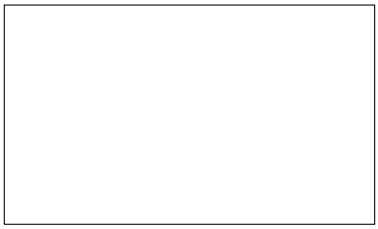
Graphical Abstract

University Chatbot System using NLP

Abhigya Verma, Chandana Kuntala, Pragya Khatri, Sristi, Sukhmani Kaur, A K Mohapatra, Shweta Singhal



Highlights

University Chatbot System using NLP

Abhigya Verma, Chandana Kuntala, Pragya Khatri, Sristi, Sukhmani Kaur, A K Mohapatra, Shweta Singhal

- This paper aims at experimenting with chatbots made from various similar metrics to check their performance on a small dataset.
- The key focus is to understand whether this approach can be used to build basic chatbots.
- Another key research scope is to identify the best metric out of the commonly used ones for the chatbot.

University Chatbot System using NLP

Abhigya Verma[a], Chandana Kuntala[a], Pragya Khatri[a], Sristi[a], Sukhmani Kaur[a], A K Mohapatra[a], Shweta Singhal[a]

[a]*Department of Information Technology, Indira Gandhi Delhi Technical University for Women, IGDTUW, Kashmere Gate, New Delhi, 110006, New Delhi, India*

Abstract

Artificial intelligence-based chatbots use programmed software instructions to mimic human speech (AI). They are made for a variety of uses, including operating electrical equipment in a smart home, serving as a personal virtual assistant, providing entertainment, responding to commonly asked queries, obtaining driving directions, etc. Due to its ability to manage several customers at once and help cut down on customer care costs, chatbots have gained popularity in the corporate sector. To make the chatbots as effective as possible, numerous activities still need to be completed. We designed a college FAQ chatbot to address this issue in the academic environment. This chatbot uses Machine Learning (ML) and Natural Language Processing (NLP) to answer questions about a particular institution in an effective and precise manner.

Keywords: NLP, Machine Learning, Chatbot

1. Introduction

Nowadays, practically all industries, including those in the domains of health, banking, e-commerce, and education, use technology and computers in their day-to-day operations. Conversational Interfaces were developed to facilitate human-computer interaction and slowly they have tried to make

Email addresses: abhigyaverma27@gmail.com (Abhigya Verma), chandanakuntala21@gmail.com (Chandana Kuntala), pragyakhatri2002@gmail.com (Pragya Khatri), sristi0108@gmail.com (Sristi), sukhmani078@gmail.com (Sukhmani Kaur), akmohapatra@igdtuw.ac.in (A K Mohapatra), miss.shweta.singhal@gmail.com (Shweta Singhal)

those interfaces smarter to improve human communication with computers. According to ([1]) Platforms known as Conversational User Interfaces (CUI) or Conversational Agents "allow individuals to connect with smart gadgets using spoken language in a natural way" by initiating conversations with humans.

([2]) identifies voice assistants and chatbots as two different categories of conversational agents. The former enables voice commands to be used to interact with the user interface; examples of popular voice assistants are Apple's Siri and Google Voice Assistant. The latter, known as chatbots, are web- or mobile-based user interfaces that provide information retrieval and question-and-answer functionality. Chatbots come in a variety of varieties. Some are created for e-commerce, information retrieval, commercial, or educational objectives according to ([1]).

A chatbot is an automated response system that responds to enquiries in a natural, conversational manner.([3]). By offering a powerful natural language interface, chatbots allow users to execute activities without having to directly contact human agents via phone calls or emails [4]. Chatbots are employed in a variety of industries, such as e-commerce, education, and therapy since they often use straightforward questions and commands. [5] to replace traditional customer service. [6] In this study, we investigate whether chatbots can save labour expenses in administrative offices by responding to students' frequently asked questions. Artificial intelligence-based chatbots may speak with users like real people and offer ideas depending on what the user says throughout the discussion. Numerous chatbots were created to allow consumers to converse with computers. These chatbot services are now accessible on Facebook Messenger, Telegram, LINE, Kik, and Viber, and they will soon be available on Twitter and Google Assistant.[7]

Chatbots have a lot of potential for usage in university administration functions. In order to increase the effectiveness of administrative officers' job and the quality of the student experience, several institutions have adopted the usage of chatbots to answer inquiries from students [7]. Every day, the administrative staff at the institution gets a number of emails, calls, and messages. The job of these assistants may be impacted by having to answer every inquiry, which is typically repetitious. Instead than spending their time answering the same questions repeatedly, administrative officers may focus on

other important activities by automating frequently asked questions (FAQ). [7] We looked at chatbot effectiveness in university administrative services and whether or not they would lighten employee workload.

We believe a chatbot can make it simple for students to locate information, making it very useful for them as they navigate campus life. Therefore we have implemented a chatbot using NLP techniques involving basic distance metrics to measure the similarity between questions.

2. Literature Review

According to ([8]), natural language processing is a system that identifies human-natural languages using artificial intelligence. It can transform human language into a form that can be understood by computers.

The need for a chatbot to cater to queries of students is eminent, studies of ([9]) show that even though the students require help; they tend not to approach for help. These authors found that in order to maintain a positive social image, college students may avoid seeking help online. As a result, students tend to look for FAQs to find answers and help in their own way. FAQ contains repeated questions together with their answers which were first used in Usenet groups in 1982.

A chatbot is a computer program that can converse with humans via text or voice. As per ([10]), a bot can basically be of two types: (i) rule-based, in which there are certain principles already created. The bot can respond only according to the principles and its intelligence is confined to the specified rules. This bot is incapable of responding to queries that are not already mentioned. (ii) Based on Artificial Intelligence (AI) techniques, which keeps on learning based on the interactions.

The first, most famous, chatbot was created in 1964 and was named ELIZA by ([11]), which can analyze the question entered by the user and then provide the required response by decomposing the input and applying certain rules.

([12]) have developed a curious chatbot that finds the missing information in inquiries and tests the queries on clients to gather the required information to answer the question. For the purpose of providing an accurate response, they have identified and queried the missing data.

The team led by ([13]) studied the use of response and recommendation tools to answer students' questions in the forums of their online class on

Artificial Intelligence. The assistant has been named Jill Watson, based on the IBM Watson platform, perhaps best known for beating two Jeopardy champions. Jill was specifically developed to handle the large number 23 of forum posts by students enrolled in an online course that is a requirement to obtain a master's degree in science in the Georgia Tech computer program. Jill responds to FAQs in the forum. It checks for the latest updates every 15 minutes and responds to students. It basically deals with student introductions. These are teaching assistants, not FAQ bots.

([14]) have shown that the existing work on a Natural Language Processing Bot for a query-based response generation system exists for Frequently Asked Questions.

([15]) has reviewed the existing chatbots in education which are basicaly two types: FAQ bot and short response quiz. The FAQ bot discussed deals with questions that are related to the course and used for knowledge retention 25 and help the tutor. This FAQ bot doesn't cover administrative issues or other questions during the initial days of the student in the university.

([16]) the research found that a significant role of the Cosine Similarity metric is present in seeking relevant queries and their responses for the customer.

3. Methodology

We divide this section into two parts; first, we explain our steps to curate our dataset. Then, in the second part, we describe the working of our chatbot and the various similarity metrics used.

3.1. Data Collection

Our key data collection goal is to obtain any queries students have in mind related to the university, curriculum, or in general any questions that they would like the chatbot to be able to answer, and in this section, we describe our collection methodology. We created our own dataset i.e. a novel dataset of questions and answers with respect to our university. In order to do so, a survey form was circulated amongst students of the university to fill out questions/queries they had regarding any aspect of college. The form was distributed to every batch of each branch in the university. Students were given a week to fill out the survey and the results were stored in an excel sheet.

3.2. Dataset Preparation

We have collected the queries from students now the next step will be to prepare a dataset that contains both questions and answers that we can use to build our chatbot. The steps we followed for dataset preparation are explained below.

3.2.1. Data Appending

After collecting questions the next step was to curate appropriate and correct answers for the questions. Our team divided the total questions into groups and questions in each group were answered and typed manually by each individual member based on their own knowledge and taking help from seniors / professors in case of any doubts.

3.2.2. Data Cleaning

The last stage of data preparation is removing the rows containing blank values. Fortunately, neither the case study dataset include any missing values. Additionally, the columns with a standard deviation of 0 are eliminated because they are not essential to the data analysis and may hinder some analyses.
The stop words and punctuation's were removed before feeding the data to get the similarity scores. The stop words are generally eliminated in Natural Language Processing as they are so commonly used and carry very little useful information.

3.2.3. Tokenizing and Stemming

Tokenization is the process of breaking up a long block of text into tokens. The most frequent method of processing the raw text occurs at the token level because tokens are the fundamental units of Natural Language. By reducing word inflection to its root forms, the natural language processing approach known as stemming helps to prepare text, words, and documents for text normalisation.

3.2.4. Data Vectorisation

We use the CountVectorizer() method on the list of questions in our database so that we can apply various similarity metrics to find the most

S.No	Before Cleaning	After Cleaning
1	do college grades matter in placement	college grades matter placement
2	How many societies will i be able to manage in my first year?	How many societies able manage first year
3	and most importantly HOW will I know about the Scholarship opportunities , their applications and their deadlines	importantly HOW I know Scholarship opportunities applications deadlines
4	is the scholarship form open?	is the scholarship form open
5	can i ask seniors for their books?	ask seniors books

Table 1: Data after cleaning

S.No	Before Stemming	After Stemming
1	do college grades matter in placement	colleg grade matter placement
2	How many societies will i be able to manage in my first year?	how mani societi abl manag first year
3	and most importantly HOW will I know about the Scholarship opportunities , their applications and their deadlines	importantli how i know scholarship opportun applic deadlin
4	is the scholarship form open?	scholarship form open
5	can i ask seniors for their books?	ask senior book

Table 2: Data after stemming

similar items to our user input i.e. question asked by the user. Text is converted to vector using the CountVectorizer. Tokenizing the questions allows it to create a lexicon of the words found in the corpus and track how frequently each word appears in each and every question. A vector of the same size as the vocabulary is used to represent each question, and entries in the vector for a given question display the number of words in that question.

3.3. Answer Processing

We describe the working of our chatbot in this section. To give answers to the input question we find the similarity between the question asked by

user and the questions present in our database and return the answer corresponding to the most similar question. We use three similarity metrics and compare the results obtained from each of them:

Manhattan Distance:

Manhattan distance compares two points by calculating the distance between them as the sum of absolute differences of their cartesian coordinates. To put it in a simpler way it refers to the total sum of difference between the x and y coordinates. It cases where the dimension of data increases the Manhattan Distance is preferred over the Euclidean Distance.

$$d = \sum_{i=1}^{n} |x_i - y_i|$$

Euclidean Distance:

Euclidean distance is a similarity metric used to find how far two vectors are from each other. It merely calculates the length of straight line that connects the two vectors.

$$Euc(A, B) = \sqrt{\sum_{i=1}^{n} (A_i - B_i)^2}$$

Cosine Similarity:

The Euclidean Distance metric has a drawback that it is biased by the size difference in representations. Hence , cosine similarity metric may be used that determines similarity using the inner angle.

$Vector\ Norm = ||\overrightarrow{V}|| = \sqrt{\sum_{i=1}^{n} {v_i}^2}$

$DotProduct = \overrightarrow{V}.\overrightarrow{W} = ||\overrightarrow{V}||||\overrightarrow{W}||cos(\beta)$

$Therefore,$

$$cos(\beta) = \frac{\overrightarrow{V}.\overrightarrow{W}}{||\overrightarrow{V}||||\overrightarrow{W}||}$$

The cosine of angle between the two vectors is calculated to find similarity.

If the cosine value is 0 this means that the vectors have no similarity and are orthogonal to one another. A cosine value close to 1 indicates that the two values are more similar since the angle between them is less.

4. Results

We used three Distance Similarity Measures, namely, Cosine Similarity, Manhattan Similarity and Euclidean Similarity.

The results of all Similarity measures are given as follows:-

4.1. Cosine Similarity

In Table 3, we can see the Cosine Similarity Scores of the questions entered by us with the Questions and Answers from the Dataset.

S.No	Question	Answer	Similarity Score
1	Hi	Hey there, tell me how can I help you?	1.000000
2	Do college grades matter in placements?	Yes , your cgpa is important for placements	1.000000
3	Where is the library?	The library is located in the building directly behind the field. It is located on the first floor	0.707107
4	How to study for exams?	The syllabus given by college has recommended books , those books can be used to study the syllabus and resources given by your teachers can be used	0.408248
5	How strict is attendance?	75% attendance is mandatory	1.000000

Table 3: Cosine Similarity Scores

statistics.mean() was used to calculate the average of the Cosine Similarity Scores that are shown in Table 3 which came out to be equal to 0.831780264690994.

Figure 1: Cosine Similarity Scores from Figure 1 v/s Dataset Question Number

S.No	Question	Answer	Similarity Score
1	Hi	Hey there, tell me how can I help you?	0.000000
2	where is library located?	Don't hesitate to approach your seniors either through societies or LinkedIn to know about different opportunities. Focus on learning and improving your skill set, achievements will follow on their own. Participate and keep leaning towards a growth mindset.	22.000000
3	bye	Bye, Chat with you later	0.000000

Table 4: Manhattan Similarity Scores

4.2. Manhattan Similarity

In Table 4, we can see the Manhattan Similarity Scores of the questions entered by us with the Questions and Answers from the Dataset.

statistics.mean() was used to calculate the average of the Manhattan Similarity Scores that are shown in Table 4 which came out to be equal to 8.0000000.

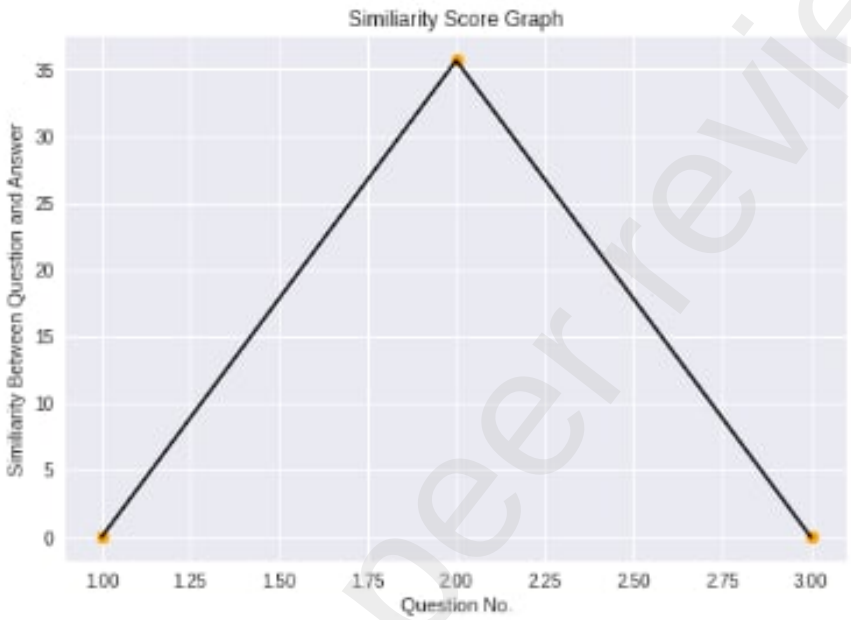


Figure 2: Manhattan Similarity Scores from Figure 3 v/s Dataset Question Number

4.3. Euclidean Similarity

In Table 5, we can see the Euclidean Similarity Scores of the questions entered by us with the Questions and Answers from the Dataset.

S.No	Question	Answer	Similarity Score
1	Hi	Hey there, tell me how can I help you?	0.000000
2	How many societies will i be able to manage in my first year?	That depends on how you manage your time but 1 to 3 should be fine	0.563576
3	How to prepare for college exams?	Make notes, be attentive, keep up with the class and revise the topics regularly.	0.457053
4	How to develop our skills?	You can explore different fields to expand your skill set. IGDTUW has really good societies like GDSC or LeanIN for students interested in Tech. In your first year apart from your DSA, you can learn web development, app development, AR-VR, ML-AI etc. Options are endless.	0.417482
5	How to be good at coding?	Making proper schedules and timetable is the key. Devote 2-4 hours to coding daily as per your capacity and the rest to college studies.	0.459884

Table 5: Euclidean Similarity Scores

statistics.mean() was used to calculate the average of the Euclidean Similarity Scores that are shown in Table 5 which came out to be equal to 3.2141863470119345.

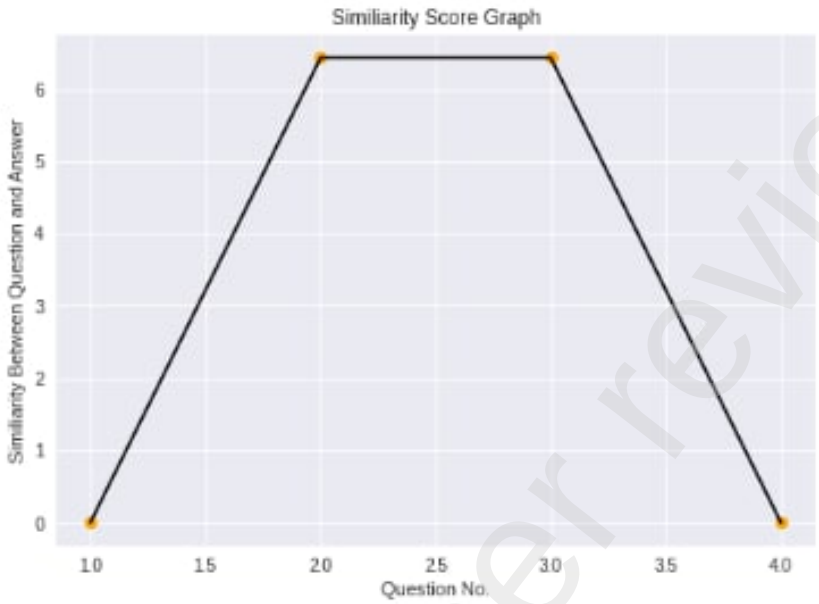


Figure 3: Euclidean Similarity Scores from Figure 5 v/s Dataset Question Number

5. Conclusion and Future Scope

The relationship of different distance similarity metrics with the performance of the chatbot was compared. The inference derived from the study pointed out that the Cosine Similarity metric showed comparatively higher accuracy for the set of test input questions ranging over topics similar to those present in the original dataset, over other distance similarity metrics like Manhattan Similarity and Euclidean Similarity with a score equal to 83.17%. The score obtained was low owing to the size and complexity of the dataset.

The result obtained can be improved further by implementing different machine learning and deep learning algorithms to effectively fit the data. This can enable the system to be intelligent enough to handle new queries and generate appropriate responses by itself, hence increasing the performance of the chatbot. The novel dataset can be further extended to include an even larger set of FAQs to make the chatbot responses robust.

References

[1] M. McTear, Z. Callejas, D. Griol, The conversational interface: Talking to smart devices (2016).

[2] B. B. R. G. S. D. Valtolina, S., P. Diliberto, Chatbots and conversational interfaces: Three domains of use (2018).

[3] S. Carayannopoulos, Using chatbots to aid transition (2018).

[4] R. Dale, The return of the chatbots (2016).

[5] I. Serban, A deep reinforcement learning chatbot (2017).

[6] A. E. Abu Shawar, B.A., Chatbots: are they really useful? (2007).

[7] K. Lee, J. Jo, J. Kim, Y. Kang, Can chatbots help reduce the workload of administrative officers? - implementing and deploying faq chatbot service in a university (2019) 348–354.

[8] S. A. Abdul-Kader, J. Woods, Question answer system for online feedable new born chatbot, Intelligent Systems Conference (IntelliSys) (2017) 863–869.

[9] K.-T. Er, E., M. Orey, Exploring college students' online helpseeking behavior in a flipped classroom with a web-based help-seeking tool, Australasian Journal of Educational Technology 3 (2015) 537–555.

[10] M. Schlicht, The complete beginners guide to chatbots (2019).

[11] J. Weizenbaum, Eliza, Communications of the ACM (1966).

[12] S. Reshmi, K. Balakrishnan, Implementation of an inquisitive chatbot for database supported knowledge bases, Sadhan 41 (2016) 1173–1178.

[13] . P. L. Goel, A. K., Jill watson: A virtual teaching assistant for online education, Technical report, Georgia Institute of Technology (2016).

[14] N. R. B. R. Ranoliya, S. Singh, Chatbot for university related faqs, International Conference on Advances in Computing, Communications and Informatics (ICACCI) (2017) 1525–1530.

[15] B. W. W. T. L. . M. E. Cunningham-Nelson, Samuel, A review of chatbots in education: Practical steps forward, Australasian Association for Engineering Educationn (2019).

[16] L. Hidayatin, F. Rahutomo, Query expansion evaluation for chatbot application, International Conference on Applied Information Technology and Innovation (ICAITI) (2018).