Multilingual Healthcare Chatbot Using Machine Learning

Conference Paper · May 2021

DOI: 10.1109/INCET51464.2021.9456304

Sagar Badlani
Georgia Institute of Technology

Sheetal Chaudhari
Sardar Patel Institute of Technology

2021 2nd International Conference for Emerging Technology (INCET)
Belgaum, India. May 21-23, 2021

Multilingual Healthcare Chatbot Using Machine Learning

Sagar Badlani
Department of Information Technology
Sardar Patel Institute of Technology
Mumbai, India
sagarbadlani2@gmail.com

Tanvi Aditya
Department of Information Technology
Sardar Patel Institute of Technology
Mumbai, India
tanviaditya1@gmail.com

Meet Dave
Department of Information Technology
Sardar Patel Institute of Technology
Mumbai, India
meetdave324@gmail.com

Sheetal Chaudhari
Department of Information Technology
Sardar Patel Institute of Technology
Mumbai, India
sheetal_chaudhari@spit.ac.in

***Abstract*— The healthcare sector is one of the largest focus areas in the world today. Health problems are becoming increasingly common. India faces a huge challenge in terms of managing rural healthcare. Early diagnosis and treatment of diseases can play an instrumental role. Physical consultation, particularly in the rural areas, is costly and time consuming. The solution is adopting healthcare chatbots. The proposed solution describes a multilingual healthcare chatbot application that can perform disease diagnosis based on user symptoms. It also responds to user queries by calculating sentence similarity by using TF-IDF and Cosine Similarity techniques and choosing the most appropriate response from its knowledge database. The multilingual capabilities of the chatbot system make it highly suitable for use in rural India. The chatbot application currently supports three languages namely English, Hindi and Gujarati. The chatbot application converses with the user using concepts of Natural Language Processing and also supports speech to text and text to speech conversion so that the user can also communicate using voice. Five different Machine Learning algorithms have been analyzed for disease prediction. The Random Forest Classifier produces the best results and gives an accuracy of 98.43%. Thus, it is used as the system's core classifier.**

Keywords*— *Healthcare, Chatbot, Disease Prediction, Natural Language Processing, Multilingual, Speech to Text, Text to Speech, Cosine Similarity, Random Forest Classifier, TF-IDF

I. INTRODUCTION

The healthcare sector is one of the largest focus areas in the world today. Individuals are becoming increasingly susceptible to lifestyle diseases. Early diagnosis and treatment of diseases becomes a key factor in this regard. The Indian healthcare sector is facing a steep challenge in the form of rural healthcare. This is a major impediment that faces the Government of India. As per the 2011 census[1], about 68.8% of the Indian population lives in rural areas having low levels of healthcare facilities and high mortality rates due to diseases. In many Indian States, rural people have to travel large distances to get proper health diagnosis and treatment[7]. The healthcare industry is facing an acute shortage of medical professionals. Physical consultation with medical professionals is time consuming and expensive. Particularly in rural areas, consultation with qualified professionals is not easily available[10]. The affordability is also a concern for the rural folks. As a result, many people avoid visiting healthcare professionals and even treatable diseases/ailments become life-threatening, contributing to the high mortality rates.

The solution is adopting healthcare chatbots[7]. They ensure initial assistance over one-on-one conversations economically and on-demand. The healthcare chatbots can be used by doctors for monitoring their patients.

This century belongs to computer intelligence. Advanced technologies to aid humans are being developed. An example of these advancements is chatbots based on Artificial Intelligence (AI). An AI-based[8] chatbot is a software application that is primarily used for imitating human conversations and providing the solution for queries provided by the user. As new businesses are flourishing with the use of the latest technology, the use of chatbots in the daily life of consumers is increasing rapidly. They can be used for various purposes like customer servicing, request routing or information fetching. One such application of chatbots is in the healthcare sector. The author in the paper [11] has discussed the importance of chatbots in the healthcare sector. They have mentioned the fact that healthcare chatbots can have wide applications ranging from booking appointments to setting reminders and consuming medicines.

The proposed solution focuses on a multilingual healthcare chatbot application that analyses the user's symptoms through a conversation with the user and maps these symptoms to the available dataset. The system can converse with the user via text or speech. The user can choose the language he wishes to communicate in. Natural Language Processing (NLP) is applied to the user input to determine the symptoms. The symptoms are then passed to a Machine Learning (ML) algorithm that has been trained to diagnose diseases based on symptoms.

The proposed solution incorporates multilingual text and speech capabilities which are indispensable for rural India. The literature survey suggests that research on healthcare chatbots has been limited to the English language or a particular language. The proposed solution supports multiple Indian languages in addition to English. The solution can answer user queries using sentence similarity in addition to disease prediction.

A chatbot cannot be expected to give a formal diagnosis. However, it can be used to provide useful information if provided with the symptoms. The chatbot can make a

978-1-7281-7029-9/21/$31.00 ©2021 IEEE

predictive diagnosis. This can assist in providing the initial response as well as guide the individual to a specialised healthcare professional. Healthcare chatbots can be used as healthcare assistants for doctors as well as patients[11].

II. Related Work

In the paper by Moshiur Rahman et al., they have proposed a chatbot for healthcare purposes. The chatbot is based on the concept of ML. This chatbot can converse with the user in the Bengali language only. They have explored six different types of supervised ML approaches. They have experimented with AdaBoost and Decision Tree algorithms. They have also worked with Support Vector Machine(SVM) and Random Forest Classifier in addition to Multinomial Naive Bayes (MNB) and K-Nearest Neighbors (KNN). Out of the above algorithms, SVM gave them the highest accuracy and Multinomial NB gave the least accuracy. For prediction in Bengali language, they have converted their dataset in Bengali language. They have used Named Entity Recognition for extracting various user fields like Name, Blood group, and age from unstructured user input. They have used TF-IDF for converting the Bengali text into vectors and the Cosine Similarity technique has been used for generating the similarity between texts. Information about diseases is being fetched by their system using Cosine Similarity measure[2].

The paper by R. B. Mathew et al. focuses on building a chatbot system where users can talk about their personal health-related issues. Their system is trained on a disease dataset consisting of many symptoms. Using NLP the system predicts the disease through analysis of symptoms provided by the user. They have used 75% of the dataset for training and the remaining 25% for testing purposes. KNN Machine Learning algorithm has been used to predict disease from symptoms. For text preprocessing using NLP, they have used the NLTK toolkit available in python. After analysis of symptoms, it provides useful links for treatment and medicine details[3].

The paper[4] has demonstrated a medical chatbot system that provides medical aid to the customers through human interactions using natural language diagnosis. The author has trained the dataset on three different models - SVM, KNN and Naive Bayes algorithm and has obtained the accuracy of 94.66%, 88.66% and 80% respectively.

The paper[5] implements a chatbot application in the medical sector for disease prediction and as a question-answer forum for health-related queries. It aims to replicate human-like conversation. The paper suggests the use of word order similarity while formulating a vector. The system makes use of Google API for text to voice and vice versa conversion. It makes use of a medicine API to fetch medicine-related information. It makes use of the SVM algorithm for disease classification.

The research paper[6] provides an overview of chatbot solutions for easier communication between customers and sellers. The system aims to deal with frequently asked questions using query expansion mechanisms with NLP. The user input passes through the query expansion module and then Cosine Similarity is used to return the most appropriate response to the user query.

The author, Urmil Bharti et al. in their paper have proposed a conversational bot named ‘Aapka Chikitsak’ for getting information regarding the COVID-19 pandemic, developed using Dialog Flow and a serverless architecture. This application is developed with the intention to interact with the patients in the form of virtual doctors. It principally aims to extend the patient's access to knowledge in the field of healthcare with multilingual support [7].

Lekha et al. have created a chatbot to answer user queries and provide basic details related to diseases. The user input is pre-processed using tokenization and stop-word removal. N-gram and TF-IDF techniques are used for keyword extraction. The keywords are weighed using a Cosine Similarity algorithm and the most appropriate response is returned from the chatbot’s knowledge database[8].

III. Proposed Methodology

The proposed system consists of a user-friendly chat interface through which a user can communicate with the system. The user can either enter symptoms which he is experiencing or enter some health-related queries. Depending on the user input, the chatbot will predict the disease or provide relevant information about his queries.

The system has pre-processed the disease dataset, converting the categorical values into a suitable structured numerical dataset for training of the Machine Learning model. The system has trained and compared five different Machine Learning classification models - SVM, KNN, Decision Tree, MNB and Random Forest Classifier - and got the best accuracy using the Random Forest Classifier of 98.43%. Thus, the primary classifier which has been used in the system is the Random Forest Classifier.

Initially, the user must select their preferred language of communication. Currently, the system supports three languages,i.e., English, Hindi and Gujarati. The next step for the user is to select their preferred mode of communication, i.e., voice or text.

The input entered by the user will first be converted into text by the system if the user is communicating via speech. The speech to text conversion has been accomplished using the SpeechRecognition library available in python. The input text will be converted into English if it is in some other language. This language translation has been achieved using the Googletrans python library. The translated input is passed on to the NLP module.

The NLP module performs tokenization which is splitting the sentence into words, it then converts the bag of words into a lower case format, and later removes the stop words i.e. commonly used words. After the removal of stop words, those words are converted into their root form using stemming. The system performs keyword extraction on the processed corpus.

The system checks if the obtained keywords correspond to a health-related user query or they correspond to symptoms that the user is experiencing. If the keywords are user symptoms, the system performs disease diagnosis. The trained Machine Learning model is used for this purpose. For user symptoms, there is a threshold of four symptoms for better disease prediction. If a user enters less than four symptoms, the accuracy of the prediction will be less as many diseases have common symptoms. Thus, a threshold of four has been decided. If the chatbot receives a lesser number of symptoms than the threshold value, it prompts the user to enter more symptoms. When the disease has been predicted,

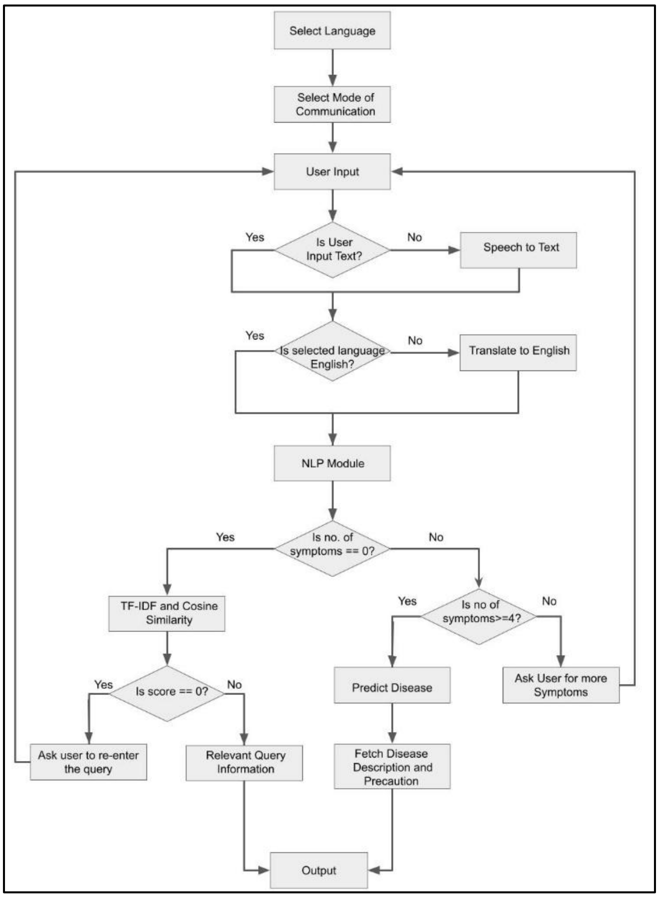

Select Language
Select Mode of Communication
User Input
Is User Input Text?
Yes
No
Speech to Text
Is selected language English?
Yes
No
Translate to English
NLP Module
Is no. of symptoms == 0?
Yes
No
TF-IDF and Cosine Similarity
Is no of symptoms>=4?
Yes
No
Is score == 0?
Yes
No
Predict Disease
Ask User for more Symptoms
Ask user to re-enter the query
Relevant Query Information
Fetch Disease Description and Precaution
Output

Fig. 1. Flow diagram for the proposed system

the result is sent to the user along with the corresponding disease description and precautions or steps which the user can take. The disease description and precaution information is obtained from the dataset. If the system is not able to extract any symptoms from the user input, it classifies the input as a health-related query. The system then applies the TF-IDF and Cosine Similarity techniques to find the most appropriate response to the user query from the knowledge database that has been provided to it. The final response is then converted back to the user's preferred language using the Googletrans python library. The output is presented to the user as text if their preferred mode of communication is text. Otherwise, if the user's preferred mode of communication is voice, then the system reads out the output to the user with the help of the gTTS and playsound libraries. Fig. 1 describes the proposed system in detail.

A. Data Preprocessing

The dataset obtained from Kaggle [9] was raw data with 4920 records and 41 unique diseases, including the mapping of disease with the corresponding symptoms. Dataset also included a description of each disease and corresponding precautions. The dataset was checked for inconsistencies and then the count vectorizer technique was applied to convert unstructured categorical data into structured numerical data. The final dataset consists of each column as a unique symptom and row with a disease. If a symptom belongs to any disease, then the corresponding cell has a value of 1, otherwise the value is 0. Thus, for any disease-symptom pair, a value of 1 indicates the presence of that particular symptom for that disease while a value of 0 indicates that the disease does not exhibit the corresponding symptom.

B. Voice to Text and Language Translation

For the implementation of language translation, speech to text and text to speech, the following mentioned libraries were used: Googletrans, gTTS, SpeechRecognition, Playsound. Googletrans is fast and reliable, supports automatic language detection and bulk translation. The system is using this library for translating the user's preferred language into English for NLP processing. gTTS stands for Google Text-to-Speech. It is a library which is supported by python. It converts text to a spoken mp3 file. SpeechRecognition creates a new Microphone instance through which the user speech is recorded and converted to the corresponding text for further processing. Playsound libraries play the sound file. In this case, the mp3 file which is saved using the gTTS library is played.

C. NLP text pre-processing

NLP is used to give users a human chat-like and easy communication experience. The user input text has to be pre-processed so as to derive symptoms and proper keywords from it. Pre-processing methods – tokenization[8], stemming[6], TF-IDF[12] and Cosine Similarity[13] have been used to generate an appropriate response for the user query.

1) Tokenization: User input text is first converted into lower case which is known as case-folding. The lower case user input is raw text which is transformed into a bag of words using tokenization. This will separate the text into the terms that compose it. Tokenization is useful for dealing with and analyzing each word independently. Further, all the punctuations are removed and the final bag of words is obtained.

2) Stop words removal: In order to extract important keywords, stop words such as 'a', 'an', 'the', etc are removed from the bag of words obtained from the previous pre-processing step. Removal of stop words is necessary as they take valuable pre-processing time and space.

3) Stemming: The bag of words is then iterated and the root form of each word is generated by removing suffixes or prefixes. This process is known as stemming in NLP. The input of the stemmer is the tokens that are generated. Here, the system is using the Porter Stemmer algorithm which gives the best output as compared to the other stemmers.

4) TF-IDF: Term Frequency (TF) is an indicator of the number of occurrences of a term in the sentence. Inverse Document Frequency (IDF) measures the uniqueness and hence the significance of a given term in the document. Weight of the term in the document is obtained by combining TF and IDF values.

$$Wi = tf * idf \qquad (1)$$

tf = number of times term occurred in a sentence
idf = inverse document frequency

Using the above formula, weight of each term from the user input is calculated and the resultant vector is obtained. Similarly, after the same pre-processing and weight calculation of terms in the list of questions database, another vector is formulated.

5) Sentence similarity: Similarity between two sentences is the distance between two vectors formed by those sentences. Cosine similarity[13] method is used to find the distance between those vectors. If the cosine angle between two sentence vectors is greater than 0, corresponding text is returned to the user in the form of chatbot response else the user is asked to enter proper information.

D. Classification Algorithms

The system has compared the following classification algorithms for disease classification:

1) Random Forest Classifier[14]
2) K-Nearest Neighbors(KNN)[15]
3) Support Vector Machine(SVM)[5]
4) Decision Tree[16]
5) Multinomial Naive Bayes(MNB)[17]

The above classifiers are supervised learning algorithms. Supervised learning is a type of machine learning where the models are trained by example. In this case, the training set consists of attributes that are mapped to labels or target values. The algorithm correlates the inputs and the outputs in the training data and learns from them.

The Random Forest Classifier is an ensemble of Decision Trees. It can be compared to a forest. A forest having more trees is said to be more robust. Similarly, a random forest makes use of multiple decision trees. It trains the Decision trees on subsamples of the original training set. This improves the accuracy of the Random Forest Classifier and reduces over-fitting.

KNN works on the idea of similar things existing in close proximity. The KNN algorithm uses the similarity between the features for the purpose of classification. The new data point is assigned a value based on its similarity with the training data points. It is one of the simplest classification algorithms that is used.

The main aim of the SVM algorithm is to discover a hyperplane in an n-dimensional space. For any classification problem, multiple hyperplanes are possible. The hyperplane which is chosen must have the maximum margin, i.e. the distance between the data points of the different classes must be maximized. The hyperplanes act as boundaries between different classes.

A Decision Tree is a classification tool that is in the form of a tree. The training data is continuously split according to the attributes.

The MNB algorithm is a probabilistic learning method. The Bayes theorem is based on the simple assumption that every feature is independent of the other features. It uses this assumption to classify the given sample.

IV. Experiments and Results

The system has been created as a web application. All the experiments for the machine learning model were performed using Python 3.8.8 as the core language along with scikit-learn Machine Learning library. The results for the disease classification algorithm have been obtained by comparing five different classification algorithms which have been mentioned in the proposed methodology.

The dataset used has 4920 records. Training and testing split has been performed on the dataset. The training data has 80% of the records while the testing data has been assigned a 20% split.

The approach involves testing the different algorithms on the dataset and evaluating the results using the K-fold Cross Validation score. The 10-fold Cross Validation has been used for validating the different classifiers in the experiment. The results obtained from the 10-fold Cross Validation are presented in in Table I.

The K-fold Cross Validation scores in Table I indicate that the Random Forest Classifier performs the best among the five classification algorithms tested on the dataset. The algorithms have been further evaluated using three different performance evaluation metrics. The experimental results for the algorithm evaluation have been summarized in Table II.

Fig. 2 depicts a graph comparing the accuracies of the different algorithms which have been trained and tested on the same train and test datasets.

Table II and Fig. 2 clearly indicate that the Random Forest Classifier has the highest accuracy and performance for the system. The experimental results indicate an accuracy of 98.43% for the Random Forest algorithm. MNB gives the worst accuracy for the system. Thus, Random Forest has been used as the primary classification algorithm for the system.

The implementation of the multilingual healthcare chatbot system is shown in Fig. 3. Fig. 3 shows the user the communicating with the chatbot system which analyses user’s symptoms and predicts the disease using the Random Forest Classifier.

In Fig. 3, the user is entering symptoms corresponding to jaundice.

The chatbot accurately analyses these symptoms and gives the user the appropriate disease diagnosis.

Fig. 4 depicts a similar conversation between the user and the chatbot where the chatbot performs disease diagnosis. However, the conversation is shown in the Hindi language demonstrating the multilingual capability of the chatbot system. The user enters symptoms in Hindi and the chatbot analyses the user text and accurately predicts Gastroesophageal reflux disease (GERD) and returns the

TABLE I. RESULTS OF K-FOLD CROSS VALIDATION

Classification Algorithm	K-Fold Cross Validation Score
Random Forest	0.9715
Decision Tree	0.9641
SVM	0.9600
MNB	0.9471
KNN	0.9688

TABLE II. ALGORITHM EVALUATION METRIC RESULTS

Classification Algorithm	Accuracy	Precision	F1-Score
Random Forest	0.9843	0.9774	0.9781
Decision Tree	0.9712	0.9693	0.9697
SVM	0.9622	0.9547	0.9582
MNB	0.9539	0.9440	0.9454
KNN	0.9788	0.9731	0.9749

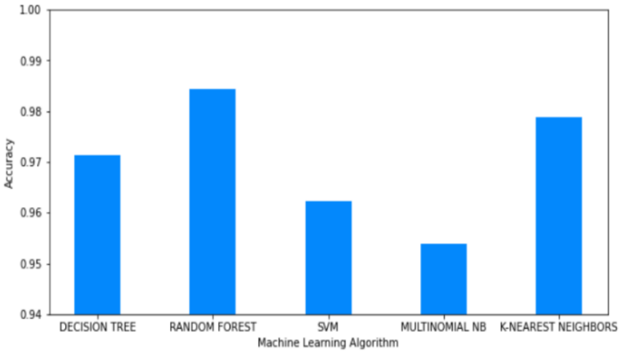
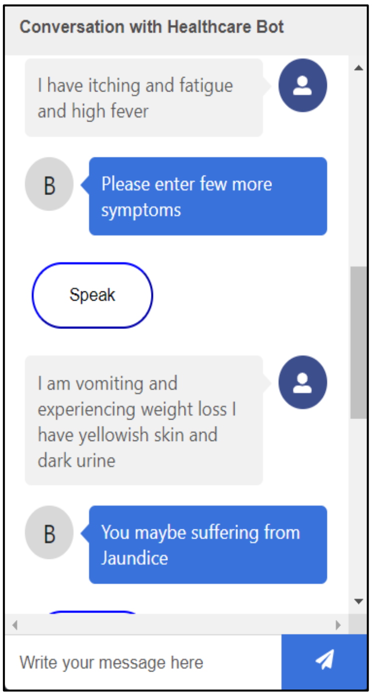


Fig. 2. Graph depicting Test Accuracy of different Classification Algorithms response in Hindi language.

Fig. 4 depicts a similar conversation between the user and the chatbot where the chatbot performs disease diagnosis. However, the conversation is shown in the Hindi language demonstrating the multilingual capability of the chatbot system. The user enters symptoms in Hindi and the chatbot analyses the user text and accurately predicts Gastroesophageal reflux disease (GERD) and returns the response in Hindi language.



Conversation with Healthcare Bot

I have itching and fatigue and high fever

B Please enter few more symptoms

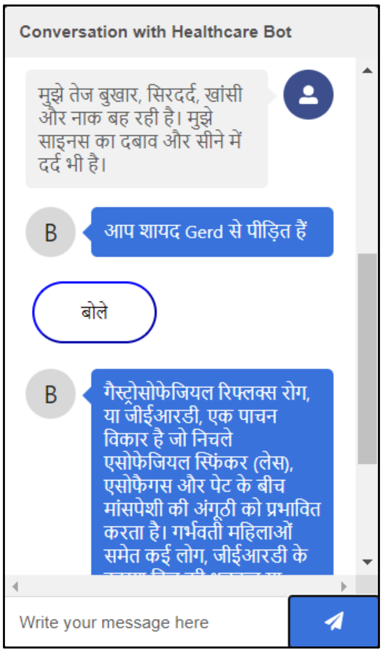
Speak

I am vomiting and experiencing weight loss I have yellowish skin and dark urine

B You maybe suffering from Jaundice

Write your message here

Fig. 3. Conversational Disease Diagnosis in English language



Conversation with Healthcare Bot

मुझे तेज बुखार, सिरदर्द, खांसी और नाक बह रही है। मुझे साइनस का दबाव और सीने में दर्द भी है।

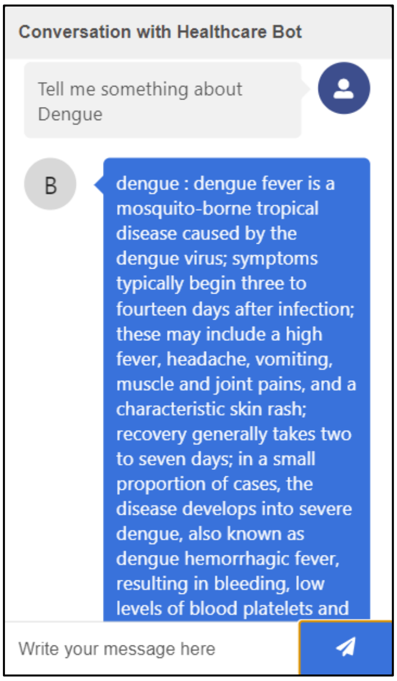
B आप शायद Gerd से पीड़ित हैं

बोले

B गैस्ट्रोसोफेजियल रिफ्लक्स रोग, या जीईआरडी, एक पाचन विकार है जो निचले एसोफेजियल स्फिंकर (लेस), एसोफैगस और पेट के बीच मांसपेशी की अंगूठी को प्रभावित करता है। गर्भवती महिलाओं समेत कई लोग, जीईआरडी के

Write your message here

Fig. 4. Conversational Disease Diagnosis in Hindi language



Conversation with Healthcare Bot

Tell me something about Dengue

B dengue : dengue fever is a mosquito-borne tropical disease caused by the dengue virus; symptoms typically begin three to fourteen days after infection; these may include a high fever, headache, vomiting, muscle and joint pains, and a characteristic skin rash; recovery generally takes two to seven days; in a small proportion of cases, the disease develops into severe dengue, also known as dengue hemorrhagic fever, resulting in bleeding, low levels of blood platelets and

Write your message here

Fig. 5. User Query Response using TF-IDF and Cosine Similarity

Fig. 5 illustrates a user query instead of a conversation where the user inputs symptoms. Thus, the chatbot system does not perform disease diagnosis. Rather, the system applies the TF-IDF and cosine similarity techniques to determine sentence similarity and returns the most appropriate response to the user query from the knowledge database that has been provided to it.

V. CONCLUSION

This paper demonstrates the implementation of a multilingual healthcare chatbot system. The chatbot system supports multilingual text and speech to be particularly useful for the rural population of India which uses regional languages. The system can address the health-related queries of the users in addition to its primary function of disease diagnosis based on user symptoms. The system also provides the user with disease description and disease precautions along with the disease diagnosis.

The paper has provided a comparative analysis between five Machine Learning Classification algorithms among which the Random Forest Classifier exhibits the highest accuracy of 98.43%. The proposed system implements TF-IDF and Cosine Similarity to find the most appropriate response to the user query.

One of the limitations of this multilingual chatbot system is the lack of data. The system can be trained with a larger and more comprehensive dataset to obtain better results. Future work can involve expanding the system to incorporate a larger number of languages. Deep Learning algorithms can be used to make the disease classification more accurate and provide better results. Natural Language Generation can be implemented by training a model on various conversational datasets to produce superior chatbot responses.

References

[1] https://censusindia.gov.in/2011-prov-results/paper2/data_files/india/Rural_urban_2011.pdf

[2] M. M. Rahman, R. Amin, M. N. Khan Liton and N. Hossain, "Disha: An Implementation of Machine Learning Based Bangla Healthcare Chatbot," 2019 22nd International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 2019, pp. 1-6, doi: 10.1109/ICCIT48885.2019.9038579.

[3] R. B. Mathew, S. Varghese, S. E. Joy and S. S. Alex, "Chatbot for Disease Prediction and Treatment Recommendation using Machine Learning," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2019, pp. 851-856, doi: 10.1109/ICOEI.2019.8862707.

[4] P. Srivastava and N. Singh, "Automatized Medical Chatbot (Medibot)," 2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC), Mathura, India, 2020, pp. 351-354, doi: 10.1109/PARC49193.2020.236624.

[5] Mrs. Rashmi Dharwadkar, Dr.Mrs. Neeta A. Deshpande "A Medical ChatBot". International Journal of Computer Trends and Technology (IJCTT) V60(1):41-45 June 2018. ISSN:2231-2803.

[6] L. Hidayatin and F. Rahutomo, "Query Expansion Evaluation for Chatbot Application," 2018 International Conference on Applied Information Technology and Innovation (ICAITI), Padang, Indonesia, 2018, pp. 92-95, doi: 10.1109/ICAITI.2018.8686762.

[7] U. Bharti, D. Bajaj, H. Batra, S. Lalit, S. Lalit and A. Gangwani, "Medbot: Conversational Artificial Intelligence Powered Chatbot for Delivering Tele-Health after COVID-19," 2020 5th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2020, pp. 870-875, doi: 10.1109/ICCES48766.2020.9137944.

[8] L. Athota, V. K. Shukla, N. Pandey and A. Rana, "Chatbot for Healthcare System Using Artificial Intelligence," 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2020, pp. 619-622, doi: 10.1109/ICRITO48877.2020.9197833.

[9] https://www.kaggle.com/itachi9604/disease-symptom-description-dataset

[10] S. Fernandes, R. Gawas, P. Alvares, M. Femandes, D. Kale and S. Aswale, "Survey on Various Conversational Systems," 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), Vellore, India, 2020, pp. 1-8, doi: 10.1109/ic-ETITE47903.2020.126.

[11] M. Bates, "Health Care Chatbots Are Here to Help," in IEEE Pulse, vol. 10, no. 3, pp. 12-14, May-June 2019, doi: 10.1109/MPULS.2019.2911816.

[12] Qaiser Shahzad, Ali Ramsha, "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents," International Journal of Computer Applications, 2018.

[13] A. W. Qurashi, V. Holmes and A. P. Johnson, "Document Processing: Methods for Semantic Text Similarity Analysis," 2020 International Conference on INnovations in Intelligent SysTems and Applications (INISTA), Novi Sad, Serbia, 2020, pp. 1-6, doi: 10.1109/INISTA49547.2020.9194665.

[14] Tin Kam Ho, "Random decision forests," Proceedings of 3rd International Conference on Document Analysis and Recognition, Montreal, QC, Canada, 1995, pp. 278-282 vol.1, doi: 10.1109/ICDAR.1995.598994.

[15] L. Jiang, Z. Cai, D. Wang and S. Jiang, "Survey of Improving K-Nearest-Neighbor for Classification," Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007), Haikou, China, 2007, pp. 679-683, doi: 10.1109/FSKD.2007.552.

[16] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," in IEEE Transactions on Systems, Man, and Cybernetics, vol. 21, no. 3, pp. 660-674, May-June 1991, doi: 10.1109/21.97458.

[17] Kaviani, Pouria and Dhotre, Sunita, "Short Survey on Naive Bayes Algorithm," International Journal of Advance Research in Computer Science and Management 04, 2017.