Natural Language Processing Role in Medical Chatbot

Dr. Sankhya N Nayak[1], Mr. Devaraj F V[2], Mr. Mohan H G[3], Mr. Nandish M[4]

[1] *Dr. Sankhya N Nayak, Associate Professor, CSE, JNNCE Shimoga, Karnataka, India*
[2] *Mr. Devaraj F V, Assistant Professor, CSE, PESITM Shimoga, Karnataka, India*
[3] *Mr. Mohan H G, Assistant Professor, CSE, JNNCE Shimoga, Karnataka, India*
[4] *Mr. Nandish M, Assistant Professor, CSE, JNNCE Shimoga, Karnataka, India*

ABSTRACT

A Chatbot is considered a new way to converse between machines and humans. Hence this is called a Conversational Chatbot. Traditionally, it is a software that uses the question-answer type of conversation between each individual and the chatbot. Therefore, a chatbot allows the user to interact with it by asking questions and in the same way revert it with an answer just like human conversations. The most well-known chatbots in recent times are Alexa and Siri. Chatbots use the concept of Natural Language Processing (NLP) for conversation. Since the use of chatbots is increasing day by day with the increase in technology, Industries are investing in it making it a viable option for organizations. The improvement in Chatbots and the concept of NLP has started additional research as well as advancements in the upcoming years.

Keyword : - *NLP ,Natural Language Processing, Medical Chatbot and Machine Learning.*

1. Introduction

In this technologically advanced era, people are using computers for a wide variety of purposes, including in the medical field, banking, and research. With natural language processing (NLP), we can create chatbots that allow humans to communicate with machines using natural language. This can be used to help tourists learn about the history and culture of different places they visit. It can also be used to help people with disabilities communicate more easily with others. Some people with conditions like autism find it difficult to communicate with others, but they may be able to use a communication aid to help them express their thoughts and feelings.

The main objective of this paper is the use of NLP in chatbot as it is needed for the conversation. The usage of NLP is there since the beginning of chatbot and is still being used. Advantages of NLP:

1. NLP provides users to ask questions. Based on the question quick answers are generated.
2. NLP gives correct answers to question asked rather not give unwanted or unreliable answers.
3. NLP is designed to help user to communicate to user in their language.

2. Related Work

JSON full form is java script object Notation is an export or import data file format. It is simple for humans to write and read as well as it helps machines to parse and generate which makes the JSON file versatile. Hence using of JSON in Python in extensive and also helps to measure the cost operation on the model [1].

NLP is defined as Natural Language Processing which is being used more in recent days. It predominantly deals with interaction with machine and human. The main parts of NLP are natural language understanding and language processing. [2] Before the process of NLP, the data should be cleaned to have a meaningful sentences or words. The

data might be having missing values, duplicate values, etc. These types of data should be handled by cleaning the data to make the data meaningful before proceeding to the next step that is pre- processing.

The two main NLP techniques used in this concept structuring the text, tokenization and lemmatization. Once we have structured data, the sentences are too tokenized. Tokenization is a major part of NLP technique where the sentences are broken down into words or expressions.[3] Lemmatization is the process of identifying the base root word from the complete word. This helps in identifying the correct word which will used in the process of training the model. Lemmatization identifies the correct base word from the complete word which being the reason it will be used predominantly in pre-processing over Stemming. [3]

2.1 Techniques used in Data Pre-processing

The various studies have shown that there are different pre- processing techniques which can be used. The techniques are given in the Table 1. [3]

Table -1: Different pre- processing techniques

No	Technique Name	Description
1	Lower text	By lowering the text, it will be treated as same word as well as it reduces the number of words in dictionary [4].
2	Removal of Unicode Characters	The removal of Unicode characters will reduce the unwanted characters which will help in reducing the text.
3	Removal of URL	This is done to remove any kind of URL present in the text vocabulary.
4	Removal of punctuation	Removing punctuation and the alphanumeric characters.
5	Removing whitespace	Unwanted white spaces is not required in the text. Therefore, whitespace should be removed. This will reduce the Text.
6	Tokenization	The sentence is separated to words separated by comma.
7	Stop word removal	The words like “a”,” an”, “the” are considered to be stop words as they don’t give any meaning to the sentence hence, they should be removed [5].

2.2 Predominantly Used Pre-Processing Technique

Table -2: Predominantly Used Pre-Processing Technique

No	Technique Name	Description
1	Tokenization	The sentence is separated to words separated by comma.
2	Lemmatization	Lemmatization is the process of identifying the base root word from the complete word.
3	Stemming	Removing of suffix from the word. Remaining is the root word.

2.3 Components of NLP

1. Natural Language Understanding (NLU)
2. Natural Language Generation (NLG)

Natural Language Understanding (NLU) is a type of Artificial Intelligence that understands sentences using text or speech. NLU uses algorithms to analyze data to form structured ontology. NLU helps extracting the information from content like ideas, entities, keywords, emotion, and relations.

Natural Language Generation (NLG) is software process that automatically transforms structured data to human readable text. It principally involves Text designing, Sentencedesigning, and Text Realization.

2.4 Relevance and Significance

In general, AI aims to make computers smarter, listening, talking, and understanding devices. Create new things, solve problems, and research above the method .NLU is relevant to many aspects of AI. The NLU's responsibilities include reading, interpreting, and categorizing. This requirement requires the creation of material systems that can answer questions after reading a document. It requires human- like language in the paragraph or document. Machine reading comprehension abilities. It can also be employed in virtual assistants so that after reading, a user can addition to assisting with documents, these assistants can also answer customer questions. Additionally, it can be used at work to read emails, process large business papers, and summarize pertinent information. In-home automation also utilizes voice-activated assistants to communicate with various appliances in meaningful ways.

3. Methodology

In this paper we discuss about NLP main steps used in this project that is Tokenization and Lemmatization. As in Related works of Table 1 we mentioned about tokenization which is breaking down the sentences into words. Now Let's get a deeper dive into how these tokens are being generated. From the below Fig1, we can understand that Tokenization of an NLP text are broken down to words.

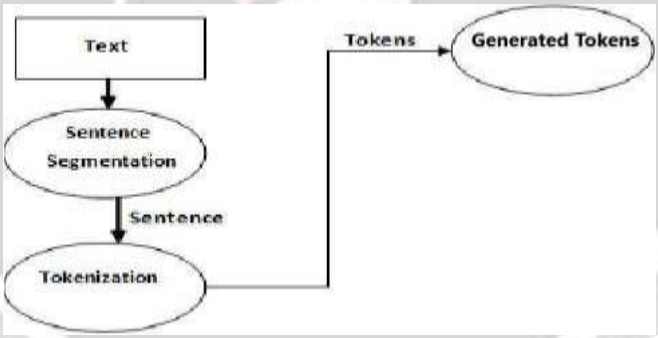

Fig-1. Tokenization.

First, the NLP text into given into Sentence Segmentation. This means that the text is broken down into sentences. This is done using morphological analysis on NLP text [6]. For Example, Chatbot is a software that can help people for conversing which is automated. Chatbot can converse via text to speech or text to text with the user. In Text Segmentation these are done by segmenting each and every sentence [6][7]. Hence the result being 1. Chatbot is a software that can help people for conversing which is automated. 2. Chatbot can converse via text to speech or text to text with the user. From these sentences we perform Tokenization of each sentence which means sentence is broken down into words called as tokens. From the above Chatbot/can/converse/via/text to speech/or/text to text/with the user. Etc. which will be converted into lower text for purpose of uniformity. For Tokenization there is an inbuilt function in python given in the library nltk called as word tokenize () which will tokenize each sentence to words [7][8].

Lemmatization is the process of grouping words together by their type. This is similar to stemming, but lemmatization takes into account the context of the words. This allows different words with similar meanings to be linked together as one word. For example, the word "walk" can be lemmatized as a verb or a noun. When it is lemmatized as a verb, it becomes "walks". When it is lemmatized as a noun, it becomes "walk".

Text pre-processing includes Stemming and Lemmatization. Lemmatization is a more accurate technique than stemming because it does morphological analysis. The output of Lemmatization is called a 'lemma', which is the root word. From the below Fig 2, we see that by using Lemmatization as technique we get the root word of 'playing', 'plays', 'played' as 'play'. This gives us the full meaning of the word, unlike stemming. Fig 2: Pre-processing technique- Lemmatization

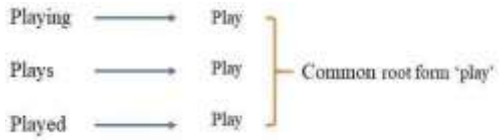

Fig-2. Lemmatization

The word "play" has the following definition: to engage in activity for enjoyment and recreation. This is the definition that we would see if we did not use any pre-processing techniques. By using Lemmatization as our pre-processing technique, we are able to see the full meaning of the word "play".

A vital part of natural language processing (NLP) is stemming, the method of manufacturing morphological variants of a root/base word. Stemming programs are called stemming algorithms or stemmers. For example, the words "chocolates", "chocolatey" are reduced to base word "Choco". Similarly, "retrieval", "retrieved", "retrieves" is reduced to the stem "retrieve". Hence stemming is necessary to reduce words to their roots for further analysis. The Tokenized words are the input to the stemming process. There are many algorithms for stemming. The most popular algorithms are the Porter Stemmer and the Lancaster Stemmer. We looked at a few different natural language processing techniques and how they could be used in chatbots. Lemmatization and tokenization are two of the most common techniques used to classify text into unique words. These words are then fed into the model used for training the chatbot. The bot's natural language processing algorithm can also be used to determine the sentiment of text. This can help the chatbot respond in a way that is appropriate for the sentiment of the text. For example, if a user's text is angry, the chatbot could respond with an apology.

Finally, the chatbot's natural language processing algorithm can also be used to determine the intent of the text. This can help the chatbot respond in a way that is appropriate for the intent of the text. For example, if a user's text is about making a reservation, the chatbot could respond with a message about making a reservation.

3.1. Errors in Stemming

Over-stemming and under-stemming are two phenomena that can often occur when stemming words. Over-stemming is when two words are stemmed from the same root, but have different stems. This can often lead to false-positives or instances where the wrong word is chosen because it is incorrectly stemmed. Under-stemming is when two words are stemmed from the same root, but have the same stem. This can often lead to false-negatives or instances where the correct word is not chosen because it is incorrectly stemmed.

3.2. Application of Stemming

1. Stemming in NLP are used in Search Engines.
2. Used in finding domain vocabularies in domain analyses.

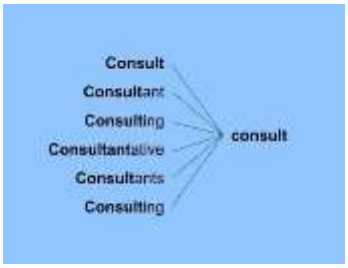

Fig-3. Stemming

Stemming reduces them to a common root word. Unlike lemmatization, stemming does not involve lexical operation. It isn't even needed that the stem be a legitimate word or clone of its morphological root. The objective is to reduce the given words into same stem.

4. CONCLUSION

From this paper we came know about the different techniques used in the Natural Language Processing Techniques. Hence in chatbot we use predominantly Lemmatization and Tokenization techniques for classifying the text into unique words. Therefore, those words are further fed to the model used for training the chatbot.

5. REFERENCES

[1] Aayush Goyal, Curtis Dyreson,"Temporal JSON", 2019 IEEE 5th International Conference on Collaboration and Internet Computing (CIC)

[2] Reshma E U and Ramya P C,"A REVIEW OF DIFFERENT APPROACHES IN NATURAL LANGUAGE INTERFACES TO DATABASES", Proceedings of the International Conference on Intelligent Sustainable Systems (ICISS 2017).

[3] Yasir Ali Solangi, Zulfiqar Ali Solangi, Samreen Aarain, Amna Abro, Ghulam Ali Mallah, Asadullah Shah3,"Review on Natural Language Processing (NLP) and Its Toolkits for Opinion Mining and Sentiment Analysis",2018 IEEE 5th International Conference on Engineering Technologies and Applied Sciences, 22-23 Nov 2018, Bangkok Thailand.

[4] L. Batista and L. Alexandre,"Text Pre-processing for Lossless Compression", Data Compression Conference, Dhaka, Bangladesh, 2008.

[5] B. Savaliya and C. Philip, "Email fraud detection by identifying email sender", 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), 2017.

[6] Jing Li, Billy Chiu, Shuo Shang and Ling Shao Senior Member, IEEE,"TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING", VOL. XX, NO. XX, 2019.

[7] Eman Btoush and Mustafa Hammad,"Generating ER Diagrams from Requirement Specifications Based On Natural Language Processing",April2015.

[8] Abhinav Nagpal and Goldie Gabrani,"Python for Data Analytics, Scientific and Technical Applications",2019 IEEE.