DESIGN AND DEVELOPMENT OF PYBOT USING NLP*

BY

ANSHULEE GAUTAM♣

Student, School of Computer Science and Engineering, Ajeenkya D Y Patil University, Pune

anshulee.gautam@adypu.edu.in

ARUNA VERMA

Assistant Professor, School of Computer Science and Engineering,

Ajeenkya D Y Patil University, Pune, India

aruna.v@inurture.co.in

ABSTRACT

A Pybot is basically a chatbot that answers all the queries related to python. It breaks downs the internet barrier and gives the information offline. As per the new education policy coding will be taught to the students from 6th standard so it will be very beneficial for students living in rural areas and even for the students residing in urban areas as it is an interacting method of clearing doubts and learning new things. It will also help in developing and building mini projects. It is built using NLP (Natural Language Processing) in order to maintain smooth interaction between human and computer. For better interaction text to speech option is available, different themes and fonts are also available. The Pybot also shows notes and displays the date and time of the last interaction. There is a text to speech choice, as well as different themes and fonts, for better interaction. The Pybot often presents notes and the last interaction's date and period.

KEYWORDS

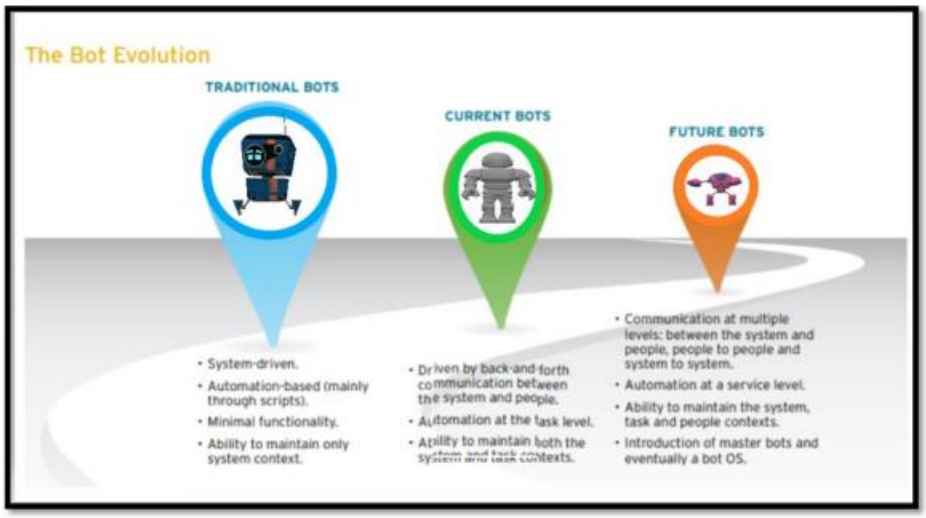
Pybot, Chatbot, NLP, Python Chatbots, Text to speech, Python queries.

1Introduction

A chatbot is a piece of artificial intelligence software that works on a system (Alexa, Google Assistant, Siri, etc.), application, website, or other network to assess students' needs and then assist them in completing a task such as a business transaction, ordering food, booking hotels, filling out forms, and so on. Today, almost every company uses a chatbot to connect with students. The history of chatbots dates from 1966, when Weinbaum invented a computer program named ELIZA. Out of only 200 lines of code, it imitated the language of a psychotherapist.

* Received 22 September 2021, Accepted 09 October 2021, Published 24 October 2021

♣ Corresponding Author



NLP (Natural Language Processing)

The field of study about that centers on the interactions between human dialect and computers is called Natural Language Processing, or NLP for brief. It sits at the crossing point of computer science, counterfeit insights, and computational linguistics. NLP may be a way for computers to analyze, get it, and determine meaning from a human dialect in a keen and valuable way. By utilizing NLP, engineers can organize and structure information to perform assignments such as programmed summarization, interpretation, named substance acknowledgment, relationship extraction, estimation analysis, speech recognition, and topic division.

Natural Language Processing with Python provides a practical introduction to programming for language processing.

Text to Speech

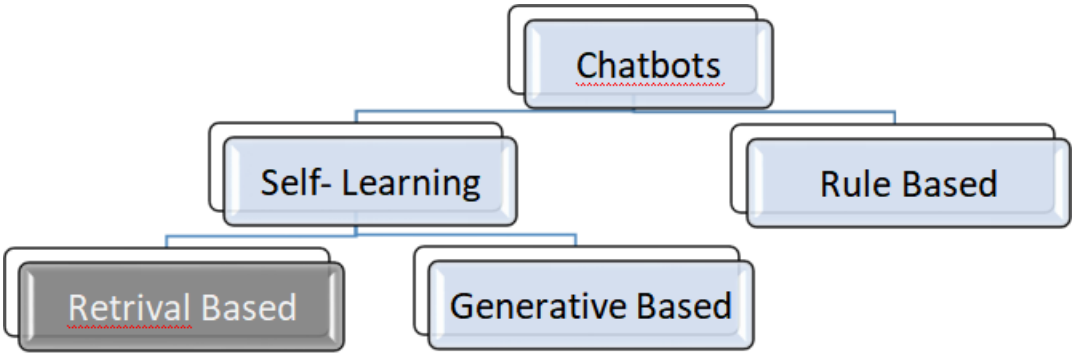
Text-to-speech is the process of converting written content into spoken words (TTS). Python is a programming language that may be used to create TTS apps. The discourse engine determines the substance of the voice. A text-to-speech (TTS) application reads out automated messages. It can convert text to audio on laptops, smartphones, and tablets. It is possible to read aloud any sort of text file, including Word, paper pages, and internet web pages. TTS will help youngsters who are struggling with reading. There are several APIs in Python that may be used to convert text to speech. Google Text to Speech, often known as the gTTS API, is one such API. The library that turns typed text into a storable audio file. The library that converts text into an audio file that can be stored as an mp3 file is extremely easy to use. Speech may be

conveyed fast or slowly at either of the two possible audio rates, and other languages are supported.

NLTK (Natural Language Toolkit)

NLTK is a popular forum for Python programmers that want to deal with human language data (Natural Language Toolkit). It provides easy-to-use interfaces for over 100 businesses and lexical tools like WordNet, in addition to a suite of text processing libraries for the processes including grouping, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and wrappers for industrial-strength NLP libraries.

1.1Approaches



Chatbots are divided into two types: rule-based and self-learning.

In a rule-based approach, a bot responds to questions based on a set of rules that it has learned. The rules defined can range from extremely simple to extremely complex. The bots are capable of handling simple questions, but they fall short when it comes to handling more complicated ones.

1.Self-learning bots are more powerful than rule-based bots since they use a few Machine Learning-based approaches. These bots can be one of two types: Generative or retrieval-based.

2.A chatbot uses a few heuristics to select a reaction from a library of predefined reactions in retrieval-based models. The chatbot uses the conversation's message and setting to choose the best response from a pre-programmed selection of bot messages. A present location inside the exchange tree, all previous messages inside the conversation, and already spared elements can all be included in the configuration (e.g., username). From rule-based if-else conditional

logic to machine learning classifiers, heuristics for picking a reaction may be constructed in a variety of methods.

3.Generative bots may generate responses rather than responding to questions with one of a series of answers. This makes them smarter so they build the answers word for word from the inquiry.

2 Literature Review

2.1Literature Review Table

Sr no	Chatbot	Input/Output	Technique	Drawback
1.	CHARLIE[1]	Self learning and response based on previous chats	AIML, NLP	Poor response on web browser
2.	ALICE[3,14]	Pattern matching in order to represent the input & output	Uses Recursive Techniques, AIML	Poor grammatical analysis to frame sentences.
3.	CLEVERBOT[3]	Keywords matching for input and response based on previous chats	Uses Rule Based Technique	Without context responses
4.	ViDi[4,5]	One-Match All-Match Category (OMAMC)	AIML, NLP	Web browser based with inappropriate output
5.	DOLLY[10]	Pattern Matching and previous chat based	Self- Learning	Incorrect translation of sentences.
7.	ELIZA[15]	In order to generate responses it deal with basic pattern matching with the templates.	Uses Template Based	Inaccurate Responses & No logic behind responses.
8.	WATSON[16]	Recognizes the feature values in order to generate responses based on the obtained score	Uses Rule Based technique, UIMA , NLP	Works without Relational Databases, cannot process structured data.
9.	CHAT SCRIPT[17]	Pattern Matching	Uses Script Based technique	Hard to learn & web page
10.	CHATFUEL[18]	Mapping of input sentences to output	Uses Rule Based technique	Flow of Conversation is Inflexible

2.2Literature Review Description

1.Chatbots may also be utilised to create educational apps. The goal is to react to students' inquiries or to test for an evaluation by presenting questions and assessing the replies. The creators of [1] are concentrating their efforts on the CHARLIE Chatbot (Chatter Learning Interface Substance). The platform is a Brilliantly Instructive Framework AIML Chatbot consolidated platform (INES). CHARLIE discusses his execution and drive in his essay, as well

as his ability to develop a shared discussion with understudies; he can show them the fabric of the courses they are contemplating, and he is willing to ask questions about the fabric presented. Instructive applications of discourse frameworks are especially valuable and are exceedingly intelligently. They can be moved forward and upgraded effectively since they are used in an scholarly environment.

2.The use of Chatbots to Incapacity Care necessitates the creation of packages and frameworks in order to provide disabled people with hitherto untapped technologies. The authors of [2] presented a question-and-answer instructional system for people with disabilities that takes into account natural

language conversation and constrained word conversation. The system is set to use an AIML knowledge base with a limited lexicon that counts voice recognition or "phoneme groups and words." The AIML question-answer framework is used to generate responses to questions, and then 2500 words of preparation data are used to test it. The examination had 250 words of information, six of which were recognised; As a result, the framework precision was 76 percent. The idea was to include it into English dialect lesson software that disabled people could easily access. This type of highlight in E- learning systems can benefit those who are blind or have hand lack of mobility. The test contained 200 words of information, and 156 of them were identified, resulting in a framework accuracy of 78 percent. The idea was to include it into English dialect lesson software that disabled people could easily access. This type of highlight in E-learning systems can benefit those who are blind or have hand lack of mobility.

3.Despite the fact that developing a new type of Chatbot is a commitment to the industry, the software architect only has a few options. The developers of [3] constructed knowledge bases for Chatbots by integrating the qualities of two different Chatbots. To avoid shielding, detect personal inquiries, and remove undesired terms or phrases from the data bases, the authors utilised three filters. The corpus consists of an ALICE basic type Chatbot, which might be a QA frame, and another Chatbot, such as CLEVERBOT or JABBERWACKY, that excels at managing conversational talk. The chatbot will either trade or QA play according on the cumulative contact ordering. Based on the appropriately processed discussion, they built a Chat corpus of roughly 7500 and more encounters in total. Their curiosity sprang from a drive to improve Chatbot planning approaches. The Chatbot's authors configured it to replace or QA suit according on the cumulative touch ordering. They created a Chat corpus of roughly 7800 sets of

interactions based on the managed conversation. Their curiosity was sparked by a desire to enhance Chatbot preparation methods.

4.Design coordinating techniques may also be integrated in the Chatbot plan environment, which can help with recovery precision. Using ViDi as a study scenario, the authors of [4] proposed a new methodology for watchword coordination ([5] and upgraded in [6]). The recommended strategy is to use one coordinate or both coordinates to match categories (OMAMC). On the basis of a single example statement, OMAMC is used to assess the development of prospective catchphrases. In the same test sentence, the results are compared to other watchwords developed by a prior Chatbot. OMAMC is thought to be making progress in catch-up coordination when compared to earlier techniques. This modern approach is likely to be found in future instantiations of Chatbots.

5.The concept in [7] became a reality when the authors in [8] displayed their novel approach to Chatbot design. The methodology combines ordering and query matching techniques with template collaboration, as well as Information Recovery (IR) algorithms, to construct a new sentence from old ones. Existing sentences became the GA's beginning population, according to their theory, and then the swap and hybrid administrators were coupled to generate the unused sentence as an unused GA era. The Chatbot's exploratory evaluation before and after utilising the sentence mix methodology may be viewed. The goal of the method was to expand the gaps in Chatbot answers. The consider's two key obligations are to i) combine two sentences into one and ii).Applying information retrieval methods to Chatbots.

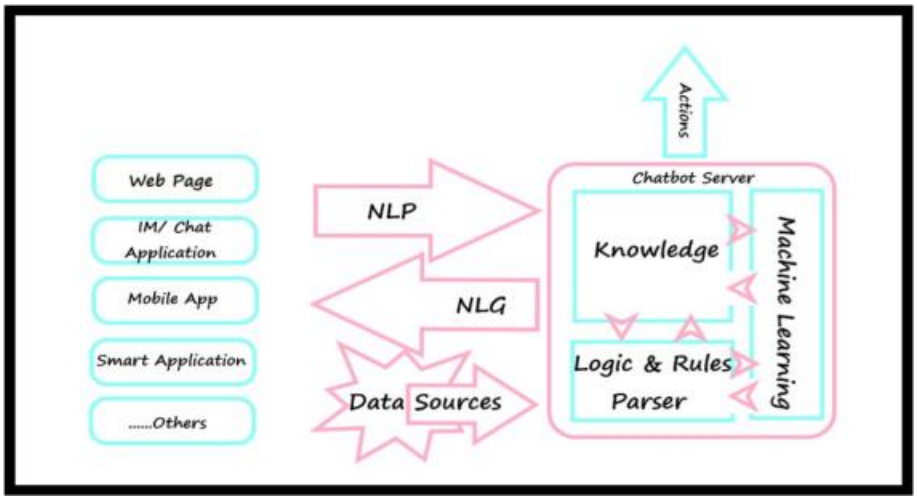
6. The study [10] covers The Dolly, a Bengali-language-based chatbot. This logical term paper might be a technique for a computerised system called "Dolly" to respond to a user enquiry without the need for a human to learn Bengali dialect. A typical AI-powered Chatbot is built mostly on machine learning algorithms and Bengali Dialect Preparation (BNLP). As a response to questions, this information is implanted into the computer so that it can distinguish between the provided statements and make an independent judgement. Education, internet investigation, account administration, and other divisions all use English Chatbots. They have made the prepare work connector to prepare the Doly by encoding (encoding="utf8") our corpus from bot information.

7.Simon Hoermann[11] studies the feasibility and efficacy of text-based synchronous chat-based online one-on-one mental health therapies. Synchronous textual dialogues (or "chats")

are becoming more popular as a type of Web-based mental health intervention. This poll is based on an assessment of people's synchronous Web-based chat developments. In the most recent proof of functioning of this technology, the conditional back for manner of intercession may be shown. Interventions employing text-based synchronous connection had somewhat enhanced impacts when compared to Waitlist settings, with generally comparable outcomes when compared to Treatment settings. They worked just as well as the control treatments, if not better. However, future research will look at whether or not these technologies are cost-effective in clinical settings. However, future research will look at whether or not these technologies are cost-effective in clinical settings.

3 Proposed Methodology

Anatomy of proposed system,



We will build a simple retrieval based pybot - a chatbot that answers all python queries based on NLTK library in python.

Chatbots are available 24 hours a day, 7 days a week: Chatbots are available to solve students' problems 24 hours a day, 7 days a week, regardless of the time of day or night! They don't need to sleep with all of this! Using human students as support is much more difficult to accomplish because it will require revolving teams and will be more difficult to manage. This suggests that chatbots will respond to python- based student inquiries when consumers have requests that tend to improve student satisfaction.

Normalization is added to each input, which includes removing all punctuation, dividing into two or three words, and converting to upper case. As an example, Is it possible for you to do this, or

is it impossible? Will YOU OR CAN YOU NOT DO THIS is how conversions are expressed. The process of interpreting spoken language. Using semantic portrayal, the module can comprehend the significance of the topic. It is accomplished in such a way that the software is not affected. The device should recognise the user's true needs. The representation of a question will aid the computer in interpreting the query's connotation and denotation. Semantic representation is done using one of two approaches: orthographic similarity or phonetic similarity.

Put into arrays, confidence levels are assigned to queries to determine whether the customer communicated and the computer understood. The scale of trust ranges from 0 to 1, with 1 being the most favourable. The refined form is then fed into the module that handles dialogue. This module would make use of the data that has already been fed into the knowledge base, as well as the various web resources available to evoke a response.

The knowledge base will be fed with a compilation of previous information on sport- related issues collected using AIML. To match a query, the pattern matching algorithm uses DFS in the solution generation module, and it continues until a match is found. The interpreter analyses and generates the output for the template that belongs to that particular category. The computer adds the constructed output as an input to the text-to-speech synthesis module. As an input, the speech response will be fed to the recipient in the form of sentences. The conversion of text back to speech is implemented using Google text-to-speech synthesis We use Google TTS synthesis by importing the e kit android. Kit for speech.tts The system asks the user to see if there is any query that needs to be answered. If the user has another query, the computer will either replay the whole operation or interrupt the execution.

Flowchart

Steps involved in building a Pybot:

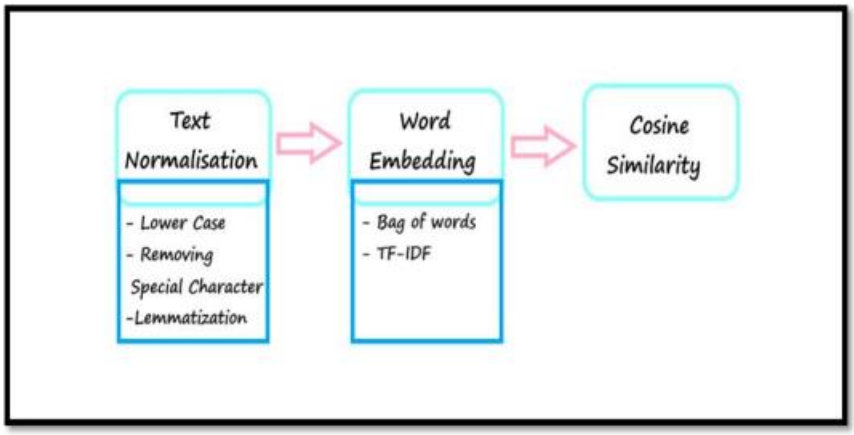
1.Importing the necessary libraries

2.Importing the data

3.Pre- processing the raw data

4.Keywords matching

5.Generating responses



With NLTK, you can pre-process text.

The fact that text data is in text format is the main source of concern (strings). Machine learning algorithms, on the other hand, require some kind of computational feature vector in order to complete the task. So, before we start any NLP project, we need to pre-process it to make it workable. Pre-treatment Simple Text contains the following features: convert the whole text to upper or lower case so that the algorithm does not interpret the same terms differently in different contexts.

The major concept of building a bot includes:

1.Bag of words: We need to convert the text into a meaningful vector (or array) of numbers after the initial pre-processing stage. A text representation that defines the presence of words in a document is the bag-of-words. This covers two things:

•A dictionary of words known.

•A test of the existence of terms understood.

Why do you call it a "bag" of words? This is because all detail in the text is ignored on the order or arrangement of words and the model only concerns where the known words appear in the document, not where they occur in the document.

2.Approach from TF-IDF: A issue with the Bag of Words technique is that in the text, very common words tend to dominate (e.g. greater score), but do not provide as much "information content." Also, longer papers would be given more weight than shorter documents. One solution

is to rescale the number of terms for how frequently they occur in all documents in order to penalize the ratings for repeated words such as "the" that are also prevalent in all documents.

Tokenization is the phrase used to describe the process of converting regular text strings into a token set, which is a list of terms that humans find appealing. To discover a collection of sentences, use the sentence tokenizer. Use the Term tokenizer to obtain a list of terms in the strings.

Noise Reduction: Anything that isn't a conventional number or letter is considered noise reduction. The words "Stop" have been removed. Certain extremely known keywords that appear to be of little value in attempting to identify texts that satisfy a user's expectations are sometimes eliminated entirely from the lexicon. Stop language refers to statements like this.

Stemming: The act of reducing inflected (or occasionally derived) words to their stem, nucleus, or root form, which is often a written word form, is known as stemming. For example, combining the terms "Stemmings," "Stemmed," "Stems," and "and Stemtization" would get a single "stem" term.

Lemmatization is a type of stemming that is a subset of stemming. The main distinction is that stemming will also yield non-existent terms, whereas lemmas are genuine words. Instead of searching up the root stem in a dictionary, which is what you'll end up with, search up a lemma. For example, "run" is a simple type of word, and "run" is a simple kind of word, and "run" is a simple kind of word, and "run" is a simple kind of word, and "run" is a simple kind of word, and "run" is a simple kind of word. The fact that "run" is a fundamental form of words like "rolling" or "ran," or that the terms "best" and "lovely" are in the same lemma and so regarded the same, are examples of lemmatization.

Since the refining process has so many distinct subsets, cosine similarity is used. And, using localization, it can be fine-tuned. We need to use cosine similarity to measure two non-zero vectors when converting a large portion of the interpreted data

Expected Outcome

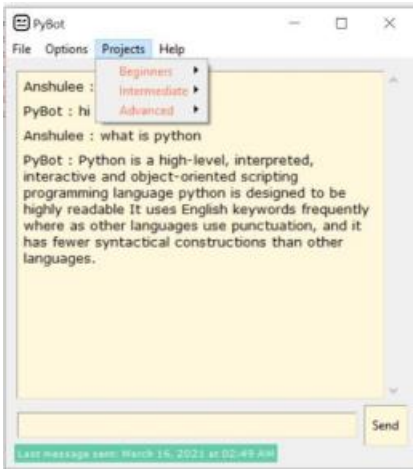
Understand the context of small talk: Each chatbot, in my opinion, should standardise itself in order to understand and react to basic small talk. You don't need to be an NLP or programming expert to have small talk with your chatbot. Simply familiarise yourself with one of the many

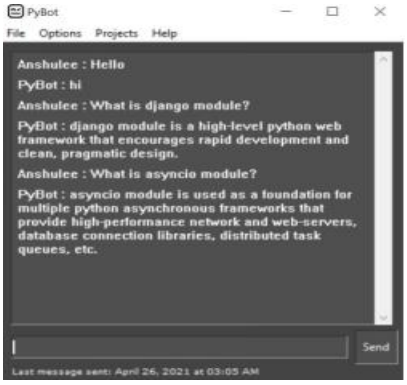
excellent third-party alternatives available, such as Api.ai, Wit.ai, Lex, and so on. Out of the box, these technologies include fast small talk solutions. For example, if a user asks, "What can you do?" you can quickly catch the question's intent and respond appropriately using third- party APIs.

Keep the conversation rationale loose: Any chatbot relies on conversation states, so good job if you're heading in that direction. However, do not build your flow logic in such a way that you require exact and rigid user responses, as this can cause your bot to fail miserably. Alternately, loosen the logic and take into account the possibility that users will choose to deviate from the flow you've developed. To do so, you must reverse the internal logic.

Redirect various intentions to what it knows: Finally, I'd like to consider ambiguous motivations. Unknown intents are messages that the chatbot has not acknowledged or is not expected to respond to. I discovered that between 70% and 80% of all customer feedback was based on ambiguous motivations. I may write multiple blog posts on how to improve the bots' reasoning in this case, but for now, I'll focus on one: redirecting the consumer to what your bot can understand.

Result





A pybot has been developed that answers all the python queries along with voice assistance. From the file option that chat can be cleared and we can exit the pybot. In options menu we can change the font and theme of the pybot. There are lot of fonts and themes present in the options menu. From the project section we can get the codes of different projects based on their levels, it can be Beginner, Intermediate and Advanced. From the menu users can get the information about the developers and the process to reach out to them.

Conclusion

Like Pybot we can built different bots for different scripting and programming language. In AI chatbots, a combination of natural language processing and interpretation (NLP and NLU) is already found. The development of natural language (NLG) enables bots to respond to the user in a similar tone of voice to create a relationship. Apps such as attitude monitoring and a deeper view of the individual can also help deliver an optimized experience for them. Persistent bots can recall past debates and will equate them with alternatives and user choices for feedback. Chatbots continued by sticking to a template, mostly replicating an internal FAQ support list that provided questions and easier responses. Modern bots can help understand the customer, defining keywords that lead the conversation in their reviews. Current and potential bots correctly understand the entire syntax of a user's feedback and can provide each part of a complicated query with precise and appropriate facts. Person chats and the general trend in all chats may be tracked through sentiment analysis to better inform an organization about how clients feel about what the business can do to maximize or improve positivity.

References

[1]F. A. Mikic, J. C. Burguillo, M. Llamas, D. A. Rodríguez, and E. Rodríguez, "CHARLIE: An AIML-based Chatterbot which Works as an Interface among INES and Humans." pp. 1-6, 2009.

[2]V. Bhargava, and N. Maheshwari, “An Intelligent Speech Recognition System for Education System,” 2009.

[3]M. J. Pereira, and L. Coheur, “Just. Chat-a platform for processing information

to be used in chatbots,” 2013.

[4]A. S. Lokman, and J. M. Zain, “One-Match and All-Match Categories for Keywords Matching in Chatbot,” American Journal of Applied Sciences, vol. 7, no. 10, pp. 1406, 2010.

[5]A. S. Lokman, and J. M. Zain, "An architectural design of Virtual Dietitian (ViDi) for diabetic patients." pp. 408-411, 2009.

[6]A. S. Lokman, and J. M. Zain, “Extension and prerequisite: An algorithm to enable relations between responses in chatbot technology,” Journal of Computer Science, vol. 6, no. 10, pp. 1212, 2010.

[7]D. Vrajitoru, "Evolutionary sentence building for chatterbots." pp. 315- 321, 2003.

[8]A. M. Galvao, F. A. Barros, A. M. Neves, and G. L. Ramalho, "Personaaiml: An architecture developing chatterbots with personality." pp. 1266-1267, 2004.

[9]Abdul-Kader, S.A. and Woods, J.C., 2015. Survey on chatbot design techniques in speech conversation systems. International Journal of Advanced Computer Science and Applications, 6(7).

[10]Kowsher, M., Tithi, F.S., Alam, M.A., Huda, M.N., Moheuddin, M.M. and Rosul, M.G., 2019, May. Doly: Bengali Chatbot for Bengali Education. In 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT) (pp. 1-6). IEEE.

[11]SimonHoermann, Kathryn L McCabe, David N Milne, Rafael A Calvo1,“ Application of Synchronous TextBased Dialogue Systems in Mental Health Interventions: Systematic Review”, Journal of Medical Internet Research ,volume: 19 , issue 8 , August 2017.

[12]Divya, S., Indumathi, V., Ishwarya, S., Priyasankari, M. and Devi, S.K., 2018. A self-diagnosis medical chatbot using artificial intelligence. Journal of Web Development and Web Designing, 3(1), pp.1-7