

AUTOMATED CLOSED DOMAIN CHATBOT FOR DIU

BY

TIPU SULTAN

ID: 152-15-5599

AND

MST. RUBAYA RUMI

ID: 152-15-6051

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

Supervised By

Sheikh Abujar

Lecturer

Department of CSE

Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY


DHAKA, BANGLADESH

APRIL 2019

APPROVAL

This Project titled "Automated Closed Domain Chatbot for DIU", submitted by Tipu Sultan, ID No: 152-15-5599 and Mst. Rubaya Rumi ID No: 152-15-6051 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 03/05/2019.

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



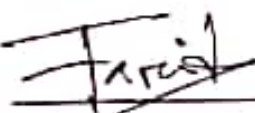
Dr. Md. Ismail Jabiullah
Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Dr. Sheak Rashed Haider Noori
Associate Professor & Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



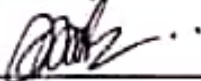
Dr. Dewan Md. Farid
Associate Professor
Department of Computer Science and Engineering
United International University

External Examiner

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Mr. Sheikh Abujar**, Lecturer, Department of CSE Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

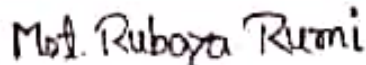


Mr. Sheikh Abujar
Lecturer
Department of CSE
Daffodil International University

Submitted by:



Tipu Sultan
ID: 152-15-5599
Department of CSE
Daffodil International University



Mst. Rubaya Rumi
ID: 152-15-6051
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

We really grateful and wish our profound our indebtedness to **Mr. Sheikh Abujar, Lecturer**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “*Field name*” to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to **Mr. Sheikh Abujar, Lecturer**, Department of CSE Daffodil International University, Dhaka, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

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

ABSTRACT

We want to provide a question answering chatbot working on Bangla language. This type of chatbot is already implemented, but this type of work is not yet properly carried out in Bangla language. This chatbot works on educational fields and answers the university FAQ in Bangla. Anyone can ask any questions in Bangla concerning the university our chatbot answers this question. We have used Naive Bayes classifier to implement our chatbot in python. We are working on the questions previously asked by the students or other peoples. By using the Naive Bayes method, we wanted to show the accuracy for the Bangla text classification. We worked on two types of questions and classify them by Naive Bayes method in python and tried to show how Naive Bayes method on a short data set as well as Bangla language.

TABLE OF CONTENTS

CONTENS	PAGE
Board of examiners	ii
Declaration	iii
Acknowledgements	iv
Abstract	v
List of Figures	ix
List of Tables	x
CHAPTER	
CHAPTER 1: INTRODUCTION	1-2
1.1 Introduction	1
1.2 Motivation	1-2
1.3 Rationale of the Study	2
1.4 Research Questions	2
1.5 Expected Output	2
CHAPTER 2: BACKGROUND	3-5
2.1 Introduction	3
2.2 Related Works	3-4
2.3 Research Summary	4
2.4 Scope of the Problem	4
2.5 Challenges	4-5

CHAPTER 3: RESEARCH METHODOLOGY	6-14
3.1 Introduction	6
3.2 Research Subject and Instrumentation	6
3.2.1 Research Subject	7
3.2.2 Instrumentation	7
3.3 Data Collection Procedure	7
3.3.1 Data Collection	7
3.3.2 Data Processing	7-9
3.4 Statistical Analysis	9-10
3.4.1 Naive Bayes algorithm	10-11
3.4.2 Multinomial Naive Bayes algorithm	11
3.5 Implementation Requirements	11-14
3.5.1 Creating BOW	12
3.5.2 Preparing training and testing data	12-13
3.5.3 Creating Model	13-14
CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION	15-18
4.1 Introduction	15
4.2 Experimental Results	15
4.3 Descriptive Analysis	15-17
4.4 Summary	18

CHAPTER 5: CONCLUSION AND IMPLICATION FOR FUTURE RESEARCH	19-20
5.1 Conclusions	19
5.2 Implication for Further Study	19
REFERENCES	20
APPENDICES	21
Appendix A: Research Reflection	21
Appendix B: Related Issues	21

LIST OF FIGURES

FIGURE	PAGE NO
Figure 3.1: Block Diagram of Chatbot	6
Figure 3.2: Sample Questions	8
Figure 3.3: Answer of Figure 3.2 Questions	9
Figure 3.4: BOW sample	10
Figure 3.5: Part of the BOW	12
Figure 3.6: Frequency distribution matrix table (feature vector)	13

LIST OF TABLES

TABLE	PAGE NO
Table 1: Input-Output table	15-17

CHAPTER 1

Introduction

1.1 Introduction

A chatbot is a kind of AI or a system which can continue a conversation as like as a human. There are two types of chatbot based on their knowledge- open domain and closed domain. An open domain chatbot can answer almost all questions perfectly. This type of chatbot needs to be very clever to answer any questions and should have a good common sense. On the other hand, close domain chatbot works on a fixed knowledge and generally cannot perform to answer all types of questions. Close domain chatbot generally used to manage a specific or fixed work. On this view, our chatbot should also be a close domain chatbot because it will work on the FAQ of the university. Though this type of chatbot can be implemented by using decision tree easily but we want to show how machine learning works on Bangla text.

Every day many students need much information about our University. Not only the students but also their parents and he who wants to admit to the university needs much information about the university. For this reason, we want to build a chatbot which will answer all of their questions easily. And we build this chatbot in Bangla so that, everyone can use this chatbot. Generally, it is an FAQ chatbot and its main purpose is answering the question asked by the users about the university. We collected data (questions asked by the students or the guardians) from different social media, admission office and from the department of the university. The main purpose of our paper is to find out the accuracy result for the data set and show how Naive Bayes classifier work on Bangla language.

1.2 Motivation

As a student when we want to get admission in a new university, we have many questions to fix our decision and to clear our confusion. Not only that after getting admission and the duration of our completing study we have to face many more questions and we need so much information about the university, it's rules, regulation and etc.

So, we want to make our all confusion clear, helps us to make our decision fix (in time) by giving all the question answer and all the information clearly in a smart way.

Again, Bangla is spoken by approximately 250–300 million people over the world. But there are a very small number of works is done in Bangla. Even there is no automated chatbot in Bangla like Siri or Cortana or Google Assistant. Even there is no closed domain chatbot too. So, we want to work with Bangla so that it could be beneficial for further research work.

1.3 Rationale of the Study

To choose this project there are two big reasons. First reason is there are a few works done on Bangla language. These works are not enough for doing something good. So, we also want to contribute on this wave.

The second reason is, there are many closed domain chatbot worked well in different sector such as online customer care, giving information about any company or selling product. But most of them are in English. Yet there is no closed domain chatbot use professionally in any sector. So, we decide to work with this project so that, it will be a new horizontal in Bangla chatbot history.

1.4 Research Questions

From this paper we can find out these questions answer:

1. Can we classify Bangla text as English text?
2. How Naive Bayes classifier give result on small dataset?
3. What are the problems in Bangla text processing?
4. Can we use raw data which is asked by the people normally?

1.5 Expected Output

By this project, we want to answer FAQ automatically about the university in Bangla. We also want to publish a research paper on Bangla text classification.

CHAPTER 2

Background

2.1 Introduction

On the road of Bangla text processing many work have done but that is not enough yet. Still we don't get a good accuracy result from Bangla text. Still there is no big implementation of Bangla language. That's prove these work is not enough for future challenges. There are many online shop using chatbot as online customer care. But on that state still we cannot reached. Still there is no morphological analyzer, Part of speech tagger are not available for Bangla text processing [9]. Bangla became International mother language day 17th November on 1999 [10]. But we have done nothing to protect our language and to introduce and make easier to the world. In this chepter we will talk about the work that have done before on Bangla language and also talk about the future scope of this project. Actually if we can process Bangla text well then it would have a great commercial value too. We will also talk about the main challenges on the way of processing Bangla text.

2.2 Related Works

The very first work on Bangla language processing was started on the 1980s in Bangladesh [11]. But as far as we know there are a very few works on Bangla chatbot. Among them, a notable work is Golpo chatbot [1], which is implemented by a conversational dialogue engine named ChatterBot [2]. Again, in a paper, the Bangla text normalization technique has shown and they got a good accuracy value [3]. Md. Nizam Uddin and Shakil Akter Khan worked on Bangla text summarization [4]. A Bangla sentence correction and auto complete work has done in 2018 using sequence to sequence neural network [8].

Beside these there are many works on chatbot in English. In chatbot history, ELIZA [7] was the first chatbot created by Joseph Weizenbaum in 1966 passed the Turing test and able to fool some user to think that it was a human they are talking to. In 1995 A.L.I.C.E. was the language processing bot performed well though it was failed the Turing test. In 2001

Smarterchild came with the features among quick data access and conversation like a human. In 2010 Apple introduced their AI bot SIRI which can do multiple tasks for the user. In 2012 Google brings Google Now for searching something with voice on google. Besides China's we chat brings an hottest trend in chatbot history [12]. Then in 2015 Amazon brings Alexa and Microsoft introduce their Cortana [6].

2.3 Research Summary

Here we work with Multinomial Naive Bayes to classify the Bangla text. We choose this classifier because this can be performed though there is a small amount of data.

From this research we want to show the output result after processing the Bangla text. We will also show how machine learning algorithms work on Bangla language. For that we use Naive Bayes text classification algorithm to process Bangla text. Here we use FAQ asked by the students and parents about the university as our dataset. We will create bag of words from that dataset and then make the data set into two parts. One is train data another is test data. Here we use 80% of the data set as train data and rest of 20% used as test data. Next, we convert string data as numeric form with the help of BOW.

2.4 Scope of the Problem

If we can solve this problem a new era of Bangla language processing will be opened. This research can be the light bringer in Bangla text processing area. It can also open a new window in chatbot history.

Again, the business value of this project is very high. To keep pace with the digital world every university need a digital assistant. In Bangladesh there are 140 university in total. So, this project has a great business value in future.

2.5 Challenges

Our main challenge will be getting a good accuracy as result. We find out some problems that are mainly responsible for poor accuracy result. Collecting data is one of the main challenges of our research. We collect data from various social media, admission office of the university and also from the department of the university. Then the next challenge is to

pre-process that data. We sorted out the same value data. And also sorted the whole data into different documents based on the answer of the question.

Beside these we need to create a dictionary that could have all the words along with the number of appearing through the whole documents. We also need to create such language model that could be turn the string words into numeric form. We called them as vectorization. As well as the model will process the input text and give some output based on the train model. We decided to create the model with the help of Naive Bayes algorithm.

If we could overcome these problems hope that we can get better accuracy for Bangla text.

CHAPTER 3

Research Methodology

3.1 Introduction

This chatbot can take input from user and give the output based on the classifier algorithm. But in here we only worked with two types of data set. And we found a poor accuracy result. We use python 3.6 to implement this project. And choose PyCharm as IDE. We collect data from different social media and from the admission office and from the department of CSE. In near future we want to do more work to improve this chatbot so that it can be chat like human.

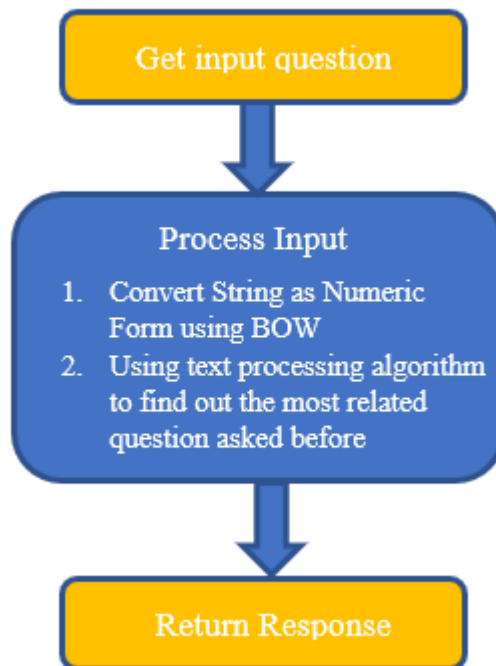


Figure 3.1: Block Diagram of Chatbot

We calculate similarities between two sentence using Naive Bayes algorithm. We convert each word into numeric form then we match them with the BOW and find out the maximum similar sentence.

3.2 Research Subject and Instrumentation

Before done any work first we need to have a clear idea about the research topic. Then we need to know which instruments we to implement this project. To do a good work its really important to have a clear idea about what we want to do and how we want to do that.

3.2.1 Research Subject

In this project our main target will be to find out a way to process Bangla text. This is a part of Natural Language Processing (NLP). This research can be the light bringer in Bangla text processing area. It can also open a new window in chatbot history.

3.2.2 Instrumentation

To build this project we use PyCharm IDE. We use python 3.6 as the base language.

3.3 Data Collection Procedure

Towards of researching any thing data collection and data pre-processing is very important. The result of any research mostly depend on which data is chosen and how the data is processed for using. Data procedure is important because we need to remove duplicate data, illogical data, noise data ect from the main dataset.

3.3.1 Data collection

We worked on the FAQ asked by the students or the guardians. We collect data from different social media specially from Facebook. Many students asked many questions in different Facebook group for the answer. We collect data from those group as well as the answer also. We also collect data from admission office. When students or guardians come to ask any question in admission office, we collect that data along with the answers.

3.3.2 Data Processing

We write down the data as text file. Then we sorted them in some group based on the question meaning. We keep that question in a group which answer is same. We also sorted out and rejected the duplicate data from the documents. We also make a simple format for the department name. Such as: বি.বি.এ. সি.এস.ই ত্রিপুরাই সিভিল so that it would be always same for each sentences.

কোন ওয়েভার নাই
সিভিল ইঞ্জিনিয়ারিং ডিপার্টমেন্টে ডিসকাউন্ট আছে কোন
সিভিল ইঞ্জিনিয়ারিং ডিপার্টমেন্টে এডমিট হলে কি রকম ডিসকাউন্ট পাব
সিভিল ইঞ্জিনিয়ারিং ডিপার্টমেন্টে অর্থনৈতিক সুবিধা কি রকম পাবো রেজাল্ট নিয়ে
সিভিল ইঞ্জিনিয়ারিং এ পার্মানেন্ট ক্যাম্পাস এ কি কি সুবিধা পাব
সিভিল ইঞ্জিনিয়ারিং এ মেইন ক্যাম্পাস এ কি কি সুবিধা পাব
৬.১২লাখ থেকে কি কোন % বা ওয়েভার পাবো না
সিভিল এ স্কলারশিপ কেমন
আপনারা কি গ্রুপ ওয়েভার দেন না
কোন ওয়েভার আছে
স্কলারশিপ আছে
টোটাল কত পারসেন্ট ওয়েভার পাওয়া যাবে
আগুলিয়া ক্যাম্পাসে কি ২০% ওয়েভার পাওয়া যাবে
কোন ইস্কলারশিপ নাই

Figure 3.2: Sample Questions

This is the sample questions showed on figure 3.2. Here we keep each question in a single line. Each question in a document has the same answer. We sorted each document based on the answer. All the questions that have same answer kept in the same document. We give a unique number of each documents so that, in future work it will be identified easily.

আপনার সেমিস্টার রেজাল্ট এর উপর নির্ভর করবে।

এইচ এস সি ও এস এস সি এ+(গোল্ডেন) =১০০% (শুধুমাত্র ১ম সেমিস্টার এর জন্য পরের সেমিস্টার থেকে পেতে হলে রেজাল্ট ৩.৫ থাকতে হবে)। সকল ক্যাম্পাস জন্য।

এইচ এস সি এ+(গোল্ডেন) =৭৫% (শুধুমাত্র ১ম সেমিস্টার এর জন্য পরের সেমিস্টার থেকে পেতে হলে রেজাল্ট ৩.৫ থাকতে হবে, মাইন ক্যাম্পাস ও উত্তরা ক্যাম্পাস এর জন্য)।

এইচ এস সি এ+(গোল্ডেন) =৯০% (শুধুমাত্র ১ম সেমিস্টার এর জন্য পরের সেমিস্টার থেকে পেতে হলে রেজাল্ট ৩.৫ থাকতে হবে, পার্মানেন্ট ক্যাম্পাস এর জন্য)।

এইচ এস সি ও এস এস সি এ+ =৫০%(শুধুমাত্র ১ম সেমিস্টার এর জন্য, পরের সেমিস্টার থেকে পেতে হলে রেজাল্ট ৩.২৫থাকতে হবে, মাইন ক্যাম্পাস ও উত্তরা ক্যাম্পাস এর জন্য)

এইচ এস সি ও এস এস সি এ+ =৬০%(শুধুমাত্র ১ম সেমিস্টার এর জন্য, পরের সেমিস্টার থেকে পেতে হলে রেজাল্ট ৩.২৫থাকতে হবে, পার্মানেন্ট ক্যাম্পাস এর জন্য)

এইচ এস সি এ+ =৩০%(শুধুমাত্র ১ম সেমিস্টার এর জন্য, পরের সেমিস্টার থেকে পেতে হলে রেজাল্ট ৩.০থাকতে হবে, মাইন ক্যাম্পাস ও উত্তরা ক্যাম্পাস এর জন্য)

এইচ এস সি এ+ =৫০%(শুধুমাত্র ১ম সেমিস্টার এর জন্য, পরের সেমিস্টার থেকে পেতে হলে রেজাল্ট ৩.২৫থাকতে হবে, পার্মানেন্ট ক্যাম্পাস এর জন্য)

এইচ এস সি ৪.৯০-৪.৯৯ =২০%(শুধুমাত্র ১ম সেমিস্টার এর জন্য, পরের সেমিস্টার থেকে পেতে হলে রেজাল্ট ৩.০থাকতে হবে, মাইন ক্যাম্পাস ও উত্তরা ক্যাম্পাস এর জন্য)

এইচ এস সি ৪.৯০-৪.৯৯ =৪০%(শুধুমাত্র ১ম সেমিস্টার এর জন্য, পরের সেমিস্টার থেকে পেতে হলে রেজাল্ট ৩.০থাকতে হবে, পার্মানেন্ট ক্যাম্পাস এর জন্য)

মুক্তিযুদ্ধ কোটা =১০০%%(শুধুমাত্র ১ম সেমিস্টার এর জন্য, পরের সেমিস্টার থেকে পেতে হলে রেজাল্ট ৩.০থাকতে হবে)

Figure 3.3: Answer of Figure 3.2 Questions

Here on figure 3.2 we gave the answer of the questions of figure 3.1. We will process the asked question and try to detect in which documents it would be matched. Then based on that decision we will provide the answer like figure 3.2.

3.4 Statistical Analysis

We proposed a simple method by which the user can ask questions as an input text. Then the machine will process the text and output the question's answer. From the input text, every word will be extracted and make a feature vector for each word using Bag of Words

(BOW). BOW will be created from the dataset and it is a kind of dictionary of all words along with its appearance.

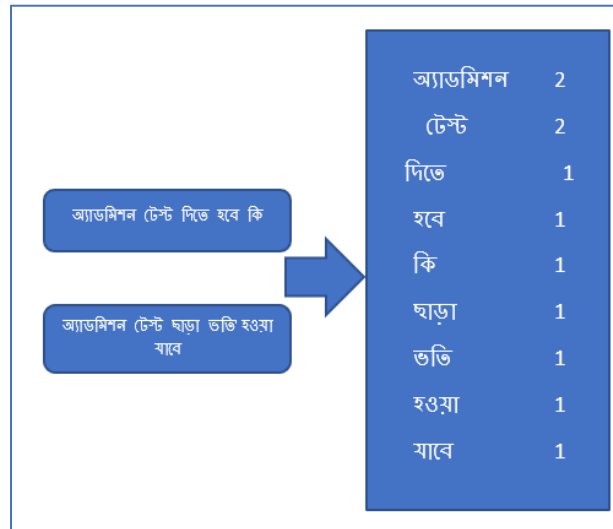


Figure 3.4: BOW sample

Then the feature vector will be classified by the text classified algorithm and give a prediction about the question's answer. Here we used Naive Bayes text classification algorithm.

3.4.1 Naive Bayes algorithm

Naive Bayes algorithm is a very popular probabilistic algorithm which is vastly used in text classification and disease type classification. Generally, Naive Bayes is very effective on small size data set.

There are five types of Naive Bayes method in the scikit-learn library (Blondel, et al., 2011):

1. Gaussian Naive Bayes
2. Multinomial Naive Bayes
3. Complement Naive Bayes
4. Bernoulli Naive Bayes
5. Out-of-core naive Bayes model fitting

Among these methods, we use Multinomial Naive Bayes because Multinomial Naive Bayes used in text classification [5].

The basic equation of Naive Bayes is:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \quad (1)$$

In general English it is written as:

$$posterior = \frac{prior \times likelihood}{evidence}$$

3.4.2 Multinomial Naive Bayes algorithm

In this project we use Multinomial Naive Bayes to process Bangla text. The equation of Multinomial Naive Bayes is depending on the support vector θ . Here θ represented

$$\theta = (\theta_1, \theta_2, \dots, \theta_n) \quad [y = class] \quad (2)$$

The main equation of Multinomial Naive Byes is:

$$\theta_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha n} \quad (3)$$

Where $N_{yi} = \sum_{x \in T} x_i$ is the number of i appear in a class y from the documents T and $N_y = \sum_{i=1}^n N_{yi}$ is the total number of all feature for class y .

3.5 Implementation Requirements

For implementing this project by Naive Bayes at first, we separate this into three different part. They are:

1. Creating BOW
2. Preparing training and testing data
3. Creating model

NB: here we use only two types of data. First one about the weaver and the second one is about admission test.

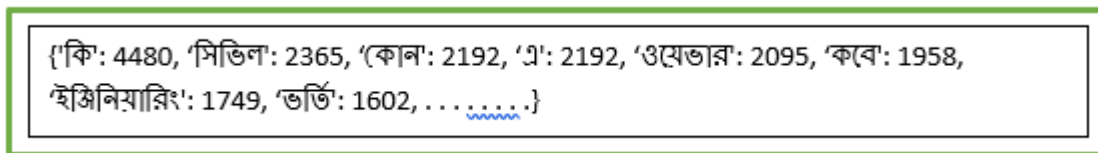
3.5.1 Creating BOW

To create the BOW we need to go through all the documents (data). From the documents, words should be separated. To do that we use **word_tokenize** from **nlTK** library and store them in a list. Then using the **Counter** function, we make the BOW from the list as a **dictionary**. The dictionary has the specific word and the number of how many times it appears in the whole documents. The pseudocode is here:

Import tokenizer

```
1.def make_dictionary:
2.directory = "path_of_the_data"
3.all_data = readAllData()
4.sentence = extractAllSentenceWith("\n") #if there is any \n program
   will take
5.that point as start of new sentence
6.words = tokenizer(sentence)
7.dictionary = counter(words) #creating Dictionary
8.return dictionary
```

And the BOW would look like that:



```
{'কি': 4480, 'সিভিল': 2365, 'কোন': 2192, 'এ': 2192, 'ওয়েভার': 2095, 'কবে': 1958, 'ইঞ্জিনিয়ারিং': 1749, 'ভর্তি': 1602, .....}
```

Figure 3.5: Part of the BOW

Here the first index data of the dictionary (figure 3.5) is the word and the second index data are how many times the word appears in the whole documents. This dictionary is necessary for preparing training and testing dataset.

3.5.2 Preparing training and testing data

For preparing data for training and testing we need to convert text data into a numeric form. For that, we need to convert each word into a feature vector. There are two types of features in the dataset. One of them is the frequency of any words and another one is from which document the word is come from. Here we mark the level of the documents as 1 and 2

respectively for number 1 document and number 2 document. Here is an example of the vectorize form of the string data:

	ভর্তি	তারিখ	স্কলারশিপ	কবে	হবার	কেমন	থেকে	শেষ
0	1	0	0	1	0	0	1	0
1	0	0	1	0	0	1	0	0
3	1	1	0	1	1	0	0	1

Figure 3.6: Frequency distribution matrix table (feature vector)

Here is the pseudocode of creating text to feature vector:

1. *def make_dataset(dictionary):*
2. *files = getAllFiles("file path")*
3. *for each file:*
4. *sentence = extractAllSentenceWith("\n") #if there is any \n program will take that point as start of new sentanse*
5. *words = tokenizer(sentence)*
6. *featureSet = matchEatchWord with dictionary*
7. *label = from which file the word is collected*
8. *return featureSet, label*

From this **make_dataset** function we will get the two different features:

1. Feature set and
2. Label

3.5.3 Creating model

To create training model, we use **sklearn** library. From this library to use multinomial Naive Bayes, we need to import **MultinomialNB** from **sklearn.naive_bayes**. To separate

training and test data, we need to import **train_test_split** from **sklearn.model_selection**.

The code is given below:

1. *featureSet, labels = make_dataset(dictionary)*
2. *trainData, testData = tts(features, lbels, test_size = 0.2)*
3. *classifier = MultinomialNB()*
4. *classifier.fit(trainData)*
5. *predict = classifier.predict(testData)*
6. *save(result)*

From this code, the train model will save as **.mdl** format.

CHAPTER 4

Experimental Results and Discussion

4.1 Introduction

In every project the most important thing is the output result. If the result is not good then the whole work will be ruined. A good accuracy will be the leading step towards the path of any work. In this research we got a good accuracy result. But we also face some problems here too. Despite of all problems we got 92% accuracy result. It could be better if we could overcome those problems.

4.2 Experimental Results

The accuracy of our chatbot is quite good. Though there is a chatbot in Bangla named Golpo [1] but it is implemented using chatterbot. In this project we use Multinomial Naive Bayes and we got 92% accuracy.

4.3 Descriptive Analysis

Here is the input-output table:

Table 1: Input-Output table

Input	Output	Correct/wrong answer
৬.১২লাখ থেকে কি কোন % বা ওয়েভার পাবো না	আপনার সেমিস্টার রেজাল্ট এর উপর নিভ'র করবে। এইচ এস সি ও এস এস সি এ+(গোল্ডেন) =১০০% (শুধুমাত্র ১ম সেমিস্টার এর জন্য পরের সেমিস্টার থেকে পেতে হলে)	correct
িভিল ইঞ্জিনিয়ারিং ডিপার্টমেন্টে এডমিট হলে কি রকম ডিসকাউন্ট পাব	আপনার সেমিস্টার রেজাল্ট এর উপর নিভ'র করবে। এইচ এস সি ও এস এস সি এ+(গোল্ডেন) =১০০% (শুধুমাত্র ১ম সেমিস্টার এর জন্য পরের সেমিস্টার থেকে পেতে হলে)	correct
ি বি এ তে অ্যাডমিশন টেস্ট কবে	আমাদের বছরে ৪মাস করে ৩ টি সেমিস্টার। স্প্রিং (জানুয়ারি, ফেব্রুয়ারি, মার্চ, এপ্রিল) সামার(মে, জুন, জুলাই, আগস্ট)	correct

	ফল(সেপ্টেম্বর, অক্টোবর, নভেম্বর, ডিসেম্বর) আপনাকে, সামার এর জন্য ডিসেম্বর- জানুয়ারি স্প্রিং এর জন্য এপ্রিল-মে ফল এর জন্য আগস্ট- সেপ্টেম্বর এ অফিসে যোগাযোগ করতে হবে।	
ঠিক কত তারিখ এর মধ্যে অ্যাডমিট হতে পারবো	আমাদের বছরে ৪মাস করে ৩ টি সেমিস্টার। স্প্রিং (জানুয়ারি, ফেব্রুয়ারি, মার্চ, এপ্রিল) সামার(মে, জুন, জুলাই, আগস্ট) ফল(সেপ্টেম্বর, অক্টোবর, নভেম্বর, ডিসেম্বর) আপনাকে, সামার এর জন্য ডিসেম্বর- জানুয়ারি স্প্রিং এর জন্য এপ্রিল-মে ফল এর জন্য আগস্ট- সেপ্টেম্বর এ অফিসে যোগাযোগ করতে হবে।	correct
স্প্রিং এর ভর্তি কবে থেকে	আমাদের বছরে ৪মাস করে ৩ টি সেমিস্টার। স্প্রিং (জানুয়ারি, ফেব্রুয়ারি, মার্চ, এপ্রিল) সামার(মে, জুন, জুলাই, আগস্ট) ফল(সেপ্টেম্বর, অক্টোবর, নভেম্বর, ডিসেম্বর) আপনাকে, সামার এর জন্য ডিসেম্বর- জানুয়ারি স্প্রিং এর জন্য এপ্রিল-মে ফল এর জন্য আগস্ট- সেপ্টেম্বর এ অফিসে যোগাযোগ করতে হবে।	correct
নেক্সট অ্যাডমিশন কবে	আমাদের বছরে ৪মাস করে ৩ টি সেমিস্টার। স্প্রিং (জানুয়ারি, ফেব্রুয়ারি, মার্চ, এপ্রিল) সামার(মে, জুন, জুলাই, আগস্ট) ফল(সেপ্টেম্বর, অক্টোবর, নভেম্বর, ডিসেম্বর) আপনাকে, সামার এর জন্য ডিসেম্বর- জানুয়ারি স্প্রিং এর জন্য এপ্রিল-মে ফল এর জন্য আগস্ট- সেপ্টেম্বর এ অফিসে যোগাযোগ করতে হবে।	correct
আশুলিয়া কাম্পাসে কি ২০% ওয়েভার পাওয়া যাবে	আমাদের বছরে ৪মাস করে ৩ টি সেমিস্টার। স্প্রিং (জানুয়ারি, ফেব্রুয়ারি, মার্চ, এপ্রিল) সামার(মে, জুন, জুলাই, আগস্ট) ফল(সেপ্টেম্বর, অক্টোবর, নভেম্বর, ডিসেম্বর) আপনাকে, সামার এর জন্য ডিসেম্বর- জানুয়ারি স্প্রিং এর জন্য এপ্রিল-মে ফল এর জন্য আগস্ট- সেপ্টেম্বর	wrong

	এ অফিসে যোগাযোগ করতে হবে।	
কোন ইন্সলারশিপ নাই	আপনার সেমিস্টার রেজাল্ট এর উপর নির্ভর করবে। এইচ এস সি ও এস এস সি এ+(গোল্ডেন) =১০০% (শুধুমাত্র ১ম সেমিস্টার এর জন্য পরের সেমিস্টার থেকে পেতে হলে	correct
ভিল এ স্কলারশিপ কেমন	আপনার সেমিস্টার রেজাল্ট এর উপর নির্ভর করবে। এইচ এস সি ও এস এস সি এ+(গোল্ডেন) =১০০% (শুধুমাত্র ১ম সেমিস্টার এর জন্য পরের সেমিস্টার থেকে পেতে হলে	correct

From that table, we can say that it is not a good chatbot at all. But it could be a milestone on the way of Bangla text classification. We find out some problems that are mainly responsible for this kind of poor accuracy result. Here are the problems:

1. In python, there is no library like word_tokenize for Bangla language. For that, some words like “খাইতেছি, খাবো, খাওয়া” will be counted as different word, though they are the same word in different tense.
2. There is no word net for Bangla language. So, it is very difficult to differentiate similar words. For example: “পৃথিবী, দুনিয়া, ধরিত্রী” these three words are different but have same meaning.
3. We can not use stopword for Bangla text. As a result, many unexpected words will appear in the BOW.
4. Nowadays we are writing English words using Bangla font. As a result, the same meaning words will be counted as a different word. For example, “পরীক্ষা” words written as “অ্যাডমিশন”.
5. We worked on the FAQ asked by the students or the guardians. We collect data from different social media and admission office. As a result, the amount of data is not enough for a good result.

If we could overcome these problems hope that we can get better accuracy for Bangla text.

4.4 Summary

Here we work with Multinomial Naive Bayes to classify the Bangla text. We choose this classifier because this can be performed though there is a small amount of data.

This chatbot can take input from user and give the output based on the classifier algorithm. But in here we only worked with two types of data set. And we found a poor accuracy result. We use python 3.6 to implement this project. And choose PyCharm as IDE. We collect data from different social media and from the admission office and from the department of CSE. In near future we want to do more work to improve this chatbot so that it can be chat like human.

CHAPTER 5

Conclusion and Implication for Future Research

5.1 Conclusion

Though the accuracy of this project is good but this work is not enough. The main target of this project is to show how accurately text processing algorithm works on Bangla language. From this project result, we can say that we still need to do a lot of work on Bangla text processing to create a better way to process Bangla text and make a standard model on that purpose. As high configuration computer, GPU then the accuracy would have been better as this model gave the result. Apart from, can produce a better output of this research work by using an efficient dataset.

5.2 Implication for Further Study

Until we cannot solve the mentioned problem in Chapter 4, it is quite difficult to get a good result. So, in future, we want to work on How to improve Bangla text tokenizer. Beside this, we also want to make a comparison research paper in which we will compare different text processing algorithm's result based on Bangla text. Beside this we want to do more work to improve this chatbot. We will try to make an open domain chatbot in future.

REFERENCES

- [1] T. Orin Dewan, “*Implementation of a Bangla chatbot*”. BRAC University, 04-Apr.-2017.
- [2] “chatterbot-corpus/chatterbot_corpus/data at master · gunthercox ...”. [Online]. Available: <https://github.com/gunthercox/chatterbot-corpus/>
- [3] F. Alam, S. Habib, and M. Khan, “*Text normalization system for Bangla*”. 2006.tree/master/chatterbot_corpus/data. [Accessed: 14-Apr.-2019].
- [4] M. N. Uddin and S. A. Khan, "A study on text summarization techniques and implement few of them for Bangla language" 2007 10th international conference on computer and information technology, Dhaka, 2007, pp. 1-4 [Accessed: 14-Apr.-2019].
- [5] “1.9. Naive Bayes — scikit-learn 0.20.3 documentation”. [Online]. Available: http://scikit-learn.org/stable/modules/naive_bayes.html. [Accessed: 14-Apr.-2019].
- [6] “The History of Chatbots [INFOGRAPHIC]”. [Online]. Available: <https://futurism.com/images/the-history-of-chatbots-infographic>. [Accessed: 14-Apr.-2019].
- [7] and J. Weizenbaum, “Eliza: a Computer Program for the Study of Natural Language Communication Between Man and Machine”, 1965.
- [8] M. Islam, Research on Bangla language processing in Bangladesh: progress and challenges. 8th international language \& development conference, 2009.
- [9] Z. Islam, A. Mehler, and R. Rahman, Text Readability Classification of Textbooks of a Low-Resource Language. Faculty of Computer Science, Universitas Indonesia, 2012.
- [10] “International Mother Language Day - Wikipedia”. [Online]. Available: https://en.wikipedia.org/wiki/International_Mother_Language_Day. [Accessed: 15-Apr.-2019].
- [11] S. Islam, M. Sarkar Farhana, T. Hussain, and M. Hasan Mehedi, “*Bangla Sentence Correction Using Deep Neural Network Based Sequence to Sequence Learning*”. 2018.
- [12] “What’s All The Fuss About WhatsApp? China’s WeChat Is a Worthy ...”. [Online]. Available: <http://time.com/8873/whats-all-the-fuss-about-whatsapp-chinas-wechat-is-a-worthy-rival/>. [Accessed: 15-Apr.-2019].

Appendices

Appendix A: Research Reflection

Here we work with Multinomial Naive Bayes to classify the Bangla text. We choose this classifier because this can be performed though there is a small amount of data.

This chatbot can take input from user and give the output based on the classifier algorithm. But in here we only worked with two types of data set. And we found a poor accuracy result. We use python 3.6 to implement this project. And choose PyCharm as IDE. We collect data from different social media and from the admission office and from the department of CSE. In near future we want to do more work to improve this chatbot so that it can be chat like human.

Appendix B: Related Issues

1. In python, there is no library like word_tokenize for Bangla language. For that, some words like “খাইতেছি, খাবো, খাওয়া” will be counted as different word, though they are the same word in different tense.
2. There is no word net for Bangla language. So, it is very difficult to differentiate similar words. For example: “পৃথিবী, দুনিয়া, ধরিত্রী” these three words are different but have same meaning.
3. We can not use stopword for Bangla text. As a result, many unexpected words will appear in the BOW.
4. Nowadays we are writing English words using Bangla font. As a result, the same meaning words will be counted as a different word. For example, “পরীক্ষা” words written as “অ্যাডমিশন”.
5. We worked on the FAQ asked by the students or the guardians. We collect data from different social media and admission office. As a result, the amount of data is not enough for a good result.

Turnitin Originality Report

Processed on: 01-Apr-2019 13:41 +06
ID: 1103645506
Word Count: 2239
Submitted: 1

Similarity Index

6%

Similarity by Source

Internet Sources: 4%
Publications: 4%
Student Papers: 5%

Automated Closed Domain
Chatbot for DIU By Tipu
Sultan

1% match (student papers from 09-Apr-2016)

[Submitted to University of Maryland, University College on 2016-04-09](#)

1% match (publications)

["Social Transformation – Digital Way", Springer Nature America, Inc, 2018](#)

1% match (student papers from 15-Jul-2015)

[Submitted to Malaviya National Institute of Technology on 2015-07-15](#)

1% match (Internet from 26-Feb-2019)

http://orca.cf.ac.uk/111799/1/2_22_2019_Ciw_%20An%20op.pdf

1% match (student papers from 21-Mar-2019)

[Submitted to Birkbeck College on 2019-03-21](#)

1% match (student papers from 17-May-2006)

[Submitted to University of Bristol on 2006-05-17](#)

< 1% match (Internet from 26-Mar-2019)

https://upcommons.upc.edu/bitstream/handle/2099.1/11094/PFC_mem%c3%b2ria.pdf?isAllowed=y&sequence=1

< 1% match (student papers from 25-Mar-2010)

[Submitted to Baccalaureate School of Global Education on 2010-03-25](#)

< 1% match (Internet from 10-Jun-2011)

<http://impact.asu.edu/cse534fa06/projects/Yinghui.ppt>

< 1% match (Internet from 25-Apr-2013)

http://www.daff.gov.au/_data/assets/pdf_file/0014/1005071/ns13_15es.pdf

< 1% match (publications)

[Samir Puuska, Matti J. Kortelainen, Viljami Venekoski, Jouko Vankka. "Instant message classification in Finnish cyber security themed free-form discussion", 2016 International Conference On Cyber Situational Awareness, Data Analytics And Assessment \(CyberSA\), 2016](#)

CHAPTER 1 Introduction A chatbot is a kind of AI or a system which can continue a conversation as like as a human. There are two types of chatbot based on their knowledge- open domain and closed domain. An open domain chatbot can answer almost all questions perfectly. This type of chatbot needs to be very clever to answer any questions and should have a good common sense. On the other hand, close domain chatbot works on a fixed knowledge and generally cannot perform to answer all

types of questions. Close domain chatbot generally used to manage a specific or fixed work. On this view, our chatbot should also be a close domain chatbot because it will work on the FAQ of the university. Though this type of chatbot can be implemented by using decision tree easily but we want to show how machine learning works on Bangla text. Every day many students need much information about our University. Not only the students but also their parents and he who wants to admit to the university needs much information about the university. For this reason, we want to build a chatbot which will answer all of their questions easily. And we build this chatbot in Bangla so that, everyone can use this chatbot. Generally, it is an FAQ chatbot and its main purpose is answering the question asked by the users about the university. We collected data (questions asked by the students or the guardians) from different social media, admission office and from the department of the university. The main purpose of our paper is to find out the accuracy result for the data set and show how Naive Bayes classifier work on Bangla language.

1.1 Project Objective

The main objective of this project is to solve the problems of student's confusion and indecision. Many students have many questions about that varsity he/she wants to get admit or already admitted. By this project, we try to collect those questions and analysis that question and try to answer in a smart way in Bangla. This project also attaches parents of the students with University by answering their questions in Bangla. Besides all of these, we want to publish a research paper on Bangla text classifications.

1.2 Motivation

As a student when we want to get admission in a new university, we have many questions to fix our decision and to clear our confusion. Not only that after getting admission and the duration of our completing study we have to face many more questions and we need so much information about the university, it's rules, regulation and etc. So, we want to make our all confusion clear, helps us to make our decision fix (in time) by giving all the question answer and all the information clearly in a smart way. Again, Bangla is spoken by approximately 250–300 million people over the world. But there are a very small number of works is done in Bangla. Even there is no automated chatbot in Bangla like Siri or Cortana or Google Assistant. Even there is no closed domain chatbot too. So, we want to work with Bangla so that it could be beneficial for further research work.

1.3 Project Goal and Outcome

By this project, we want to answer FAQ automatically about the university in Bangla. We also want to publish a research paper on Bangla text classification.

CHAPTER 2 Chatbot

According to Wikipedia, a chatbot is such an artificial intelligence system or a computer program that can drive a conversation by listening and talking or by texting. In chatbot history, ELIZA was the first chatbot created by Joseph Weizenbaum in 1966 passed the Turing test and able to fool some user to think that it was a human they are talking to. In 1995 A.L.I.C.E. was the language processing bot performed well though it was filed the Turing test. In 2001 Smarterchild came with the features among quick data access and conversation like a human. In 2010 Apple introduced their AI bot SIRI which can do multiple tasks for the user. In 2012 Google brings Google Now for searching something with voice on google. Then in 2015 Amazon brings Alexa and Microsoft introduce their Cortana.

2.1 Related Works

The very first work on Bangla language processing was started on the 1980s in Bangladesh (Islam, 2009). But as far as we know there are a very few works on Bangla chatbot. Among them, a notable work is Golpo chatbot (Orin, 2017-04), which is implemented by a conversational dialogue engine named ChatterBot. Again, in a paper, the Bangla text normalization technique has shown and they got a good accuracy value (Alam, Murtoza, & Khan, 2008). Md. Nizam Uddin and Shakil Akter Khan worked on Bangla text summarization (Uddin & Khan, 2007).

CHAPTER 3 Methodology and Implementation

3.1 Proposed Method:

We proposed a simple method by which the user can ask questions as an input text. Then the machine will process the text and output the question's answer. Figure 3.1: Block

Diagram of Chatbot From the input text, every word will be extracted and make a feature vector for each word using Bag of Words (BOW). BOW will be created from the dataset and it is a kind of dictionary of all words along with its appearance. Figure 3.2: BOW sample Then the feature vector will be classified by the text classified algorithm and give a prediction about the question's answer. Here we used Naive Bayes text classification algorithm. 3.2 Naive Bayes algorithm: Naive Bayes algorithm is a very popular probabilistic algorithm which is vastly used in text classification and disease type classification. Generally, Naive Bayes is very effective on small size data set. There are five types of Naive Bayes method in the scikit-learn library (Blondel, et al., 2011): • [Gaussian Naive Bayes](#) • [Multinomial Naive Bayes](#) • Complement [Naive Bayes](#) • Bernoulli Naive Bayes • [Out-of-core naive Bayes model fitting](#) Among these methods, we use [Multinomial Naive Bayes](#) because Multinomial [Naive Bayes](#) used [in](#) text classification (Blondel, et al., 2011). The basic equation of Naive Bayes is: $P(c|x) = \frac{P(x|c)P(c)}{P(x)}$. In general English it is written as: $n_{rrerir} = n_{rinr} \times k_{ikckihnn} c_{viccncc}$ In this project we use Multinomial Naive Bayes to process Bangla text. The equation of Multinomial Naive Bayes is depending on the support vector θ . Here $\theta = (\theta_1, \theta_2, \dots, \theta_n)$ [$x = clarr$] The main equation of Multinomial Naive Byes is: $N_{xi} + \alpha \theta_{xi} = N_x + \alpha n$ Where $N_{xi} = \sum_{x \in T} x_i$ is the number of i appear in a class y from the documents T and $N_x = \sum_{i=1}^n N_{xi}$ is the total number of all feature for class y.

3.3 Implementation: For implementing this project by Naive Bayes at first, we separate this into three different part. They are: 1. Creating BOW 2. Preparing training and testing data 3. Creating model NB: here we use only two types of data. First one about the weaver and the second one is about admission test.

3.3.1 Creating BOW: To create the BOG we need to go through all the documents (data). From the documents, words should be separated. To do that we use word_tokenize from nltk library and store them in a list. Then using the Counter function, we make the BOG from the list as a dictionary. The dictionary has the specific word and the number of how many times it appears in the whole documents. The code is here: Figure 3.3: Creating BOW And the BOG would look like that: Figure 3.4: Part of the BOW Here the first index data of the dictionary is the word and the second index data are how many times the word appears in the whole documents. This dictionary is necessary for preparing training and testing dataset.

3.3.2 Preparing training and testing data: For preparing data for training and testing we need to convert text data into a numeric form. For that, we need to convert each word into a feature vector. There are two types of features in the dataset. One of them is the frequency of any words and another one is from which document the word is come from. Here we mark the level of the documents as 1 and 2 respectively for number 1 document and number 2 document. Figure 3.5: Frequency distribution matrix table (feature vector) Here is the code of creating text to feature vector: Figure 3.6: Creating data set for training and testing From this make_dataset function we will get the two different features: 1. Feature set and 2. Label

3.3.3 Creating model: To create training model we use sklearn library. From this library to use multinomial Naive Bayes, we need to import MultinomialNB from sklearn.naive_bayes. To separate training and test data, we need to import train_test_split from sklearn.model_selection. The code is given below: Figure 3.6: Creating model From this code, the train model will save as .mdl format.

3.4 Getting input and showing Output: In this portion, we take input from the user then using the train model and try to provide the answer as output. Here is the code: Figure 3.7: Code of getting input

CHAPTER 4 Result Analysis The accuracy of our chatbot is not so good. Though there is a chatbot in Bangla named Golpo (Orin, 2017-04) but it is implemented using chatterbot. We got 56% accuracy. Here is the [input-output table:](#)

Input	Output	Correct/wrong
correct	correct	correct
wrong	wrong	wrong
correct	wrong	wrong
wrong	correct	correct

correct ' From that table, we can say that it is not a good chatbot at all. But it could be a milestone on the way of Bangla text classification. We find out some problems that are mainly responsible for this kind of poor accuracy result. Here are the problems: 1. In python, there is no library like word_tokenize for Bangla language. For that, some words like " , , will be counted as different word, though they are the same word in different tense. 2. There is no word net for Bangla language. So, it is very difficult to differentiate similar words. For example: " , , these three words are different but have same meaning. 3. We can not use stopword for Bangla text. As a result, many unexpected words will appear in the BOW. 4. Nowadays we are writing English words using Bangla font. As a result, the same meaning words will be counted as a different word. For example words written as " ". 5. We worked on the FAQ asked by the students or the guardians. We collect data from different social media and admission office. As a result, the amount of data is not enough for a good result. If we could overcome these problems hope that we can get better accuracy for Bangla text.

CHAPTER 5 Conclusion and Future work

5.1 Conclusion: Though the accuracy of this project is not so good but this work can be inspired to move forward the research on Bangla text processing. The main target of this project is to show how accurately text processing algorithm works on Bangla language. From this project result, we can say that we still need to do a lot of work on Bangla text processing to create a better way to process Bangla text and make a standard model on that purpose.

5.2 Future work: Until we cannot solve the mentioned problem in Chapter 4, it is quite difficult to get a good result. So, in future, we want to work on How to improve Bangla text tokenizer. Beside this, we also want to make a comparison research paper in which we will compare different text processing algorithm's result based on Bangla text. Beside this we want to do more work to improve this chatbot.

APPENDIX Here we work with Multinomial Naive Bayes to classify the Bangla text. We choose this classifier because this can be performed though there is a small amount of data. This chatbot can take input from user and give the output based on the classifier algorithm. But in here we only worked with two types of data set. And we found a poor accuracy result. We use python 3.6 to implement this project. And choose PyCharm as IDE. We collect data from different social media and from the admission office and from the department of CSE. In near future we want to do more work to improve this chatbot so that it can be chat like human.

References

Alam, F., Murtoza, S., & Khan, M. (2008). Text normalization system for Bangla. semantic scholar.

Blondel, M., Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., . . . Duchesnay, E. (2011). 1.9. Naive Bayes. Retrieved from Scikit-learn: Machine Learning in Python: https://scikit-learn.org/stable/modules/naive_bayes.html

Islam, M. (2009). Research on Bangla language processing in Bangladesh: progress and challenges. 8th international language \& development conference, (pp. 23--25).

Orin, T. (2017-04). Implementation of a Bangla chatbot. BRAC University. Uddin, M. N., & Khan, S. A. (2007). A Study on Text Summarization Techniques and Implement Few of Them for Bangla Language. 2007 10th international conference on computer and information technology. (pp. 1-4). IEEE.