CAPSTONE PROJECT REPORT ON PREDICTION OF DISEASE BASED ON SYMPTOMS

Submitted to

Praxis Business School, Kolkata

Post Graduate Program in Data Science
May 2022

BY

Mriganka Paul(A21019)

Adarsh Raj(A21001)

Manpreet Singh(A21017)

Avirup Saha(A21008)

K V V S Krishna prasad(A21016)

Mohammad Shaan(A21018)

Guided by : Prof. Dr. Subhasis Dasgupta

Department of Data Science Academic Year: Jul 2021 – May 2022

Acknowledgement

We are profoundly grateful to Prof. Dr. Subhasis Dasgupta, Head of Machine Learning & Analytics for his expert guidance and continuous encouragement throughout to see that this project meets its target from its commencement until completion.

Lastly, we would like to express our sincere heartfelt gratitude to all the staff members of the Data Science Department who helped us directly or indirectly during this course of work.

Regards,

Adarsh Raj

Avirup Saha

Krishna Prasad

Manpreet Singh

Mohammad Shaan

Mriganka Paul

ABSTRACT

Applications in the field of machine learning and artificial intelligence have been in great demand over the recent decade. Now it has various applications in the field of health industry. With the help of machine learning algorithms, prediction of diseases has been made easier. Now doctors can concentrate only on treatment with the help of technology. Technology is accelerating innovations in the healthcare domain which has increased people's standard of living over the years. Here in our project we are making a healthcare chatbot with help of Natural language processing and machine learning algorithms to predict disease. User interacts with the chatbot just like one interacts with his doctor and based on the symptoms provided by users and the chatbot will identify the symptom and predict the disease.

INTRODUCTION

It is important to maintain health if one wishes to be happy. Only a healthy body can have a healthy mind. Nowadays, people are less aware of their health. In their busy life, they forget to take suitable measures to maintain their health and are less aware of their health status. Such a problem can be avoided by using the symptoms driven disease prediction application. This research intends to apply the concepts of natural language processing and machine learning to create a chatbot application. People can interact with the chatbot just like they do with another human and through a series of queries; chatbot will identify the symptoms of the user and thereby, predicts the disease using a machine learning algorithm. This system can be of great use to people in conducting daily check- ups, makes people aware of their health status and encourages people to make proper measures to remain healthy. According to this research, such a system is not widely used and people are less aware of it. Executing this proposed framework can help people avoid the time consuming method of visiting hospitals by using this free of cost application, wherever they are. The disease prediction chatbot can make a great change in the health of our society. It is more reliable and less prone to human errors. People avoid visiting hospitals over small issues which can be a major problem in the future. Idea is to focus on the solution which is free of cost and available throughout the day. In this way people can be more aware about their health. A user can ask questions at any time of the day even with his busy

schedule and keep a check on his own health without visiting any specialized doctor just for consultation.

PROBLEM STATEMENT

Hospitals are the most widely used means by which a sick person gets medical check-ups, disease diagnosis and treatment recommendation. This has been a practice by almost all the people over the world. People consider it as the most reliable means to check their health status. Normally users are not aware about all the treatment or symptoms regarding the particular disease. For small problems a patient has to go personally to the hospital for check-up which is more time consuming. Consequently, prediction systems using Machine Learning and deep learning have recently been a significant effort to lessen doctor's workload and help develop the overall competence of the health care method with machine learning.

PROJECT OVERVIEW

The main idea was to build a chatbot that would be able to predict common diseases using symptoms provided by the user. For this project we took an NLP approach.Starting from data extraction from various healthcare websites using python and selenium, then this data was tagged using manually so that machine will be able to understand different entities and then using the keyword extracted from our Custom Ner model we created the machine learning model that was able to predict particular disease on give set of symptoms.

STEPS FOLLOWED IN THE PROJECT

- **Data extraction:** Using selenium to extract data from different websites related to healthcare/common diseases namely tata 1mg, Hopkins medicine, Healthline, Everyday1, emedicine using web scraping.

- **Keyword tagging:** Tagging the keywords manually using custom annotator tool (UIBAI) and saving it as IOB format

- **Training transformer model**

    - Extracting IOB format files of the manually tagged data
    - Converting IOB format files to spacy format for usage in training transformer model
    - Creating train & validation set for spacy format files
    - Using CUDA & Spacy's model pipeline that wraps HuggingFace transformer package to train the model
    - The model is then tested on the validation set
    - Extracting the model and using it to predict on unseen custom text

- **TFIDF vectorizer**
    - Cleaned the unwanted texts
    - Created the pandas dataframe with 2 features, where one feature is the disease name and the other feature is the texts of the disease that were extracted.
    - Converted all the words to small caps
    - Created the tfidf vectorizer using the vocabulary where vocabulary contains all the keywords that we received from the Transformer Model.
    - Then passed the corpus through this tfid vectorizer

- **Logistic Regression**
    - Trained the model using the corpus that was passed through the tfidf vectorizer which was the independent variable and name of the diseases were the dependent variable
    - Calculate the kappa score

How does it work :

- User is asked to give his/her name age gender and symptoms
- The important keywords are extracted and matched with the symptoms to predict the possible disease on the basis of probability score
- To improve the predictability the user is again asked to verify other possible symptoms on the basis of the initial predicted disease to improve the prediction probability
- Finally the predicted disease is given

Future scope:

This project will surely help a person who has less knowledge on the medical science and can save a little from his/her hard earned money by getting the initial direction. By this application one can get an idea what one might be suffering from. This will also save a huge amount of time of the patients and can visit the right doctor.

Presently due to lack of time we could not build the front end and will be showing the demo in the command prompt. But this is a future scope where we can have user friendly front end where the user can interact with the application Also in our present work, we are not displaying names or details of doctors which can also be a part of future scope.

And one more important enhancement that we have as our future scope is that for each of the diseases we will display the name of the tests that the patient can go for which will in turn help the patient in investing money only on those tests which are required for the disease predicted by the application or the doctor.