AUTOMATIC EMAIL RESPONSE TO USER QUERIES

by

SAI MAGESHVAR	**2013103047**
MATHAN PANDI	**2013103550**
SIDHARTH YATISH	**2013103564**

A project report submitted to the

FACULTY OF COMPUTER SCIENCE AND ENGINEERING

in partial fulfilment of the requirements for the award of the degree of

in

COMPUTER SCIENCE AND ENGINEERING

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ANNA UNIVERSITY, CHENNAI – 25

APRIL 2017

BONAFIDE CERTIFICATE

Certified that this project report titled **AUTOMATIC EMAIL RESPONSE TO USER QUERIES** is the *bonafide* work of **SAI MAGESHVAR (2013103047)**, **MATHAN PANDI (2013103550)** and **SIDHARTH YATISH (2013103564)** who carried out the project work under my supervision, for the fulfilment of the requirements for the award of the degree of Bachelor of Engineering in Computer Science and Engineering. Certified further that to the best of my knowledge, the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or an award was conferred on an earlier occasion on these or any other candidates.

Place: Chennai	**T.V. Geetha**
Date:	Senior Professor
	Department of Computer Science and Engineering
	Anna University, Chennai – 25

COUNTERSIGNED

Head of the Department,
Department of Computer Science and Engineering,
Anna University Chennai,
Chennai – 600025

ACKNOWLEDGEMENT

Foremost, we express our deep gratitude to our course coordinator **Dr. T.V. Geetha** for guiding us throughout the project. We appreciate her tolerance and ability to share her knowledge with us. We thank her for being supportive and encouraging to complete the project successfully. Without her support we could have never done this. We are extremely grateful to **Dr.D.Manjula**, Head of the Department of Computer Science and Engineering, Anna University, Chennai 25, for extending the facilities of the Department towards our project and for her unstinting support.

We are extremely grateful to our panel members **Dr.S.Valli**, Professor, Department of Computer Science and Engineering, Anna University, **Dr.S.Sudha**, Senior Assistant Professor, Department of Computer Science and Engineering, Anna University and **Dr.B.L.Velammal**, Assistant Professor, Department of Computer Science and Engineering, Anna University for their immense support and valuable suggestions for making the project successful. We thank our parents and all our friends for helping us throughout the project work.

SAI MAGESHVAR MATHAN PANDI SIDHARTH YATISH

ABSTRACT

Email remains to be one of the important means of communication till date. Several problems falling under various domains are put up as queries in the form of emails. Customer support plays a vital role in every company's growth. It employs huge human resource to satisfy every customer. These emails are read and are responded manually by individuals. Since the number of emails sent on a day is predominantly large, responding all such mails manually takes much time and labour. Human errors and delays makes the problem more complex than resolving it. Moreover the replies vary depending on stress factors and mood levels. This can reduce the quality of the customer support.

Automatic Email Response replies to user queries falling under a specific domain automatically. The query is provided as an email. It is then analysed and appropriate responses are provided back to the user. The user then continues the thread by providing further queries or information about the previous queries. In such case, the context of the previous query is kept track and response is given accordingly.

Automating email responses saves more time and labour and is less prone to errors. For an organisation relying on manual customer support, it can prove to be a solution to minimise cost and time. Questions need to be domain specific to provide an appropriate solution as response.

ABSTRACT

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 OBJECTIVE

Users specify their queries in the form of emails. These emails have to be replied manually that consumes more time and is error prone. An automatic email response system can solve this problem by providing appropriate responses automatically. The response must be appropriate and in accordance to the context of the query. The time taken to provide appropriate response for user queries in any company providing customer support is large. Our objective is to reduce the manual labour and to reduce the time and errors that might occur when email replies are given manually in such customer support systems. The response volume for customer support is 63%. This has to be improved by automating the process. There are chances of biased responses based on the mood of the individual replying for that mail. This error can be minimised by providing automatic responses.

1.2 PROBLEM STATEMENT

Given a user query specifying a problem on a specific domain, the system must be able to choose appropriate solution for the given problem using cosine similarity[1] and provide the solution as an email. This solution must be provided with respect to the context. The system must keep track of the subject using Long Short term memory (LSTM)[4] when repeated or similar queries are given. Some queries may not contain a definite subject but may be reference of the previous query. In this case, the stored memory is used to find out what the query is actually about and the necessary response is given accordingly. Any repeated or empty emails have to be ignored. The generated response must be sent as in the form of email to the sender. The sender may then close or continue the thread.

1.3 NEED FOR THE SYSTEM

There is a need to answer millions of user queries relating to various domain everyday. This process involves manual labour and is time consuming. Manual finding of the solution may not be appropriate or exact. This can cause severe problems if the domain dealing with the problem can affect many people. Automating this process can help in reducing the manual labour and provide appropriate answers to the queries specified by the user. This system provides response for given medical queries. Almost every customer product comes with customer care. Nowadays, every customer query is responded with a set of FAQs and they seem to be irrelevant in most of the cases. To get a human support, it seems to be a tedious process and even after getting in contact with a support staff, explaining the problem seems challenging. Moreover, a support staff responds based on his/her training levels, mood swings, verbal ability and stress factors. Our system presents a method to solve this problem.

1.4 CHALLENGES

Response generation for an email only on the basis of template matching cannot give appropriate solutions. The system must be able to analyse the content and reply accordingly. The generated response must be suitable and in accordance to the context if the query. The response is mainly based on semi-supervised machine learning. This states the importance of training data. Right now, we have concentrated only on a specific domain, restricting the user input. This reduces the challenges faced regarding training data set. To improve the robustness of the system, huge data will be needed which isn't available right now. Moreover, filtering unwanted queries is a tedious task. Even though training has been done, edge cases keep growing and it greatly affects the response volume, spending time for unnecessary queries. Linguistic factors pose a huge challenge. Handling queries of improper format and sending a reply which is rich in language grammar has to be dealt seriously.

1.5 ORGANISATION OF THESIS

Chapter 2 discusses the existing approaches in generating automatic replies to email. Chapter 3 gives the requirements analysis of the system. It explains the functional and non-functional requirements, constraints and assumptions made. Chapter 4 explains the overall system architecture and the design of various modules. Chapter 5 elaborates on the results of the implemented system and gives an idea of its efficiency. It also contains information about the observations made during testing. Chapter 6 concludes the thesis and gives an overview of its criticisms.

CHAPTER 2

RELATED WORK

Introduction

This chapter gives a survey of the previous work done in providing automatic response to email. The replies are short and provides satisfying reply to the user email. Despite giving satisfying replies, they fail in identifying the context and provide a solution for the query specified in the email.

2.1 USING LSTM AND CLUSTERING

LSTM stands for long short term memory. This deals with choosing what to rememeber and what not. Feed forward neural networks take long time since they remember everything. This issue has been dealt by LSTM. The system is used to autocomplete mails. Whenever an user starts typing a reply, it suggests next phrase/word based on the typed context. This involved supervised learning and clustering. Relevant words were suggested based on clustering. Initially, clusters were created using supervised training. Once the user types in, it is classified and responses are generated based on the cluster it falls into. A system for providing smart replies to emails using clustering and LSTM[5]. This system used clustering to determine closest reply possible for a given email. These are termed smart replies. This system uses supervised learning. Response scoring was introduced to find out the most suitable response from the list of available ones. Convincing reply was given for a given email. The replies where short with no more than a few phrases. The reply was also given in reference to the context. Inappropriate queries were handled but there were problems of specificity.

2.2 USING GAUSSIAN DENSITY CLASSIFICATION

A system to respond to emails using machine learning[7]. It involved classification of FAQs and filtering. FAQ was classified using Gaussian Density classification. It is a non parametric classification. It is based on a Bayesian methodology. It assumes some prior distribution on the underlying probability densities that guarantees some smoothness properties. The final classification is then determined as the one that provides a good fit for the observed data, while at the same time guaranteeing smoothness. Based on training and classification, the category is identified and relevant responses are taken. Gaussian method proves to be effective incase of huge data with disturbances. Since it is parameter independent, it ignores the disturbances and works effectively. This helped in identifying the categories the query falls under and also helped in filtering the query. It helped in narrowing down the query to identify the categories. There is a need to identify the category quickly to categorise a query. To classify, the distribution had to be done previously. This had problems if assumption of the need of parameters was wrong.

2.3 USING TEMPLATE MATCHING

A system for automatic answering with template matching for natural language questions[6]. Natural Language Processing[2] was used to identify the template which matches for the current email.

A question may not exactly match with a given quer. Instead, it conveys the same meaning but with the use of different words or phrases. Naive string matching cannot find this. Disemvoweling[3] and synonym matching with naive template matching could match more set of questions for a given query. Any given query is checked for template by replacing with its synonyms. Once match is obtained, the response for the given template was given as answer. User queries in naive languages was also dealt. This invloved increasing the synonyms set including frequently used phrases for checking for match in template. Since this was naive template matching, the accuracy was less and was limited to a specific domain. Questions outside the domain was considered error.

Conclusion

All the previous works on providing a reply to an email focuses on giving a short reply to the email. The email is similar to an acknowledgement for the query given by the user. Our project focuses on providing a solution to the query rather that just acknowledgement.

CHAPTER 3

SYSTEM DESIGN

3.1 SYSTEM ARCHITECTURE

The block diagram of the automatic email response system is shown in Figure 3.1. The system aims in providing appropriate response to given user query. Naive Bayes classifier is used to find the category under which the query falls. LSTM (Long Short Term Memory)[4] is used to keep track of the subjects in multiple queries. From the given set of responses, most appropriate one is selected using Cosine similarity. The questions and their responses are entered into a database and used accordingly.

The input is preprocessed to remove any exceptions. Then SMS abbreviations are handled then it is given to FAQ mapper after removing stop words. FAQ is extracted and stored in database. It is then searched for exact query match. In case of its non-availability the classifier identifies the category under which the query falls. Queries similar to the user query are found out using cosine similarity[1]. In case of matching answer is found, the response mail is generated. Otherwise, keyword ranking is used to find the category and past response are kept track using LSTM to identify the context. In case no solution is available, that is notified to the user as a response itself. Finally header and footer is attached and it is sent as an email to the sender.

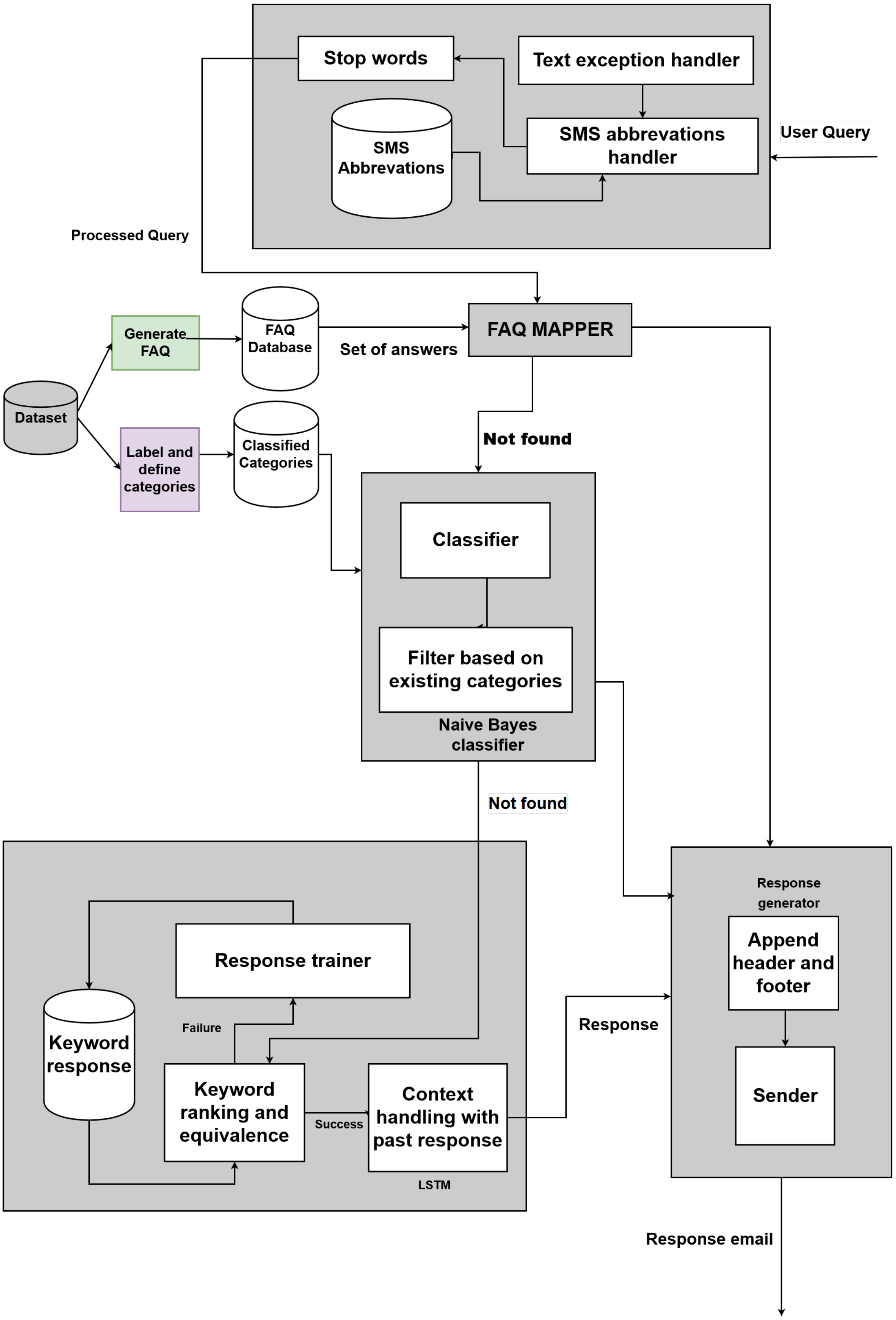


Figure 3.1: Block diagram of automatic email response system

3.2 SYSTEM MODELS

3.2.1 Use case diagram

The Overall usecase diagram is shown in Figure 3.2. A query is given in the form of email. The category must be found. A proper response must be generated if there is an appropriate answer for the query. Or else the response must be generated accordingly.
Pre-Condition: User query in the form of email.
Post-Condition: Response mail containing the solution for the given query.

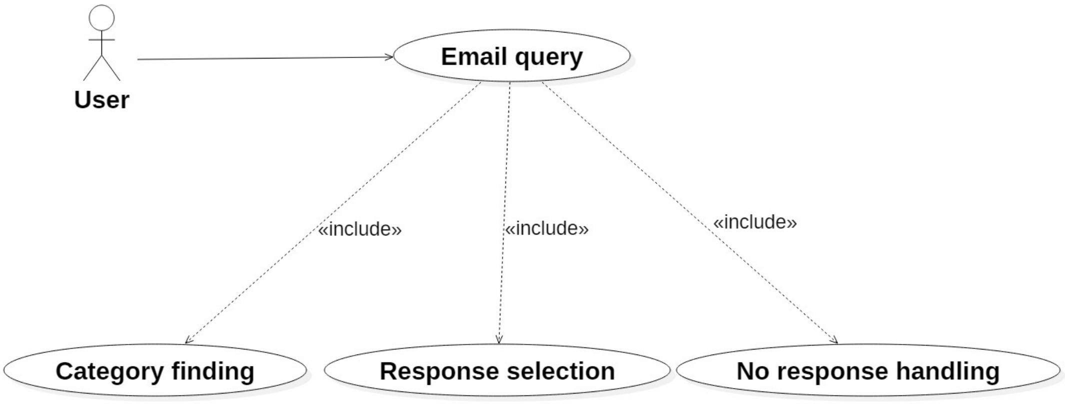


Figure 3.2: Overall use case diagram

The use case diagram for context handling block is shown in Figure 3.3. If the query does not fall under any specific category, keyword ranking has to be done to identify the subject. Context can be identified from the past responses. In case no past response is available, then it has to be trained accordingly.
Pre-Condition: Un-classified user query.
Post-Condition: User query with identified context.

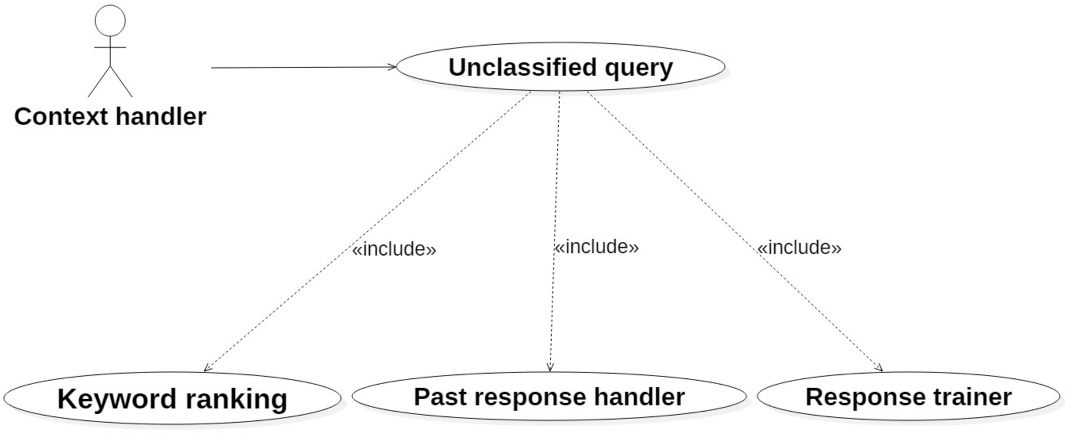


Figure 3.3: Context handler use case diagram

3.2.2 Sequence diagram

The sequence diagram and steps involved are shown in the the Figure 3.4. The input text enters the preprocessor where SMS abbreviations are handled and then searched for exact match in the FAQ. If found, it is sent to the response generator. Or else it is sent to the content filter to find the appropriate category. In case no category is found,

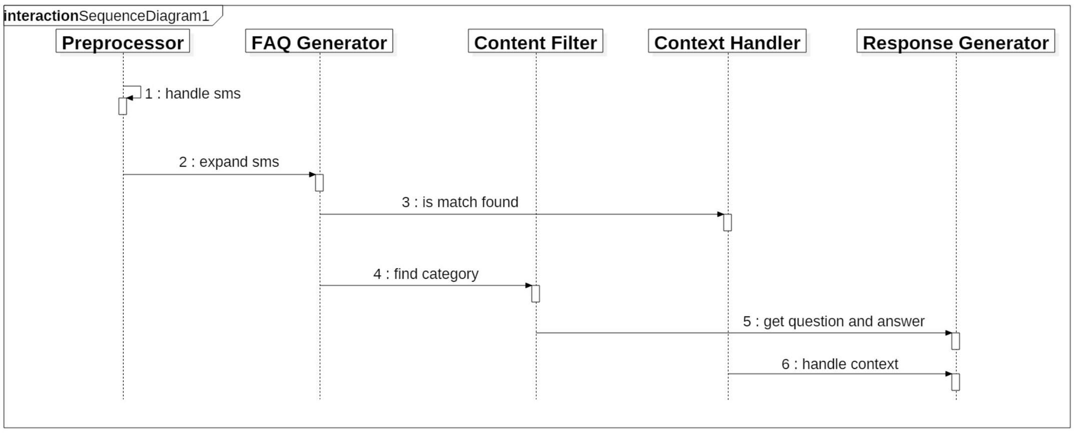


Figure 3.4: Sequence diagram

3.3 LIST OF MODULES

Automatic email response system consists of modules to perform specific functions listed below:

1. Preprocessing
2. FAQ Extraction
3. Content Filtering
4. Context Handling
5. Response Generation

3.4 MODULE DESIGN

3.4.1 Preprocessing

When it comes to large amounts of text, preprocessing seems to be a must-perform activity. Nowadays, since email allows humongous volume of text for free of cost, people tend to make us of them. They either send paragraphs of empty spaces or lines of meaningful text. It is upto the user. Processing the entire string without any idea about what it has is time consuming and not preferable. Preprocessing plays a vital role in this context. In our system, we deal with three cases: empty/repeated queries, SMS abbreviations, stop words. The query given by the user might be a repeated query or sometimes empty. These queries need to be ignored. This is done by Text Exception Handler that handles any such empty or repeated query and throws exception. Any SMS abbreviations must be handled. The list of SMS abbreviations are stored in a file. This is referred to substitute with their expanded form and replaced in the query. The SMS list consists of standard abbreviations and can be updated according to the user requirements. Final processing is removal of stop words. Removing stop words gives us more simpler query consisting of keywords that help us to classify them quickly. The output from the pre processing module is given to the content filter module where it is classified.

Input: User query with SMS conventions and/or with errors.

Output: User query with full form of SMS or Exception.

Algorithm

```
1. Q←Input Query
2. if(Q not NULL)
3.      W←Words from Q
4.      A1,A2,A3An←Abbreviations
5. if(W in Ai)
6.           Expand_ Abbreviations()
7.      S1,S2,S3Sn←Set of stop words
8. if(W in Si)
9.           Remove_ Stop_ Words()
```

3.4.2 FAQ extraction

As stated formerly, the system is mainly dependent on its training. Being non parametric, to handle disturbances and to attain stability, the system needs huge training set. Such data has been represented as set of FAQs. We refer drugs.com to get help with this. It has lot of questions and each question has a lot of answers. The entire website has been parsed to get required text. The questions are grouped based on medical conditions and they serve as category. Each question has several answers. This gives us an entity-relationship based on which we designed a database schema. Foreign integrity constraints are established and categories, questions and answers are stored in tables.

As said in objective, to make it faster and error free, this process has been made semi-autonomous. Once the path is set, the module parses the file and stores the category, question and answers along with their relationships. This saves a lot of time and reduces manual error that arises due to confusion. Medical queries are available as questions and answers referred to as Frequently Asked Questions of FAQ. These are available in various websites. For our dataset involving medical queries, questions from drug.com is parsed and stored in the form of relational tables. Data is extracted by parsing the web pages. The categories, questions as well as their corresponding answers are stored in separate tables.

Input: HTML Files of all the web pages of website containing data.

Output: Database tables with data populated from the parsed web pages.

Algorithm

```
F←each file in folder
    Q ←each question in F
    A1, A2, A3 ... An ← set of answers for Q
    N ← Answers size
    For i←1 to N
        Pair(Q,A)
    T←threshold Difference
    Q←Question from the query
    F1, F2, F3..., Fm ←FAQ set
    m←FAQ Size
    For j←1 to m
        V← Matching_ threshold(Fj,Q)
        If(V<=T) return Response(Q)
```

3.4.3 Content filter

The input from the pre processing module is given to the content filter block. The input query is directly matched with the list of available FAQs. In case a match is found the corresponding answer is given. If not the query has to be classified from the list of available categories. Naive Bayes Classifier is used to train the data with their respective categories and find the category from the learned data. The dataset containing the questions for various categories is trained with their respective category using Bayes Classification. Once trained, the trained set can be used to give a question as an input that returns the category under which it falls. So the question or the user query is given to this block and it returns the category. After finding the category, the possible matching questions are found out. Each category has several matching questions. Selecting all the questions as the response will be inappropriate. So the questions have to be sorted according to which they are closely related to the user query. From the list of questions nearest matching questions are found out using cosine similarity. The questions are sorted according to cosine similarity values and top three are taken into the response. The answers for these questions are also taken from the database. The questions and answers are part of the response to be generated to send to the user. The response list is sent to the response generator block.

Input: Pre processed user query.

Output: Category under which the query falls.

Algorithm

```
1. Q← Input query
2. C← classify (Q)
```

```
3. F1, F2, F3..., Fm ←FAQ set
4. S1, S2, S3..., Sn←Classified set
5.n←Classified set size
6. For i← 1 to n
7.      If(C = Si)
8.          T←threshold Difference
9.              For j←1 to m
10.               V← Matching_ threshold(Fj,Q)
11.               If(V<=T)
12.                   return Response(Q)
13.               Else
14                   (Si=synonym(Si))
15.      Repeat Step 8
16. Else return
```

3.4.4 Context handler

Context Handler plays an important role in the functioning of the system. A user may not be satisfied with a single response. Never ending doubts is human trait. He/She shouldn't be forced to refer to context every time. He/She should be given a feeling that the response is given by a human. To ensure this, the system should be able to identify the context that the user is referring to. This is achieved using LSTM. Main purpose of using LSTM is to decide what to remember and what not. It is purely constructed using states. Each state represents a user query and states are interrelated. Based on user's response, navigation can be done among different states. Each state can convey a message as well as give a response. There are states that are used to handle exception too. Users' every query is stored to get an idea of what he/she is seeking for and responses are chosen accordingly. There may be several responses for a query and in such case, response ranking must be done to give appropriate responses.If category is not found, then it is possible that the subject is unknown. This may be because of a subsequent thread not mentioning what the subject is about since it might have already been mentioned on previous threads. In such cases, keyword ranking is used to rank the keywords to identify the subject. Using Long Short Term Memory, the subject of the previous queries are stored. States are maintained for this purpose. For each thread under this subject is asked, a different state is attained. By identifying in which state the user is the subject can be identified. This helps in sending a response for a long thread with reference to the context.

Input: Unclassified query.

Output: Query with context identified.

Algorithm

```
Q1, Q2, Q3, Qn←Previous queries of current user
Q←Current Query
For i←n to 1
    If Qi in FAQ_ DB
        S←FAQ subject
    Else if Qi in categories_ FAQ
        S←Categories subject
    Else if Qi in keyword_ response
        S←keyword_ subject
```

3.4.5 Response generator

After the response is obtained in the form of question and answers, it has to be sent in the form of email to the intended user. This is done in response generator block. The response is put in the form of email and sent to the respective users using the necessary mailing protocol. The protocol, email store and email id of the user is obtained and the email is put into the existing thread. Every new mail of the same subject is grouped under similar threads as one mail. The previous mail and current mail is sent to the user. It must be made sure that same subject name is used so that it is grouped as similar thread. After the response has been appended with appropriate header and footer, it is send as an email.

Input: Response containing query and possible solutions

Output: Response with header and footer in the form of an email.

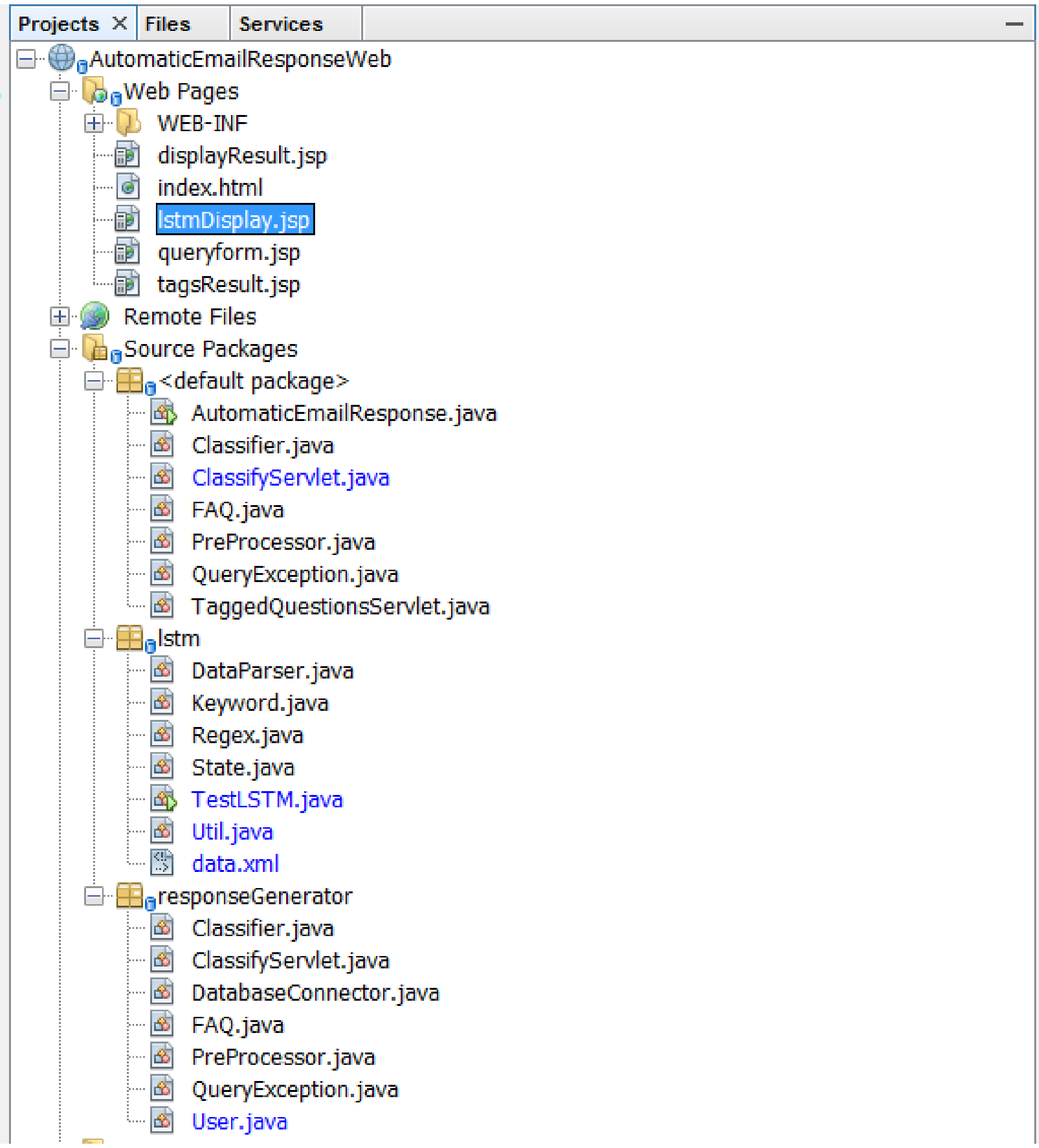
Algorithm

```
$U \leftarrow$ User
$C \leftarrow$ Conversations
for M Messages in C
    sort(M,TimeStamp)
    addToThread(M)
addHeader()
setEmailProtocol()
setPortNumber()
setStoreAccess()
$I \leftarrow$ mailId(U)
addFooter()
pushToServer()
```

CHAPTER 4

SYSTEM DEVELOPMENT

The system described consists of various packages like Classifier,LSTM, FAQ generator. The overall code overview showing the organisation of these various packages of the Automatic email response to user queries can be seen in Figure 4.1. The web pages include forms to get the email address and the query. Once the query has been submitted, the displayResult.jsp shows the possible matching questions and their answers. Clicking on a question displays the answers. The tags associated with the questions are also shown. Servlet packages include the ClassifyServlet. The LSTM packages include java files to store states using the Long Short Term Memory.



```
AutomaticEmailResponseWeb
    Web Pages
        WEB-INF
        displayResult.jsp
        index.html
        lstmDisplay.jsp
        queryform.jsp
        tagsResult.jsp
    Remote Files
    Source Packages
        <default package>
            AutomaticEmailResponse.java
            Classifier.java
            ClassifyServlet.java
            FAQ.java
            PreProcessor.java
            QueryException.java
            TaggedQuestionsServlet.java
        lstm
            DataParser.java
            Keyword.java
            Regex.java
            State.java
            TestLSTM.java
            Util.java
            data.xml
        responseGenerator
            Classifier.java
            ClassifyServlet.java
            DatabaseConnector.java
            FAQ.java
            PreProcessor.java
            QueryException.java
            User.java
```

Figure 4.1: List Of packages

An overview of the algorithm of entire system is shown below.

```
I ← Input Query
Q ← Every Query(F)
PROCESS(Q)
    1. C←CLASSIFY_ QUERY(W)
    2. L←FIND_ RESPONSE(C)
    3. R ← RELEVANT_ RESPONSES(L)
    4. Return R
```

4.1 PROTOTYPE ACROSS THE MODULES

The input and output to each module of the system is described in this section.

- **Preprocessor:** This module takes input query, handles exceptions and SMS abbreviations produces output as stop word removed query.
- **FAQ Extractor:** This module extracts question and answers related to specific domain from various web pages. It also aids in exact mapping of given query.
- **Content Filter:** In case exact match is not found, the category under which the query falls is found out using Naive Bayes Classification of the query and closest queries are found out using Cosine similarity[1]
- **Context Handler:** In case it is unable to identify the category, the subject is identified using keyword ranking. Previous queries are considered to keep track of the context using LSTM.
- **Response Generator:** Questions and set of answers are put together, appended header and footer and sent in the form of email.

4.2 RESPONSE GENERATION ALGORITHM

The algorithm for generation of response for given user query is :

```
I $\leftarrow$ Input Query
PROCESS(I)
1. E $\leftarrow$ handleException(I)
2. S $\leftarrow$ handleSMS(E)
```

```
Q ← removeStopWord(S)
if(exactMatch(Q))
    A ← getAnswers(Q)
else
    C ← classify(Q)
    K ← getQuestions(Q)
    S ← cosineSimilarity(K)
    Q’ ← sort(K,S)
    for(i←1 to 3)
        A ← getAnswers(Q’)
return A
```

4.3 DEPLOYMENT DETAILS

Any IDE like Netbeans can be used to successfully deploy the system. Internet connection is required. Glassfish server was used for hosting the servlet pages.

CHAPTER 5

IMPLEMENTATION AND RESULTS

5.1 DATASET DESCRIPTION

Dataset consists of medical related questions and their corresponding answers put in a database. Questions falling under 79 categories having more than 84,000 questions and more than 1,90,000 answers are present in the database. Categories are kept in a separate table named categories. Questions and answers are kept in separate tables. The database schema is shown in Figure 5.1. The tables in the database are answers, categories, questions, tags and users.

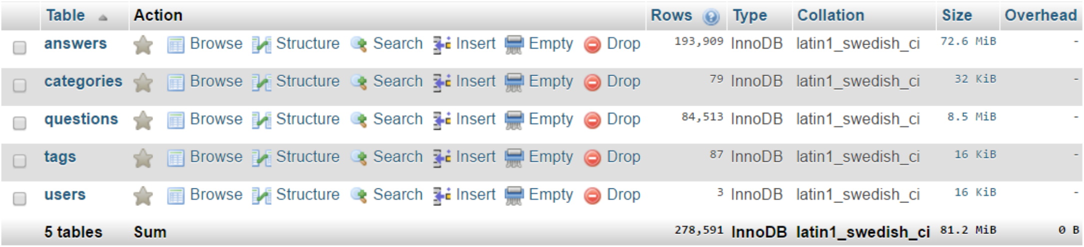
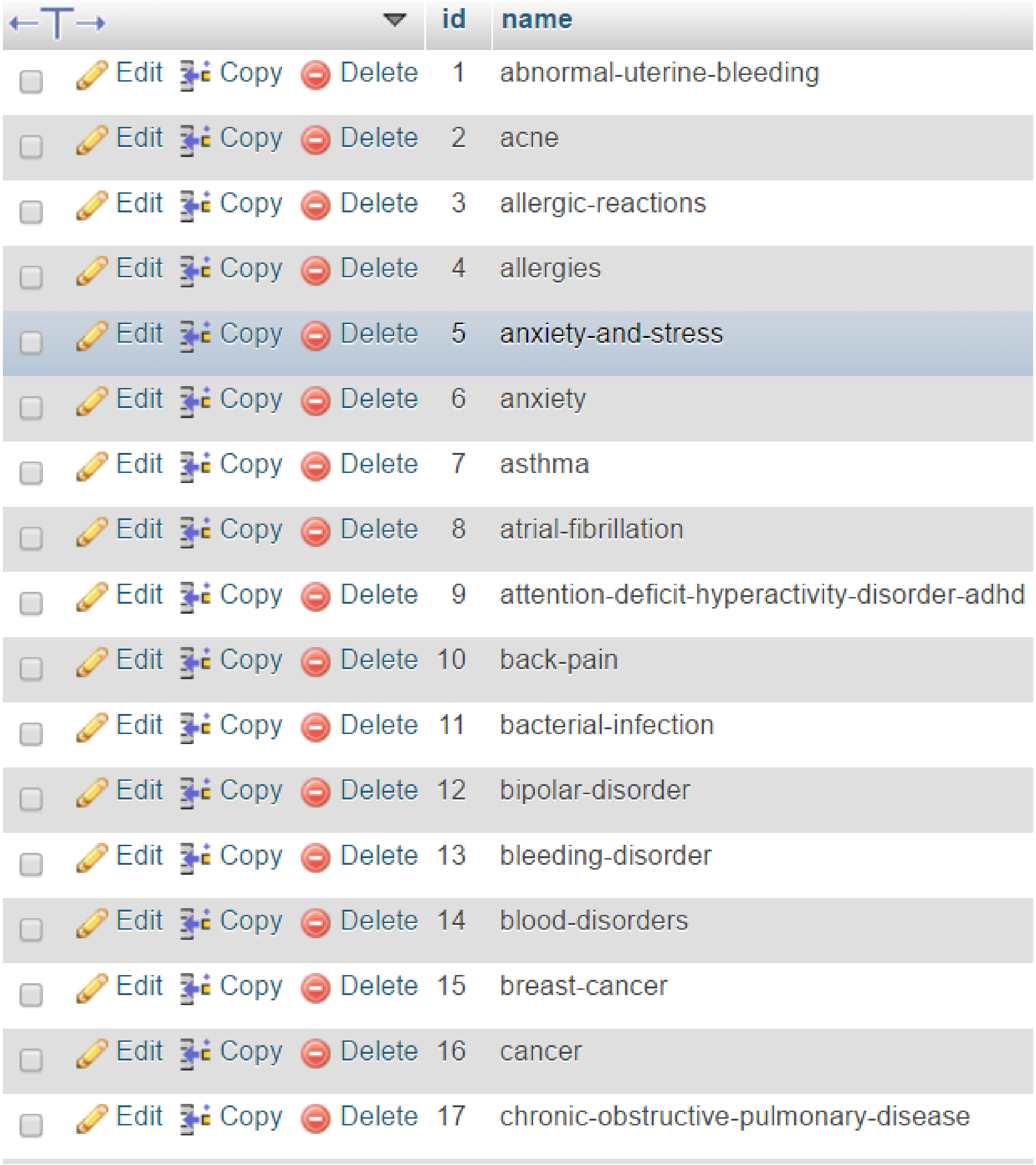


	Table ▲	Action							Rows	Type	Collation	Size	Overhead
□	answers	★	Browse	Structure	Search	Insert	Empty	Drop	193,909	InnoDB	latin1_swedish_ci	72.6 MiB	-
□	categories	★	Browse	Structure	Search	Insert	Empty	Drop	79	InnoDB	latin1_swedish_ci	32 KiB	-
□	questions	★	Browse	Structure	Search	Insert	Empty	Drop	84,513	InnoDB	latin1_swedish_ci	8.5 MiB	-
□	tags	★	Browse	Structure	Search	Insert	Empty	Drop	87	InnoDB	latin1_swedish_ci	16 KiB	-
□	users	★	Browse	Structure	Search	Insert	Empty	Drop	3	InnoDB	latin1_swedish_ci	16 KiB	-
	5 tables	Sum							278,591	InnoDB	latin1_swedish_ci	81.2 MiB	0 B

Figure 5.1: Database schema

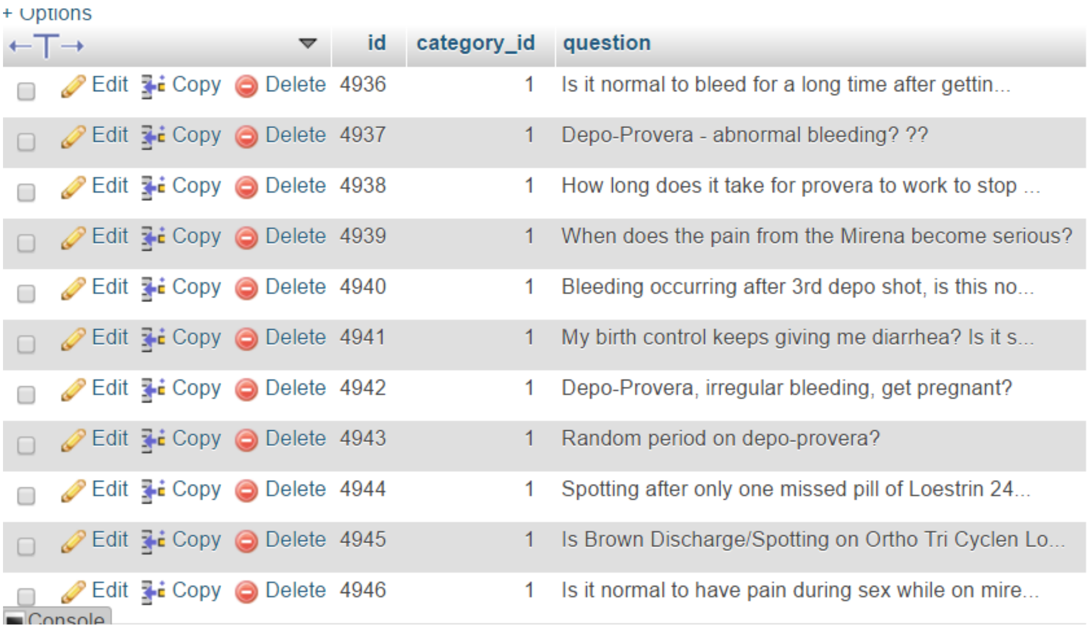
The categories table consists of list of categories along with their id. This is shown in Figure 5.2. Various medical categories are stored in this table along with an id.



←T→	▾	id	name
□	Edit Copy Delete	1	abnormal-uterine-bleeding
□	Edit Copy Delete	2	acne
□	Edit Copy Delete	3	allergic-reactions
□	Edit Copy Delete	4	allergies
□	Edit Copy Delete	5	anxiety-and-stress
□	Edit Copy Delete	6	anxiety
□	Edit Copy Delete	7	asthma
□	Edit Copy Delete	8	atrial-fibrillation
□	Edit Copy Delete	9	attention-deficit-hyperactivity-disorder-adhd
□	Edit Copy Delete	10	back-pain
□	Edit Copy Delete	11	bacterial-infection
□	Edit Copy Delete	12	bipolar-disorder
□	Edit Copy Delete	13	bleeding-disorder
□	Edit Copy Delete	14	blood-disorders
□	Edit Copy Delete	15	breast-cancer
□	Edit Copy Delete	16	cancer
□	Edit Copy Delete	17	chronic-obstructive-pulmonary-disease

Figure 5.2: Categories table

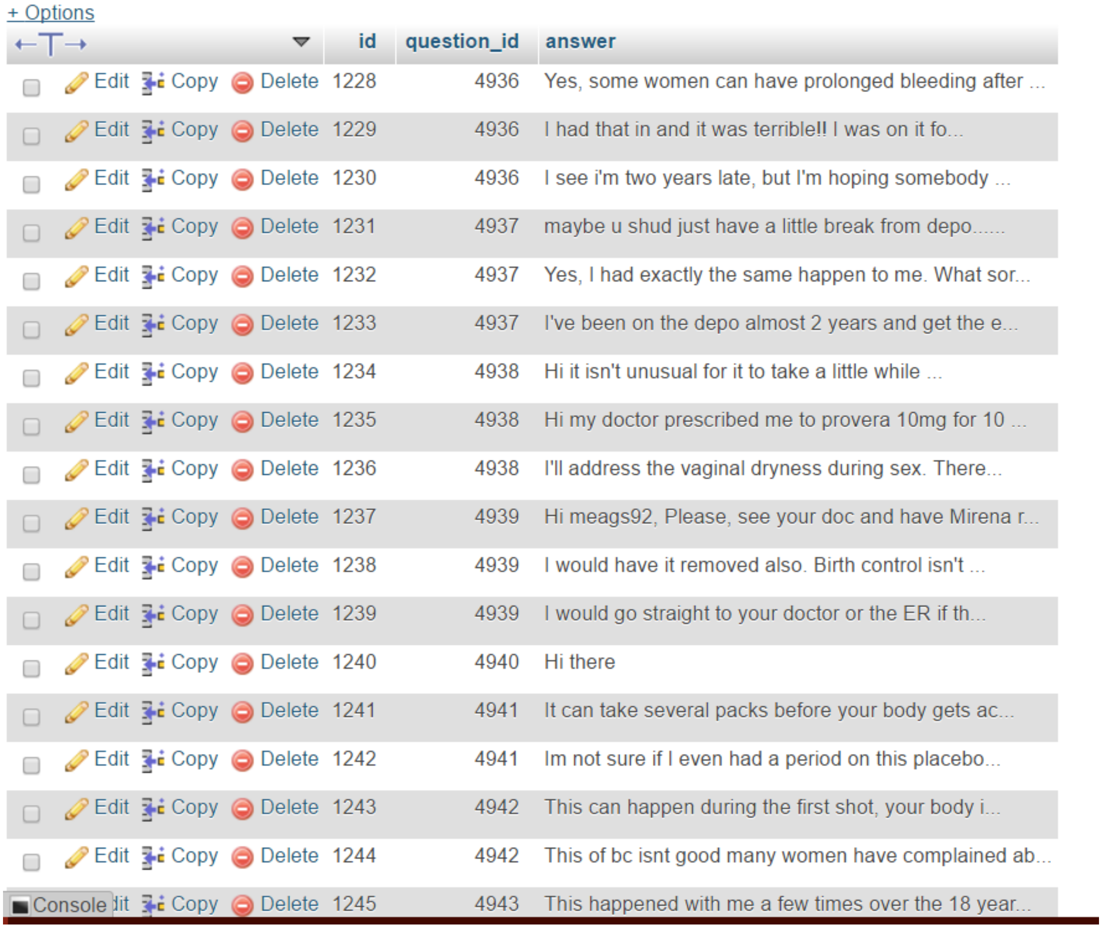
The questions table consists of question id, category id and question. Each question is associated with a category and this is shown in Figure 5.3.



		id	category_id	question
□	Edit Copy Delete	4936	1	Is it normal to bleed for a long time after gettin...
□	Edit Copy Delete	4937	1	Depo-Provera - abnormal bleeding? ??
□	Edit Copy Delete	4938	1	How long does it take for provera to work to stop ...
□	Edit Copy Delete	4939	1	When does the pain from the Mirena become serious?
□	Edit Copy Delete	4940	1	Bleeding occurring after 3rd depo shot, is this no...
□	Edit Copy Delete	4941	1	My birth control keeps giving me diarrhea? Is it s...
□	Edit Copy Delete	4942	1	Depo-Provera, irregular bleeding, get pregnant?
□	Edit Copy Delete	4943	1	Random period on depo-provera?
□	Edit Copy Delete	4944	1	Spotting after only one missed pill of Loestrin 24...
□	Edit Copy Delete	4945	1	Is Brown Discharge/Spotting on Ortho Tri Cyclen Lo...
□	Edit Copy Delete	4946	1	Is it normal to have pain during sex while on mire...

Figure 5.3: Questions table

The answers table consists of answer id, question id and answer. One question may have more than one answer. This is shown in Figure 5.4

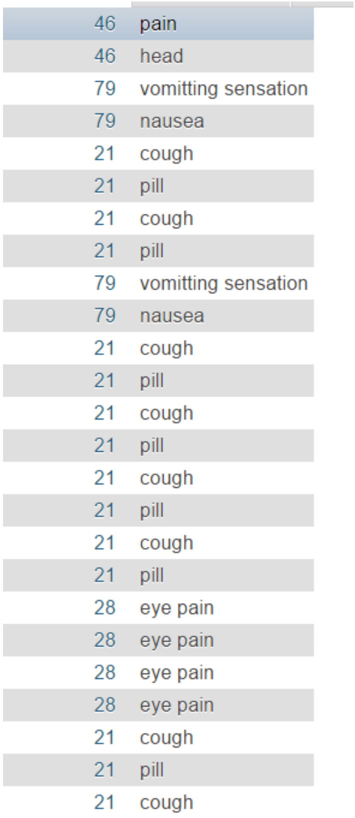


+ Options

←T→			▼	id	question_id	answer
□	Edit	Copy	Delete	1228	4936	Yes, some women can have prolonged bleeding after ...
□	Edit	Copy	Delete	1229	4936	I had that in and it was terrible!! I was on it fo...
□	Edit	Copy	Delete	1230	4936	I see i'm two years late, but I'm hoping somebody ...
□	Edit	Copy	Delete	1231	4937	maybe u shud just have a little break from depo......
□	Edit	Copy	Delete	1232	4937	Yes, I had exactly the same happen to me. What sor...
□	Edit	Copy	Delete	1233	4937	I've been on the depo almost 2 years and get the e...
□	Edit	Copy	Delete	1234	4938	Hi it isn't unusual for it to take a little while ...
□	Edit	Copy	Delete	1235	4938	Hi my doctor prescribed me to provera 10mg for 10 ...
□	Edit	Copy	Delete	1236	4938	I'll address the vaginal dryness during sex. There...
□	Edit	Copy	Delete	1237	4939	Hi meags92, Please, see your doc and have Mirena r...
□	Edit	Copy	Delete	1238	4939	I would have it removed also. Birth control isn't ...
□	Edit	Copy	Delete	1239	4939	I would go straight to your doctor or the ER if th...
□	Edit	Copy	Delete	1240	4940	Hi there
□	Edit	Copy	Delete	1241	4941	It can take several packs before your body gets ac...
□	Edit	Copy	Delete	1242	4941	Im not sure if I even had a period on this placebo...
□	Edit	Copy	Delete	1243	4942	This can happen during the first shot, your body i...
□	Edit	Copy	Delete	1244	4942	This of bc isnt good many women have complained ab...
Console	dit	Copy	Delete	1245	4943	This happened with me a few times over the 18 year...

Figure 5.4: Answers table

The tags contain multiple tags associated with the categories. So they include category id and tag name. This is shown in Figure 5.5

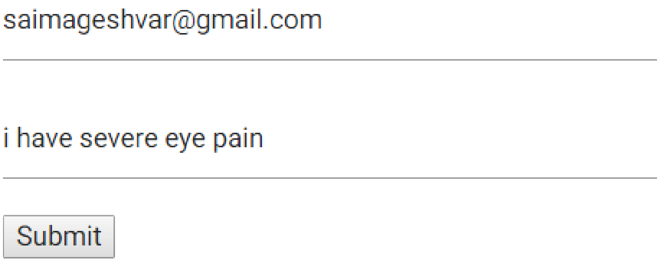


46	pain
46	head
79	vomitting sensation
79	nausea
21	cough
21	pill
21	cough
21	pill
79	vomitting sensation
79	nausea
21	cough
21	pill
21	cough
21	pill
21	cough
21	pill
21	cough
21	pill
28	eye pain
28	eye pain
28	eye pain
28	eye pain
21	cough
21	pill
21	cough

Figure 5.5: Tags table

5.2 EXPERIMENTAL RESULTS

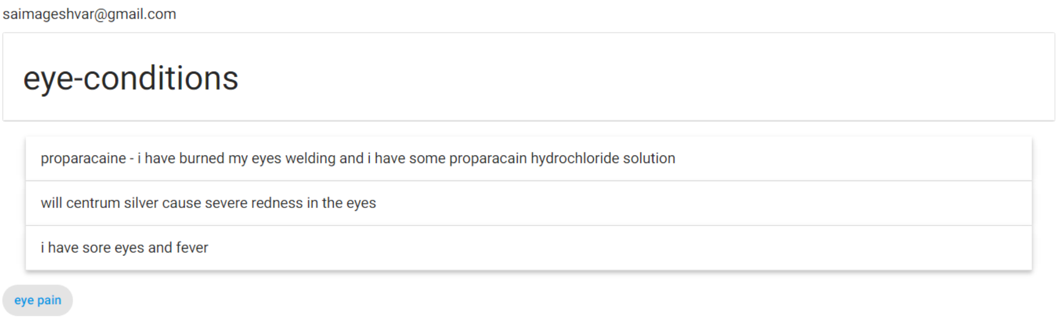
This section shows the results obtained during module testing. Various queries were given belonging to several categories. Their results are summarised here.



saimageshvar@gmail.com

i have severe eye pain

Figure 5.6: Input



saimageshvar@gmail.com

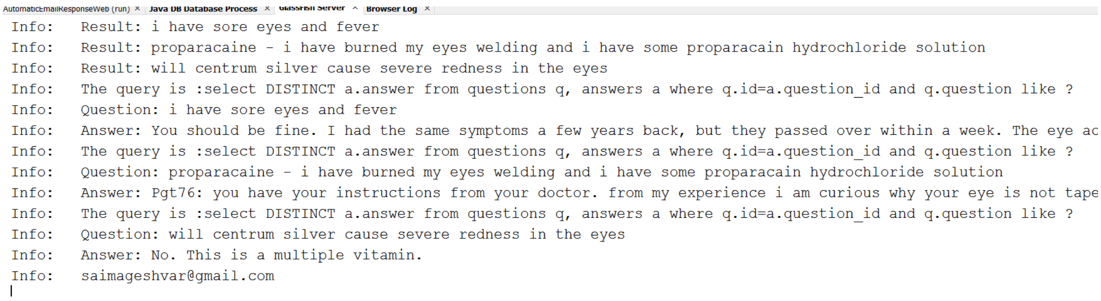
eye-conditions

proparacaine - i have burned my eyes welding and i have some proparacain hydrochloride solution

will centrum silver cause severe redness in the eyes

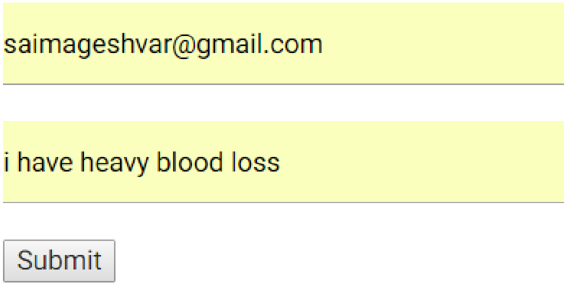
i have sore eyes and fever

Figure 5.7: Output



```
Info:   Result: i have sore eyes and fever
Info:   Result: proparacaine - i have burned my eyes welding and i have some proparacain hydrochloride solution
Info:   Result: will centrum silver cause severe redness in the eyes
Info:   The query is :select DISTINCT a.answer from questions q, answers a where q.id=a.question_id and q.question like ?
Info:   Question: i have sore eyes and fever
Info:   Answer: You should be fine. I had the same symptoms a few years back, but they passed over within a week. The eye ac
Info:   The query is :select DISTINCT a.answer from questions q, answers a where q.id=a.question_id and q.question like ?
Info:   Question: proparacaine - i have burned my eyes welding and i have some proparacain hydrochloride solution
Info:   Answer: Pgt76: you have your instructions from your doctor. from my experience i am curious why your eye is not tape
Info:   The query is :select DISTINCT a.answer from questions q, answers a where q.id=a.question_id and q.question like ?
Info:   Question: will centrum silver cause severe redness in the eyes
Info:   Answer: No. This is a multiple vitamin.
Info:   saimageshvar@gmail.com
```

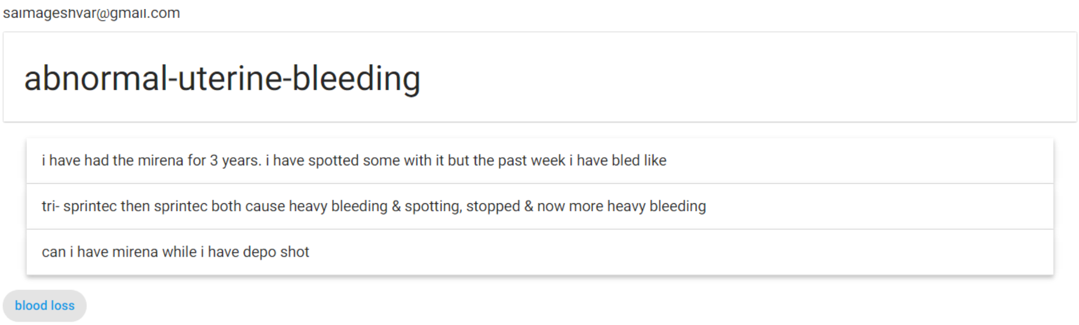
Figure 5.8: Log results



saimageshvar@gmail.com

i have heavy blood loss

Figure 5.9: Input



saimageshvar@gmail.com

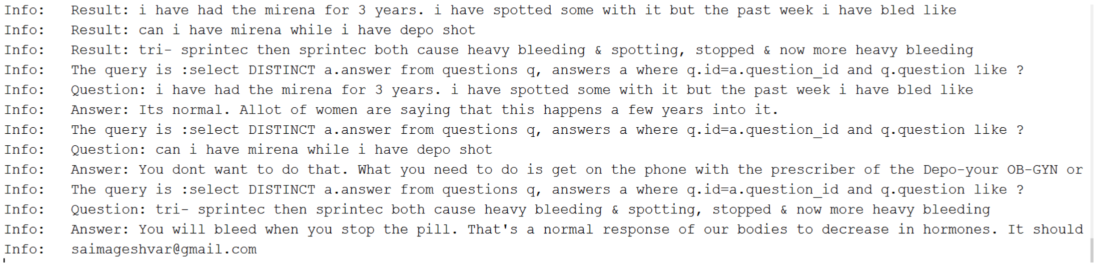
abnormal-uterine-bleeding

i have had the mirena for 3 years. i have spotted some with it but the past week i have bled like

tri- sprintec then sprintec both cause heavy bleeding & spotting, stopped & now more heavy bleeding

can i have mirena while i have depo shot

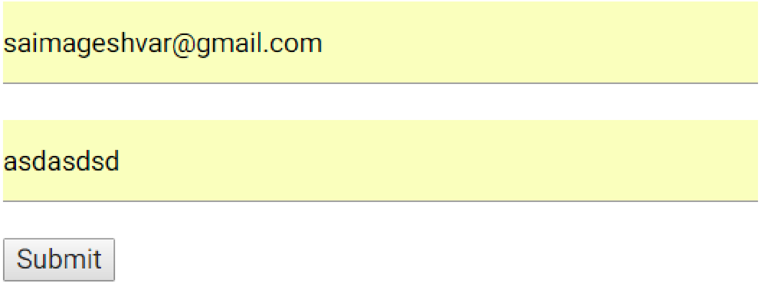
Figure 5.10: Output



```
Info:   Result: i have had the mirena for 3 years. i have spotted some with it but the past week i have bled like
Info:   Result: can i have mirena while i have depo shot
Info:   Result: tri- sprintec then sprintec both cause heavy bleeding & spotting, stopped & now more heavy bleeding
Info:   The query is :select DISTINCT a.answer from questions q, answers a where q.id=a.question_id and q.question like ?
Info:   Question: i have had the mirena for 3 years. i have spotted some with it but the past week i have bled like
Info:   Answer: Its normal. Allot of women are saying that this happens a few years into it.
Info:   The query is :select DISTINCT a.answer from questions q, answers a where q.id=a.question_id and q.question like ?
Info:   Question: can i have mirena while i have depo shot
Info:   Answer: You dont want to do that. What you need to do is get on the phone with the prescriber of the Depo-your OB-GYN or
Info:   The query is :select DISTINCT a.answer from questions q, answers a where q.id=a.question_id and q.question like ?
Info:   Question: tri- sprintec then sprintec both cause heavy bleeding & spotting, stopped & now more heavy bleeding
Info:   Answer: You will bleed when you stop the pill. That's a normal response of our bodies to decrease in hormones. It should
Info:   saimageshvar@gmail.com
```

Figure 5.11: Log results

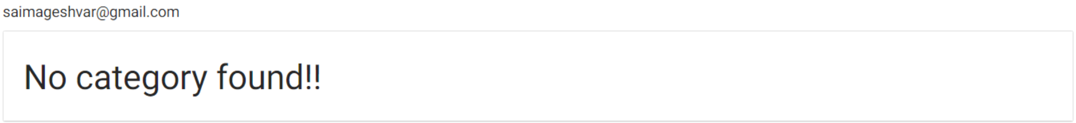
The following input is for a query that does not belong to any category.



saimageshvar@gmail.com

asdasdsd

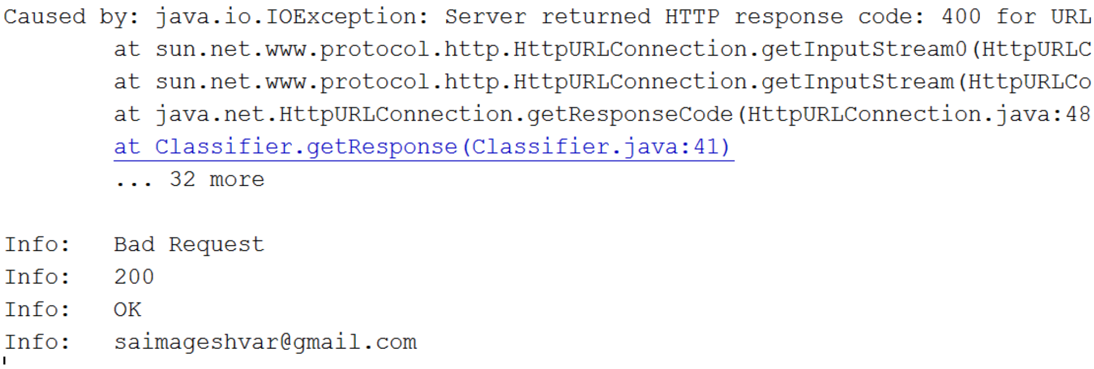
Figure 5.12: Input



saimageshvar@gmail.com

No category found!!

Figure 5.13: Output



```
Caused by: java.io.IOException: Server returned HTTP response code: 400 for URL
        at sun.net.www.protocol.http.HttpURLConnection.getInputStream0(HttpURLC
        at sun.net.www.protocol.http.HttpURLConnection.getInputStream(HttpURLCo
        at java.net.HttpURLConnection.getResponseCode(HttpURLConnection.java:48
        at Classifier.getResponse(Classifier.java:41)
        ... 32 more

Info:   Bad Request
Info:   200
Info:   OK
Info:   saimageshvar@gmail.com
```

Figure 5.14: Log results

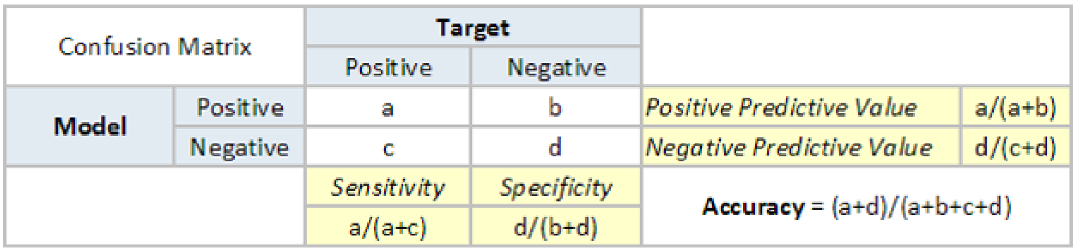
5.3 EVALUATION METRICS

5.3.1 Confusion matrix

Definition

A confusion matrix is an N X N matrix, where N is the number of classes being predicted.

Formula



Confusion Matrix		Target			
		Positive	Negative		
Model	Positive	a	b	*Positive Predictive Value*	a/(a+b)
	Negative	c	d	*Negative Predictive Value*	d/(c+d)
		Sensitivity	*Specificity*	**Accuracy** = (a+d)/(a+b+c+d)	
		a/(a+c)	d/(b+d)		

Figure 5.15: Confusion matrix

- **Accuracy:** The proportion of the total number of predictions that were correct. It can be calculated using the formula given by the equation 5.1.

$$Overallaccuracy = \frac{a+d}{a+b+c+d} \tag{5.1}$$

- **Positive Predictive Value:** The proportion of positive cases that were correctly identified.It can be calculated using the formula given by the equation 5.2.

$$Positivepredictivevalue = \frac{a}{a+b} \tag{5.2}$$

- **Negative Predictive Value:** the proportion of negative cases that were correctly identified.It can be calculated using the formula given by the equation 5.3.

$$Negativepredictivevalue = \frac{d}{c+d} \tag{5.3}$$

- **Sensitivity:** The proportion of actual positive cases which are correctly identified.It can be calculated using the formula given by the equation 5.4.

$$Sensitivity = \frac{a}{a+c} \tag{5.4}$$

- **Specificity:** The proportion of actual negative cases which are correctly identified.It can be calculated using the formula given by the equation 5.5.

$$Specificity = \frac{d}{b+d} \tag{5.5}$$

Calculations

For 20 test cases:

	Predicted rightly	Predicted wrongly	total
Proper Query	12	2	14
Improper Query	1	5	6

$$Sensitivity = \frac{12}{12+1} = 0.923$$

$$Specificity = \frac{5}{5+2} = 0.714$$

$$Positive\ predictive\ Value = \frac{12}{12+2} = 0.857$$

$$Negative\ predictive\ Value = \frac{5}{5+1} = 0.833$$

$$Overall\ accuracy = \frac{12+5}{12+2+1+5} = 0.85$$

5.3.2 Area under ROC curve

Definition

This is one of the popular metrics used in the industry. The biggest advantage of using ROC curve is that it is independent of the change in proportion of responders.

Formula

The Specificity and sensitivity is plotted as shown in Figure 5.16

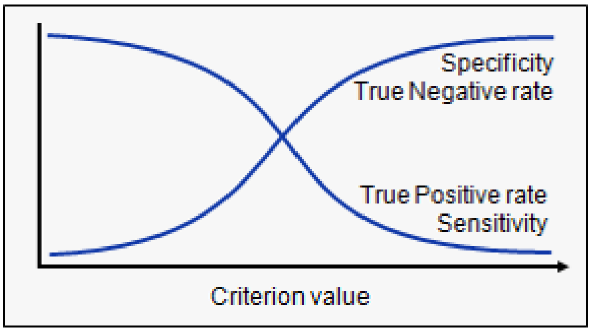


Figure 5.16: Negative and positive rate curves

The ROC Curve for a model and random data looks as shown in Figure 5.17 Figure 5.17

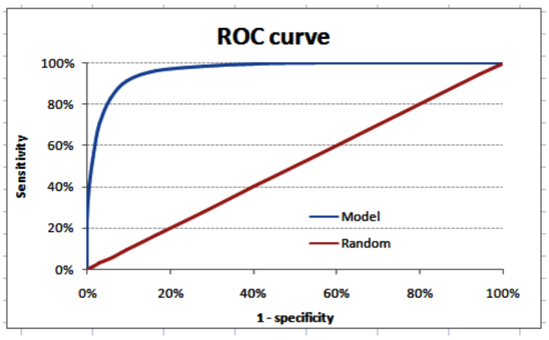


Figure 5.17: ROC sample curve

Calculations

The ROC Curve plotted for the tested data set is as shown in Figure 5.18

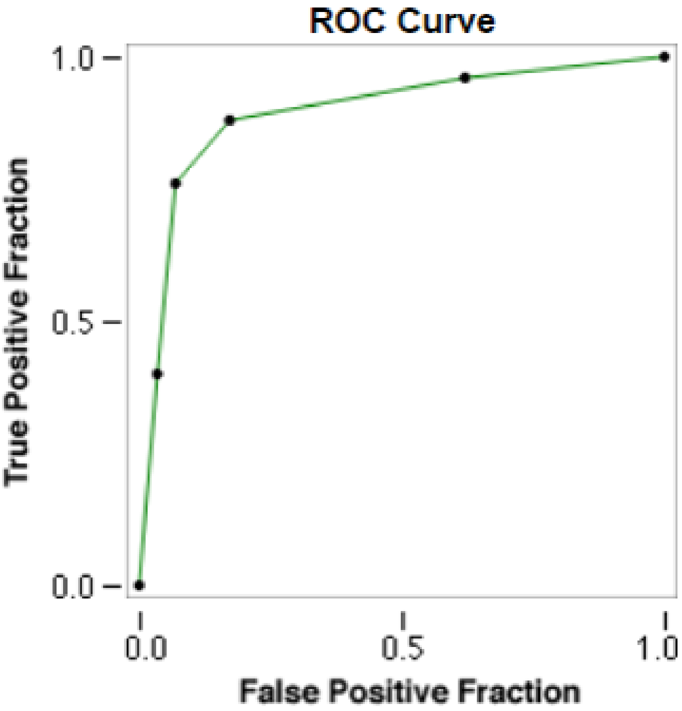


Figure 5.18: ROC curve

5.3.3 Gini Coefficient

Definition

Gini coefficient is used in classification problems. Gini is the ratio between area between the ROC Curve and the diagonal line and the area above the triangle.

Formula

Gini = 2*AUC-1

Where AUC : Area Under Curve

Calculations

Fitted ROC Area: 0.91

Empiric ROC Area: 2*AUC = 0.91 * 2 = 1.896

Gini Coefficient = 1.82 1 = 0.82

5.3.4 Problem resolution time

Definition

The time taken from the query was sent to the response received.

Formula

P.R.T = F.T - R.T

F.T - Time when final response was sent

R.T - Time when query was received

5.3.5 Average resolution time

Average Resolution Time is the time taken to give a final solution to the user query. The time taken for various iterations is given in Table 5.1

Table 5.1: Resolution time

S.No	Time taken(s)
1	77
2	72
3	69
4	74
5	70
6	70
7	79
8	82
9	76
10	78

$$Average\ resolution\ time = \frac{747}{10} = 74.7s$$

The graph showing the resolution times for various iterations (in seconds) is shown in Figure 5.19. Resolution times for ten iterations are noted and are plotted against the iterations. The average resolution times turns around 74.7 seconds.

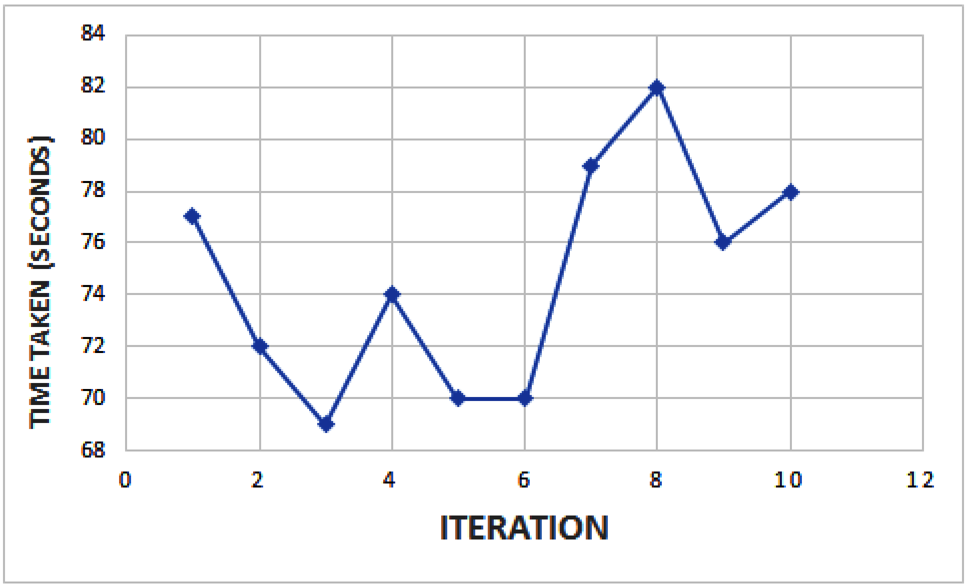


Figure 5.19: Resolution time

5.4 RESULTS AND EVALUATION

5.4.1 Accuracy

Accuracy is the percentage of correct results obtained during each iteration. It shows how correct output is given by the system. The accuracy for ten iterations are noted down in Table 5.2 as follows:

Table 5.2: Accuracy results

Iteration	Accuracy (%)
1	84
2	87
3	84
4	80
5	91
6	90
7	82
8	84
9	83
10	89

The plot of accuracy against iterations gives us the accuracy graph shown in Figure 5.20

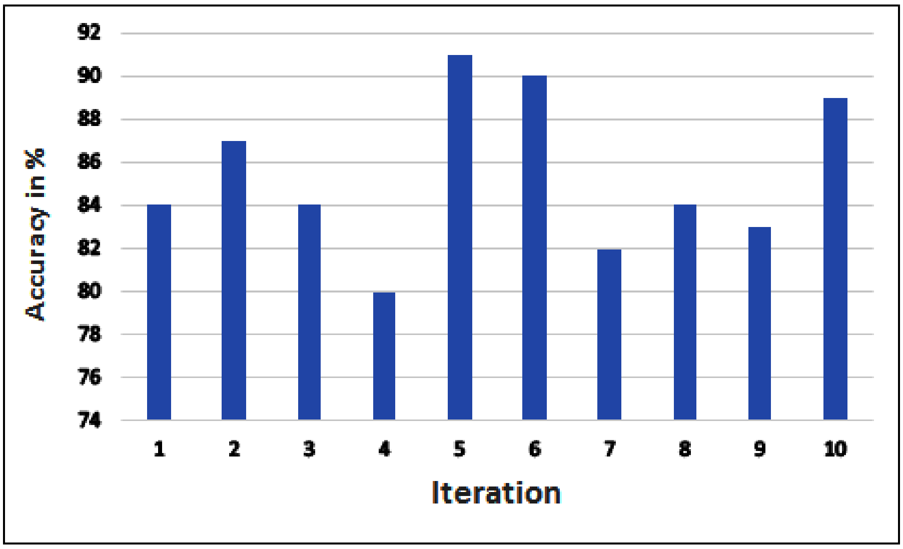


Figure 5.20: Accuracy for all iterations

Figure 5.20 infers that greater the accuracy, more stable the system will be. Two iterations had an accuracy over 90% and nine iterations had an accuracy over 80%.

CHAPTER 6

CONCLUSIONS

6.1 CONCLUSION

An Automatic Email Response system to answer medical queries is designed with an overall accuracy of 85%. Naive Bayes Classifier and Cosine Similarity is used to find the matching queries to that of the user. The current work is limited to a specific domain-Medicine. The query that user enters is checked for similarity from the available dataset and is provided an appropriate response. The response is then sent back to the user in the form of email. This process effectively reduces the manual work necessary to answer all the queries in a customer support center. Moreover subsequent threads that needs to gather more information on the query is also dealt in this by using Long Short Term Memory. States are maintained to identify which state the current user is in and the response is generated accordingly. This is the first attempt in providing a complete response for a specific domain rather than mere phrases.

6.2 FUTURE WORKS

Since our project is limited to a specific domain and depends on the available dataset, the accuracy can be increased by expanding the dataset. Sensitivity of the query can be identified using NLP[2] techniques. Using deep learning, the information provided as response can be made further accurate and detailed.

REFERENCES

[1] https://https://en.wikipedia.org/wiki/Cosine_similarity.

[2] http://www.mind.ilstu.edu/curriculum/protothinker/natural_language_processing.php.

[3] http://www.macmillandictionaries.com/wordoftheweek/archive/070813-disemvowelling.htm.

[4] http://people.idsia.ch/~juergen/lstm/.

[5] F Sujith Ravi Anjuli Kannan, F Karol Kurach, "Smart reply: Automated response suggestion for email", *International Conference on Information and Communication Technology*, vol. 35, pp. 297–301, 2016.

[6] Nishara Pathirana Tilani Gunawardena, Medhavi Lokuhetti, "An automatic answering system with template matching for natural language questions", *Journal of Computer Aided Communication*, vol. 47, pp. 353–358, 2015.

[7] Linchi Kwok Weiven Yang, "Improving the automatic email responding system for computer manufacturers", *Journal of Computational Intelligence*, vol. 27, pp. 487–491, 2015.