Sentiment Analysis and Machine Learning AlgorithmImplementation on Health Care System

Mrs.R.Nancy Deborh[1], Mr.S.Alwyn Rajiv[2], Gunasree G G[3], Iswarya S[4], Jerusha Judith J[5]

[1]Assistant Professor, IT Department, Velammal College of Engineering and Technology,Madurai, Tamilnadu, India
[2]Assistant Professor, ECE Department, Kamaraj College of Engineering and Technology, Madurai, Tamilnadu, India
[3]UG Student, Dept. of Information Technology, Velammal College of Engineering and Technology, Madurai, Tamilnadu, India
[4]UG Student, Dept. of Information Technology, Velammal College of Engineering and Technology,Madurai, Tamilnadu, India
[5]UG Student, Dept. of Information Technology, Velammal College of Engineering and Technology,Madurai, Tamilnadu, India

---***---

Abstract:- *Recommendation System (RS) has been utilized in a variety of fields and it is used as an efficient tool to overcome information overload as the amount of information on the internet is getting increased day by day. In recent years, the appliance of Recommendation System for health has become a growing research topic and helps the people to make the right decisions for their health. In the Proposed Methodology, a recommendation system is developed to help patients to find the best hospital and doctors for a particular treatment based on their symptoms in the field of Health care. The system aims to provide accurate analysis of hospitals by taking into account the reviews and ratings by thousands of patients, which were written by the patients themselves in various online forums and also provides doctor's names related to the disease given by the patients. The Proposed System performs sentiment analysis on the reviews of various patients using Natural Language Processing Technique to classify them as Positive and Negative review and then applying the Machine Learning Algorithms to predict the diseases for the given symptoms. After that based on the polarity score, the system predicts the hospital name and doctor's name for certaindiseases.*

***Keyword*:- NLP, Classification, Clustering, SVM, Random Forest, Gaussian NB, LogisticRegression, XG Boost.**

1. INTRODUCTION:

It is mostly necessary to make a choice without having prior knowledge (or) Personal experiences about something. In our lifestyle, we rely on recommendations driven by people either by normal words (or) the reviews of general survey. Suppose, many of the days you walk to restaurant, maybe you are there for the first (or) second time, since you don't have any idea what to order so you undergo the menu and take some time to choose the food and therefore the waiter asks "what would you wish to have?", and you are not certain and you have not decided what to order. So, during this case, you ask the waiter to recommend something to you, then on the idea of provided recommendations, you select your choice. Similarly, people often use a recommendation system over the online to form decisions for the things associated with their choice. The goal of the recommendations is to recommend a set of users for their items (or) products which may interest them. In other words, recommendation systems are a part of the knowledge filtering domain, where the goal is to filter the abundant information from the website to be more specific and meaningful. Recommendation systems are implemented during a sort of application and became really useful in recent years. The foremost famous areas where the concept of recommendation system is implemented are movies, music, news, books, social tags, products, restaurant, financial services, life insurances, etc.

However, in spite of all these advances, the system still requires further necessary improvements to make recommendation approaches more effective and also challengesfaced by ma ny people are looking for health care information regarding treatments, diagnosis and hospitals. In the Proposed Methodology, a recommendation system that would select doctors and hospitals by people from anywhere and anytime which in turn reduce the time. The Proposed Methodology is designed to support the organization's needs and help them improve their customer experience.

A Sentiment analysis-based system for helping the layman make informed choices is very scarce. The Existing system which uses sentiment analysis of comments uses the comments written in a particular forum. This might result in a biased result and will produce less accurate results. So, In the Proposed Methodology, we overcome this issue by developing a system which analyses the large amount of user discussion comments and reviews from multiple user discussion websites. The Proposed Methodology is aimed to gather the comments and reviews written by patients from numerous discussion forums and analyzes the sentiment of those comments and gives an informed and accurate result to the user.

The Proposed Methodology helps the patientsto find a hospital and doctor which can do the best during their period of illness.

2. LITERATURE SURVEY

Alexandra Fanca developed a method of fuzzy clustering, which improves the accuracy of predicting ratings of products. The method of association rules mining is developed, which solving the problem of searching associative items for recommendations. The method of categorical clustering is developed, which solving the problem of new user and new item [1].

Farhin Mansur uses the approaches of recommending a movie include a group affinity based social trust model for intelligent movie recommender systems as proposed. The proposed method evaluates sentimental similarities between users based on the ratings of movies and makes the results into abstract trust values. It also extracts user profiles and analyzes collective tendencies among the profiles. These two user tendencies are synthesized to form trust models and are applied to recommender systems [2].

Kwanghee Hong proposed a system tries to utilize the recommendations by friends and family (collaborative approach) and the content of the movie and purpose for watching it (content-based approach) for recommendation of the movies [3].

Marwa Hussien Mohammed introduces a survey about recommendation systems, and the technique they have used and the challenges faced by the recommendationsystems [4].

Madhuri Kommineni have used a User Based Collaborative Filtering (UBCF) approach and measured the performance of similarity measures in recommending books to a user [5].

Augusto Pucci proposed a research paper recommending algorithm based on the Citation Graph and random-walker properties. This paper presents a random–walk based scoring algorithm, which can be used to recommend papers according to a small set of users selected relevant articles [6].

Shreya Agrawal proposed a new recommender algorithm based on multi- dimensional users' behavior and new measurements. It is used in the framework of our recommendation system that uses knowledge discovery techniques to the problem of making product recommendations during a live user interaction [7].

S. Swarnalatha aimed at hospital recommendations such as in the system aims to extract important information and encoded in free-text patient comments. The system determines the most common topics in patient comments, design automatic topic classifiers, identify comments' sentiment, and find new topics in negative comments. This system could be used by the hospital to make better their services unlike other recommendation systems aimed at helping the customers make informed purchases. The drift towards utilizing the sentiments in the comments of customers is an evolving trend for building recommendation systems[8].

M. Viswa Murali used a collection of historical rating data of m users on n products as input, which are collected by asking users to input the rating of the products as numerical values. Text mining techniques are employed to extract useful information from review comments. Ontology has also been defined to translate the review information into a form suitable for utilization by the recommender system. A ranking mechanism for prioritizing that information with respect to the consumer level of expertise in using that product has been developed [9].

3. EXISTING WORK

Existing recommendation systems use many different approaches such as collaborative and content-based approach. One of the most important drawbacks of existing recommendation systems is that they usually use only rating matrix as useful information and not fully consider contextual information such as attributes (reviews, ratings, location, doctor name) for improving recommendation. In the Proposed Methodology, we consider multiple attributes of user-preference for improving the accuracy of the recommendation.

4. METHODOLOGY

The Proposed Methodology provides the information about the best hospital and doctors related to the disease to the patients based on the public comments. This is achieved through the Natural Language Processing (NLP) approach using Sentimental Analysis by considering the polarity factor. Polarity is used to determine the score (Positive, Negative) of a sentence. The Proposed System involves the following steps:

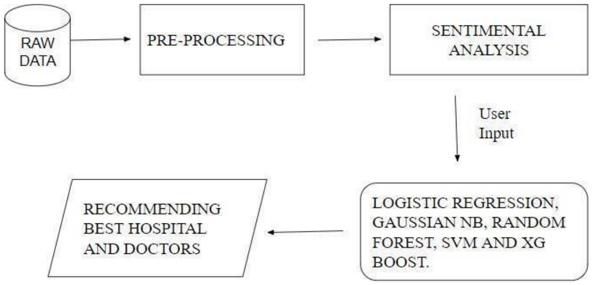

RAW DATA
PRE-PROCESSING
SENTIMENTAL ANALYSIS
User Input
LOGISTIC REGRESSION, GAUSSIAN NB, RANDOM FOREST, SVM AND XG BOOST.
RECOMMENDING BEST HOSPITAL AND DOCTORS

Fig 1. Architecture Diagram

a. The Proposed System gathers the comments on various hospitals from numerous public discussion and review forums periodically.

b. Performs sentiment analysis of the comments collected.

c. The disease would be predicted based on the symptoms given by the user by applying the algorithms.

d. The System suggests the best hospital for curing a particular ailment based on the polarity score.

a. Data Collection

Information about hospitals and doctors is obtained from comments and reviews posted by people across the world in different public forums of the world wide web and pre- processing is applied on a dataset which will remove all the unnecessary data. Then, the data is fed to the algorithms for further processing

b. NLP

Natural Language Processing or NLP may be a field of Artificial Intelligence (AI) that makes the machines to read, understand and derive meaning from human languages.

Companies like Yahoo and Google filter are using natural language processing to classify the emails by analyzing the text flow through their servers and stopping spam before they even enter your inbox.

c. Sentimental Analysis

Sentiment Analysis is a Process and knowledge

Extraction task that aims to get writer's feelings expressed in positive or negative comments, questions and requests, by analyzing large numbers of documents. Sentimental Analysis helps in determining whether a text is positive, negative or neutral, using various classification techniques. This analysis often aggregated over large sets of knowledge and therefore the resulting information is often helpful in several contexts. For example, positive –Looking very good and taking as a family care; negative – The service is not good; Here the words from the sentence are appropriately classified. So as to perform sentimental analysis, it's necessary to coach a classifier so to classify the unknown samples accurately. The various steps are illustrated in the diagram.

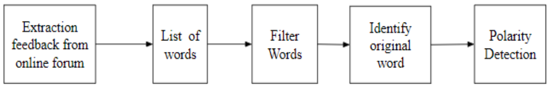


Fig 2. Sentimental Analysis Process

The feedbacks are first extracted from a web forum, these feedbacks are nothing but sentences that are posted by users that give an insight into the hospital's different aspects.

For example, "The hospital features a good service". First, clean the extracted data by removing slangs and emotions, using string operations and stop words using Natural Language Tool Kit (NLTK) English stop words dictionary. For example, words like an, the, are, has, removed. After cleaning the feedback, the results "Hospital good service". The subsequent step is to convert the sentence into an inventory of words and perform POS (Parts of Speech) tagging, this will be done by simply using the split() method which returns an array of terms within the sentences. For example, [hospital, service, good]

Polarity is the main factor estimated for sentimental analysis of the user's review and comments. The sentiment polarity may be a verbal representation of the sentiment. It is often "negative" or "positive". The sentimental score may be a more precise numerical representation of the sentiment polarity. The typical polarity of all the words during a sentence gives the polarity of an entire sentence.

d. Logistic Regression:

Logistic Regression has become a crucial tool within the discipline of machine learning. A logistic regression model predicts a dependentdata variable by analyzing the connection between one or more existing independent variables. For example, a logistic regression will predict whether a political candidate will win or lose an election or whether a high school student is going to be admitted to a particular college.

The purpose of logistic regression is to estimate the possibilities of events, including determining a relationship between features and therefore the probabilities of particular outcomes. One example is predicting if a student will pass or fail an exam when the amount of hours spent studying is provided as a feature and therefore the variables for the response has two values: pass and fail.

e. Gaussian Naive Bayes:

Naive Bayes may be a simple but surprisingly powerful algorithm for predictive modeling. Naive Bayes are often extended to real-valued attributes, most ordinarily by assuming a normal distribution . The extension of naive Bayes is known as Gaussian Naive Bayes. Other functions will not be able to estimate the distribution of the information, but the Gaussian (or Normal distribution) is the easiest to figure with because you simply got to estimate the mean and therefore the variance from your training data.

f. Random Forest:

Random forests are a supervised learning algorithm. It is often used both for classification and regression. It is also the foremost flexible and straightforward to use algorithm. A forest is comprised of trees. It is said that if there are more trees then the robustness is more. Random forests create decision trees on randomly selected da ta samples, gets prediction from each tree and selects the simplest solution by means of voting. It also provides a reasonably good indicator of the feature importance. Random forests feature a sort of applications, like recommendation engines, image classification and have selection. It classifies loyal loan applicants, identify fraudulent activity and predict diseases. It lies at the bottom of the Boruta algorithm, which selects important features during a dataset.

g. SVM:

Support Vector Machine (SVM) is a group of supervised learning methods used for classification, regression, and outlier's detection. This can be used to detect cancerous cells supported many images otherwise this can use them to predict future driving routes with a well-fitted regression model. There are specific sorts of SVMs you'll use for particular machine learning problems, like Support Vector Regression (SVR) which is an extension of Support Vector Classification (SVC). The main thing to stay in mind here is that these are just math equations tuned to offer you the foremost accurate answer possible as quickly as possible. SVMs are different from other classification algorithms due to the way they choose the choice boundary that maximizes the space from the closest data points of all the classes.

h. XG BOOST:

From predicting ad click-through rates to classifying high energy physics events, XG Boost has proved its mettle in terms of performance and speed. The beauty of this powerful algorithm lies in its scalability, which drives fast learning through parallel and distributed computing and offers efficient memory usage. XG Boost is an ensemble learning method. Sometimes, it may not be sufficient to rely upon the results of just one machine learning model. Ensemble learning offers a systematic solution to combine the predictive power of multiple learners. The resultant is a single model which gives the aggregated output from several models. The models that form the ensemble, also known as base learners, could be either from the same learning algorithm or different learning algorithms. Bagging and Boosting are two widely used ensemble learners. Though these two techniques can be used with several statistical models, the most predominant usage has been with decision trees.

Input for the five algorithms are the Symptom s, Reviews, Ratings and Doctors, based on the Symptoms, the Disease is predicted and suggesting the hospital using the algorithm which have the highest accuracy.

Table 1. Performance Comparison of algorithms

Algorithm	Accuracy
Logistic Regression	80
Gaussian NB	72
Random Forest	80
SVM	80
XG BOOST	87.5

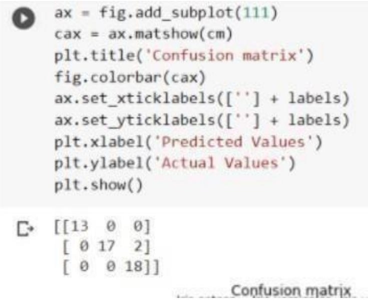
The accuracy for each algorithm as shown in Table 1 are 80,72,80,80,87.5. Among the five algorithms, XG Boost has the highest accuracy. So, the proposed system predicts the outcome by using the XG Boost algorithm.

5. EVALUATION METRIC

Evaluation metric is used to calculate the quality of the machine learning model. These metrics are essential and many types of evaluation metrics available to test a model. A confusion matrix provides us a matrix as output and the complete performance of the model.

5.1 Confusion Matrix

Confusion matrix is simply measuring performance of ML classifications.



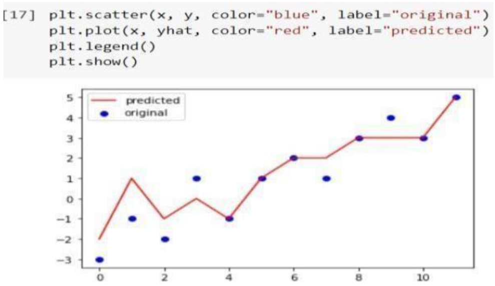
```
ax = fig.add_subplot(111)
cax = ax.matshow(cm)
plt.title('Confusion matrix')
fig.colorbar(cax)
ax.set_xticklabels(['' ] + labels)
ax.set_yticklabels(['' ] + labels)
plt.xlabel('Predicted Values')
plt.ylabel('Actual Values')
plt.show()
```

```
[[13  0  0]
 [ 0 17  2]
 [ 0  0 18]]
```

Fig 3. Confusion Matrix

The performance for the Classification Algorithm is shown in Fig 5.1

5.2 Root Mean Squared Error (RMSE) RMSE is one of the evaluation metrics used in regression problems.



```
[17] plt.scatter(x, y, color="blue", label="original")
     plt.plot(x, yhat, color="red", label="predicted")
     plt.legend()
     plt.show()
```

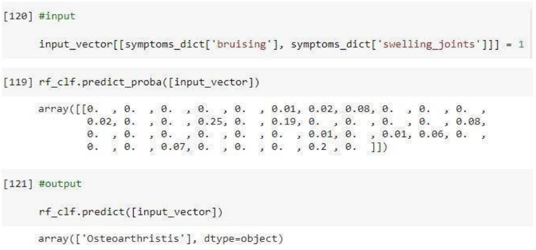
Fig 4. RMSE Metric

The RMSE Metric graph shows the predicted value and original value, the predicted value is drawn as a line and the original value is marked in dots shown in Figure 5.2

6. IMPLEMENTATION AND RESULT

Google Colab is a strong platform for learning and quickly developing machine learning models in Python. It supports the Jupyter notebook and supports collaborative development. Colab supports many popular ML libraries like PyTorch, TensorFlow, Keras and OpenCV. The restriction as of today is that it does not support R or Scala yet.There is also a limitation to sessions and size.

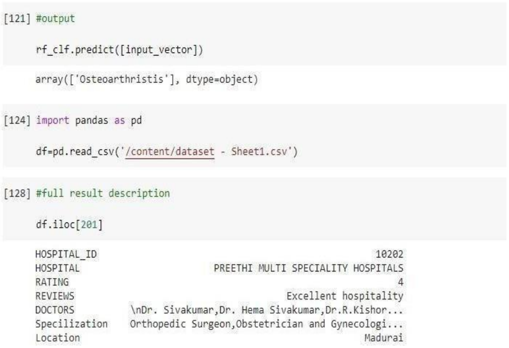
The interface asks the user for the symptom in which they would like to find the best hospital.



```
[120] #input

      input_vector[[symptoms_dict['bruising'], symptoms_dict['swelling_joints']]] = 1

[119] rf_clf.predict_proba([input_vector])

      array([[0.  , 0.  , 0.  , 0.  , 0.  , 0.01, 0.02, 0.08, 0.  , 0.  , 0.  ,
              0.02, 0.  , 0.  , 0.25, 0.  , 0.19, 0.  , 0.  , 0.  , 0.  , 0.08,
              0.  , 0.  , 0.  , 0.  , 0.  , 0.  , 0.01, 0.  , 0.01, 0.06, 0.  ,
              0.  , 0.  , 0.07, 0.  , 0.  , 0.  , 0.2 , 0.  ]])

[121] #output

      rf_clf.predict([input_vector])

      array(['Osteoarthristis'], dtype=object)
```

Fig 5. Disease Prediction

After giving the input, the system automatically detects the type of disease asshown in Fig 6



```
[121] #output

      rf_clf.predict([input_vector])

      array(['Osteoarthristis'], dtype=object)

[124] import pandas as pd

      df=pd.read_csv('/content/dataset - Sheet1.csv')

[128] #full result description

      df.iloc[201]

      HOSPITAL_ID                                                 10202
      HOSPITAL                          PREETHI MULTI SPECIALITY HOSPITALS
      RATING                                                          4
      REVIEWS                                      Excellent hospitality
      DOCTORS          \nDr. Sivakumar,Dr. Hema Sivakumar,Dr.R.Kishor...
      Specilization    Orthopedic Surgeon,Obstetrician and Gynecologi...
      Location                                                  Madurai
```

Fig 6. Recommended Hospital forOsteoarthritis

The Proposed System recommends the hospitalname and doctor name along with rating, review and location as shown in Fig 6.2.

REFERENCES:

1) Alexandra Fanca, Adela Puscasiu,Dan-Ioan Gota, Honoriu Valean. "Recommendation Systems with Machine Learning". 2020 21th International Carpathian Control Conference (ICCC).

2) Farhin Mansur, Vibha Patel, Mihir Patel. "A review on recommender systems". 2017 International Conference on Innovations in Information, Embedded and CommunicationSystems (ICIIECS).

3) Kwanghee Hong, Hocheol Jeon, Changho Jeon. "UserProfile-based personalized research paper recommendation system". 2012 8th International Conference on Computing andNetworking Technology (INC, ICCIS andICMIC)

4) Marwa Hussien Mohamed, Mohamed Helmy Khafagy MohamedHasan Ibrahim. “Recommender Systems Challenges and Solutions Survey” 2019 International Conference on Innovative Trends in Computer Engineering(ITCE)

5) Madhuri Kommineni, P.Alekhya, T. Moha na Vyshnavi, V.Aparna, K Swetha, V Mounika. “Machine Learning based Efficient Recommendation System for Book Selection using User based Collaborative Filtering Algorithm”. 2020 Fourth International Conference on Inventive Systems and Control (ICISC).

6) Marco Gori, Augusto Pucci. “Research Paper Recommender Systems: A Random-Walk Based Approach”. 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06)

7) Shreya Agrawal, Pooja Jain. “An improved approach for movie recommendation system” 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analyticsand Cloud) (I-SMAC).

8) S.Swarnalatha, I.Kesavarthini, S. Poornima,N. Sripriya. “Med-Recommender System for Predictive Analysis of Hospitals and Doctors” Second International Conference onComputational Intelligence in Data Science (ICCIDS-2019).

9) M Viswa Murali, T G Vishnu, Nancy Victor. “A Collaborative Filtering based Recommender System for Suggesting New Trends in Any Domain of Research”. 2019 5th InternationalConference on Advanced Computing & Communication Systems (ICACCS).