MEDBOT: A CHATBOT FOR DETERMINING THE PROBABLE DISEASES BASED ON THE USER’S SYMPTOMS

Karan Gori*1, Yasir Ahmed*2, Shubham Chikane*3, Amrita Mathur*4

*1,2,3Student, Information Technology Department, St. Francis Institute Of Technology, Mumbai, Maharashtra, India.
*4Asst.Professor, Information Technology Department, St. Francis Institute Of Technology, Mumbai, Maharashtra, India.

ABSTRACT

Many times, a patient might underestimate an underlying health condition and neglect it, which may worsen over time and cause a life-terminating disease and sometimes a person might irrationally worry about having a serious medical condition leading to unnecessary health anxiety. Patients in both these scenarios can be helped by having an interface that can provide preliminary diagnostics regarding if he/she has a serious medical condition or it’s just health anxiety. A chatbot can be one such interface that can be deployed to gain information of a patient’s symptoms and determine the underlying medical conditions if any, based on the information given by the patient. Chatbots can be deployed using various technologies such as Deep Learning, Natural Language processing and long-short term memory networks.

Keywords: Medical chatbot, Deep Learning, SVM, KNN, ELMo Model, Universal Sentence Encoder.

I. INTRODUCTION

In the era of smartphones, every person has access to applications which helps in performing any required task. A chatbot can be one such application that can be easily accessible due to this advancement. As they offer a user-friendly UI it can be used in many fields. This project is an attempt to introduce a chatbot in the medical field, by making use of various algorithms, datasets and deploying the same as a web app which is a medical chatbot. Health is a matter of great concern in today's world as the lifestyle has evolved to such an extent that things have gone virtual and artificial, thus neglecting the health. The existing healthcare infrastructure expects a patient to get an appointment from a doctor and either be in a clinic physically or virtually in some exceptions for diagnosis. In this fast paced digital world many people choose to ignore potentially life threatening symptoms to avoid such an infrastructure. Due to the advancements in Artificial Intelligence it is possible to predict any life threatening diseases early in just a matter of time. These systems cannot replace a doctor but can act as a prerequisite for a patient. The proposed medical chatbot expects a user to input descriptions of symptoms that he/she is experiencing and the chatbot will give the closest diseases as output. For extracting important features of symptoms Named Entity Recognition (NER) is used and for the prediction of diseases, Support Vector Machine (SVM) algorithm is used. For pandemic situations, technical support that ensures some assurance to health is always fruitful so a chatbot in the form of a medical assistant can be helpful. It will detox all the stressful thinking and reduce mental strain to some extent. In remote places where travelling and the availability of doctors is an issue, it can be used as a backup medical resource.

II. RELATED WORKS

In this paper we have introduced a system which is trained on sentences consisting of various symptoms and later by using the dataset consisting of disease and the set of symptoms they possess the most probable disease the user may be suffering from is determined.

Use of a chatbot in medical field

The first paper of our reference title of which is “A novel approach for medical assistance using trained chatbot” gives us the brief idea about how this technology can be proven useful in the medical field, by allowing the person having some symptoms to know the disease he/she might be suffering from by giving their symptoms as an input to the chatbot.

Importance preprocessing of the data

The paper titled "Designing Disease Prediction Model Using Machine Learning Approach" [2] and "Disease Prediction and Early Intervention System Based on Symptom Similarity Analysis" [4] highlights the importance
of preprocessing of the dataset before passing it to the classification algorithm. According to [4] the preprocessing can help in reducing the computation power needed for the prediction and can improve the efficiency of the working of the model.

**Algorithm required for the working of the proposed system**

The paper titled "Chatbot for Disease Prediction and Treatment Recommendation using Machine Learning" [3] shows the use of KNN algorithm to predict the disease by converting the symptoms entered by the user into a bag of word and later passing it to KNN to form map the symptoms to the most probable disease. [5] gives the idea of use of LSTM along with the sentence embeddings in order to make the symptoms entered by the user into to the form acceptable by the algorithm.

### III. SYSTEM IMPLEMENTATION

Model and Material which are used is presented in this section. Table and model should be in prescribed format.

![System Block Diagram](image)

**Figure 1: System Block Diagram.**

**Creating the dataset**

List of symptoms keywords is made by extracting them from the disease-symptoms dataset. The tweets containing the words which are part of this list are fetched. Some of these fetched tweets contain the words which are the symptoms but in context they don’t talk about the disease, so such tweets are eliminated in order to avoid inconsistency. The emojis present in the tweets, along with retweet keywords are cleaned by using the python's regular expression function regex() by passing the various parameters to it in accordance to the inconsistency which is to be eliminated, which in our case is the hashtags, emojis, hyperlinks of retweets, etc. The dataset of these tweets are made sentence wiz where each sentence is a tweet. Each sentence contains words which represent a symptom or continuation of symptom or non symptom. Each symptom in this dataset is given a label 'b-sym' indicating the starting of the symptom and 'c-sym' indicating the continuation of the symptom, rest of the other words of the sentence are given the tag 'o'. These labels are assigned in order to train our symptoms recognition model to differentiate the words in these three categories, so that when an end user explains his suffering to the medical chatbot the chatbot is able to differentiate the words in the above mentioned categories and proceed with the further phases which are the part of the implementation.

**Recognizing Symptoms through Named Entity Recognition (NER)**

The dataset is made of sentences extracted from twitter, where each sentence consists of the symptoms keywords that were pre-decided on the basis of disease-symptoms dataset. These sentences are passed through
Embeddings from Language Models (ELMo) in order to assign the mathematical value to each word, thus generating the vector of dimension 512. The ELMo model uses embeddings from a language model that is trained on the 1 Billion word benchmark. The ELMo model computes contextualized word representations using character based representations and bidirectional LSTM network. These vectors and their labels are passed through the Long Short Term Memory network, in order to train the model to recognize the symptoms present in the sentence entered by the user. Here transfer learning was used by using weights of ELMo Model as the twitter dataset was small in context of Neural Network. This LSTM network consists of three bi-directional LSTM layers each having 256 neurons and a 512 neuron dense layer. Softmax activation is used here as there are multiple classes (b-sym, c-sym, o).

$$\text{softmax}(z_j) = \frac{\exp(z_j)}{\sum \exp(z_j)}$$

(1)

Here ‘z’ is the value of the neuron from the output layer.

After the training of the model, this completes Named Entity Recognition consisting of ELMo and LSTM is able to recognize the symptoms present in the sentence entered by the user. These symptoms are passed to Universal Sentence Encoder which is the next phase of the system.

**Processing the recognized symptoms via Universal sentence encoder**

The recognized symptoms received from the LSTM network are pre-processed. Lemmatization was used in this step. Using lemmatization, a word can be reduced to its base form or dictionary form. For example: Words like changed, changes would be reduced to “change” after lemmatization. In the pre-processing step stop words as well as pronouns were removed from the recognized symptom text.

These pre-processed symptoms were passed through Universal Sentence Encoder, which encodes these symptoms into a vector of dimension 512. Universal Sentence Encoder comes in two variations i.e. Deep Averaging Network (DAN) and Transformer encoder. We have used Deep Averaging Network variation of Universal Sentence Encoder as it is computationally less expensive. These help in mapping the group of symptoms to their respective disease present in the disease-symptoms dataset. The vector of dimension 512 generated by the Universal Sentence Encoder is then passed to Support Vector Machine.

**Mapping the unique vector obtained by universal sentence encoder to a disease in the dataset**

There are many methods to perform the task of mapping the set of symptoms entered by the user to the disease based on the training of the model. They are Cosine Similarity, KNN, Manhattan Distance & Support Vector Machine.

With the help of SVM we are able to give three top most probable diseases, the user may be suffering from based on the symptoms entered by him/her.

**IV. RESULTS AND DISCUSSION**

![Figure 2: Medical Chatbot.](image-url)
NER Model
The NER model which consists of ELMo and LSTM network when tested on the twitter dataset gave an f1 score of 0.84 for picking symptoms phrases.

This NER model was also tested on the NCBI-disease dataset. This dataset consists of text sentences where diseases and other words were labelled in BIO formatting. After testing it on our NER model the f1 score for picking disease phrases came out to be 0.87.

SVM Model
The SVM model was tested on AG news dataset for classifying which category the news article belongs to on the basis of the news article's title. This dataset consists of article's category, title, description. Universal Sentence Encoder was used here to encode the titles of new articles to a vector of 512. The SVM model was then trained on these vectors for classification of the article's category. The accuracy of this SVM model came out to be 83.08% for correctly classifying the category of the new article.

V. CONCLUSION
The project is an approach of introducing the machine learning in the medical field. The chatbot can give a brief idea of the disease a person is suffering from based on the symptoms he/she is having. This project can be made better by adding features such as video call with a doctor and integrating this application with nearby hospitals. The disease-symptoms dataset can be improved by validating it with an experienced personnel in medical field. This would make this dataset more consistent and make the predictions more appropriate. In future the twitter dataset fetched for NER model could be replaced by actual transcripts of patient-doctor conversation. By doing this dataset will be more refined and NER model may give even higher accuracy. the main points of the research work are written in this section. Ensure that abstract and conclusion should not be same. Graph and tables should not use in conclusion.

VI. REFERENCES


