IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 22, Issue 1, Ser. III (Jan - Feb 2020), PP 50-56
www.iosrjournals.org

Chatbot for Medical Treatment using NLTK Lib

Dinesh Kalla[1], Fnu Samaah[2]
[1]*Department of Computer Science/ Colorado Technical University, United States of America*
[2]*Department of Computer Science/ Northeastern Illinois University, United States of America*

Abstract:*The project involves creating a chatbot to be used in healthcare treatment. The app uses Artificial Intelligence and can help in diagnosing various diseases and providing necessary details about the patient's disease or illness. It helps in reducing the cost of healthcare and improving access to healthcare services. It also enhances the user to chat and know about their medical status and issues.*

Keywords*:Chatbot; Cosine similarity; Artificial Intelligence, Machine Learning; Stemming;NLTK Lib; Medical Chatbot; Virtual Assistant; Natural Language Processing; Medbot.*

Date of Submission: 11-02-2020 Date of Acceptance: 26-02-2020

I. Introduction

The system is trained on Medical information gathered from different online sources and databases. It's an application that takes queries from the patient related to diseases and gives them answers based on their questions and symptoms. The reason I selected this topic is to make the search process user friendly for patients who are suffering from medical illness or who want to know the reason behind their symptoms and causes or also to get immediate answers for user questions posed as input query. It uses techniques from natural language processing, which deals with human language to process the data and to provide a response. The goal here is to understand and prepare user queries based on some defined rules, which are explained further to perform the task and to show results based on generated output.

The system that is being developed is meant to interact with users using human's natural language. The internet provides a lot of information, thus allowing the chatbot to provide adequate and accurate information depending on the requirements of the user. The chatbot works in various ways, such as virtual assistants, customer support, and online training[i]. As a result, the user will be in position to get a realistic experience to chat with medical professionals besides asking questions. It helps in retrieving critical words from the user, thus knowing the medical problems may be having based on their input[3]. In short, our medbot helps in providing a QA forum for the users. The app answers similar questions that medical professionals may have answered. Currently, the chatbot created has only been designed to conduct conversations using textual methods[6]. All the same, it can be improved in the future to take care of oral conversations. This project explains how the app has been created and how the users will use it.

II. Material and Methods

In this study, a mixed-method approach was used that helped in generating a multi-layered issue. We did online surveys and conducted face to face interviews to know the motivations behind the medibot usage in the healthcare industry. A total of 50 subjects, both male and female, were interviewed. More so, we explored various other similar studies and found ten reliable articles from 2015 to 2019.

We used thematic analysis on the qualitative data to identify the common trends and patterns. Two of our partners were involved in familiarizing themselves with the data. As such, they had to read the transcript multiple times to enhance understanding. Using NVIVO software, the analyses were done independently. Through the software, the data got recoded and coded before being categorized into themes and subthemes that are understandable. It is through these results that we agreed to have the final set of findings. The validation of the themes was by the two partners who compared the quotes with the identifiable themes. Quantitative data helped in conducting inferential and descriptive statistics. In these ones, there was the exclusion of neutral values and dichotomizing all the variables. Therefore, this enhanced logistic regressions, thus determining the correlations of the chatbot acceptability. Since the model could not meet the statistical assumption, it could not accordingly be adjusted.

My query answering system's intention is to take ainput from the patient or user, process it using certain methods, and provide results to the user based on their symptoms. In beginning phase we process the input text (query input from the user) by converting the entire sentence into lowercase so that the algorithm does not treat the same kind of words which appear in the upper case different from the same word, which appears in

the lower case. This may result to a decline in the accuracy of the result search process. The query is been processed by dividing and breaking up strings into tokens, which are small words or units that can be used further. It helps to break a compound sentence into words and understand the significance of the terms concerning the sentence to result a structural description of an input query.

Table no 1: Questions and Breaking Words

Question	**After breaking it into words**
['What is breast cancer']	['What', 'is', 'breast', 'cancer']
['Why do I feel pain in my Knee']	['Why', 'do', 'I', 'feel', 'pain', 'in', 'my', 'hands']
['How cancer comes']	['How', 'cancer', 'comes']

Now we categorize each word based on its parts of speech and label them accordingly, which are usually used to describe whether the word is a noun, pronoun, verb, adverb, or is it a symbol, etc.

Table no 2: Questions and POS Tagging

Question	**POS tagging**
['What is breast cancer']	[('What', 'WP'), ('is', 'VBZ'), ('breast', 'JJ'), ('cancer', 'NN')]
['Why do I feel pain in my Knee']	[('Why', 'WRB'), ('do', 'VBP'), ('I', 'PRP'), ('feel', 'VB'), ('pain', 'NN'), ('in', 'IN'), ('my', 'PRP$'), ('hands', 'NNS')]
['How cancer comes']	[('How', 'WRB'), ('cancer', 'NN'), ('comes', 'VBZ')]

Some parts of the speech tag list:

- NN noun, singular 'desk.'
- JJ adjective
- VB verb, base form take
- VBD verb, past tense, took
- VBG verb, gerund/present participle taking
- VBN verb, past participle taken
- VBP verb, sing. present, non-3d take
- VBZ verb, 3rd person sing. present takes
- WDT wh-determiner which
- WP wh-pronoun who, what
- WRB wh-adverb where, when
- PRP personal pronoun I, he, she
- PRP$ possessive pronoun my, his, hers
- NNS noun plural 'desks'
- IN preposition/subordinating conjunction

Now the question is directed to different methods based on their tagging. And a partial answer is then generated based on order of the tagging. Different types of questions trigger different methods based on their tagging (what, why, how).

III. Tree for Different Types of Queries

Below are the structures of queries or types of queries starting from what, how, why, etc. and based on their sentence structure and parts of speech, the partial answer is generated.

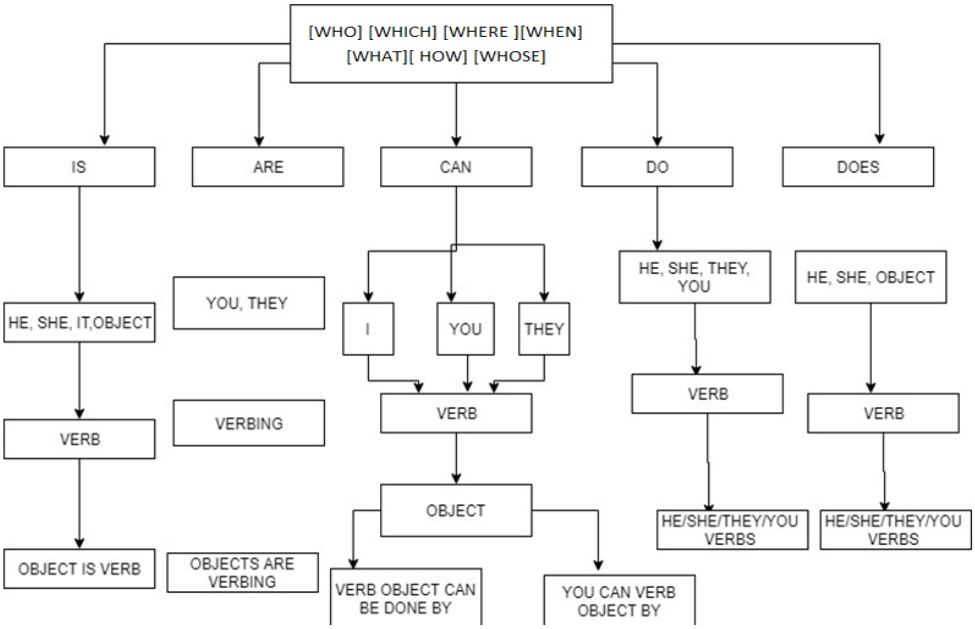

Fig. no 1: Different Types of Query Tree

Based of the query tree it will take questions from user to generate partial answers to make meaningful phrases.

Table no 3: Questions and Partial Answers

Questions	Partial Answer
What is breast cancer	Breast cancer is
How do I manage pain in my bone	You manage pain in the bone
Why do I feel pain in my hands	You feel pain in hands because

Then we remove all stop words from our partial answer. It's a process of choosing the required words form meaningful phrases. A stop word is a commonly used word (such as "the", "a", "an", "in"). They are useful in the formation of sentences and without which the sentences won't even make sense, but these do not provide any help in Finding our exact answer[5]. Once we remove these stop words from our partial answer, we only have important words to search for our answers or output.

Table no 4: Questions and Partial Answers without Stop Words

Questions	Partial Answer
breast cancer is	breast cancer
you manage pain in bone	manage pain bone
you feel pain in hands because	feel pain hands

3.1 Case 1

A text extraction is a process of retrieving relevant content from a text file or a text paragraph. While extracting a sentence from a given text file, we break the paragraph into multiple sentences. We search eachand everybroken sentence from that file, which covers all the words in the partial answer. Once all the words match, we append that sentence into a new string array created. We follow the same process for all the broken strings. These arrays of strings are then compared with a partial answer and filtered later using cosine similarity to find the exact final answer. Cosine similarity is a metric used to ascertain how resembling the documents are irrespective of their data size. It calculates the cosine of the angle between two vectors projected in a multi-dimensional space. The two vectors are integers arrays containing the count of wordsin provided two documents. The lesser the angle, the larger the cosine similarity. The cosine similarity is useful because even if the two similar word documents are different and far apart because of the size and length, they could still have a minute smaller angle between them. Smaller the angle, the higher the similarity. We compare each sentence retrieved using string extraction to our partial answer and compare the cosine similarity between them. The two strings or sentences with higher cosine similarity will be our answer, which is displayed.

Example:

d1: “The pain continued over the knee.”

d2: “My painstarted inthebroken knee."

d3: "The broken knee started to pain.”

Table no 5: Count of Words Occurred in Each Sentence

Doc	The	pain	continued	Over	to	knee	my	started	in	broken
D1	2	1	1	1	0	0	0	0	0	0
D2	1	1	0	0	0	1	1	1	1	1
D3	1	1	0	0	1	1	0	1	0	1

and now we use the dot product to measure similarity.

+

$$cos\,\theta = \frac{\vec{a}\cdot\vec{b}}{\|\vec{a}\|\|\vec{b}\|} = \frac{\sum_1^n a_i b_i}{\sqrt{\Sigma_i^n a_i^2}\sqrt{\Sigma_1^n b_i^2}} \qquad (1)$$

Where, $\vec{a}\cdot\vec{b} = \sum_1^n a_i b_i = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$ is the dot product of the two vectors.

3.2 Case2:

In case where all the words in the partial answer doesn't exist in the text file or text paragraph, We then do query reformulation where, Instead of finding the exact words in the sentence(text file) that match words from our partial answer, we also try to find stemming word of partial answer's words. Stemming is the method of normalizing the words of a sentence into its base root form. For example, the words "likes," "liked," "likely," "liking," all these words originate from a single root word that is like. Stemming programs are mainly referred to as stemming algorithm works by removing off the last or beginning of the word considering a list of common prefixes and suffixes. Apart from stemming, we also try to find synonyms of all partial answer's words that can be in our text file. Wordnet is an NLTK collection reader, a word database for English. It can be used to find the meaning of words, synonym (words having the same name). One can define it as a dictionary of English. **Synset** is called a synonym set or collection of synonym words. It is very much useful in artificial intelligence while using text analysis. We convert the partial answer by adding all the noun synonyms of the words also use stemming for each word as per requirement while searching for an answer string [1].

Table no6: Partial answer without top Words and including Stemming and Synonyms

Partial answer without top Words	Partial answer including Stemming and Synonyms of above words
Will cancer come back	{'vertebral column', 'stake', 'plump for', 'back', 'indorse', 'number', 'backwards', 'Cancer_the_Crab', 'Will', 'cancer', 'come in', 'amount', 'ejaculate', 'semen', 'volition', 'descend', 'bookbinding', 'genus Cancer', 'cover', 'backbone', 'Cancer', 'cum', 'come', 'Crab', 'rearwards', 'malignant_neoplastic_disease', 'binding', 'game', 'backup', 'backward', 'hind', 'seed', 'spine', 'support', 'hinder', 'rearward', 'hail', 'punt', 'rear', 'come up', 'bet on'}
the risk factors be controlled are	{'factor', 'ingredient', 'endangerment', 'controlled', 'master', 'operate', 'jeopardy', 'danger', 'chance', 'contain', 'gene', 'take_a_chance', 'factor_out', 'factors', 'insure', 'constituent', 'lay_on_the_line', 'peril', 'command', 'divisor', 'factor_in', 'hold in', 'hazard', 'risk', 'check', 'broker', 'cistron', 'element', 'take_chances', 'risk_of_exposure', 'agent', 'control', 'curb', 'risk_of_infection'}

IV. Result

In the medical chatbot that we have created, the dialogue uses the linear design, where it shows extraction symptoms toward the mapping symptom. In this case, it helps to identify the question that the user is asking or the symptoms he or she is having. If the user is having significant symptoms, he or she may be referred to a doctor [8].

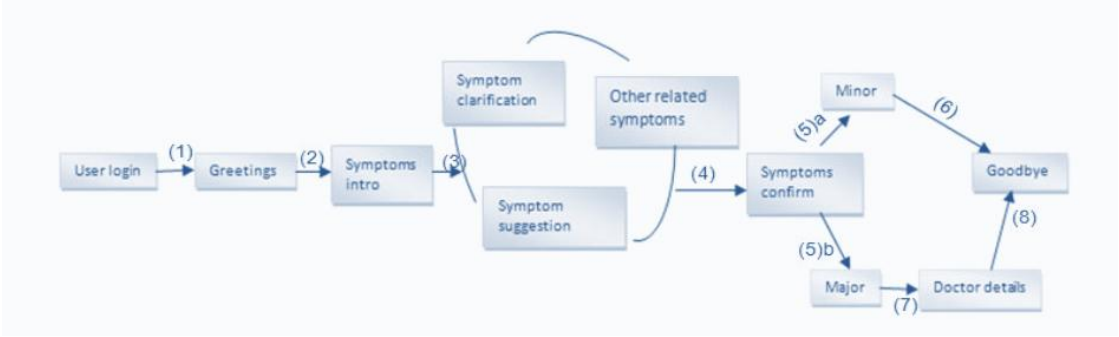

Fig. no2: Chatbot Dialogue Design

The app will also provide users logins where their details will be stored in the chatbot database. As a result, these details are preserved for future reference. The chatbot will then enhance clarification of the symptoms of the users with the series of questions. As such, it will be able to determine the kind of disease that the patient is suffering from [9][8]. It starts by validating the user's login details. After that, it extracts the symptoms by use off String Searching Algorithm[1]. For instance, it will be more comfortable when the user states the symptoms of breast cancer.

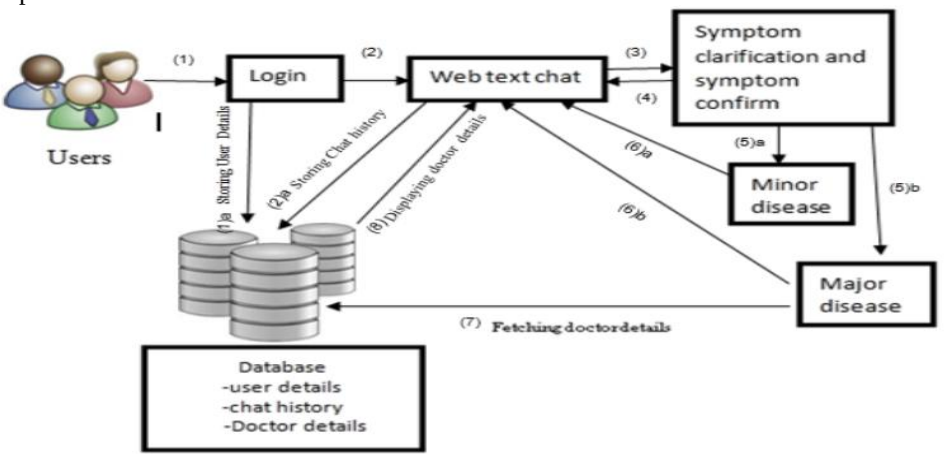

Fig. no3: Functional Architecture

Table no 6: Example Question and Answers

Question	Answer
What is breast cancer?	breast cancer is an uncontrolled growth of breast cells. To better understand breast cancer, it helps to understand how any cancer can develop.
What are benign tumors?	a tumor can be benign (not dangerous to health) or malignant (has the potential to be dangerous). Benign tumors are not considered cancerous: their cells are close to normal in appearance, they grow slowly, and they do not invade nearby tissues or spread to other parts of the body. Malignant tumors are cancerous. Left unchecked, malignant cells eventually can spread beyond the original tumor to other parts of the body.
myth: early-stage breast cancer rarely recurs.	Fact: even with early-stage breast cancer, there is always some risk the cancer will return. Many people believe that early-stage breast cancer — cancer that hasn't moved beyond the breast and underarm lymph nodes — has almost no chance of recurring or coming back. Although most people with early-stage breast cancer won't have a recurrence, the risk never goes away completely

V. Discussion

There are some cases where the output hasvery low cosine similarity, and the answer may or may not be an exact match. For all those cases, I have decided to put a threshold of 0.20. Strings extracted with a cosine similarity of 0.20 or above are considered to be exact answers, and others are still displayed as a possible answer, which may not be the exact one.

Along the way, I explored technologies like NLTK, Django, Stemming, cosine similarity, synonyms, etc.The project will work in a way that the user will have to text, thus communicating with chatbot and getting specific disease symptoms. In addition, the user will get the previous chat history as their details will be stored in the chatbot database [10].

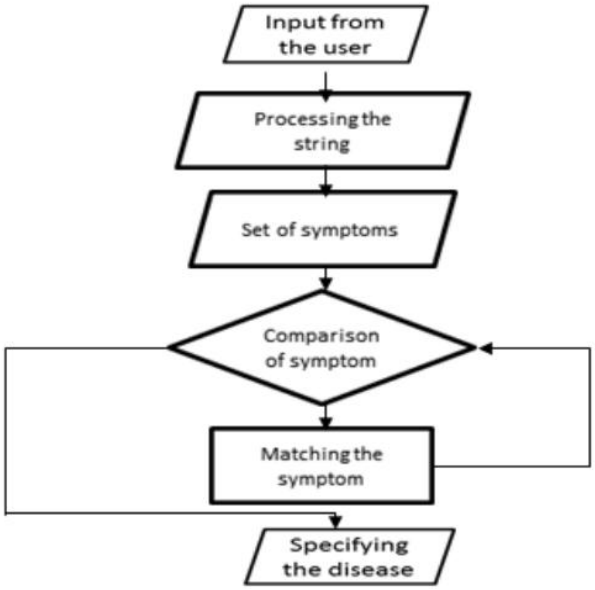

Fig. no 4: Flow Chart for Specifying the Disease

The chatbot will provide an accurate result with symptom clarification [7]. Besides, the user can view the previous chat, thus knowing what he or she was inquiring earlier.

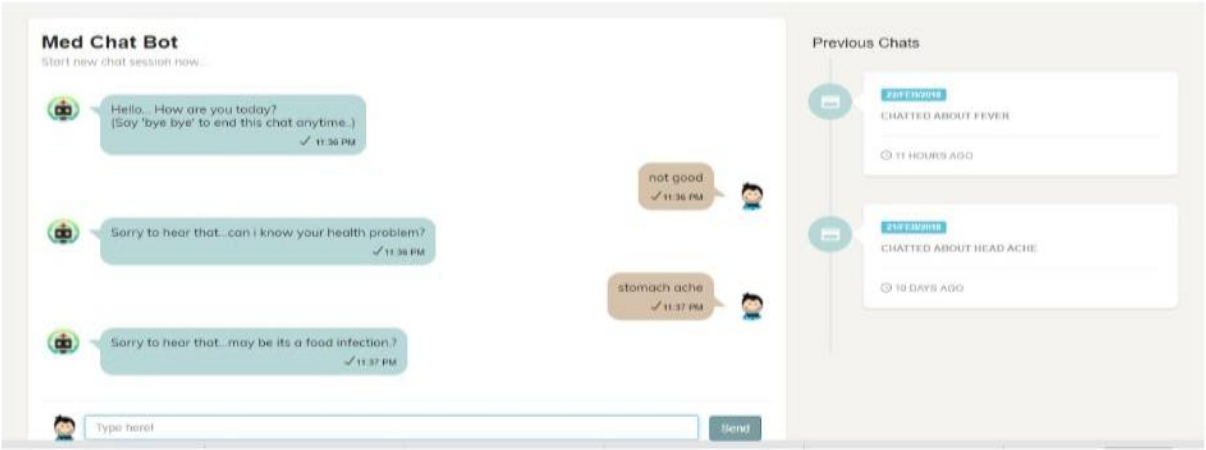

Fig. no 5: Chat Bot Window (Sample Chat)

VI. Conclusion and Limitations

It is evident that chatbot is user-friendly as it can get used by anyone who knows how to write and read. It helps in giving personalized diagnosis based on the questions and texts of the user. The app heavily relies on AI algorithms, together with training data.

A question answering system trained on disease information made in such a way that it has no conversation with humans and more of finding the textual answer to a given question. Sometimes the solutions may not be 100 percent exact as its basic application built from scratch. I would like to take it to the next level by adding greetings and conversations with the user to receive more information about what they are looking for and making the answer search more advanced, so the chances of giving unrelated answers are less.

References

[1]. Amato, Flora, Stefano Marrone, Vincenzo Moscato, Gabriele Piantadosi, Antonio Picariello, and Carlo Sansone. "Chatbots Meet eHealth: Automatizing Healthcare." In WAIAH@ AI* IA, pp. 40-49. 2017.

[2]. Belfin, R. V., A. J. Shobana, MeghaManilal, Ashly Ann Mathew, and Blessy Babu. "A Graph Based Chatbot for Cancer Patients." In2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), pp. 717-721. IEEE, 2019.

[3]. Bibault, Jean-Emmanuel, Benjamin Chaix, Arthur Guillemassé, Sophie Cousin, Alexandre Escande, Morgane Perrin, Arthur Pienkowski,Guillaume Delamon, Pierre Nectoux, and Benoît Brouard. "A Chatbot Versus Physicians to Provide Information for Patients With Breast

Cancer: Blind, Randomized Controlled Noninferiority Trial." Journal of medical Internet research 21, no. 11 (2019): e15787.
[4]. Chung, Kyungyong, and Roy C. Park. "Chatbot-based heathcare service with a knowledge base for cloud computing." Cluster Computing22, no. 1 (2019): 1925-1937.
[5]. Divya, S., V. Indumathi, S. Ishwarya, M. Priyasankari, and S. Kalpana Devi. "A self-diagnosis medical chatbot using artificialintelligence." Journal of Web Development and Web Designing 3, no. 1 (2018): 1-7.
[6]. Hoermann, Simon, Kathryn L. McCabe, David N. Milne, and Rafael A. Calvo. "Application of synchronous text-based dialogue systems in mental health interventions: systematic review." Journal of medical Internet research 19, no. 8 (2017): e267.
[7]. Kavitha, B. R., and Chethana R. Murthy. "Chatbot for healthcare system using Artificial Intelligence." (2019).
[8]. Mishra, Saurav Kumar, Dhirendra Bharti, and Nidhi Mishra. "Dr. Vdoc: A Medical Chatbot that Acts as a Virtual Doctor." Research &Reviews: Journal of Medical Science and Technology 6, no. 3 (2018): 16-20…+
[9]. Nadarzynski, Tom, Oliver Miles, Aimee Cowie, and Damien Ridge. "Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: A mixed-methods study." Digital health 5 (2019): 2055207619871808.
[10]. Rarhi, K., Bhattacharya, A., Mishra, A., & Mandal, K. (2017). Automated Medical Chatbot.

Dinesh Kalla, etal. "Chatbot for Medical Treatment using NLTK Lib." *IOSR Journal of Computer Engineering (IOSR-JCE)*, 22.1 (2020), pp. 50-56.