

# **DIUbot – A CHATBOT FOR DIU ADMISSION SECTION**

**By**

**MD. MASUM REZA**

**ID: 161-15-6894**

**RUHULLAH BIN KALIM**

**ID: 161-15-7092**

**MD. ABDULLAH AL MUID**

**ID: 161-15-6999**

This Report Presented in Partial Fulfilment of the Requirements for the Degree of  
Bachelor of Science in Computer Science and Engineering

Supervised By

**Md. Sadekur Rahman**

Assistant Professor

Department of CSE

Daffodil International University

Co-Supervised By

**Md. Tarek Habib**

Assistant Professor

Department of CSE

Daffodil International University



**DAFFODIL INTERNATIONAL UNIVERSITY**


**DHAKA, BANGLADESH**

**DECEMBER, 2019**

## APPROVAL

This Project/internship titled “**DIUbot – A Chatbot For Diu Admission Section**”, submitted by **Md. Masum Reza**, ID No: 161-15-6894, **Ruhullah Bin Kalim** ID No: 161-15-7092, **Md. Abdullah Al Muid**, ID No: 161-15-6999 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 6 December, 2019.

## BOARD OF EXAMINERS

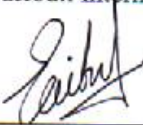
  

---

**Dr. Syed Akhter Hossain**  
**Professor and Head**

Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University

**Chairman**

---

**Saiful Islam**  
**Senior Lecturer**

Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University

**Internal Examiner**

---

**Shaon Bhatta Shuvo**  
**Senior Lecturer**

Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University

**Internal Examiner**

---

**Dr. Dewan Md. Farid**  
**Associate Professor**

Department of Computer Science and Engineering  
United International University

**External Examiner**

## DECLARATION

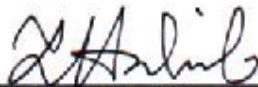
We hereby declare that, this project has been done by us under the supervision of **Md. Sadekur Rahman, Assistant Professor, Department of CSE, Daffodil International University**. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree.

**Supervised by:**



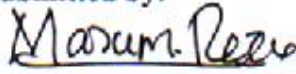
**Md. Sadekur Rahman**  
Assistant Professor  
Department of CSE  
Daffodil International University

**Co-Supervised by:**

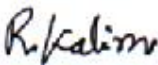


**Md. Tarek Habib**  
Assistant Professor  
Department of CSE  
Daffodil International University

**Submitted by:**



**Md. Masum Reza**  
ID: - 161-15-6894  
Department of CSE  
Daffodil International University



**Ruhullah Bin Kalim**  
ID: -161-15-7092  
Department of CSE  
Daffodil International University



**Md. Abdullah Al Muid**  
ID: -161-15-6999  
Department of CSE  
Daffodil International University

## ACKNOWLEDGEMENT

First we express our heartiest thanks and thankfulness to all-powerful Allah for his perfect gift that makes us conceivable to finish the last year venture effectively.

We extremely thankful and wish our significant and obligation to **Md. Sadekur Rahman**, Assistant Professor, Department of CSE Daffodil International University, Dhaka. Profound Knowledge and unmistakable fascination of our director in the field of utilization advancement impacted us to complete this venture. His interminable persistence, academic direction, nonstop support, steady and enthusiastic supervision, productive analysis, profitable exhortation, perusing numerous substandard draft and revising them at all stage have made it conceivable to finish this venture.

We might want to offer our heartiest thanks to **Dr. Syed Akhter Hossain**, Head, Department of CSE, Daffodil International University, Dhaka, for his thoughtful help to complete our undertaking and furthermore to other employee and the staff of CSE bureau of Daffodil International University.

We might want to thank our whole course mate in Daffodil International University, who participated in this exchange while finishing the course work.

Finally, we must acknowledge with due respect the constant support of our parents.

## **ABSTRACT**

DIUbot is a research-based project targeted to propose and implement a chatbot for DIU. At present, it's key focus is to provide support for DIU admission section. System take a question as an input and return an answer relevant to that question. For the sake of the research frequently asked questions were collected from DIU website and were used as the dataset of the project. Questions and their relevant answers were tokenized first and then stop words were removed. After that each sentence was converted into vector using word2vec method. Various similarity algorithms were studied and finally cosine and TF-IDF were implemented to match the best similar question with the given input. Index of the answer from selected question was defined and return as an output. Whole project was done by using Google Colab and no interface has been developed till now.

## **TABLE OF CONTENTS**

<b>CONTENT</b>	<b>PAGE NO</b>
Board of examiners	i
Declaration	ii
Acknowledgements	iii
Abstract	iv
 <b>CHAPTER</b>	
<b>CHAPTER 01: INTRODUCTION</b>	<b>1-3</b>
1.1 Introduction	01
1.2 Motivation	01
1.3 Rationale of the Study	02
1.4 Research Questions	02
1.5 Expected Outcome	02
1.6 Report Layout	
 <b>CHAPTER 02: BACKGROUND</b>	<b>4-8</b>
2.1 Introduction	04
2.2 Related Works	04
2.3 Research Summary	06
2.4 Scope of the Problem	06
2.5 Challenges	07
 <b>CHAPTER 03: RESEARCH METHODOLOGY</b>	<b>9-17</b>
3.1 Introduction	09
3.2 Research Subject and Instrumentation	09
3.3 Data Collection Procedure	10
3.4 Methodology	15
3.5 Implementation Requirements	16

<b>CHAPTER 04: EXPERIMENTAL RESULTS AND DISCUSSION</b>	<b>18-17</b>
4.1 Introduction	18
4.2 Experimental Results	18
4.3 Descriptive Analysis	15
<b>CHAPTER 05: SUMMARY, CONCLUSION, RECOMMENDATION AND             IMPLICATION FOR FUTURE RESEARCH</b>	<b>18-30</b>
5.1 Summary of the Study	18
5.2 Conclusions	20
5.3 Recommendations	30
5.4 Implication for Further Study	30
<b>REFERENCES</b>	<b>32</b>
<b>APPENDIX</b>	<b>33</b>

## **LIST OF FIGURES**

<b>FIGURES</b>	<b>PAGE NO</b>
Fig-3.1: Proposed Methodology	14
Fig 3.2: Elaborated Methodology	15
Fig 4.1: Raw data	18
Fig 4.2: Script file after pre-processing.	19
Fig 4.3: Input File Creation graph.	20
Fig 4.4: Tokenized data	20
Fig 4.5: Excluded words removal	21
Fig 4.6: Coding for Vectorization	21
Fig 4.7: Outputs file generate after Vectorization	22
Fig 4.8: 10 Similarity matching	22
Fig 4.9: Answer Selection	23
Fig 4.10: Answer Selection	23
Fig 4.11: Frequency distribution of questions	23
Fig 4.12: Frequency distribution of users' feedback about the chatbot	24
Fig 4.13: Sample responses of the users	25



# CHAPTER 1

## INTRODUCTION

### 1.1 Introduction

A chatbot is a piece of software that conducts a conversation via auditory or textual methods. Domain specific chatbot is a chat bot for only one specific domain like selling products or communication with customers. We are in the age of 4th IR (Industrial revolution) which is the current and developing environment in which disruptive technologies and trends such as the Internet of Things (IoT), robotics, virtual reality (VR) and artificial intelligence (AI). Chatbot is one of the implementations of AI. In other way chatbot can be defined as an alternative of a reception employee. Today many international organizations are trying to upgrade the concept of automatic replying digital agent. Google, Facebook, Amazon are very much familiar for 4th Industrial revolution. Such big companies are trying to do the domain specific chatbot for different purposes to make our day to day life easier. In this thesis we are going to make a domain specific chatbot for the specific domain “University Admission Information”.

### 1.2 Motivation

Daffodil International University is using “Pure Chat” which is not perfect for students and guardians. We asked some questions from different user’s device and the result is quite depressing. The chatting system replies and gives us website link rather than straight and accurate answer. Being a user, this is not expectable to any query-answer system from where I need to go to a website and find my answer on my own. During office hour chatting system reply within some seconds, but after office hour chatting system delays the reply. In this situation, any user won’t be interested anymore to know the answer to the queries. 4 test case shows that the average replying time is  $(3+2+3+(3600*10))/4 = 9002$  seconds. This is a very slow average response time.

### **1.3 Rationale study**

The problem can be solved by using a chatbot system which can make real time response more accurately. chatbot system can save a lot of budget for any university. We decided to develop the chatbot for admission sector of Daffodil International University. There is no requirement for any heavy weight device to develop the system. So, this is going to be much economic and useful for our University.

### **1.4 Research Question**

For the purpose of the research following research question have been observed:

1. What are the methods used for chatbots?
2. Is it possible to make domain specific chatbot?

### **1.5 Expected Outcome**

Our expected outcome is to make a chatbot system which will be able to recognize the user's question and reply with accurate and specific information regarding the university. This chatbot is not 100% perfect but we are trying to make this chatbot perfect. Response time is very low that means the chatbot is very fast. The chatbot can reply 24/7, which can make more opportunities for any university admission sectors.

### **1.6 Report Layout**

This report is divided into 5 chapters.

Chapter 1 is all about the introduction part. Where we discussed the motivation of this project and why we are doing this project. Besides we are discussed about the expected outcomes and the report layout of this report. In a word chapter 1 is the introduction of this project.

In Chapter 2, we discuss about the related works and scopes of the project including summarizing related work. The challenges we face in this research are briefly discussed here.

In Chapter 3, we directly focused on the research subject and Instrumentation.

Data collection and their statistical analysis are presented with proper diagram.

Implementation Requirements are represented here with exposition.

In Chapter 4, descriptive analysis with experimental results are the key topic in this chapter. We tried to represent all of the test data and its result with easy understandable figures.

In Chapter 5, we discussed the future work and limitations of the project. Full project summarization and recommendation are mentioned here also.

## CHAPTER 2

### BACKGROUND

#### 2.1 Introduction

In this chapter, we are going to describe several important facts like related work with our project. There is some research work related to our research. We will review those papers in 4 parameters. Algorithm, accuracy, data collection, and the domain in which the research based on. Then we will summarize those researches in order to find the most related work here. After that we will describe some scopes for our research. As there are many scopes in our research work and we are researching a new topic we had some challenges. Various types of challenges we had to deal with. At the end of this chapter we will give a small brief of those challenges.

#### 2.2 Related Work

**Liu et. al.**, presented in their work a novel way of designing an agent based chatbot which is domain specific and can learn from mobile. In this work for extending the DeepQA on any domain establish a domain-specific gate using k-mean algorithm. The DOGDeepQA is designed as the combination of both subjective and objective perspectives and constructing a domain-specific corpus to accelerating the effectiveness of chatbot. Dataset included with 150 queries and expected answers. The queries were inputted by participants from a particular course exam of an educational institute. Answer was selecting through the process of discrimination using K-means clustering based on domain-specific definitions where the response rate was approximately 96% for 50 users [1].

**Sinha et. al.**, discussed in their work a novel way that how to make knowledge of a chatter robot from documents. Dataset contains around 1000 pairs of questions and answers (800 pairs training data and 200 pairs testing data) from different data source and also manually which was pre-processed under many different packages like regex. That was employed a well-known unsupervised clustering approach namely K-means algorithm where the response rate almost satisfactory. Its responses with 60.10% accuracy [2].

**Ghose. & Barua.,** presented in their study about the implementation of topic-specific Undergraduate Advisor chatbot based on NLP for a particular purpose of information student information desk using the semantics of Alice CodeBase. Dataset was collected with a survey of students of particular course. The evaluation of the bot's performance presented with two parameters; Satisfactory and Unsatisfactory. For different Conversation Context there are various Accuracy, for Admission info 70% Satisfactory and 30% Unsatisfactory, for Course info 80% Satisfactory and 20% Unsatisfactory and for Faculty info 60% Satisfactory and 40% Unsatisfactory [3].

**Ahmed Fadhil** presented the challenges based on a question that can a chatbot determine my diet for a chatbot application. This system designed with some proper frequent steps to meet the goal. A study on BotAnalytics [4] shows that only 60% of users are continuing their conversation for the second message and another 75% are continuing their conversation for further. Based on this study, an investigation has been done for meal recommendation and also for lifestyle promotion on the performance of chatbots activity [5].

**Kar. & Haldar.,** presented their works on the application of chatbot to the internet of things that specially described about the opportunities and architectural elements. Study shows that there some chat interfaces are being used in Instant Messaging (IM) platforms like Slack, Facebook Messenger, Kik, Telegram etc those are very popular and growing rapidly. Moreover, this study shows how the top ten messaging platforms alone account for about 4 billion users. So, the worldwide demand of chatbot accelerating the invention with the more user-friendly applications and specific purpose chatbots [6].

**Arruda et. al.,** presented in their work which was about a chatbot for goal-oriented requirements modelling where described about a possible solution by supporting requirements elicitation for sudden requirements engineers with the implementation of NLP with the context of the KAOS goal-oriented requirements engineering method within a chatbot. KAOSbot services through a graphical interface which identifying the proper Goals and trying to fulfil expectations based on provided requirements with the process of NLP techniques. In the process of transforming used syntactic tree, the

Stanford CoreNLP library and the parsed tree for the identification of syntactic patterns to convert the finalized tree. The result of quasi-experiment is about 0.862, that response is satisfactory [7].

### **2.3 Research Summary**

So, it has been noticed that a great deal of research has been done on point displaying for making a domain specific chat bot using various types of Natural Language Processing (NLP) related techniques and algorithms including unsupervised clustering approach namely K-means algorithm, syntactic tree, the Stanford CoreNLP library, parsed tree and syntactic patterns. The vast majority of them utilized the greatest measure of foundation data as training data and the remainder of the data as test data. Finally, the two sorts of data yield a definitive outcome is produced with the help of proposed classifier model and most of the result is in satisfactory level.

### **2.4 Scope of the Problem**

Chabot is used in several domains but education. Education is one of the basic needs of anyone. All over the World the education sector is focused and so many organizations are trying to improve the quality of education. Technologies help us to increase the efficiency of education. Chatbot will also boost up technological use of robots in education. Again, there is very few chatbot available what are able to work in a domain specific way. As there are very few domains specific research, this is a significant and worthy scope of the problem.

On the other hand, there is no complete Chatbot developed in Bangladesh. There are many universities in Bangladesh and they are using traditional methods to communicate with new clients. Already it is mentioned that it takes a lot of time to reply to the client. It takes almost 3 hours which is very depressing for any client. Every corner of our day to day is enriched by using technology in communication what is very good news but we did not develop any chatbot for education sector. Being more specific, admission sector is an important part for every student and universities. It is not possible for everyone to go to all university and collect information about his/her admission. Though websites made easy this task, it is hard to find all necessary information in a very short time. Regarding this problem chatbot is the only solution. Admission sector

will be more accurate and time saving for both students and universities. As this is a domain specific chatbot and there is no work about domain specific chatbot we are very much focused to solve this problem by chatbot.

## **2.5 Challenges**

### **Data collection**

Data collection is a tough task for every research work. Sometimes, data collection become the base of research work. It was a very challenging work to fix what type of data is appropriate for this research work. Questions and its answers collection is a difficult task for us, because we collected it from different websites types of question here. The whole research depends on this data set so if we make the data set with some wrong data it would effect on the final result of this research.

### **Making Language Compatible with system**

The most difficult part was to make words of the questions and answers compatible with the system. As we are working with English language, there are some similar types of word sets in the database. The meaning of a single word and the meaning of the whole line is quite different. We have searched a lot to find a way out. But English has many similar meaning words. English also has a very critical grammatical structure. All of this makes the research difficult. But we take help from our teacher and some NLP book where we got to know how the system works, such as how the system know that clients want to know about any specific course. At the day's end we made the language compatible with system accurately.

### **Model Selection**

Model selection was confusing, because we do not about the appropriate model for this research work. We had to test lots of model to find the appropriate one. In any research project success depends on the model selection and its data sets. The accurate and right choice will lead you to your goal very fast and the wrong choice will ruin it, which is why it was very sensitive task for us from the beginning. We test again and again with different types of model in order to find the best model, which means the model will

work as we wanted. Most of the research recommended an optimal model for a specific task. After lots of test cases we find the model working with my project.



## **CHAPTER 3**

### **RESEARCH METHODOLOGY**

#### **3.1 Introduction**

In this chapter, we will describe our research methodology and working procedures. Tools for research project are described here also. Data collection, data pre-processing and their implementation will be described all over the session. Topic, will be discussed in this session. One of the important parts of this project is research instrumentation that means the methods of data collection and the key factors collection for the research. We have described it step by step with proper diagram included.

#### **3.2 Research Subject and Instrumentation**

##### **Data for the domain**

Every domain has its own types of data. Like that, our domain is also having some data types. We needed a dataset. The dataset must contain questions and answers. Our domain is education, so the fact is to make the dataset useful. We need to collect lots of questions because our this chatbot must be ready to answer all types of questions which are related to admission sector, such as questions about teachers, questions about students, departmental questions, questions related to any course provided by the university. The university offers BSc. program every semester. There are also MSc program and diploma program offered by the university. So, we tried to make data sets what has all the information needed to answer any question of a BSc. MSc. and diploma candidate. For this reason, we think that the best way to collect questions and answers to use the FAQ question from different subpages of the official daffodil website. Daffodil has different FAQ question sets for different courses offered. Then we collected each and every questions and answers from there. From the point of view of a new student, these questions are not enough so we go to the admission office and tried to observe about the interest of a new student. New students are interested about their security inside the campus. Some of the students and guardians are interested in student hostel. From those conversations we had made some more questions and find the

answers from the university and admin panel. And finally, we made the data set. Now the data set contains right question and accurate answers.

## **Similarity**

Similarity between two things indicates how much one matches with another. In this research we tried to find the similarity between two words and then find the similarity between two sentences where one sentence is the question provided by the user of the system and another sentence is from dataset's question column from which we will find the accurate answer. This description is about the similarity to find the best one. Assume that as a user of the system we asked a question which is similar to 5-6 question from the database's question then how our system will find the answer. We will find the similarity value of those semi matched question, then find the highest value of similarity checking. We will discuss it later in another chapter where will show how we make the machine able to read the file and find the similarity of words and sentences.

## **3.3 Data Collection Procedure**

In the examinations, we gathered English text as data sets. Data sets are mainly combined with various questions and answers that uses as text. Those texts are included in the data set considering various highlights and volumes of data. Very nearly 500 questions and answers were collected as datasets where 350 of them are preparing as the training data set and 150 used as the test data set.

## **Data Pre-processing**

The pre-period of handling datasets is considering as Data pre-preparing. Most of the case, raw data sets are not ready to perform all activities and create disagreed results. Therefore, data pre-processing becomes part and parcel for both training data and testing data. Furthermore, it is viewed as one of the most significant pieces of exploration.

For text similarization, total collected data from survey more than 500 questions and answers. To similarities the desired question into dataset before similarization it is necessary to pre-process the raw data into CSV format dataset. Therefore, we have to follow some steps sequentially including text (questions and answers) segmentation,

sentence segmentation, word segmentation, removes stop words, tokenization, vectorization, stemming, Calculate Sentence Score, Calculate Word Score, Filtering/Smoothing etc.

### **1) Text (questions & answers) segmentation:**

Text segmentation is the way toward isolating composed content into significant units, for example, words, sentences, or subjects. The term applies both to mental procedures utilized by people when understanding text, and to artificial procedures executed in digital devices, which are the output of Natural Language Processing (NLP). At first, it is needed to separate queries and answers from raw data using text segmentation. Then segmentation applied for dividing the whole text into meaningful unit with eliminating unnecessary parts [8].

### **2) Sentence Segmentation:**

Sentence Segmentation is the way toward deciding the more drawn out handling units comprising of at least one words. This assignment includes distinguishing sentence limits between words in various sentences. It's very much more difficult task to determine the similarity from such a large data set. Firstly, matching the most similarity in questions data sets then again searching for the most relevant answer. Using NLTK toolkit for this purpose to separate sentences from given datasets [9].

### **3) Word segmentation:**

Segmenting a lump of content into words is normally the initial step of handling text, however its need has once in a while been investigated. In this paper, we pose the essential inquiry of whether word segmentation (WS) is vital for deep learning-based Natural Language Processing (NLP). Generally, for word segmentation space and comma is used as separator in between one word to another word. We also follow the same process for word segmentation [10].

### **4) Sentence Tokenization:**

Tokenization is one kind of way that towards separate sensitive data with some kind distinguishing remarkable symbols which contains the basic information without trading off its security. The purpose of tokenization is to make data into pieces so that

we can find similarity or to find some correlation between one token list and another token list. This tokenization also can be implemented in the banking sector where token is like a key and footprint of every transaction. Finally, apply tokenization to all sentences [11].

#### **5) Remove stop words:**

A stop word is an ordinarily utilized word, (for example, "the", "an", "an", "in") that always introduced to ignore when searching anything through a search engine. Therefore, stop words are unnecessary in the purpose of matching similarity that's reason it always removed from the word segmentation. In this process NLTK tools applied to remove stop words [12].

#### **6) Stemming:**

Stemming is a method of decreasing a phrase to its phrase stem that affixes to suffixes and prefixes or to the roots of words recognized as a lemma. In the process of (NLU) Natural Language Understanding and (NLP) Natural Language Processing stemming is an essential method. Stemming is a section of linguistic studies in morphology and artificial brain (AI) statistics retrieval and extraction. In this process smoothing data are collected and stored in the necessary format. Data are now geared up to be analysed and further processing [13].

#### **7) Build vocabulary and generate vectors:**

We realize that the greater part of the application needs to manage a large number of datasets. Thus, a non-computationally-ideal capacity can turn into an immense bottleneck in your algorithm and can take bring about a model that takes ages to run. To ensure that the code is computationally effective, we will utilize vectorization.

Time complexity in the execution of any calculation is extremely critical choosing whether an application is dependable or not. To run a huge calculation in as much as ideal time conceivable is significant with regards to the constant utilization of yield. To do as such, Python has some standard numerical capacities for quick tasks on entire arrays of data without composing loops. Such as library which contains such capacity is NumPy.

## **8) Bag-of-words:**

The bag-of-words model is a well-known and straightforward component extraction method utilized when we work with string. It depicts the event of each word inside an archive. Raw text is not perfect for applying Machine learning algorithms directly, so we have to convert the text into vectors of numbers. This process is known as feature extraction. It describes the state of each word inside a dataset [14].

This procedure executed through two significant parts; structure a vocabulary of known words (likewise called tokens) and pick a proportion of the nearness of known words. Any data about the request or structure of words is disposed of. That is the reason it's known as a bag of words. This model is attempting to comprehend whether a realized word happens in an archive, however don't have the foggiest idea where is that word in the data.

The instinct is that comparative records have comparative substance. Additionally, from a substance, we can get the hang of something about the meaning of the data.

## **9) Similarity matching:**

The question is collected from user input and using similarity algorithm (such as cosine-similarity, TF-IDF matching) matching the similar question in dataset. In this step, the algorithm produces 10 similar questions as output from dataset [15].

## **10) Answer selection:**

From the process of similarity matching, the algorithm gives 10 similar questions as primary result. Now we have to finalize the most relevant question among 10 questions that already selected from previous output. This process finished by selecting the final question among them. Then the program finds the answer against the finalized question from the dataset.

### 3.4 Methodology

The proposed method of the research is shown in figure 3.1.

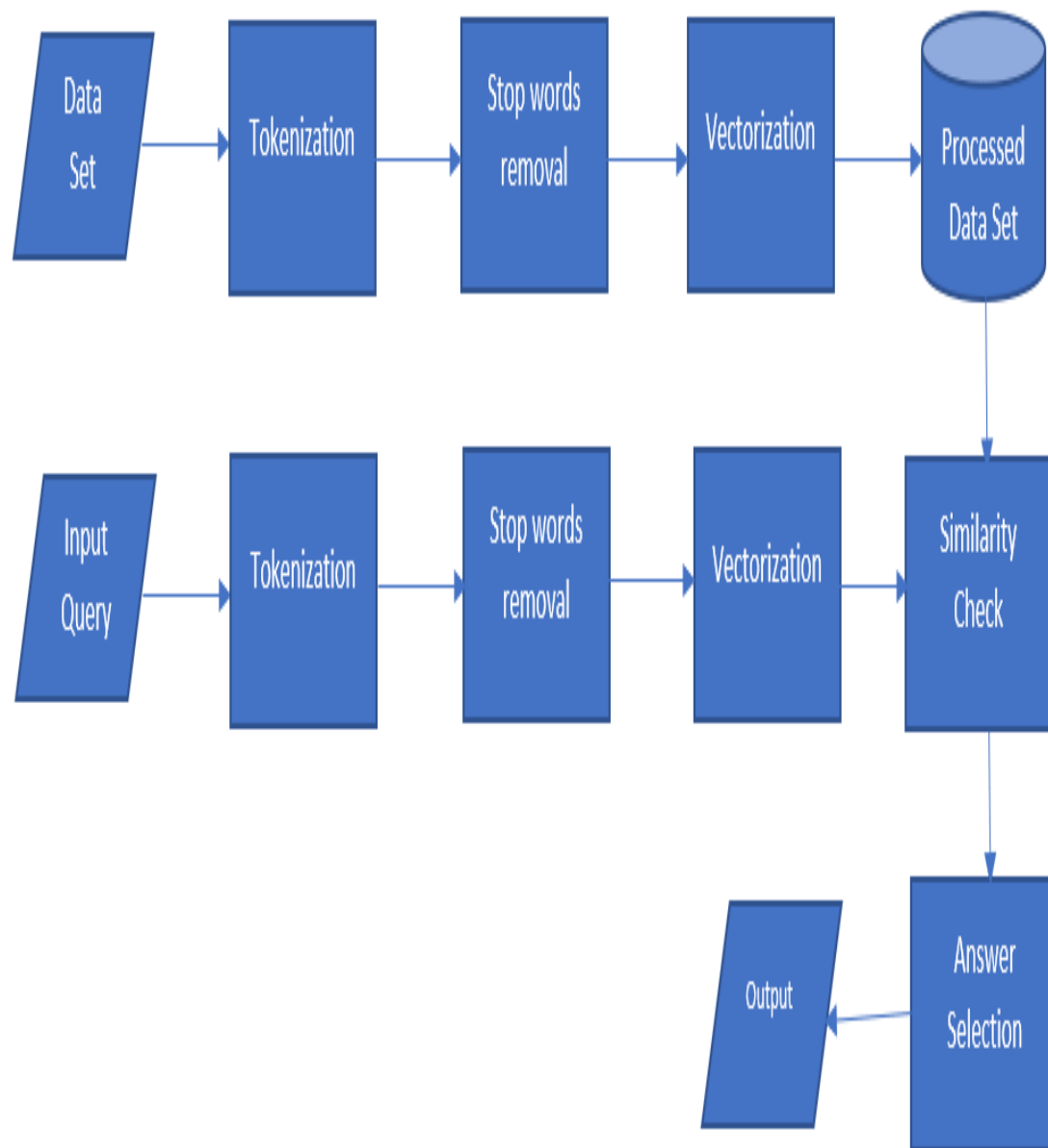


Fig 3.1: Proposed Methodology of research

However, the extended methodology follows in figure 3.2.

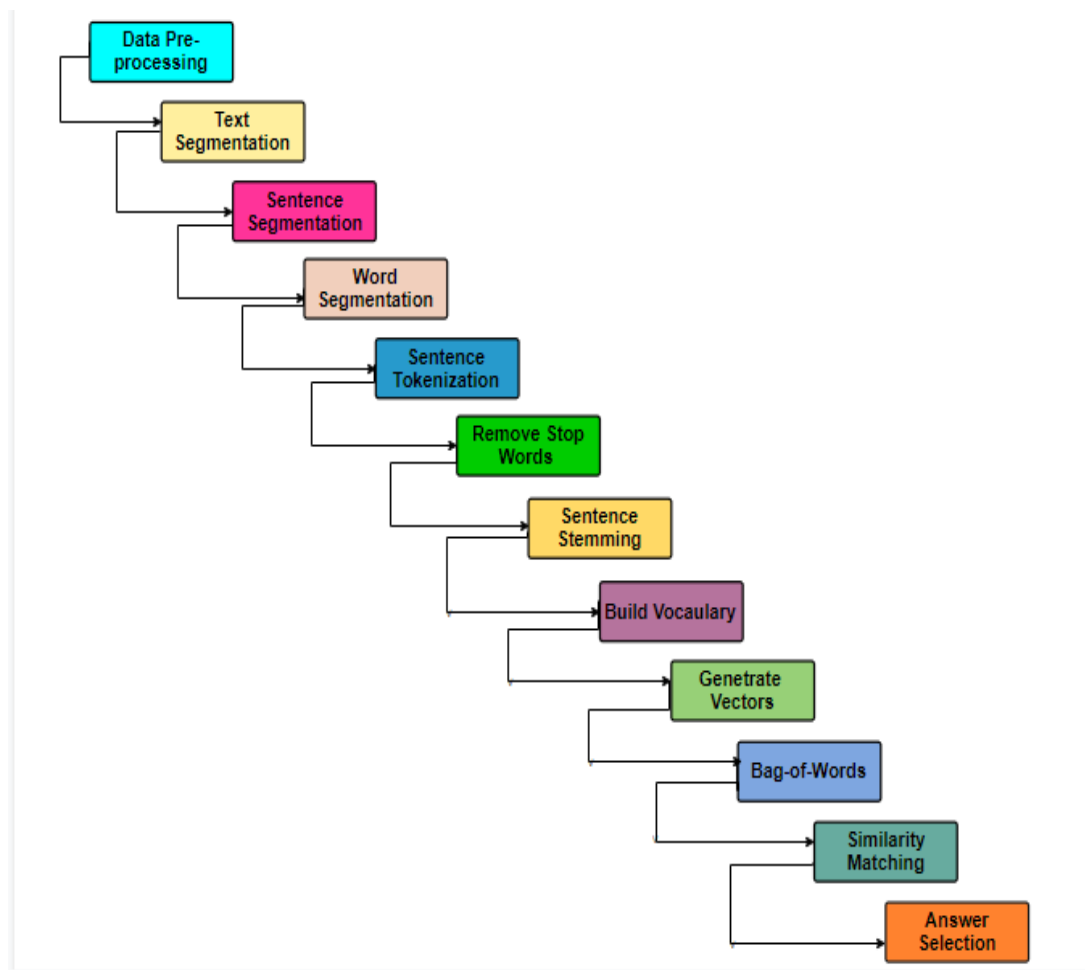


Fig-3.2: Elaborated Methodology

### 3.5 Implementation Requirements

#### Python 3.0:

Python is an interpreted language which is also known as high level language and used for general purpose programming. Use of whitespace makes python code more readable to user. For both small and large projects python is best choice because of its special coding format including object-oriented module with clear and logical distribution of code. It also introduced as dynamic language. It also contains more than one paradigm, procedural and functional capacity. It enriches with its comprehensive standard library (includes process flow control capabilities, possesses strong integration, text processing capabilities) and also consider as "batteries included" language for these features. All those advantages recognized it as one of the speedy and productive languages [16].

### Python Standard Libraries:

Python included with many standard libraries which uses purpose of extracting and semantic. It also contains few numbers of additional components that are generally contain with Python modules. The library included some of built-in modules which are written with the help of C languages and some of module are created with python also [17]. Python's standard library produces very extensive performance by providing very large scale of advantages as shown below:

- *import numpy*
- *import scipy*
- *import pandas*
- *import matplotlib*
- *import os*
- *import spacy*
- *import nltk*
- *nltk.download('punkt')*
- *from nltk import word\_tokenize*
- *from gensim.models import Word2Vec*
- *nlp = spacy.load('en')*
- *WmdSimilarity*

### Google Colab:

Google Colab is a cloud-based service provided by google which totally free. It also contains the feature of free Graphics processing unit (GPU). It is a great platform to skilled up by Python programming and also developing deep learning projects with the help of some very usable libraries including Tensor Flow, Keras, PyTorch and OpenCV [18]. Some very basic commands are given below:

- *from google.colab import drive*
- *drive.mount('/content/gdrive')*



### **Google Drive:**

Google Drive is one of the best services of Google which provided the features of storing and synchronizing files since 2012. It gives the access of storing files on server, sharing files and it is very secure because of synchronization features. It also contains apps for various Operating System including Windows, macOS computers, Android, iOS smartphones and tablets. Google Drive is a combo package of Google Docs, Google Sheets, and Google Slides where anyone can manipulate the documents, spreadsheets, presentations, drawings, forms etc [19]. All the actions can be saved in google drive after the necessary manipulation.

```
→ from google.colab import drive  
→ drive.mount('/content/gdrive')  
→ root_path = 'Path'
```

### **PyCharm:**

Integrated development environment (IDE) is essential for computer programming. PyCharm is one of the most popular IDE being used for the Python programming which released by the Czech company Jet Brains. It provides the advantages of web development with many popular frameworks especially Django and also eligible for practicing Data Science with the support of Anaconda with some analysing features including a debugger with graphical representation, a unit tester and version controlling system with integration. PyCharm supports in various operating system with Windows, macOS and Linux. PyCharm popular with two different version; The Community Edition and with some additional advantages the Professional Edition [20].

### **Text Editor Tools:**

Notepad is very popular for taking notes and saved them in it. It is mainly a text editor where anyone can manipulate the text as needed. There are many text editors available such as Microsoft Notepad with Microsoft Windows, Notepad+ is well known for the freeware feature, Notepad++ also a tool for manipulating text and an advanced version of Notepad+ and NotePad2 is a text manipulating tools where anyone can contribute to develop it more.

## CHAPTER 4

### EXPERIMENTAL RESULTS AND DISCUSSION

#### 4.1 Introduction

In terms of NLP, we have tried to develop a simple Chabot by which one we can measure the accuracy of text matching. In chapter 3, we have described about the research project and procedure of its working. Now, in this chapter, we have tried to show the output result & process of categorization.

#### 4.2 Experimental Results

For the research project, we collected pre-processed data from our university web site & admission to take the university is new students. But still, there are some redundant data of our dataset.

Now a view of raw data are given below.

175	Online Notice Board	As a Digital University, DIU publishes all notices and announcements on Online Notice Board. As a result, all		
176	DIU Medical Center	DIU established a Medical Center in the Campus for the health care of DIU students. DIU students are able to		
177	Fresherâ€™s Welcome	DIU receives its future Leaders and Scholars with a warm and fresh welcome through Orientation program. D		
178	Guardian Communication	For the betterment of the students, DIU communicates with the Guardians and informs them about their stud		
179	Special counseling for weaker students	If it is identified that few students are not getting marks up to satisfactory level, DIU faculty members suppor		
180	Religious Events	DIU observes all religious programs and arranges various religious events (Iftar party, Miladunnabi, Sarwasw		
181	Digital Class Rooms	As the first digital university in Bangladesh, DIU introduced digital class rooms for making the students famil		
182	One stop services	To solve the problems of students and to make a friendly educational environment in campus DIU started on		
183	Gymnasium	DIU has established a large gymnasium in the campus. The gymnasium is furnished with modern equipment.		
184	Financial Aid & Scholarships	<a href="https://daffodilvarsity.edu.bd/scholarship/diu-scholarship">https://daffodilvarsity.edu.bd/scholarship/diu-scholarship</a>		
		Visit the following link of DIU website		
		Officers: <a href="https://daffodilvarsity.edu.bd/administration">https://daffodilvarsity.edu.bd/administration</a>		
185	How can I find the contact numbers of a teacher/officer?	Teachers: <a href="http://faculty.daffodilvarsity.edu.bd/">http://faculty.daffodilvarsity.edu.bd/</a>		
186	What is the process of teaching evaluation?	Login to student portal and update your profile. After that, fill the teaching evaluation form and submit it.		
187	How to check payment ledger from student portal?	Login to student portal and click â€œpaymentâ€ option for student payment ledger and payment scheme.		
188	Can I meet with my course teacher other than class period	Yes, you can meet with your teacher during the counseling hours.		

Fig 4.1: Raw data

In the fig.4.1, we can see some of the invalid words.

So, we need data pre-processing to reduce data redundancy to find the exact data or information of our dataset.

##### 4.2.1 Data pre-processing

We design a Python script file to initiate the data pre-processing task. The script is basically working for:

- Replace all unnecessary spaces from text.
- Remove all Punctuation marks.
- Remove all-new line.
- Assign a document number for predefining the category of each data set.

```

What do I need to borrow items from the Library ',
'How long can I check out Library items ',
'Where do I return Library items ',
'How can I get an item that is currently checked out ',
'Can you tell me who has an item that is currently checked out ',
'What are the fines for overdue items What happens if I fail to return an item checked out in my name ',
'What happens if an item checked out in my name is lost or damaged ',
'I am a daffodil International University student in a graduate program What do I have to do to submit my thesis to the Library ',
'How can I find out when a library item is due ',
'How many books may I have checked out at one time ',
' How to solve Multimedia/ IT related Problem ',
' How can I change the Campus ',
'What are the requirements to get Laptop from university ',
' How I can transfer credit with a foreign university ',
' How I can join in International Summer Programs ',
' How can I find the contact numbers of a teacher officer ',
' What is the process of teaching evaluation ',
' What is the role of DSA Office to the students' generally asked by fresh Student ',
' Is membership needed to get service from DSA Office ',
' How can we get notification about DSA Office ',
' How can we join with club activities ',

```

Fig 4.2: Script file after pre-processing.

### 4.2.2 Input File Creation

After the successful pre-processing process, we have some categorical question files at hand. There are different types of questions like about administration, about all departments, about the departmental teacher, about all courses, about all student facilities, about all other activities, etc. Then, we have to perform Natural language processing on this question files, we must join all these files into a file. For this, we use another python script. This file takes the folder name that contains dataset.csv file as an input and produces only a file where all questions and answers contained individually being merged.

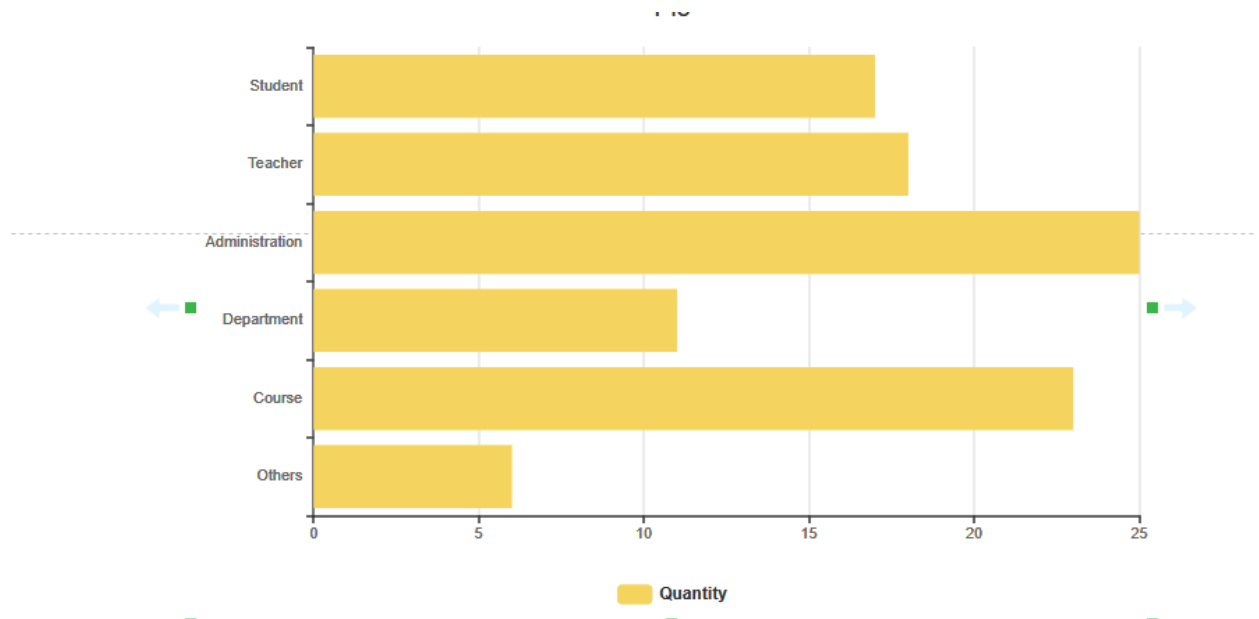


Fig 4.3: Input File Creation graph.

### 4.2.3 Tokenization

When all of our data is formatted then we sort the data from a list. Then put that listing data in a file and we applying the tokenized model. Then generate the tokenize file of all datasets. The format that came after tokenization is given below.

Q	A
Tokenize questions	Tokenize answer
'How', 'do', 'I', 'apply'\n\n']	['For', 'directions', 'on', 'how', 'to', 'apply',, 'please', 'visit', 'the', 'web', 'page', 'for', 'Graduate', 'Adm
'How', 'many', 'months', 'in', 'a', 'semester'\n\n']	['1', 'semester', 'in', '4', 'months.\n\n']
'What', 'criteria', 'do', 'you', 'use', 'for', 'acceptance', 'into', 'your', 'program'\n\n']	['evaluation', 'process', 'for', 'applications', 'is', 'extremely', 'complex',, 'taking', 'into', 'account', 'a
'How', 'much', 'extra', 'curriculum', 'fee'\n\n']	['About', '18000', 'taka.\n\n']
'How', 'do', 'I', 'choose', 'whether', 'to', 'apply', 'to', 'the', 'MS', 'or', 'the', 'PhD'\n\n']	['The', 'answer', 'to', 'this', 'question', 'depends', 'upon', 'your', 'individual', 'career', 'goals.', 'Our',
'Which', 'semester', 'how', 'amount'\n\n']	['Depend', 'on', 'your', 'credit', '.\n\n']

Fig 4.4: Tokenized data

### 4.2.4 Excluded Words Removal

The python code is developed for classifying all questions into a category. After joining all questions and answers into a file, the system gets ready for building a model. And then tokenizing all files produce some excluded words. The small size of list is created

for this purpose that contains all the words that have no meaning on their own. This list is called a list of excluded words. During checking the input file, it is also checked whether excluded words exist or not. If there is any, it must be removed from the tokenize file.

```
[ 'How',
  'can',
  'I',
  'find',
  'out',
  'when',
  'a',
  'library',
  'item',
  'is',
  'due',
  '?']

[ 'How',
  'many',
  'books',
  'may',
  'I',
  'borrow',
  'at',
  'one',
  'time',
  'from',
  'this',
  'library',
  '?
```

Fig 4.5: Excluded words removal

### 4.2.5 Vectorization

The function of this step is to do the vector convert. We work on excluded words removed at the previous step. Now we will just apply the vectorizing to the meaningful tokenizing word. Then a vector file will be generated. And a binary file will be generated. The code and the result of the figure are given below.

[illegible]

Fig 4.6: Screenshot of generated vector array.

```
[ ] import nltk
    from nltk import word_tokenize
    from gensim.models import Word2Vec

[ ] model = Word2Vec(line,min_count=3,size=32)

[ ] model.save("gdrive/My Drive/Myproject/vec4.txt")
    model.save("gdrive/My Drive/Myproject/vec4.bin")

[ ] /usr/local/lib/python3.6/dist-packages/smart_open/smart_open_lib.py:398: UserWarning: This function is deprecated, use smart_open.open instead. See t
    'See the migration notes for details: %s' % _MIGRATION_NOTES_URL
```

Fig 4.7: Outputs file generate after Vectorization

## 4.2.6 Similarity matching

This part working the best similarity matching of the user's question to the data set question. This matching model provided the 10 most similar questions. Then we find the max value of the similar question and take our system. Then we provided the most similar answer. The result of the figure is given below.

```
[ ] vec1=model1.wv['What'] + model1.wv['do'] + model1.wv['I'] + model1.wv['need'] + model1.wv['to'] + model1.wv['items'] + model1.wv['Library']
    model1.wv.most_similar([vec1])

[ ] /usr/local/lib/python3.6/dist-packages/gensim/matutils.py:737: FutureWarning: Conversion of the second argument of issubdtype from 'int' to 'np.signa
    if np.issubdtype(vec.dtype, np.int):
    [('to', 0.7664876580238342),
     ('What', 0.6936897039413452),
     ('item', 0.6796090006828308),
     ('do', 0.6566534042358398),
     ('?', 0.6476669311523438),
     ('is', 0.6264458894729614),
     ('I', 0.6074718236923218),
     ('and', 0.6053404211997986),
     ('student', 0.57569420337677),
     ('for', 0.5516664981842041)]
```

Fig 4.8: 10 Similarity matching

## 4.2.7 Answer selection

The question is selected by using the most similar words in the dataset. According to the question, 10 similar questions are primarily selected by the algorithm. After that, it finalized the best one as the desired output among 10 questions. Then the process generates the final result by matching the most relevant answers from the dataset based on the finalized question which was determined at the previous step.

```

import spacy
nlp = spacy.load('en')

search_doc = nlp(str(input()))
main_doc = nlp(str(line))
print(main_doc.similarity(search_doc))
#print(line)

... Does CSE offer a program in Computer Animation?

```

Fig 4.9: Answer Selection

```

import spacy
nlp = spacy.load('en')

search_doc = nlp(str(input()))
main_doc = nlp(str(line))
print(main_doc.similarity(search_doc))
#print(line)

Does CSE offer a program in Computer Animation?
0.7662096818931873
/usr/lib/python3.6/runpy.py:193: ModelWarning: [W007] The model you're using has no word vectors loaded, so the result of the Doc.similarity method
"__main__", mod_spec)

```

Fig 4.10: Answer Selection

### 4.3 Descriptive analysis

It is mentioned before that for the sake of the research data were collected from DIU website. More than 300 questions and answers were collected from FAQ section. Then they were checked and selected manually. Later they were classified into seven groups. Frequency distribution of the questions of different groups are shown in figure 4.11.

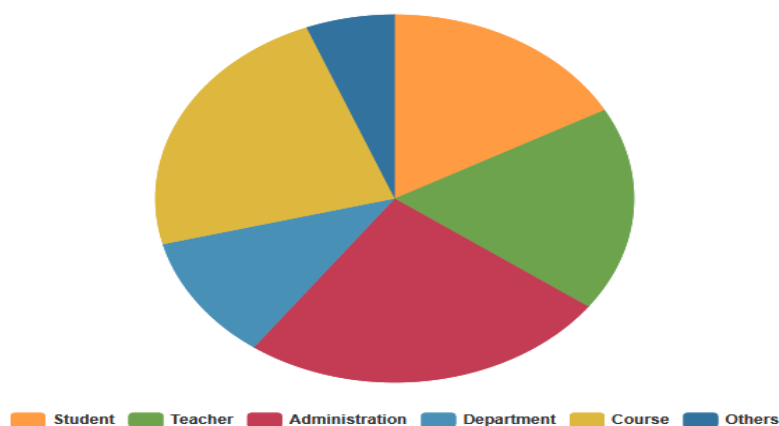


Fig 4.11: Frequency distribution of questions

#### 4.4 Performance of the chatbot

Primary data was collected from users through a survey to measure the accuracy of the chatbot. Total 72 persons took part in the survey. They all used the chatbot, asked whatever they want to know about CSE admission in Daffodil University. Then they gave feedback about their satisfaction based on a likert scale. Every respondent were asked to evaluate their each query in a scale of of 5 where 5 means highly satisfied, 4 means satisfied, 3 means neutral, 2 means dissatisfied and 1 means highly dissatisfied. Frequency distribution of the users' responses is shown in figure 4.12.

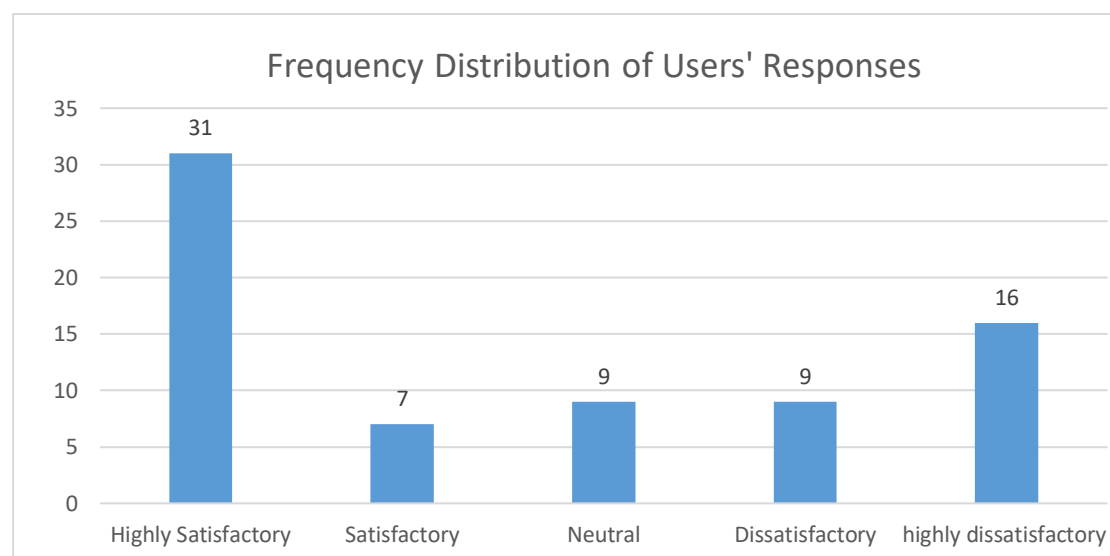


Figure 4.12: Frequency distribution of users' feedback about the chatbot

Based on these responses a weighed arithmetic mean score were calculated to evaluate performance of the chatbot. The score were calculated based on the following equation.

$$\bar{x} = \frac{\sum_i^n w_i x_i}{\sum_i^n w_i}$$

Therefore,

$$\text{Performance of the chatbot} = \frac{31 * 5 + 7 * 4 + 9 * 3 + 9 * 2 + 16 * 1}{31 + 7 + 9 + 9 + 16} = 3.38$$

Which is approximately 68%.

Sample responses of the users are shown in figure 4.13 to illustrate the idea of the responses of our chatbot generated against every user's query. Here, the first response



were evaluated highly satisfactory, the second as satisfactory, the next as neutral, the last but one is dissatisfactory and the last one as highly dissatisfactory.

```
Is there any admission test?
/usr/local/lib/python3.6/dist-packages/ipykernel_launcher.py:16: RuntimeWarning: invalid value encountered in true_divide
  app.launch_new_instance()
Reply: If your GPA is less than 3.50 then you have to test exam.

If you have no more question then please write 'No' or ask the further question.
What is the procedure of admission of cse
/usr/local/lib/python3.6/dist-packages/ipykernel_launcher.py:16: RuntimeWarning: invalid value encountered in true_divide
  app.launch_new_instance()
Reply: There is no minimum GPA to apply, and there is no automatic admission or denial based on GPA. Everyone who has met t

If you have no more question then please write 'No' or ask the further question.
What is the result category to needed a get waiver?
/usr/local/lib/python3.6/dist-packages/ipykernel_launcher.py:16: RuntimeWarning: invalid value encountered in true_divide
  app.launch_new_instance()
Reply: You can get your scholarship/waiver information from the Registrar Office.

If you have no more question then please write 'No' or ask the further question.
For a student what extracurricular activity help his future carrier?
/usr/local/lib/python3.6/dist-packages/ipykernel_launcher.py:16: RuntimeWarning: invalid value encountered in true_divide
  app.launch_new_instance()
Reply: DIU has arrangement of residence for female students. The accommodation provided upon requirements of the female stu

If you have no more question then please write 'No' or ask the further question.
how's the faculty?
/usr/local/lib/python3.6/dist-packages/ipykernel_launcher.py:16: RuntimeWarning: invalid value encountered in true_divide
  app.launch_new_instance()
Reply: Prof. Dr. S.M. Mahbub Ul Haque Majumder

If you have no more question then please write 'No' or ask the further question.
For otistic student is there any special facility?
/usr/local/lib/python3.6/dist-packages/ipykernel_launcher.py:16: RuntimeWarning: invalid value encountered in true_divide
  app.launch_new_instance()
Reply: https://www.google.com/a/diu.edu.bd/

If you have no more question then please write 'No' or ask the further question.
```

---

Figure 4.13: Sample responses of the users

## **CHAPTER 5**

### **SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH**

#### **5.1 Summary of the Study**

The research project is developed by finding a way to find similarity from a csv file from which it is possible to develop a chatbot so that it can be implemented in any website and able to reply different question accurately. We collected data from website and admission office to enlarge the question dataset and then collected data as answer from the official website of Daffodil International University. After data pre-processing, we convert it into a format which is compatible with the machine. We tried a lot of models to find the way to solution. Firstly, we tokenize database into token, remove unnecessary words. After that we vectorize data in order to find the similarity. Finally, we get 0.40-.88 % similarity between the database and the question that is given by the user.

#### **5.2 Conclusions**

In this modern technological world, we are depending on technology because technology is more reliable. Technology makes our lives easier and comfortable. Our research work just a step toward modern world. Once we have imagined that we will have a computer in our hands and now our imagination becomes true. Mobile phone is a small computer which was very bulk computer in the past. Nowadays machines are able to do job of human. Once upon a time there are thousands of workers in a production company but by the weaves of technology it become less than hundreds of workers in that company. We developed a chatbot which is basically a replacement of the receptionist in the front desk of an educational sector. Educational is open for everyone but here is the problem it was not possible to contact with a student who is far from the university. By this thesis project we believe that we can reach every corner of the world and talk to students just as like as humans without any human beings.

However, our current system has some limitations which are as follows:

1. Accuracy needs to improve more
2. Need to check by using other algorithms whether accuracy increase or decrease
3. It can't response to linked questions

### **5.3 Recommendations**

Several adaptive algorithms may develop to understand the question and answer of any text based on the dataset. Because understanding the questioning based on answer is very important.

It may change the entire concept what have been understood earlier.

### **5.4 Implication for Further Study**

- To upgrade it to a software.
- Is there any way to find similarity without tokenization and vectorization.
- Increase the accuracy of the model.
- To add more data to get better accuracy.
- To add more domains
- To implement question classification to simplify the complexity of the model.

## References

- [1] Liu, Q., Huang, J., Wu, L., Zhu, K., & Ba, S. (2019). CBET: design and evaluation of a domain-specific chatbot for mobile learning. *Universal Access in the Information Society*, 1-19.
- [2] Sinha, S., Basak, S., Dey, Y., & Mondal, A. (2020). An Educational Chatbot for Answering Queries. In *Emerging Technology in Modelling and Graphics* (pp. 55-60). Springer, Singapore.
- [3] Patil, V., Chaudhari, Y., Rohila, H., Bhosale, P., & Desai, P. S. (2019). Topic-Specific Natural Language Chatbot as General Advisor for College. *Applied Machine Learning for Smart Data Analysis*, 135.
- [4] Chatterjee, S., & Price, A. (2009). Healthy living with persuasive technologies: framework, issues, and challenges. *Journal of the American Medical Informatics Association*, 16(2), 171-178.
- [5] Fadhil, A. (2018). Can a chatbot determine my diet?: Addressing challenges of chatbot application for meal recommendation. *arXiv preprint arXiv:1802.09100*.
- [6] Kar, R., & Haldar, R. (2016). Applying chatbots to the internet of things: Opportunities and architectural elements. *arXiv preprint arXiv:1611.03799*.
- [7] Arruda, D., Marinho, M., Souza, E., & Wanderley, F. (2019). A Chatbot for Goal-Oriented Requirements Modeling. In *International Conference on Computational Science and Its Applications* (pp. 506-519). Springer, Cham.
- [8] Llopis, F., Vicedo, J. L., & Ferrández, A. (2002). Passage selection to improve question answering. In *proceedings of the 2002 conference on multilingual summarization and question answering-Volume 19* (pp. 1-6). Association for Computational Linguistics.
- [9] Treviso, M. V., Shulby, C., & Aluísio, S. M. (2016). Sentence segmentation in narrative transcripts from neuropsychological tests using recurrent convolutional neural networks. *arXiv preprint arXiv:1610.00211*.
- [10] Gambell, T., & Yang, C. (2006). Word segmentation: Quick but not dirty. *Unpublished manuscript*.
- [11] Fagan, J. L., Gunther, M. D., Over, P. D., Passon, G., Tsao, C. C., Zamora, A., & Zamora, E. M. (1991). *U.S. Patent No. 4,991,094*. Washington, DC: U.S. Patent and Trademark Office.
- [12] Zheng, G., & Gaowa, G. (2010). The selection of Mongolian stop words. In *2010 IEEE International Conference on Intelligent Computing and Intelligent Systems* (Vol. 2, pp. 71-74). IEEE.
- [13] Frakes, W. B. (1992). Stemming Algorithms.
- [14] Wallach, H. M. (2006). Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine learning* (pp. 977-984). ACM.

- [15] Santini, S., & Jain, R. (1995). Similarity matching. In *Asian Conference on Computer Vision* (pp. 571-580). Springer, Berlin, Heidelberg.
- [16] Kharkovyna, O. (2019). A Beginner's Guide to Python for Data Science. [online] Medium. Available at: <https://towardsdatascience.com/a-beginners-guide-to-python-for-data-science-60ef022b7b67> [Accessed 9 Dec. 2019].
- [17] Folkman, T. (2019). The Essential Python Libraries for Data Science. [online] Medium. Available at: <https://towardsdatascience.com/the-essential-python-libraries-for-data-science-ce55c53dfd6b> [Accessed 9 Dec. 2019].
- [18] Bonner, A. (2019). Getting Started With Google Colab. [online] Medium. Available at: <https://towardsdatascience.com/getting-started-with-google-colab-f2fff97f594c> [Accessed 9 Dec. 2019].
- [19] Ching, J. (2019). How to manage files in Google Drive with Python. [online] Medium. Available at: <https://towardsdatascience.com/how-to-manage-files-in-google-drive-with-python-d26471d91ecd> [Accessed 9 Dec. 2019].
- [20] Russo, J. (2019). 4 Tips to Get the Best Out of PyCharm. [online] Medium. Available at: <https://towardsdatascience.com/4-tips-to-get-the-best-out-of-pycharm-99dd5d01932d> [Accessed 9 Dec. 2019].