Content Extraction Through Chatbots With Artificial Intelligence Techniques

Suresh P, Ravikumar O, Hari Krishna Mahesh K, Sri Aashritha S

Abstract : In recent years, the search engine plays a major role in extracting content in real time for analysis, understanding, etc., and also needs time to process. The chatbots is another form of extracting the required content in different ways with Artificial Intelligence. The main objective of this article is to extract the required content from wiki platform and to deliver in speech and text format for better understanding. There are countless ways, were the chatbots helping in most of the environments with better performance. Even though there is lagging in delivering the required content. In this paper, the AI delivers the required information with shorter period of time in the proposed deliverable formats. The proposed chatbots will work efficiently on the existing database even in addition of more information in the same database. With the advent of these chatbots, needy people can get updated proper information on their own required proposed formats.

Index Terms: Chatbots, Speech Processing, Artificial Intelligence, Social Networking Sites, Wiki, Content Delivery, Deep Learning.

1. INTRODUCTION

Artificial Intelligence gives the incredible power to mimic the way of how humans think and behave to a computer. The chatbot is such kind of computer programs that associate with user's natural languages [1]. However, Chatbots are not only erected to mimic the conversation of humans but entertain users. Chatbots are applicable due to the following reasons like, 1. To provide people with the information they need at the right time when they want, 2. about 90% of the time we spend on chatting in day to day life. So, this helps us understand how much we need a chatbot and 3. the advent of AI allows the chatbot to easily interact with user. The chatbot works basically on Artificial intelligence, so using this ability we need to contribute some health informatics [9]. A Chatbot system [4] is a software program that associates with the user's natural language. Many investors came to know the capability of AI and they have made significant investments to tackle the technology. Investors include: Jack Ma, the founder of Alibaba [8]. He invested in an Israeli start up using AI to evolve e-commerce search technologies. Li Ka-Shing, the Hong Kong billionaire. He invested in several startups focusing on AI [9]; Dr. Kai-Fu Lee, a famous tech investor. He invested in several investments on AI start-ups that focus on the development of AI [10]. Many names have been used to call a Chatbot such as Talkbot, interactive agent and artificial conversation entity [3]. The main aim of the chatbot is to communicate with humans with conversation the Chatbot architecture merges a language model and computational algorithms [10] to follow informal chat conversation between a user and a computer using natural language processing methods. In addition to that, we have added some pharmaceutical details to the system so that chatbots delivers this information to the user whenever they need it. Text-based chatbots are more where you look right now. Voice-enabled chatbots are becoming more familiar with the advent of Google Home, Amazon Echo, etc. The chatbot is an intelligent system that can handle complex interactions with users in their natural language. What we suggest is a voice-based chatbot that associates with the user in their natural language via speech for allowing users to search for their queries. Successful execution of our system will understand that what user is saying, analyse it, and gives an acceptable response if the query is related to the user input.

Speech is one of the most important keys to enable communication between human and chatbot; hence, it is important to improve audio interaction between the human and the chatbot to simulate human to chatbot speech interaction. Speech interaction with modern networked computing devices has achieved more interest in the past few years with offering from IOS, Google and Android. Natural Language Processing is one of the finest techniques to interact with chatbots Therefore, speech interaction will play a vital role in humanizing chatbots in the near future [2]. Speech Interaction splits into speech parsing, speech recognition, NLP (Natural Language Processing), keyword identification.

2 BACKGROUND

2.1 User – Chatbot Speech Interaction

Speech recognition is one of the most natural and familiar techniques in computer and chatbot interaction with users has recently become possible with the advent of fast computing. Speech is an enlightened signal and happens at different levels such as linguistic, semantic, articulatory, and acoustic [12]. Speech is one of the most natural among the aspects of human communication, owing to copious information implicitly existing beyond the meaning of the spoken words. The speech information extraction techniques are speech to text via mining speech information and Automatic Speech Recognition (ASR) [5]. Hence the text can be easily treated to obtain the suitable meaning of the any words. Speech recognition is widely accepted as the future of interaction with chatbots; there is no need of external sources. It can help disabled people by interacting with them without any movement of hands.

- *Suresh P, Dept. of ECE, Vel Tech Rangarajan Dr.Sagunthala R & D Institute of Science and Technology, Chennai, Tamilnadu – 600062. India. E-mail: sureshp@ieee.org.*
- *Ravikumar O, Dept. of ECE, Vel Tech Rangarajan Dr.Sagunthala R & D Institute of Science and Technology, Chennai, Tamilnadu – 600062. India. E-mail: ravikumarorsu7@gmail.com.*
- *Hari Krishna Mahesh K, Dept. of CSE, Vel Tech Rangarajan Dr.Sagunthala R & D Institute of Science and Technology, Chennai, Tamilnadu – 600062. India. E-mail: ravikumarorsu7@gmail.com.*
- *Sri Aashritha S, Dept. of CSE, Vel Tech Rangarajan Dr.Sagunthala R & D Institute of Science and Technology, Chennai, Tamilnadu – 600062. India. E-mail: ravikumarorsu7@gmail.com.*

2.2 Natural Language Processing using NLTK

Suitable toolkits are essential in order to deal with and manipulate the text obtained from speech recognition and conversion of speech to text to manipulate the text into sentences and then split them into words, to generate semantic and meaning extraction. NLTK is one of the free plugins for Python, that is most widely used toolkit for speech processing. The Natural Language Toolkit (NLTK) is one of the leading text processing platforms for natural language processing (NLP) for English. Generally, this toolkit used [13] in the Python programming language. Actually, this toolkit was developed and implemented by Edward Loper and Steven Bird in the Department of CIS at the University of Pennsylvania. It includes graphical data demonstrations and sample data. The documentation regarding Natural Language Toolkit can be obtained from its cookbook. [6] The intention of NLTK is to support research and teaching in NLP, including cognitive science, artificial intelligence, empirical linguistics, information retrieval, and machine learning [11].

2.3 TF-IDF (term frequency–inverse document frequency)

For information retrieval, term frequency-inverse document frequency is the most familiar toolkit among all other text processing toolkits. In short, it can be termed as TF-IDF. The main objective of this is to form a correct sentence using the words which were more repeated in the document.it also used as a weighting factor to determine the repetition of words in the given sentences. this toolkit also provides text mining and user modelling. As the repetitions of words increases then the TF-IDF value also increases. Term frequency determines the occurrence of the word in the document. Inverse document frequency calculates the weight of the word in the document. However, this technique helps in many strategies such as checking the sentences which belong to the given title.

TF-IDF can be calculated as:

1.Term frequency (TF):

It is the ratio of occurrence of the term in the document and sum of all terms in the document.

Tf(“word”,p1)=no.of times occurred/sum of all terms (1)

2.Inverse document frequency (IDF):

It is the logarithmic ratio of total number of documents and number of documents in which the required term occurred

Idf(“word”,P)=log(d1\d2) (2)

Where, d1=total documents

d2= number of documents in which the required term occurred.

Hence TF-IDF of the word will be the product of tf and idf values i.e.,

Tf(“word”,p1)* Idf(“word”,P) (3)

The set of data regarding tf-idf will be stored in matrix by converting each row into a vector.

2.4. Cosine similarity

It is the measure to check the similarity between two non-zero vectors in space by cosine angle between them. Whenever, the two vectors with the same orientation have unit cosine similarity. similarly, vectors oriented orthogonal to each other then, the similarity will be zero. cosine similarity is mainly used in positive space in order to be bounded. This technique also used in data mining. Since it is the familiar technique because of the computations less complex. Cosine similarity can be calculated as follows,

$$\cos(\Theta)=A.B/\|A\|*\|B\| \quad (4)$$

Angular distance of the two vectors can be calculated

Angular distance is the ratio of inverse cos of cosine similarity to pi.

Hence angular similarity will be 1-angular distance.

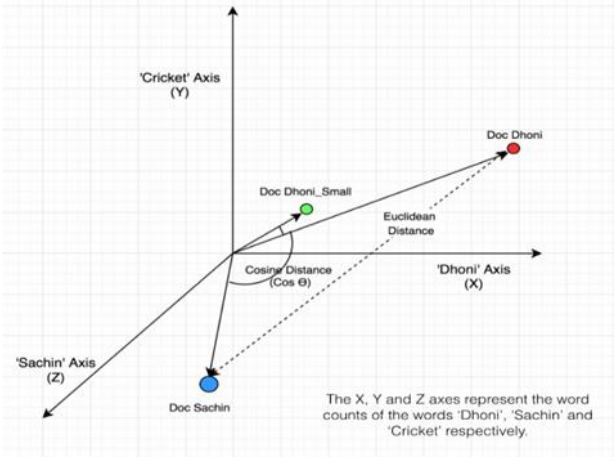

***Fig. 1**. Projection of documents in 3D space*

Consider an example with two documents which are having nearly similar Content. Let us check the occurrence of words “sachin”, “dhoni” and “cricket” in two documents using cosine similarity.

This example results the cosine similarity between two non-zero vectors.

2.5 Python Wikipedia Wrapping

Python supports the wrapping of websites using API. Among those Wikipedia is a more familiar Python library that makes the user access the desired data from Wikipedia. This provides browsing, summaries of articles, images from the required page. Even though, many search engines exist like google, yahoo, bing, etc. it is more prominent to users because of its accuracy. Wikipedia provides worldwide data with a single API. That’s why this application had been used in this chatbot. Wikipedia wrapping is the heart of this proposed idea.

2.6 Datasets of health informatics

Apart from the browsed data, this chatbot consists data collection of various diseases and their description in the form of dataset. This enables the user come to know about various

diseases. This chatbot reveals the data about diseases using user audio input. The dataset is given below:

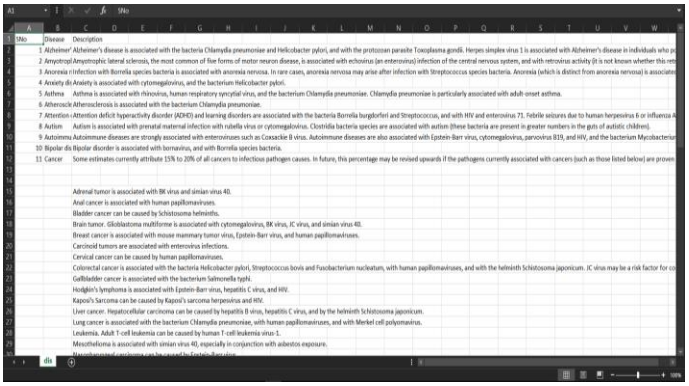

***Figure 2**. Dataset of Various diseases and their description*

3. PROPOSED SYSTEM

Our proposed chatbot takes user input in the form speech using python speech recognition package and analyses the user input using natural language processing technique to find what the user tries to ask and provides response accordingly. Apart from other chatbots, our chatbot provides output to user either from Wikipedia page or pre-loaded datasets. The output will be in audio format. It enables the user to avoid effort on typing on system.

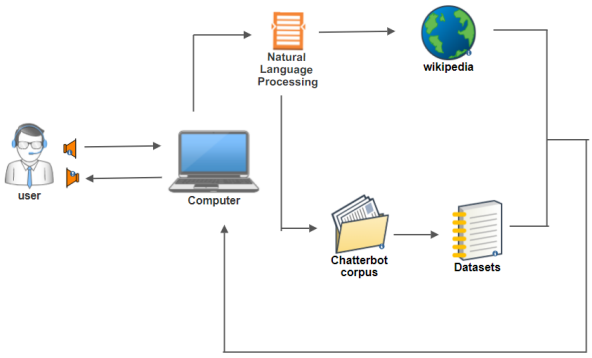

***Figure 3**. Proposed System Architecture*

4 RESULTS

The results of our proposed chatbot are more accurate which satisfies the user and successfully covered the above mentioned techniques by maintaining somewhat efficiently. Here we observe how our chatbot interacts with the users and fulfilling their queries.

Case 1: Mahendra Singh Dhoni

Let us observe how our proposed chatbot responding to the user input in audio format “Mahendra Singh Dhoni”, chatbot responded to user with more details about Mahendra Singh Dhoni.

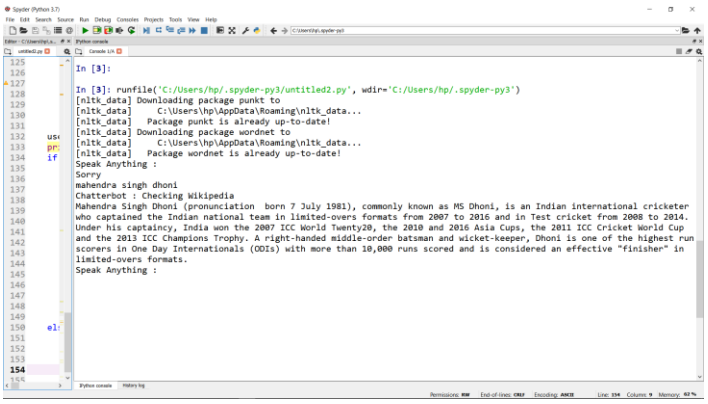

***Figure 4.** Chatbot response for Mahendra singh Dhoni*

Case 2: Mahatma Gandhi

Test with the name of our famous freedom fighter “Mahatma Gandhi”. He was well-known to all but, lets look how our chatbot will respond to the user.

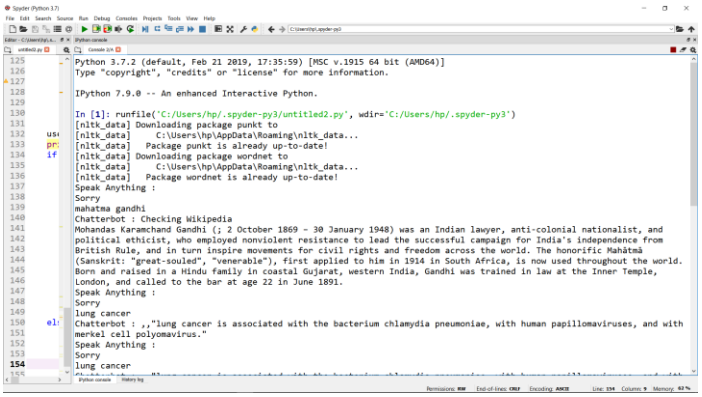

***Figure 5:** Chatbot response for Mahatma Gandhi*

Case 3: Diseases

Lets us give the input as “Lung cancer” in audio format then the output will be as shown in Figure 6

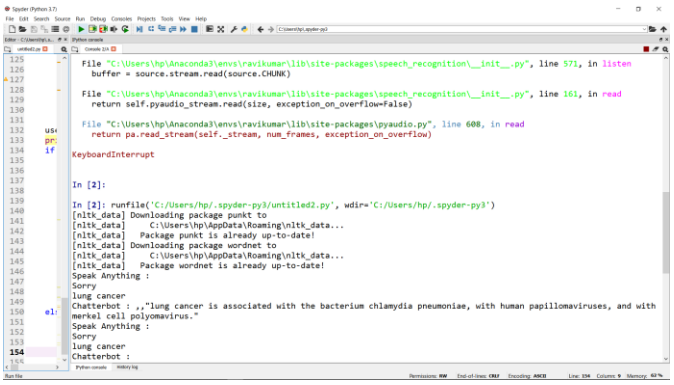

***Figure 6.** Chatbot response for lung cancer*

5 CONCLUSIONS

Chatbots are more familiar to people because of their involvement in our day to day life. So far, we have seen

chatbots for particular domain, such as business, customer services, entertainment, etc. But this proposed chatbot resolves the queries of user from different domains i.e., this chatbot used for multipurpose. The proposed chatbots delivers the required information in nano seconds by using the Artificial Intelligence techniques. The proposed system is tested for different cases and the sample is discussed above. The deliverable content is in form of speech (audio) format and on screen graphics text format, even if needed the same can delivered in printed format with addition printing equipment interfacing with it.

REFERENCES

[1] B.A. Shawar, Eric Atwell, “Chatbots: are they really useful?”, LDV Forum 2007.
[2] V. Bhargava, and N. Maheshwari, “An Intelligent Speech Recognition System for Education System,” 2009.
[3] S. J. du Preez, M. Lall, S. Sinha, “An Intelligent Web based voice chat bot”, IEEE, 2009.
[4] S.A. Kader, J. Woods, “Survey on Chatbot Design Techniques in Speech Conversation Systems”, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 6, No. 7, 2015.
[5] C.H. Lee, "From knowledge-ignorant to knowledge-rich modeling: a new speech research paradigm for next generation automatic speech recognition", 2004.
[6] S. Bird, "NLTK: the natural language toolkit." pp. 69-72, 2006.
[7] D. Vrajitoru, "Evolutionary sentence building for chatterbots." pp. 315-321, 2003.
[8] P. Galvin, “Alibaba Invests in AI Startup,” Tech Exec., 2016. [Online]. Available: http://techexec.com.au/alibabainvests-in-ai-startup/. [Accessed: May 19, 2017].
[9] S. Deveau, “Li Ka-Shing Buys Canada’s Reliance Home for $2.1 Billion,” Bloomberg, 2017. [Online]. Available: https://www.bloomberg.com/news/articles/2017-03-31/lika-shing-s-ckp-buys-canada-s-reliance-home-for-c-2-82- billion. [Accessed: May 18, 2017].
[10] R. Dillet, “Sinovation Ventures’ Dr. Kai-Fu Lee is betting big on artificial intelligence,” TechCrunch, 2016.
[11] J. Awwalu, A. Garba, A. Ghazvini, R.A. Dale, “The return of the chatbots. Natural Language Engineering, Vol. 22, Issue 5, (2016) pp.811-817.
[12] J. P. Campbell, “Speaker recognition: a tutorial,” Proceedings of the IEEE, vol. 85, no. 9, pp. 1437-1462, 1997.
[13] E. Loper, and S. Bird, "NLTK: The natural language toolkit." pp. 63-70, 2002.