IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol. 10, Issue.3, March 2022 E-ISSN: 2347-2693

Modelling a Conversational AI Chatbot for Academic Websites

T. Bhaskar[1*], Y.A. Dive[2], A.J. Gujarathi[3], S.A. Gangurde[4], N.D. Rajput[5]

[1,2,3,4,5]Department of Computer Engineering, Sanjivani College of Engineering, Savitribai Phule Pune University, Kopargaon, India

[*]*Corresponding Author: bhaskarcomp@sanjivani.org.in, Mob.:+91-9766466079*

DOI: https://doi.org/10.26438/ijcse/v10i3.812 | Available online at: www.ijcseonline.org

Received: 19/Feb/2022, Accepted: 14/Mar/2022, Published: 31/Mar/2022

***Abstract*—** This project aims in developing a conversational chatbot for college universities, which can converse in any desired language. The main purpose of the planned system is to develop a chatbot which can intelligently answer all the queries related to universities. For this purpose, machine learning algorithms are used. Primarily, NLP (Natural Language Processing) is used for the interaction between the chatbot and the user. To answer the users query, many different machine learning algorithms can be used. For this purpose, we move to cosine similarity. It, in a way, helps the chatbot to search for the answer to the user's query. Also, various libraries are used for different purposes like accepting the audio input, converting speech to text, translating to desired language, etc. The pandemic has affected a lot of fields around the world. One of them is the educational field. Many students in the urban as well as the rural parts of India were not able to visit the universities to get information about academics. This chatbot would eliminate these worries. All the information needed can be accessed online.

***Keywords*—**Chatbot, Machine learning, Naive bayes classifier, NLP, Cosine similarity

I. INTRODUCTION

Introduction Today, we live in a pandemic world. People all over the world are not able to move or travel from one place to another. In these hard times obtaining information about anything, especially information related to educational institutes, has become difficult as we cannot travel to the institute and seek information. During the admissions process or for their regular needs, students are unable to visit schools and universities to get different information, such as tuition rates and term schedules. Thus, a chatbot system for universities and colleges has been developed in order to address these issues.

This chatbot would be able to provide different information related to college and its other departments. The chatbot can be accessed at any time from the college website by multiple users. In this project, we will implement the chatbot with the help of machine learning and artificial intelligence. This chatbot will be able to take input in multiple languages and provide output in desired language.

The further paper is assembled as follows, Section II contains related work, Section III contains the design of the experiment, Section IV contain the brief explaination of proposed system , Section V explaination the methodology, Section VI explains in brief the algorithms, Section VII describes dataset and results, Section VII states the applicatios and Section IX concludes research work with future directions.

II. RELATED WORK

The traditional chatbot takes the users query in one particular language. The chatbot used by academic universities needed to be changed in such a way that every type of person, technical or non-technical, would be able to use it. The accuracy of the chatbot was one other factor to be considered. The answer provided by the chatbot had to be such that it solved the users query perfectly. If the chatbot didn't have the answer to a new query, it should be able to add the query to its database, so that the next time this query pops, it answers accordingly to it.

III. DESIGN OF EXPERIMENT

Various blocks are present in the architecture; following is the brief description of it:

Intent Classifier

Natural Language Processing(NLP) contains one branch known as intent classifier(NLP). The purpose of intent classifier is to extract intent, that is, purpose and goal from a text. For example, you are trying to classify emails received by a shopping website. One of the email says, "There is a problem with making a purchase on the website, and I'd want your assistance?". So, this mail would be classified to "interested" intent. With the help of intent classifier you can easily classify the email/query, among thousands of mails received on a daily basis, to the respective intent. The ultimate goal of the intent classifier is to identify the exact purpose of the interaction of customers. It may be orders intent, unsubscribe intent or

payment intent. Classifying the intent increases the chances of resolving the issue or closing a sale with the customer. That's why NLU's "intent classifier" (Natural Language Understanding) component is so important. (E.g. “Good Morning. How are you?” Intent-Greeting. “When was the institute established?” Intent-institute info)

Natural Language Tool Kit (NLTK)

The Natural Language Tool kit is also called as NLTK. NLTK was created by Steven Bird and Edward Loper at the University of Pennsylvania's Department of Computer and Information Science. It is a collection of Python-based tools and applications for symbolic and statistical processing of natural language of English. NLTK not just contains the toolkit, but also a text which explains the basic ideas behind language processing tasks and also a recipe book. Research and education in NLP and related disciplines, such as empirical linguistics and cognitive science, artificial intelligence, information retrieval and machine learning are supported by the NLTK. Since its formation, the NLTK has served as a teaching as well as learning tool, and a platform for modelling and developing research systems. A total of 32 colleges and institutions in the United States and 25 other countries are utilising NLTK. Semantic reasoning and tokenization are all supported by NLTK's classifiers and tokenizers.

Dialogue Manager

After understanding what the user actually meant, the chatbot needs to correctly and accurately resolve the query by giving correct reply. The correct answer to the query will depend on what is asked and what the chatbot knows. The databse of the chatbot refers to knowledge the chatbot has.

Rule based Approach

Most of the chatbot work in this way where they steer the user through several questions. Every time the user selects some question they move to the next question in the selected direction.

Input

The input will be accepted in voice or text format. If its voice it will be store in .wav or .mp3 format and converted to text using text-to-speech. If the input is in language other than English then it will be converted to English text using language translation.

Output

The chatbot will answer the query in either text form or audio form. The audio will be played on the screen and the text output will be displayed on screen.

IV. PROPOSED SYSTEM AND ALGORITHM

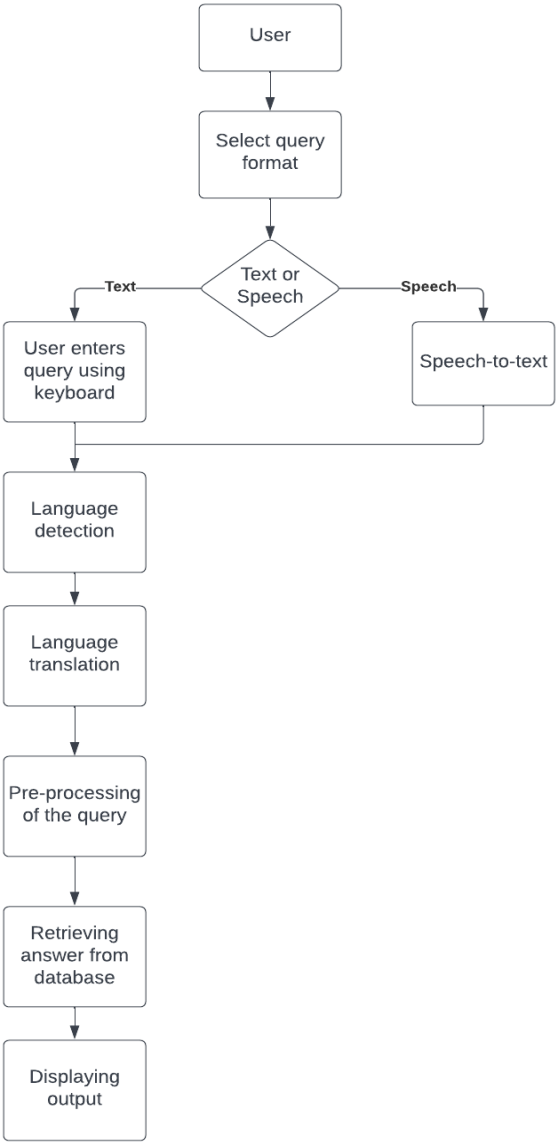

Figure 1. System Architechture

Proposed Algorithm

Step 1: Open chatbot, ask users to enter questions in text or speech.

Step 2: If the question is in speech, convert it into text.

Step 3: Detect the language of the text (eg- English, Marathi) and store the language name in a variable ‘lang’.

Step 4: If the language is not English, convert it into English.

Step 5: Pre-processing of the text into tokens using tokenization.

Step 6: Apply Cosine Similarity Algorithm on the text to find the best match in the question field.

Step 7: Store the answer to the question in variable ‘answer’.

Step 8: Display the answer to user

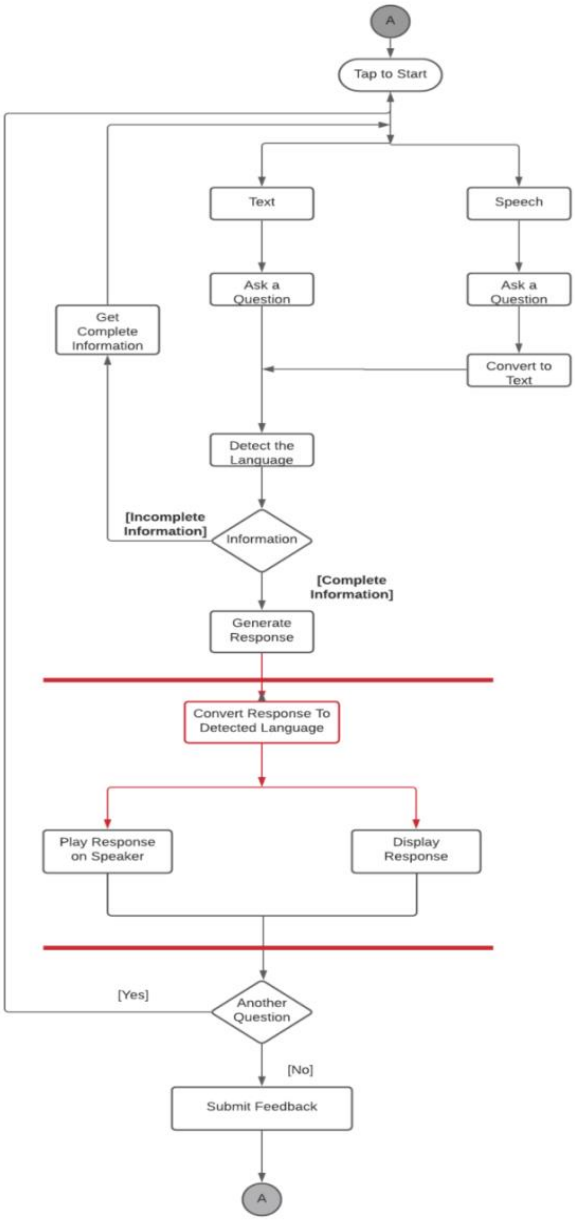

Figure 2. Activity Diagram

V. METHODOLOGY

Dataset

In the field of Machine Learning, we believe this is the most important phase in the system. We've done all we can to make the data as simple as possible, so that the model can be trained more quickly and more simply. Before feeding the model, all raw data should be cleansed. There isn't a single way to clean the data; instead, we may try a variety of approaches. Our dataset is particularly in the text or string format. The most important part of our dataset preparation is tokenizing the sentences into words.

Pre-processing

The dataset contains a lot of noise which may affect the accuracy. For this purpose we use some Natural Language Processing (NLP) techniques.

a. *Tokenization*

To make the chatbot clearly understand what the users query is, it needs to understand each word in the query. So, there is a need to convert the sentence into tokens of words. The process is tokenization. Phrases, words, syllables, and characters may all be used as tokens.

b. *Annotation*

It is simpler for the algorithm to detect words like "saw" as a verb (past tense of the word see) or noun (instrument for cutting wood) when they are annotated with their grammatical function, known as POS tagging.

c. *Filtering*

It is simpler for the algorithm to detect words like "saw" as a verb (past tense of the word see) or noun (instrument for cutting wood) when they are annotated with their grammatical function, known as POS tagging.

Dataset (Train | Test)

It is necessary to divide data into training and validation sets after the model is ready. When we partition the dataset into 80% of the training set and 20% of the testing set, we obtain this data structure.

Similarity Algorithm

Using Cosine Similarity, you may determine how similar two or more vectors are. To obtain the measurement of similarity between two documents, first the words or phrases present in a document or sentences are converted to a vectorised form. Then these vectorised representations of the documents are used along with the cosine similarity formula to obtain the measurement of similarity. If the cosine similarity of the two texts is 1, it suggests that they are identical. If the cosine similarity between two papers is zero, it signifies that there is no resemblance between the two texts. With cosine similarity we check the similarity between the user input and the dataset we have and then return the corresponding answer.

User Input

We will accept the input from the user through the microphone and store it in a .wav file. This voice data will be converted into text using Speech Recognition. If the text is not in English then it will be translated so that the chatbot can understand.

VI. ALGORITHM

Natural Language processing (NLP)

Using artificial intelligence or NLP, a chatbot may respond to a user's inquiries in a natural language. For computers, natural language processing allows them to take in information, scatter it and extract its meaning before deciding what to do with it and responding to the user in their own language. Both Natural Language Understanding and Natural Language Generation are important components of natural language processing (NLP) (NLG). Informal data is sent into the NLU, which organises it into structured data that the machine can deal with. The goal of the NLU is to understand what the user is trying to communicate. As its name suggests, Natural Language

Generation (NLG) is a tool for transforming structured data into natural language. NLP is a five-step technique for processing natural language. First, unstructured data is converted to text for analysis. An investigation of the grammatical structure of words is conducted. Tokens are used to segregate all of the text being entered. As a result, tokens are subsequently passed to syntax analysis, where the link between words may be clearly seen. After that, the data is sent to the semantic analysis engine for processing. This phase removes the meaning of the words or tokens. The description is gleaned from the input by using the taskbar objects and the map's syntactic structure. The next stage is to construct the sentence from which the meaning of the new sentence may be gleaned. Step three is a pragmatic evaluation. Many of the findings in this study are based on what was said and how it was interpreted. Finally, the user's inquiry is answered when all these processes are completed by a computer. Natural Language Generation (NLG) then produces the final response based on this answer. Most chatbots employ Natural Language Processing (NLP) to converse successfully with their users.

Naive Bayes Algorithm

Naive Bayes, a popular algorithm for chatbots, is another option. Token creation is the initial phase in this process. Tokens are created by dividing a text into tokens. Then, each token is stemmed. As an example, the phrase "have a terrific day" is marked and tagged "have," "a," "great," and "day," among other things. Training data must be provided next. Lists or dictionaries are often used to hold this information. Class and phrases are adjectives in this vocabulary. As an example, the word 'greeting' might be used in the statement above. Afterwards, each class's words are compiled into a list. Input sentences are checked and tokens are compared to determine which classes are most likely to include an input sentence. Each phrase that we offer as an input may be assigned to more than one class, and the scores for each class will be added together. Then out of two classes or more one is chosen with the highest score. In this way, this algorithm works. If the data supplied as input does not fall into one of these predetermined categories, predicting the output becomes more difficult.

Cosine Similarity

An inner product space's cosine similarity is calculated by comparing two vectors. An angle between two vectors is used to detect whether they are heading in the same general direction. In text analysis, it's often used to gauge how similar two documents are. When building a chatbot, we evaluate cosine similarity to see whether an input from the user is comparable to a query already in our database. If it is available then the answer corresponding to that question is given as output. If multiple similar questions are available in the database then the best question is selected on the basis of accuracy.

$$Cosine\ Similarity\ (CS) = (A\ .\ B)/\ (\|\ A\ \|\ \|\ B\ \|)$$

Here, A and B are the two vectors of two different documents.

VII. DATASET AND RESULTS

Dataset Details

In the dataset, there are a total of 6 intents which are general info, admission, placement, bot, facilities, dept info(department information). It consists of 407 questions along with their intent. When 0.2 test size was considered using cross-validation test-train split 325 records were trained and 82 records were tested. Similarly, when 0.15 test size was considered 70 records were tested based on training on 337 records.

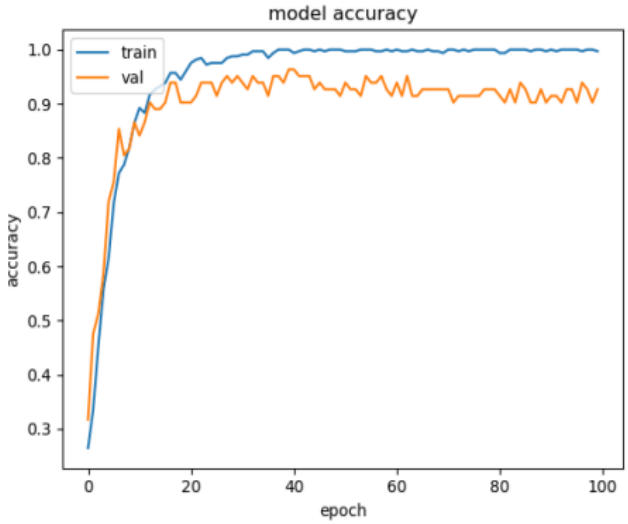

Figure 3. Model Accuracy.

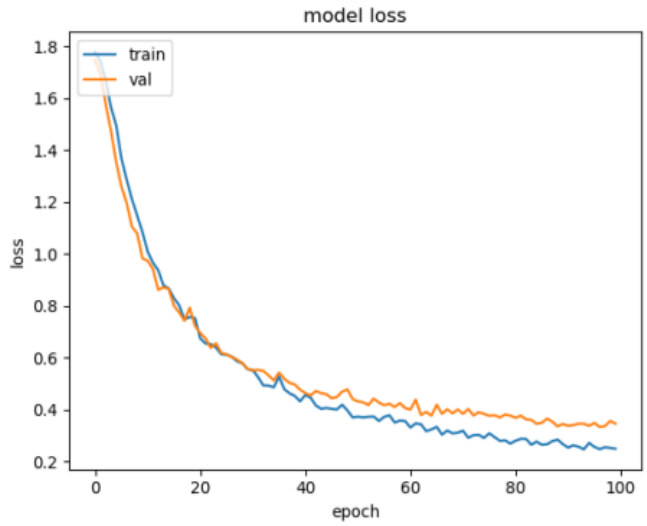

Figure 4. Model Loss

Results

- The model gives 99.69% accuracy on training dataset while 92.68% accuracy on validation dataset. This suggests that the model correctly predicted 92.68% of the total predictions.
- The model gives 25.03% loss on training dataset while 34.07% loss on validation dataset. On a single query, the percentage of bad prediction was 34.07%.

VIII. APPLICATIONS

- The chatbot will help in accessing any information related to universities and academics.
- The chatbot will eliminate the need to visit colleges now and then to inquire.
- The chatbot will be able to accept input in text and audio manner.
- The chatbot will be able to take input in any desired language.
- There will be no user limit or time restriction for the chatbot. It will be able to answer 24/7.

IX. CONCLUSION AND FUTURE SCOPE

A chatbot is AI software that simulates a conversation with the user. This academic chatbot works with goal of answering the user's query related to the university or curriculum. This conversational chatbot will be based on Artificial Intelligence and assist user to get information regarding college or university. With more data in multiple languages and hyper parameter tuning the accuracy of intent classifier can be improved. A robot can also be developed to be stationed at different places in the college and assist users. The chabot can be updated to fully communicate using speech.

ACKNOWLEDGEMENT

An AI chatbot for its many applications have long been a focus of intense study because of the enormous potential for new discoveries in the discipline. We would like to thank Dr. T. Bhaskar, (Associate Prof., Computer department) our valued mentor, for his attention and advice during this research. He has also offered opportunities for us to further our knowledge and understanding of the subject matter. We will always remember this event and use it as a motivator to accomplish our job to the best of our abilities. In addition, we would like to thank Dr. D. B. Kshirsagar (H.O.D. Computer Department). We would like to thank the whole faculty and staff of Sanjivani College of Engineering, Kopargaon's Department of Computer Engineering for their assistance and support.

REFERENCES

[1] Prof. Darshan A. Patel, Neelkumar P. Patel1, "AI and Web-Based Human-Like Interactive University Chatbot (UNIBOT)", IEEE Xplore, **pp. 309-315, 2020.**

[2] A K M Shahariar Azad Rabby, Md. Majedul Islam, "Language Detection using Convolutional Neural Network", IEEE Xplore, **pp. 1-5,2020.**

[3] A. Ansari, M. Maknojia and A. Shaikh, "Intelligent question answering system based on ArtificialNeural Network," IEEE International Conference on Engineering and Technology (ICETECH), **pp. 1793-1805, 2017**.

[4] B. R. Ranoliya, N. Raghuwanshi and S. Singh," Chatbot for university related FAQs," 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), **pp. 1-6, 2018.**

AUTHORS PROFILE

Dr. T. Bhaskar, reiceved a PhD(CSE) from SSSUTMS,Bhopal, M. Tech (CSE) from JNTU Hyderabad. He is Currently working as Associate. Professor in Computer Engineering Department, Sanjivani College of Engineering, Kopargaon and Maharashtra India. His research interest includes Machine Learning & Data Science.He has published/presented 30+ Papers in various international journals/Conferences also has 2 patents.

*Yash A. Dive***,** pursuing his Bachelors Degree in Sanjivani College of Engineering, Kopargaon. His research interest includes statistical Learning, data analytics and software development.

*Atharva J. Gujarathi***,** pursuing his Bachelors Degree in Sanjivani College of Engineering, Kopargaon. His research interest includes software development, data science and machine learning.

*Sanket A. Gangurde***,** pursuing his Bachelors Degree in Sanjivani College of Engineering, Kopargaon. His research interest includes software development, cloiud computing and machine learning.

*Nikita D. Rajput***,** pursuing her Bachelors Degree in Sanjivani College of Engineering, Kopargaon. Her research interest includes software development and machine learning.

