here, we can see that the predicted labels are more accurate than the actual labels! For instance, in the predicted output given above, the second review is actually positive but is incorrectly labelled as negative. Whereas our model correctly predicts it as positive! Which makes it a good reason to recommend a drug from our sentiment analysis than the actual threshold rating.

Conclusion and Future Scope

- As the project is not yet complete, we cannot conclude anything as of now. But from what we have done so far, it can be said that the datasets gathered are proving to be very useful for disease prediction and drug recommendation. They have a lot of information which makes this data mining project very helpful in healthcare systems.
- We are trying to combine and analyse 3 different datasets to get maximum possible features which are crucial when dealing with drugs.
- All the details and work mentioned by far in this report has already been completed as a part of Milestone 2.
- For our Final submission, we will be training all the models with the given datasets to effectively predict disease and recommend drugs accordingly.
- We will also be giving a list of side effects of particular drugs after recommending them.

Work Distribution

1. Data Gathering:

Each member worked on gathering the datasets from different sources.

Dataset 1: Rakshitha and Sughosh

Dataset 2: Khushboo and Anurag

2. Data Pre-processing and Visualization:

Each member was responsible for the pre-processing of a different dataset which helped to get a lot of work done in less time.

Dataset 1: Rakshitha and Sughosh

Dataset 2: Khushboo and Anurag

3. Model training:

We are trying out multiple models for prediction as well as recommendation. Each member is working on a different model as of now. The work which has been completed is pushed to GitHub but is yet in process.

4. Report:

It is a collaborative work of all!

References

[1] Kamaraj, K.Gomathi & Priyaa, D.Shanmuga. (2016). Multi Disease Prediction using Data Mining Techniques. International Journal of System and Software Engineering.

[2] M.A.Nishara Banu, B Gomathy. (2013). Disease Predicting System Using Data Mining Techniques. International Journal of Technical Research and Applications.

[3] Druglib.com - Drug Information, Research, Clinical Trials, News. http://www.druglib.com/