International Conference on Recent Trends in Data Science and its Applications
DOI: rp-9788770040723.102

HealthCare Chatbot Using NLP Techniques and Deep Learning Algorithm

NarayanaDarapaneni
Director - AIML
Great Learning
Bangalore, India
darapaneni@gmail.com

Deepthi V
Student – PGPAIML
Great Learning
Bangalore, India
deepthiv286@gmail.com

DivyaChowdaiah
Student – PGPAIML
Great Learning
Bangalore, India
Itsdivya17@gmail.com

Anwesh Reddy Paduri
Senior Data Scientist
Great Learning
Hyderabad, India
anwesh@greatlearning.in

Shashank P Raikar
Student – PGPAIML
Great Learning
Bangalore, India
Shashankraikar276@gmail.com

Ramyasri R
Student – PGPAIML
Great Learning
Bangalore, India
ramyasrirajasekhar@gmail.com

Sunil Bhagwat
Student – PGPAIML
Great Learning
Bangalore, India
sunilnilkod@gmail.com

Mahesh Deshpande
Student – PGPAIML
Great Learning
Bangalore, India
Maheshdeshpande2007@gmail.com

***Abstract*— The healthcare sector is one of the largest focus areas in the world today. Health related problems are becoming increasingly common. Early diagnosis and treatment of diseases can play a vital role. The solution is adopting healthcare chatbots. The proposed solution is a healthcare chatbot application that can provide solutions based on user queries. The NLP component of the chatbot allows it to understand the intent of the user's query and extract relevant information from the user's input, making it more effective at providing accurate and helpful responses. The chatbot is designed to be easy to use for patients, making it a useful tool for providing basic healthcare information and assistance. Deep Learning algorithm was used and gives an accuracy of 91%.**

Keywords—Chatbot,Health support, NLP, Deep Learning, LSTM, Seq2Seq, Document - Term Matrix, Bag of words

I. INTRODUCTION

Welcome to our healthcare chatbot, designed to assist you in finding the information and resources you need to navigate the complex healthcare system. We understand that the healthcare landscape can be confusing and overwhelming, with various challenges such as long wait times, lack of information, and difficulty in finding the right provider. This can make it difficult for people to access the care they need and make informed decisions about their health. This chatbot is here to help alleviate some of these issues. With this chatbot, you can quickly access a wide range of information about medical conditions, treatments. We also understand that cost is a major concern for many people when it comes to healthcare. That's why this chatbot is designed to help users to find the most cost-effective options available. Our chatbot is a valuable tool that can help you take control of your health and make the healthcare experience more efficient and accessible for you.Artificial Intelligence is creating an era that it is going beyond human imagination and thinking. The technological experts with their tremendous skills and expertise developed the science of AI. The study mainly focusses on bringing up Chatbots which are developed using Natural Language processing. These are the virtual assistants powered by built-in features of AI.

Michael Mauldin in 1994, formulated the term 'Chatbot' which is a combination of words 'Chat' and 'Robot'.

We are fairly familiar with Amazon's Alexa, Apple's Siri and Microsoft's Tay – these are the conversational agents that would respond and act according to the queries that users would probe.

Our capstone project focuses on the Healthcare Chatbot which is meant for early diagnosis of diseases and treatments thereof. The reason why we chose Healthcare chatbot in particular is because we all seek medical assistance at some point of time because we live a sedentary lifestyle due to which we get exposed to various diseases. Healthcare industries play a significant role when it comes to life expectancy and quality of lifestyle.

Many are the roadblocks and challenges that the healthcare industries in India are facing these days especially when it comes to the people living in rural areas who lack medical facilities. The census show high mortality rates due to deadly disease that is taking away lives of people and due to unavailability of medical professionals in the rural areas where physical consultation has become a major challenge and totally expensive. When our health goes upset we all want immediate treatment, early diagnosis to avoid any serious consequences. None of us want to wait longer to take appointments with the physician or to meet the concerned doctor just to know what went wrong and what medication we would need to take.

So, the best solution to overcome this problem is to integrate Chatbots in the Hospital's IT systems. By doing so this would serve thousands of customers or patients at a single point of time and at a very low cost. Also that the patients can get access 24/7 to their health information from anywhere and anytime that they need. The healthcare chatbot that we have come up with uses text, or instant messaging for natural interaction with the patients and all it's users. What we've developed is the QA Chatbot with the the built in AI features, NLP text processing techniques and Deep

DOI: rp-9788770040723.102
Learning Algorithm. When I say QA chatbot this would actually answer the given question that the user would type in. Meaning this would mimic human intelligence. It gives an impression as though the user is talking to the medical professional on the other end as they have meaningful conversation with the patients or users in real-time.

The healthcare chatbots can be used for making appointments, pulling out medical history of a patient, setting up reminders for consuming medicines and other proactive alerts that would inform the medical staff to let them know what's in their action for treating patients.

The usage of chatbots in every business sector is becoming increasingly popular these days. Turning the clock back to 1966 where the first chatbot named Eliza was invented. The sole purpose of Eliza was to provide responses to users based on different keywords. This was purely the role based chatbot. Then came Parry, Jabberwacky, Alice, Smarterchild, Cortona, Siri, Alexa, GoogleNow and then the latest Microsoft Tay – which is a twitter chatbot. The goal was to analyze the tweets on twitter in order to understand the conversations.

Here with the Healthcare chatbot system, the dataset or the model is preloaded is with the set of keywords, set of data. Meaning we have trained the chatbot with word cloud where it picks up keywords from questionnaires and then responds with the answers from the word cloud of answers.

For text pre-processing, NLP has been used with NLTK toolkit that's available in python. The algorithm that we've used is the Deep Learning technique with 4 layers. Dense layers along with dropouts of 20% is used and last layer, output layer contains no. of neurons = no. of intents is used to predict output intent with softmax.

Relu activation function has been used in the dense layers. Stochastic gradient descent is used as an optimizer.

Finally, GUI is created using tkinter to provide the chatbot experience where user can enter their query and the system's response is shown in the GUI.

II. Literature Survey

The proposed solution [1], focuses on predicting agent response for user's query given in the electronics e-commerce industry. They have used the dataset from the support tickets created at Robocraze platform. Initially, they have performed basic NLP operation using NLTK packages like removing stop words, identifying the most common words, bi-grams and basic EDA and word cloud analysis. During, the cleaning of the data they have divided the data in to query – response pairs. Here, they have used an encoder-decoder structure wherein the encoder is customer queries and decoder are the agent response. Further, these inputs are converted into a simple Seq2Seq LSTM model with a bi directional layer and dot product of this is sent to dense layer to generate the final model predictions.

The proposed solution[6], focuses on creating a chatbot on general knowledge-based question and answer. Here, they have compared the model performance through both transformer model and Seq2Seq learning and later they found that the one with transformer model was giving a better result. The model Architecture has i) Encoder ii) Decoder adds another layer apart from 2 sub layers which perform multi head attention. iii) Transformer which further helps to find the a) Scaled-dot product b) Multi Head attention c) Position-wise FFN with ReLU activation function and d) Position encoding. In the training, data was split at 80:20 and batch size as 28 or epoch as 120. The final result of Seq2Seq model was giving 23.5 BLEU of dataset whereas the Transformer model was giving 85 BLEU. Higher the BLEU better is the model for automatic chatbot.

In this model[7] they have used NLP techniques, first tokenizing the corpus, removing stop words, correcting any spelling mistakes, lemmatizing the corpus, vectoring the input and then applying Cosine Similarity between input and training set to find the best suitable output for the question. They tried experimenting both with cosine similarity and Euclidean distance as well. In this example, 2 examples of Dhoni's Wikipedia page, one small para and another big one is taken into account and another one is Sachin's Wikipedia page. The distance between Dhoni's similar type document is much lesser than the distance between them and Sachin's. Thus, questions can be identified nearest to the training set even though user's language for asking question is different. To improve the model even further they can train the vocabulary on Google's word embedding, which will give words similar in meaning almost identical vector. Thus when computing Cosine distances, even though the word is different from what was originally used in our question set, the difference between those words will be minimum. Thus it can give more accurate replies in Chatbot UI.

III. Dataset

A. *Overview*

This paper provides appropriate answer to the patient's health related query. The Healthcare JSON dataset is a collection of data that includes question and answer pairs related to healthcare information. The dataset is structured in JSON format with 465 records, which makes it easy to use and integrate with other systems. Each question-answer pair in the dataset contains a question that is related to healthcare and a corresponding answer that provides information about the topic of the question. The questions in the dataset cover a wide range of healthcare topics, including diseases, treatments, symptoms, and general health information. The answers provided in the dataset are accurate, up-to-date, and written in a clear and concise manner. The dataset can be used to train a healthcare chatbot or other applications that require healthcare knowledge. Additionally, the dataset can be used as a resource for healthcare professionals and researchers, providing a valuable source of information for their work.

B. *Characteristics of the Data*

The dataset features information on 465 QA pairs. We removed stop words from the dataset along with a few custom stop words. Here, We looked at the Bigram and Trigram words from the answers in the dataset.

DOI: rp-9788770040723.102

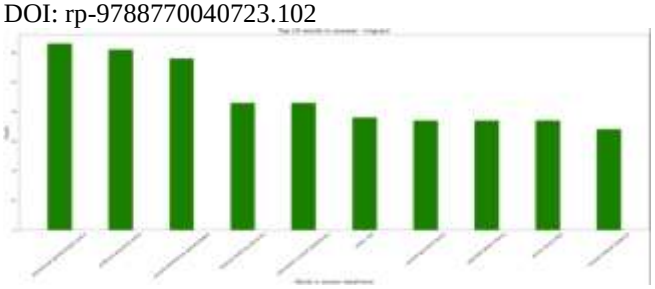


Fig. 1. Top 10 words in answer - trigram

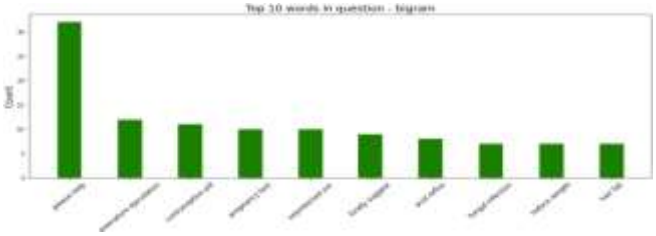


Fig. 2. Top 10 words in question - bigram

The word cloud representation of texts from QandA responses are seen as below. Fig 3 and 4

Fig. 3.Word Cloud of Questions

Fig. 4.Word cloud of Answer

IV. Materials And Methods

The proposed system consists of a user-friendly chat interface through which a user can communicate with the system. Depending on the user input, the chatbot will provide relevant information about his queries.

The system has pre-processed the question answer(icliniqQA) dataset, for training of the Deep Learning model. The system has trained and compared different number of layers and number of neurons in Deep Learning model and got the best accuracy of around 91%.

First, a dataset of healthcare-related questions and answers were collected and pre-processed. Dataset is converted from json file to pandas data frame for preprocessing. We dropped "question text" and "url" columns from the data frame which were not required for chatbot. The data frame is remained with question, answer and tag. The question and tag columns goes through text processing techniques of NLP.

They are tokenized into smaller units or words. Further the tokenized words are converted to lower case. After that we removed punctuations and stop words in question and tag columns. Also lemmatization is done to bring words to its meaningful base form. We converted variable-length texts into a fixed-length vector using the Bag-of-Words technique.

The cleaned dataset is loaded as intents. Each question is tokenized to words and then it is mapped to it's respective tags and stored as documents. Unique tags are stored in one variable. Then a document-term matrix is created against its tags and used as training data. This dataset was then used to train a deep learning model. This process allows the model to understand the meaning of the user's input and provide a relevant response.

After the model was trained, it was tested on a set of unseen data to evaluate its performance. Fine-tuning was performed to improve the model's accuracy by adjusting the hyperparameters of the model.

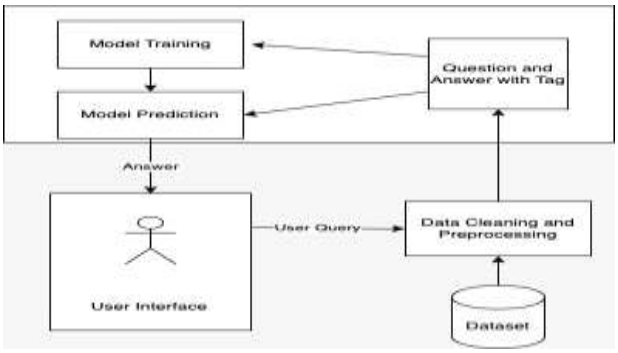


Fig. 5. Flow Model - Chatbot Architecture

Lastly, the model (Fig 3) was integrated into the chatbot interface, allowing users to interact with it through text-based inputs and receive text-based responses. The chatbot also includes a Knowledge base where it can retrieve information to answer user's queries.

A. Data Preprocessing

The dataset is retrieved from git hub [11] which contains data of 465 records, including the mapping of answers with the corresponding questions and tags.

The dataset was checked for inconsistencies and removed unwanted data. Then the missing values were imputed. The answer column contained "--><link>" at end of each record, so replaced it with empty string.

B. NLP text pre-processing

The user input has to be pre-processed in order to provide appropriate response. Pre-processing methods – punc-

DOI: rp-9788770040723.102

tuation, tokenization, stop words and lemmatization have been used to generate an appropriate response for the user query.

1) *Punctuation removal:* Python's string module contains the following list of punctuation '!"#$%&\'()*+,-./:;<=>?@[\\]^_`{|}~'. Using this list , punctuation is removed from the user's input.

2) *Tokenization:* Here the user input is converted into a bag of words using tokenization.

3) *Stop words removal:* Removal of stop words is necessary as they take valuable pre-processing time and space. eg:- 'a', 'an', 'the', etc.

4) *Lemmatization:* lemmatization considers the context and converts the word to its meaningful base form. For example, lemmatization would correctly identify the base form of 'caring' to 'care'.

Wordnetlemmatizier is used to achieve lemmatization in our case.

C. Deep Learning Algorithm

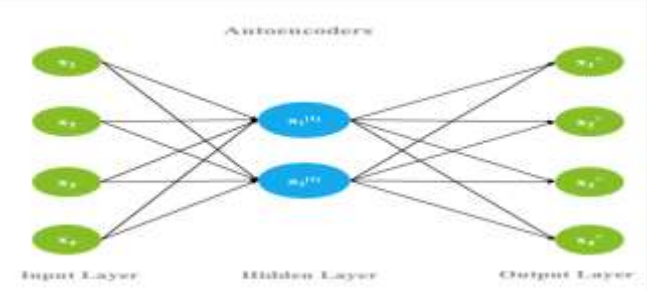


Fig. 6. Deep Learning Architecture

Components used in the model

1) *Dense Layer:* In natural language processing (NLP), a dense layer is a type of layer in a neural network that is fully connected to all the neurons in the previous layer. It is called a "dense" layer because each neuron in the layer receives input from all the neurons in the previous layer, as opposed to a "sparse" layer where some neurons do not receive input from all the neurons in the previous layer.

Dense layers are commonly used in NLP tasks such as language translation, text classification and language generation. They are used to process the input data, extract useful features, and make predictions based on the learned features. Dense layers are usually followed by an activation function such as ReLU or sigmoid, which helps introduce non-linearity into the model and allows it to learn more complex relationships in the data.

2) *Activation Function:* Activation functions are mathematical functions that are used in artificial neural networks to introduce non-linearity into the model. They are applied element-wise on the output of each neuron, allowing the network to learn and represent more complex relationships in the data.

ReLU (Rectified Linear Unit): It is defined as f(x) = max(0, x). It is commonly used in the hidden layers of neural networks and has the advantage of being computationally efficient.

Softmax: The softmax function is a popular activation function used in the output layer of neural networks for multi-class classification problems. It is a generalization of the sigmoid function, which is typically used for binary classification problems.

The softmax function is useful in multi-class classification problems because it converts the output of the neural network into a probability distribution over the classes.

The softmax function is computationally efficient and differentiable, which makes it easy to train the neural network using gradient-based optimization methods.

3) *Optimizers:* Optimizers are algorithms used to adjust the parameters of a neural network in order to minimize the error between the predicted output and the true output. They are used during the training process of a neural network to update the weights and biases of the network in order to improve its performance.

Stochastic Gradient Descent (SGD): It is one of the most basic and widely used optimization algorithm for training neural networks and other machine learning models. It updates the weights and biases of the network based on the gradient of the error with respect to the parameters.

The update rule for the parameters in SGD is given by:

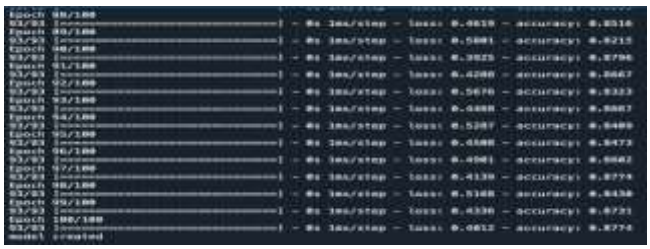
w <- w - learningrate * gradient

where w is the parameter to update, learningrate is a scalar that controls the step size.

D. Methodology

The deep learning model utilizes a multi-layer perceptron architecture with 3 layers. The input data is fed into the first layer, which consists of 128 neurons. The outcome of the first layer is then passed through the second layer, which contains 64 neurons. Dropout with a rate of 20% is applied to these layers to prevent overfitting. The outcome of the second layer is then passed through the final layer, which contains neurons equal to the number of intents. These neurons use a softmax activation function to predict the output intent.

The dense layers in this model use the Rectified Linear Unit (ReLU) activation function, which helps to introduce non-linearity and allows the model to learn more complex relationships in the data. The optimizer used for training the model is stochastic gradient descent (SGD) algorithm. The model is trained for 100 epochs with a batch size of 5 and achieved accuracy of around 88%



DOI: rp-9788770040723.102

Fig. 7. Model accuracy for 3 dense layers

It was further improvised by adding another layer with 32 neurons. So now the model utilizes a multi-layer perceptron architecture with 4 layers, where the first layer contains 128 neurons, second layer with 64 neurons, third layer with 32 neurons. Dropout with a rate of 20% is applied to these layers to prevent overfitting. The outcome of the third layer is then passed through the fourth and final layer, which contains neurons equal to the number of intents. These neurons use a softmax activation function to predict the output intent.

The model uses the same activation function, optimizer, and is trained for the same number of epochs and batch size and achieved accuracy of around 91%.

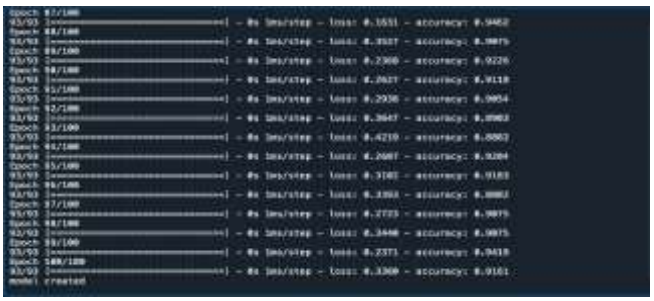


Fig. 8. Model accuracy for 4 dense layers

E. User Interface

A Chatbot GUI, or graphical user interface, is a type of interface that allows users to interact with a chatbot using graphical elements, such as buttons and images, rather than just text. This can make the chatbot more user-friendly and intuitive to interact with, as users are able to see options and make selections more easily. Some popular chatbot GUI platforms include Dialogflow, Botpress, and Microsoft Bot Framework.

Tkinter is a Python library for creating graphical user interfaces. It is built on top of the Tcl/Tk GUI toolkit and provides a simple and easy-to-use interface for creating windows, buttons, labels, and other widgets. Tkinter is included with the standard Python distribution and is widely used for creating simple GUI applications. Some of the features of Tkinter include support for events and callbacks, geometry management, and support for various types of widgets.

Tkinter library is used to view the gui for providing the chatbot experience.

Here the user input is processed with necessary techniques, then using the bag of words technique, a document-term matrix is created. Then it is sent for prediction. Error threshold is considered as 0.25. A tag is extracted after model prediction which is matched with the tag in the dataset and appropriate answer is returned to the user.

V. Evaluation and Model Results

The approach involves testing the dataset using deep learning algorithm using few dense layers and dropout to prevent over fitting. The algorithm has provided with accuracy around 91%. The model analyses user's query accurately and provides the user with appropriate answer.

Example 1 –

Fig. 9. Accuracy graph between models containing different layers

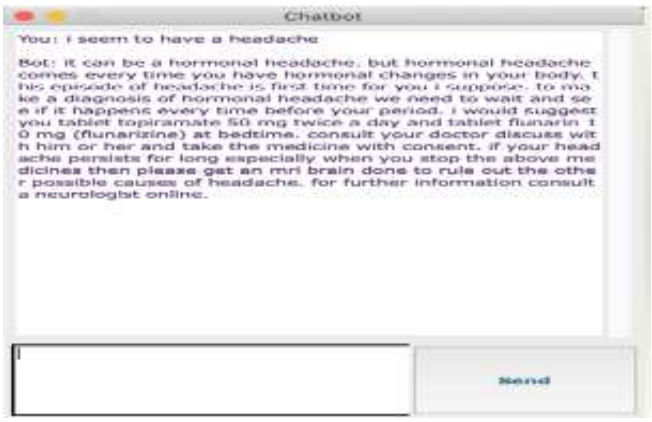


Fig. 10. Chatbot example1

Example 2

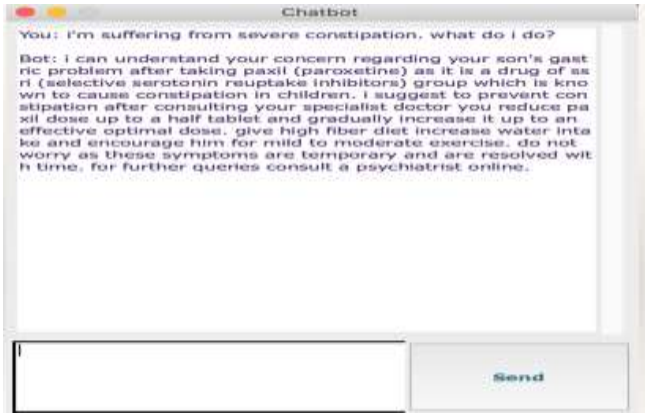


Fig. 11.Chatbot example2

VI. Discussion and Conclusion

The study was done within a very short span. All the possibilities are not considered in this study and a lot of improvement is possible. There are several features or algorithms that can be implemented for improving the overall performance of the model. By using NLP techniques such as unigrams, bigrams, and trigrams, healthcare chatbots can understand and respond to user's inquiries more accurate and appropriately.

There is a definite scope of development in this area, we can enhance the current model and work towards slowly eliminating the limitations such as :

- Currently negative scenarios are not handled. For any irrelevant query, random answer from the dataset is provided. It doesn't prompt the user to enter a relevant query.
- There are no greetings. The algorithm presented currently only works as a simple question answer healthcare chatbot.

Future work can involve improving the model by making use of optimizers, different algorithms, multilingual ability. Using different dataset system will be able to predict diseases based on the symptoms provided by the user. We can also incorporate voice to text and text to voice functionality to make it user friendly.

REFERENCES

1. N. Darapaneni et al., "Customer Support Chatbot for Electronic Components," 2022 Interdisciplinary Research in Technology and Management (IRTM), 2022, pp. 1-7, doi: 10.1109/IRTM54583.2022.9791730
2. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
3. Chenguang Zhu, Michael Zeng, and Xuedong Huang, "Sdnet: Contextualized attention-based deep network for conversational question answering,"arXiv preprint arXiv:1812.03593, 2018.
4. Rajesh, M., &Sitharthan, R. (2022). Image fusion and enhancement based on energy of the pixel using Deep Convolutional Neural Network. Multimedia Tools and Applications, 81(1), 873-885..
5. Kennedy Ralston, Yuhao Chen, HarunaIsah, and FarhanaZulkernine. A voice interactive multilingual student support system using ibmwatson. In 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), IEEE, pp. 1924–1929, 2019.
6. Urmil Bharti, Deepali Bajaj, HunarBatra, Shreya Lalit, ShwetaLalit, and AayushiGangwani, "Medbot: Conversational artificial intelligence powered chatbot for delivering tele-health after covid-19," In 2020 5th international conference on communication and electronics systems (ICCES), IEEE, pp. 870–875, 2020.
7. JashDoshi,"Chatbot user interface for customer relationship management using nlp models," In 2021 International Conference on Artificial Intelligence and Machine Vision (AIMV), IEEE , pp. 1–4, 2021.
8. SagarBadlani, Tanvi Aditya, Meet Dave, and Sheetal Chaudhari. Multilingual healthcare chatbot using machine learning. In 2021 2nd International Conference for Emerging Technology (INCET), IEEE, pp. 1–6, 2021.
9. Ming-Hsiang Su, Chung-Hsien Wu, Kun-Yi Huang, Qian-Bei Hong, and Hsin-Min Wang, "A chatbot using lstm-based multi-layer embedding for elderly care," In 2017 International Conference on Orange Technologies (ICOT), IEEE, pp. 70–74, 2017.
10. LekhaAthota, Vinod Kumar Shukla, Nitin Pandey, and Ajay Rana,"Chatbot for healthcare system using artificial intelligence," In 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO), IEEE, pp. 619–622, 2020.
11. Jintae Kim, Hyeon-Gu Lee, Harksoo Kim, Yeonsoo Lee, and Young-Gil Kim, 2018. "Two-Step Training and Mixed Encoding-Decoding for Implementing a Generative Chatbot with a Small Dialogue Corpus," In Proceedings of the Workshop on Intelligent Interactive Systems and Language Generation (2IS&NLG), Tilburg, the Netherlands, Association for Computational Linguistics, pp. 31–35, 2018.
12. M.H. Su, C.H. Wu, K.Y. Huang, Q.B. Hong, and H.M. Wang, "A chatbot using LSTM-based multi-layer embedding for elderly care," 2017 International Conference on Orange Technologies (ICOT), Singapore, pp. 70-74, 2017,doi: 10.1109/ICOT.2017.8336091.
13. D.S. Chauhan, A. Rastogi, and K. Singh, "Depression Chatbot using Deep Learning," 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, pp. 1294-1297, 2022,doi: 10.1109/ICACITE53722.2022.9823825.
14. P. Muangkammuen, N. Intiruk and K. R. Saikaew, "Automated Thai-FAQ Chatbot using RNN-LSTM," 2018 22nd International Computer Science and Engineering Conference (ICSEC), Chiang Mai, Thailand, pp. 1-4, 2018,doi: 10.1109/ICSEC.2018.8712781.
15. A. K. M. Masum, S. Abujar, S. Akter, N. J. Ria, and S. A. Hossain, "Transformer Based Bengali Chatbot Using General Knowledge Dataset," 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA), Pasadena, CA, USA, pp. 1235-1238, 2021,doi: 10.1109/ICMLA52953.2021.00200.
16. 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.
17. 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, pp. 870-875, 2020,doi: 10.1109/ICCES48766.2020.9137944.
18. https://github.com/LasseRegin/medical-question-answer-data.
19. https://hiteshmishra708.medium.com/build-your-first-chatbot-using-nltk-and-keras-6470501abe28.
20. Ivan Vasilev, Daniel Slater, Gianmario Spacagna, Peter Roelants, Python Deep Learning, Packt Second edition.
21. Pazhani. A, A. J., Gunasekaran, P., Shanmuganathan, V., Lim, S., Madasamy, K., Manoharan, R., &Verma, A. (2022).Peer–Peer Communication Using Novel Slice Handover Algorithm for 5G Wireless Networks.Journal of Sensor and Actuator Networks, 11(4), 82..