Conspiracy Detection by Real Time Email Analysis

BY

MD. RABIUL HASAN
ID: 142-15-3793

MD. TARIKUL ISLAM
ID: 151-15-5313

MOST. JANNATUL FERDOUS
ID: 151-15-5089

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

Supervised By

Anup Majumder
Lecturer
Department of CSE
Daffodil International University

Co-Supervised By

Shah Md. Tanvir Siddiquee
Senior Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

October 2018

APPROVAL

This Project titled **"Conspiracy Detection by Real Time Email Analysis"**, submitted by **Md. Rabiul Hasan**, ID No: **142-15-3793, Md. Tarikul Islam,** ID No: **151-15-5313 and Most. Jannatul Ferdous,** ID No: **151-15-5089** 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 10th December 2018.

BOARD OF EXAMINERS

_____

Dr. Syed Akhter Hossain — **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

_____

Narayan Ranjan Chakraborty — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

_____

Md. Tarek Habib — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

_____

Dr. Mohammad Shorif Uddin — **External Examiner**
Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Mr. Anup Majumder, Lecturer, and 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:

_____

Anup Majumder
Lecturer
Department of CSE
Daffodil International University

Submitted by:

_____

Md. Rabiul Hasan
ID: 142- 15- 3793
Department of CSE
Daffodil International University

_____

Md. Tarikul Islam
ID: 151- 15- 5313
Department of CSE
Daffodil International University

_____

Most. Jannatul Ferdous
ID: 151- 15- 5089
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

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

I would like to express my sincere gratitude to my honorable project supervisor **Mr. Anup Majumder, Lecturer**, Department of CSE Daffodil International University, Dhaka, for his valuable advices, constructive suggestions and sincere guidance with all the necessary facilities for assimilation, research and preparation for the project.

We would like to express our heartiest gratitude to **Anup Majumder, Lecturer**, Department of CSE, **Shah Md. Tanvir Siddiquee, Senior Lecturer**, Department of CSE, and **Professor Dr. Syed Akhter Hossain, Head,** Department of CSE, 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.

I would like to thank my family for their constant love and support. Finally, I would like to take this opportunity to express my gratitude to one and all, who directly or indirectly, have lent their hand in this venture.

Abstract

In this thesis, we have proposed a method to turn this psychological concept into a machine that can automatically detect the conspiracy among the employee by analyzing their email data in real time. Here we have proposed the design using vector based classification method for analyzing the text data. We have used TFIDF method to victimization and prioritize the frequency of conspiracy related word and concept. And also we used Logistic Regression, a prediction based classifier to classify the text sentiment. Supervised vector-based methods to sentiment can design rich lexical meanings. This method for machine learning is largely used in present days. Sentiment analysis for online text document has been a burgeoning field of text mining among researchers for the past few decade. Nevertheless and sentiment analysis on Email data, a ubiquity means of social network and communication, has been studied thoroughly. Email has become the most popular communication tools for official purpose. Almost every private company uses their own mail server for exchanging their official mail. So, it has a great significance in terms of business and communication. In the other hand conspiracy is a social concept that has also a great importance and impact over the working place. It is a pure psychological concept. It influences in the progress of any working place.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

FIGURES **PAGE NO**

Chapter 1

Introduction

1.1 Introduction

We live in the age of modern technology. Here almost every things are dependent on the technology .People are getting use to the technology to make their life easy and more comfortable. Modern technology is simply an advancement of old technology, the impact of technology in modern life is unmeasurable, we use technology in different ways and sometimes the way we implement various technologies ends up harming our lives or the society we leave in. What we call modern technology is technically not so new in most cases. For example, communication technology has evolved with years, nowadays we use email which has been an advancement of Fax.

Email is widely used as a form of business communication and overall it is a highly effective communication tool. Email is inexpensive, only requiring an internet connection that is generally already present in the business. As a result as online social network and communication is increase appealing to the public. From 2011 to 2015, statistics indicate an increase of 3% in the number of global Email users with an average of 1.7 Email accounts per user counted in 2015.Furthermore and business Email communication accounts for the majority of the total Email traffic with over 108.7 billion Emails exchange every day, and Email remains the most common way of business workspace communication.

In order to exchange the mail most of the large private company use private mail server. They provide all their employees an individual email account. And continue the communication with them. Using a private mail server, the biggest problem is to handle the spam challenge. There are some tools that can handle the problem also.

But there exist another problem that if any of the company employee is doing the conspiracy about the company, exchanging any sensitive information that can make a bad effect for the company, no way to detect it. There exist some big named company once that spiraled downward into bankruptcy due to the conspiracy between their employees. And this problem is getting increased day by day. Now mail is the most efficient way of transferring the information between the people.

In this study, we propose a system that will automatically detect the conspiracy related mail from real time mail box. As the detection of conspiracy will be fully automated, account of the sender and receiver employee will be detected and the information will be safe. We have to face some challenge in this thesis. Collecting real-time mail from mail server by customizing the POP3 protocol and at the same time analyzing them to detect conspiracy will be challenged. Also we have to first build an algorithm which will be able to detect conspiracy from textual content. It is the biggest challenge for us.

1.2 Motivation

Enron was an American energy company based on Houston, Texas created by Ken Lay. This company became bankrupted in October 2001. It was the largest bankruptcy reorganization in American history at that time. Enron was cited as the biggest audit failure.

Many executives at Enron were indicted for a variety of charges and some were later sentenced to prison. Many of them found guilty of illegally destroying documents relevant to the SEC investigation, which voided its license to audit public companies and effectively closed the farm. During the investigation Federal Energy Regulatory Commission made the email data of these employee public. After reading about the Enron case study something got to mind to make something that can automatically investigate the email data that are passing through the employees. Thus I was interested in analyzing data by classifying them into conspiracy class.

1.3 Rationale of the Study

The goal of this project is to design a system to detect conspiracy from the Email conversation between the employees of any company or farm. This system will detect the suspicious conversation between the employees against the company in deferent angle. We will detect the sentiment between the test conversations. Here we are proposing the method of designing a system that can automatically analysis the conversation and give the feedback with the related employee name to the owner or the authority. The most common approach to text sentiment analysis consists in detecting the occurrence of features (words) of known positive or negative semantic value.

There are some works on analyzing email data. Some of these tried to analysis the large data of email. They use sentimental analysis to detect positive negative sentiment.

1.4 Research Question

- Can we collect row data of Email Conspiracy?
- Can we pre-process the row data to be used for the Machine Learning approaches?
- Can LogisticRegression Classifier algorithm be used on the pre-processed data?
- Can the Machine Learning process correctly detect or identify the category of the conspiracy?

1.5 Expected Output

This project has a large prospective in the present word. It has a practical value in any organization where individual exchanges confidential information among themselves. It will audit the information exchange. It will help the management to have a good look over the employees for not being harmed. It will make sure the proper working condition in the work place. It maintain the commitment and trust between the individual and confirm the profit of the company. It provides the company to relay on their strategy and model of work properly. So we can say that this thesis and proposed model of system can make a good impact on the digital automated world.

1.6 Report Layout

This report is organized into five chapters.
Chapter one contains some introductory text and preliminary information about our work, previous works contains the similar forms of work that has been worked before, present state and contribution contains my contribution in this work ,motivation of the research specify the initial thought that makes me interested in this work .

Chapter two contains literature review about the required knowledge about the project and gives the over view of the technique and study that should be done by me.

Chapter three deals with the overall process of the system and my working method or suggested technique and procedure.

In chapter four, we have presented our implemented work, experimental results and evaluations are explained.

Finally, chapter five concludes our overall work.

CHAPTER 2

BACKGROUND

2.1 Introduction

Our goal is to implement a system using machine learning for conspiracy checking from email data. To do so we have used some efficient algorithms and tools and study. It will be described in details here.

2.2 Related Works

The most common approach to text sentiment analysis consists in detecting the occurrence of features (words) of known positive or negative semantic value. In that sense, sentiment analysis has a lot in common with classical text mining and classification [1], and one would be tempted to use statistical keyword significance metrics such as chi-square or TFIDF. Unfortunately, these do not give good results for sentiment classification. To be sure, some work has been done applying standard machine learning techniques to sentiment analysis, such as Pang and Lee [2] who used Bayesian classifiers, maximum entropy and SVM. Likewise, Turney and Littman [3] used latent semantic analysis (LSA) to measure the relationship between words observed in a text and a predefined praise word set. But the unique nature of the challenges of sentiment mining has given the rise to innovative new approache as well.

There are some works on analyzing email data. Some of these tried to analysis the large data of email. They uses sentimental analysis to detect positive negative sentiment. Sisi Liu and Ickjai Lee has proposed a framework for Email sentiment analysis using a hybrid scheme of algorithms combined with Kmeans clustering and support vector machine classifier. The evaluation for the framework is conducted through the comparison among three labeling methods, including SentiWordNet labeling, Kmeans labeling, and Polarity labeling, and five classifiers, including Support Vector Machine, Naïve Bayes, Logistic Regression, Decision Tree and OneR [4]

Feng, Wang, Yu, Yang and Yang [5] combine clustering approach with SWN lexicon for blogs sentiment analysis; Li and Wu [6] utilize Kmeans clustering for hotspot detection and SVM sentiment classification and prediction. Current research on text mining and sentiment analysis mainly addresses large scale social media data, such as Facebook data, Twitter corpus and blogs. For instance, Li and Wu [6] conduct, a study on hotspot detection through online forum.

Additionally, Balasubramanyan, Routledge and Smith [7] propose an algorithm model for prediction of poll result using public opinion mining. As for Email data analysis, most studies focus on the identification of spam mails, discarded mails, the study of social networking among Emails and priority issues. [8] [9] [10]. However, less research has been conducted on Email conspiracy analysis. Mohammad and Yang [11] study the gender difference in sentiment axis among set of sentiment labeled Email data, and Hangal, Lam and Heer [12] design a system for visualizing archived Email data with sentiment words tracking. Despite of these studies, a systematic and structured framework for conspiracy detection from Email data has not been investigated yet.

I. Conspiracy

According to the Oxford dictionary conspiracy is a secret plan by a group to do something unlawful and harmful. Ex: 'a conspiracy to destroy the government'. In another word the action of plotting or conspiring

II. Conspiracy Theory

Conspiracy theories are ubiquitous among members of modern and traditional societies. A common definition of conspiracy theory is the confidence that a group of actors meets in secret agreement with the purpose of attaining some malevolent goal

III. Psychology of Conspiracy Theories

Conspiracy theories explain event as the result of secret and deliberate actions and cover ups at the hands of malicious and powerful groups.

IV. Organizational Conspiracy Theories

We define organizational conspiracy theories as notions that powerful groups (e.g., managers) within the workplace are acting in secret to achieve some kind of malevolent objective. For example, managers may deliberately conspire to hire a preferred candidate for a job, or work together to have an employee fired.

V. Organizational Identification

Organizational identification refers to individuals' self-definition as members of a particular organization. Organizational identification has been found to uniquely predict organizational outcomes and attitudes and behaviors at work. For example, it has been associated with workers' well-being, performance, and, most relevant to the current investigation, turnover intentions.

VI. Job Satisfaction

Job satisfaction is the evaluation that employees make of their job and includes their attitudes to specific aspects of the job. Research has found that, like organizational identification and organizational commitment, job satisfaction is associated with turnover intentions, and that more satisfied workers are less likely to want to leave their jobs.

VII. Dataset

The term data sets refers to a file that contains one or more records. The record is the basic unit of information used by a program running on z/OS.

viii. Private Mail Server

A private server is a physical computer that you own and operate, and has all the operating systems, software and programs in place to provide essential services, including email. In a textbook definition *A private server is a machine or virtual machine that is privately administrated. As servers need an adequate Internet connection, power and can be noisy, they are often located in a colocation center.*

A private email server would be the email system that's offered by the private server.

In other words, with a private email server one can have his own email system, from computers to programs. One can run it, use it, manage it and limit (allow and prevent) access to it.

2.3 Research Summary

The above discussion done on various types of research works from different research teams, it is being appeared to us that recently, research work on email conspiracy is increasing day by day. Some good outcomes already prove this statement well. Though, enough resources are not present, but hope is that this field is becoming more resourceful each after passing a single day.

2.4 Scope of the Problem

In order to exchange the mail most of the large private company use private mail server. They provide all their employees an individual email account. And continue the communication with them. Using a private mail server, the biggest problem is to handle the spam challenge. There are some tools that can handle the problem also.

2.5 Challenges

We have to face some challenge in this thesis. Collecting real-time mail from mail server by customizing the POP3 protocol and at the same time analyzing them to detect conspiracy will be challenged. Also we have to first build an algorithm which will be able to detect conspiracy from textual content. It is the biggest challenge for us.

Chapter 3:

Research Methodology

3.1 Introduction

In this chapter we will describe the architecture of Conspiracy Detection Framework. There are mainly three sections in this chapter. First section 3.1 is about the system architecture of the Conspiracy Detection Framework where different modules of the architecture and relationship among them are described briefly. Among the main modules of the architecture like Mail Data Fetching module, storage module, Data Analyzing module, alert sending module. In section 3.2 we discuss about the analytical representation of our system which gives the details of the developed system with different algorithms, necessity flowcharts and tables required for analysis. Section 3.3 is about the complexity analysis of Conspiracy Detection Framework.

3.2 Research Subject and Instrumentation

We mean by research subject is that research area that is being studied and researched for clear understandings. Not only for clear understanding, but also research subject is responsible for giving the right knowledge of various research parameters. On the other hand, Instrumentation refers to the required instruments or tools that are used by the researchers.

3.3 Data Collection Procedure

To research on specific field, the fast and foremost thing is the Data. Data is, actually, considered as the heart of the machine learning process. And for our research, there has no alternative of data. So, it has become our most challenging task for our research.we build our data set by analysing lots of journal that's are related to compiracy theory.

3.4 Data Preprocessing

Data Scientists across the word have endeavored to give meaning to Data preprocessing. However, simply put, data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in

certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues.

1. Data cleaning: fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies.
2. Data integration: using multiple databases, data cubes, or files.

3. Data transformation: normalization and aggregation.

4. Data reduction: reducing the volume but producing the same or similar analytical results.

5. Data discretization: part of data reduction, replacing numerical attributes with nominal ones. [16]

i. System Architecture

In this section of conspiracy detection model there are four module to work sequentially for analyzing the conspiracy. Data Acquisition and Refining module, Data Processing module, Training module, Testing in real time module. Here Data Acquisition and Refining module fetch the mail data from the database and then refine the data with unnecessary symbol, character and some other unwanted factors that will have no work with the detection model. Data Processing module reach the refined data to have the data with a matrix of frequency with respect to the importance in any pole. After finding the significance matrix the value of these significance goes to the classifier in this module of Training model. Now we have the trained module of conspiracy detection predictor. And the last and the most important module of this model is the testing with real time environment. In this module the data from the mail server is crawled by a crawler to give the input to the trained model. After analyzing this data model gives the answer or verdict to the management or monitoring body of any individual. The architecture is shown in below

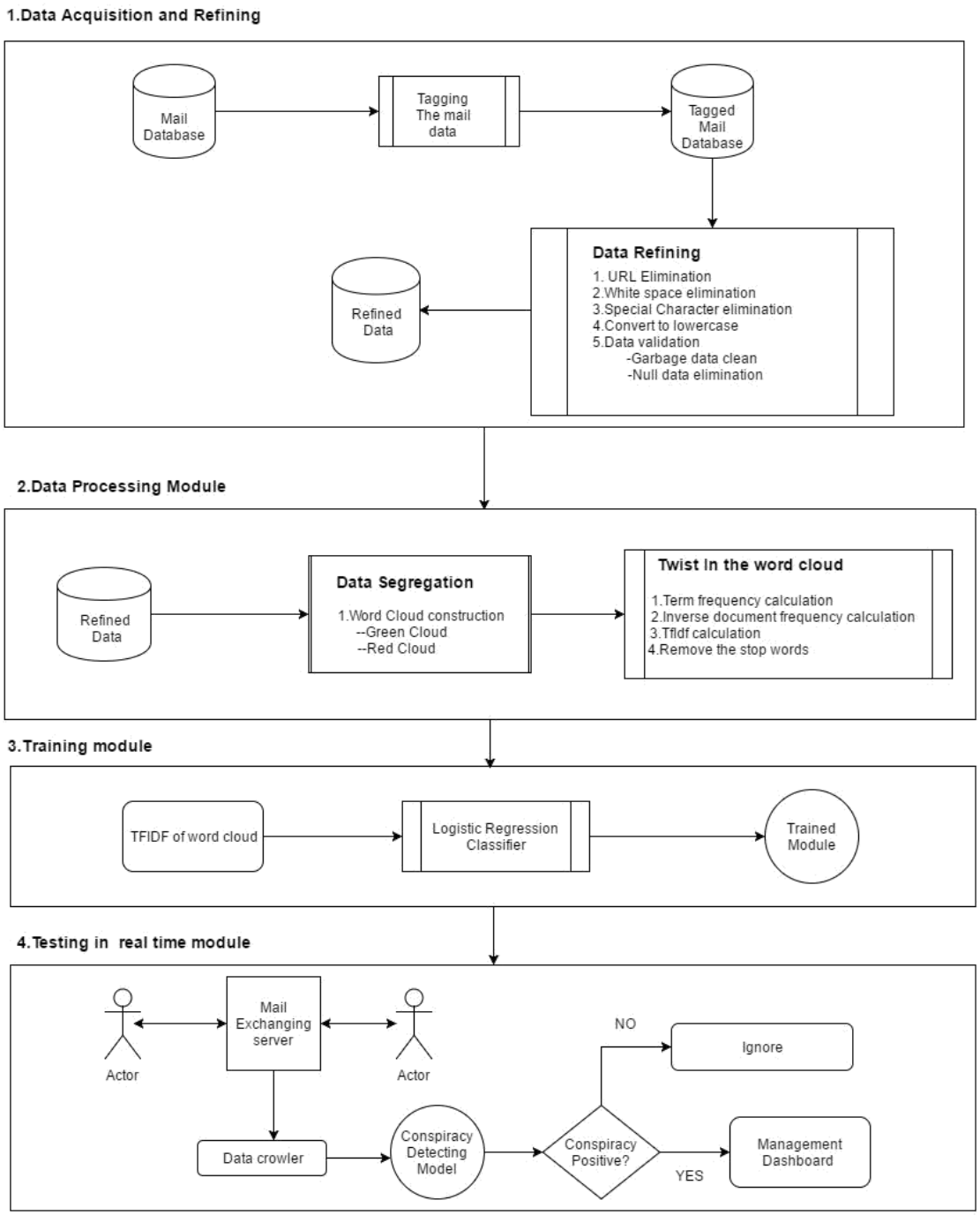


Figure 3.1: Conspiracy Predictor Model Architecture.

ii. Data Acquisition and Refining

The very first module of this proposed model is *Data Acquisition and Refining.* This model is used to level the dataset with green and red level. This tagging makes the dataset leveled properly and segregate the body section. After tagging the dataset with conspiracy infected mail and conspiracy uninfected mail, the mail data is stored in a csv file. Now we have the dataset stored in the form of a comma-separated values file with mail data and corresponding category. Category is defined by 0 and 1. 0(zero) means the Green data and 1(one) means the Red data in the dataset.

After saving the leveled data in a .csv extension file the data is then import to the module to analyze and refine. The data is the refined with some necessary steps. The dataset is a mixture of words, emoticons, symbols, URLs and references to people.

In the first category of Green set the mail body can reflect so many affair in the sentiment. Such as office affair, request mail, satisfaction mail, gratitude sentiment mail and some more documentation mail. Fig 2 shows the Green mail data in csv file

```
Message-ID: <8572706.1075855378498.JavaMail.evans@thyme>
Date: Thu, 3 May 2001 15:57:00 -0700 (PDT)
From: phillip.allen@enron.com
To: rlehmann@yahoo.com
Subject:
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: Phillip K Allen
X-To: rlehmann <rlehmann@yahoo.com>
X-cc:
X-bcc:
X-Folder: \Phillip_Allen_Jan2002_1\Allen, Phillip K.\'Sent Mail
X-Origin: Allen-P
X-FileName: pallen (Non-Privileged).pst

Reagan,

Just wanted to give you an update.
I have changed the unit mix to include some 1 bedrooms and reduced the number of buildings to 12.
Kipp Flores is working on the construction drawings.  At the same time I am pursuing FHA financing.
Once the construction drawings are complete I will send them to you for a revised bid.
Your original bid was competitive and I am still attracted to your firm because of your strong local presence and contacts.

Phillip
```

Figure 3.2: Green Mail Content

In the next category of the Red set of mail infected with office conspiracy. This category is leveled by one.This sort of mail data reflect rage, dissatisfaction, overpower intention etc that means the overall bad effect for the company. Fig 3 shows the Red mail category

```
Message-ID: <15464986.1075855378456.JavaMail.evans@thyme>
Date: Fri, 4 May 2001 13:51:00 -0700 (PDT)
From: phillip.allen@enron.com
To: john.lavorato@enron.com
Subject: Re:
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: Phillip K Allen
X-To: John J Lavorato <John J Lavorato/ENRON@enronXgate@ENRON>
X-cc:
X-bcc:
X-Folder: \Phillip_Allen_Jan2002_1\Allen, Phillip K.\'Sent Mail
X-Origin: Allen-P
X-FileName: pallen (Non-Privileged).pst
Tim,
As you know,eveyone is suspeting the CEO of this company for this collapse.
```

Figure 3.3: Red Mail content

And the overall csv file will look like these in figure 4

Ic. There are a lotta childporn cars then.	0
No I was trying it all weekend ;V	0
You know wot people wear. T shirts jumpers hat belt is all we know. We r at Cribbs	0
Cool what time you think you can get here?	0
Wen did you get so spiritual and deep. That's great	0
Have a safe trip to Nigeria. Wish you happiness and very soon company to share moments with	0
Hahaha..use your brain dear	0
Well keep in mind I've only got enough gas for one more round trip barring a sudden influx of cash	0
Yeh. Indians was nice. Tho it did kane me off a bit he he. We shud go out 4 a drink sometime soon. Mite hav 2 go 2 da works 4 a laugh soon. Love Pete x x	0
Yes i have. So that's why u texted. Pshew...missing you so much	0
No. I meant the calculation is the same. That <#> units at <#> . This school is really expensive. Have you started practicing your accent. Because its important. And hav	0
Sorry I'll call later	0
if you aren't here in the next <#> hours imma flip my shit	0
Anything lor. Juz both of us lor.	0
Get me out of this dump heap. My mom decided to come to lowes. BORING.	0
Ok lor... Sony ericsson salesman... I ask shuhui then she say quite gd 2 use so i considering...	0
Ard 6 like dat lor.	0
Why don't you wait 'til at least wednesday to see if you get your .	0
Huh y lei...	0
Will Ä¼ b going to esplanade fr home?	0
Pity * was in mood for that. So...any other suggestions?	0
The guy did some bitching but I acted like i'd be interested in buying something else next week and he gave it to us for free	0
Rofl. Its true to its name	0
Do you need money from the company work?	0
I like to ensure you that we should need some political influence to make this company suffer.destrudctive change	1
We have to have some political power practicing to collapse the structure of the company..destrudctive change	1
We have to make sure that the local political leader can transpass the administratio..destrudctive change	1

Figure 3.4: A Peek into the Dataset

iii. Data Refining

In our dataset we have the mail with some noise. Noises in the mail is natural because people send so many things in the mail to express his opinion. In mail data there are some link, URLs, emotion some unwanted symbol to refine .This raw data should be refined before training the model. To get the best out of our dataset, we applied a number of data cleaning process. At first some general cleaning are done, such as:

- Every sentence is first converted into lowercase format.
- Two or more spaces are replaces with a single space

- Quotes (" and ’), extra dots (.) and spaces are stripped from the ends of sentences.
- Null data elimination as well as the garbage data.

To handle the special component of a sentence, we have done the following pre-processing tasks:

1. **URL:** Users often sends URL in their mail. In our training, any particular URL doesn’t contain any special feature and if we kept the URLs in the sentences, that would have been leaded to sparse feature. Therefore, we remove all the URL from the sentences. To match the URLs we have used this regular expression ((www\.[\S]+)|(https?://[\S]+)).

2. **Special Cleaning:**
   - Any punctuation [’"?!,.():;] from the word is stripped. Words with three or more letter repetitions are converted to two letters.
   - Some people send their mail like I am happppppy which adds multiple characters on a certain words. Mail containing this type of words are handled by converting the word happppppy to happy.
   - To handle the words like sugar-free and our’s, we have removed - and ’. This type of words are converted into a more general form like sugarfree and ours.
   - Then we checked for valid word by checking successive alphabets, if it is not valid then we have stripped them.

3. **Contracted Word Handling:** Users often sends mails containing words in contracted form. Like are not is written as aren’t, I am is written as I’m etc. We converted the contracted word to their long form. A list of contracted word and their long form are given in Table 3.1:

Table 3.1: Contracted Word and Long Form

Contracted form	Long form	Contracted form	Long form
Aren’t	Are not	I’m	I am
Isn’t	Is not	Weren’t	Were not
Haven’t	Have not	Hasn’t	Has not
Hadn’t	Had not	Won’t	Will not
Don’t	Do not	Doesn’t	Dose not
I’d	I had	I’ll	I shall
They’ll	They will	He’ll	He will
She’ll	She will	Can’t	Can not
Mustn’t	Must not	Shouldn’t	Should not
Couldn’t	Could not	Wouldn’t	Would not

iv. Data Processing Module

In this module the cleaned mail data is being further processed with some algorithmic process. Such as vectorization, featuring and stop word removing. By following these process the mail data will be ready for using in the classifier to train our predicting model. These following steps are described below.

v. Tokenization:

Given a character sequence and a defined document unit, tokenization is the task of chopping it up into pieces, called tokens, perhaps at the same time throwing away certain characters, such as punctuation. Here is an example of tokenization:

Input: Mr .X can we meet today. Some documents should be reviewed

Output: Mr, X, can, we, meet, today, some, document, should, be, reviewed.

vi. Feature Vector:

We have produced our dataset based on the working place conspiracy theory from a pronounced journal. Here we have come to know about the details about the conspiracy and its related outcome, concept, outcome and immediate effect to detect from the work place. So a common style is found in the mail data during producing the mail data set.

Tf-idf is a simple twist on the bag-of-words approach. It stands for term frequency-inverse document frequency. Instead of looking at the raw counts of each word in each document in dataset, tf-idf looks at a normalized count where each word count is divided by the number of documents this word appears in

It is another way to convert textual data to a numeric form. The vector value it yields is the product of these two terms; TF and IDF

Relative term frequency is calculated by:

$$\text{TF (t, d)} = \frac{number\ of\ times\ term(t)\ appares\ the\ document(d)}{total\ number\ of\ term\ in\ document(d)}$$

And we need to get inverse Document Frequency, which measures how important a word is to differentiate each document by following the calculation as below

$$\text{IDF (t, D)} = \frac{total\ number\ of\ document(D)}{number\ of\ documents\ in\ term(t) in\ int}$$

Once we have the values of TF and IDF, now we can calculate TFIDF as below:

TFIDF (t, d, D) = TF (t, D).IDF (t, D)

Here if N is the total number of documents in the dataset. The fraction N/ (document……) is

What is known as the inverse document frequency? If a word appears in many documents, then its inverse document frequency is close to 1. If a word appears in just a few documents, then the inverse document frequency is much higher.
Alternatively, we can take a log transform instead using the raw inverse document frequency.
Logarithm turns 1 into 0, and makes large numbers (those much greater than 1) smaller.

vii. Stop word: There are some words which do not make any significant change in absence of them. Those words are called stop word. Our current step is to remove those words from the document.

ix. Training Module

In this module the tfidf matrix of every vector is used in our classifier to train the model. In this way first select the classifier that is best for the model. We prefer logistic regression classifier to make this happen.

i. Logistic Regression Classifier

Logistic regression is a simple, linear classifier. Due to its simplicity, it's often a good first classifier to make a model. It takes a weight combination of the input features, and passes it through a sigmoid function, which smoothly maps any real number to a number between 0 and
algorithm that performs very well on a wide range of problem [26]. Logistic regression corresponds to a linear regression where the dependent variable is binary. It is very useful for understanding or predicting the effect of one or more variable on a binary response variable.

The probability for class j with the exception of the last class is [17]:

$$P_j(X_i) = \frac{e^{X_i B_j}}{\left(\sum_{j=1}^{k-1} e^{X_{i^*} B_j}\right) + 1}$$

B: parameter matrix.

K: number of classes.

The last class has probability

$$1 - \sum_{j=1}^{k-1} P_j(X_i) = \frac{1}{\left(\sum_{j=1}^{k-1} e^{(X_i * B_j)}\right) + 1}$$

In the linear regression classification method the equation is

$$y = bo + b1 * x$$

In this equation if we use a sigmoid function: $p = \frac{1}{1+e-y}$

Then the equation will look like:

$$\ln\left(\frac{p}{1-p}\right) = b0 + b1 * x$$

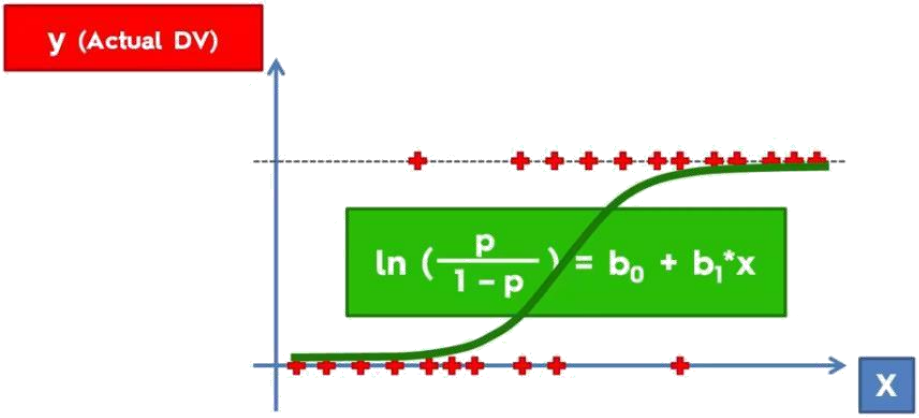


Figure 5: Graphical representation of Logistic Regression Classifier

From this graph we can say that the nonlinear data that don't have any straight line classifier to differentiate the classes. Thus use the logistic regression to classify the level. Here we can have the probability from the classifier. And from the probability we can predict the polarity of our input

Here in X axis the independent variable the word frequency is plotted and in the Y axis the probability is found.

Here we spit the dataset into training set and testing set for both X and Y axis. And thus have the accuracy of the model by testing the testing leveled data.

x. Testing In Real-time

In this module the trained model is worked for predicting the conspiracy from the users mail data. Here we have some client those can communicate with email by a mail server. A data crawling method is set to crawl the email data and store it in a database. Then the trained model analyze the data and predict a probability of conspiracy or not. If the model has the higher probability of conspiracy positive in the body of the mail, then the model give alert to the monitoring body of the company and store the data and those who exchange the mail between them. Thus the verdict of every email is monitored automatically.

3.5 Statistical Analysis

This section gives an analytical description of the system architecture given in previous sections. The system architecture above illustrates the internal and external structure of system modules integrated together in one package to form one system.

i. Labelling the Email Data

We have labeled the email data by Green data and Red data. We collect the Green email from the Enron dataset. And the Red dataset are collected from the practical field.

ii. Clean the Data

After data is labeled then we have the raw email data. We cannot use these data to classify or train. So, we have cleaned the data before using them in classifier or training. We have performed several cleaning like removing URL, removing stop words, removing multiple spacing etc.

Algorithm 3.1: Cleaning raw email

Input: raw email

Require: clean the raw email

1. Begin
2. Remove url from raw email data
3. Convert the raw email into lowercase form
4. Search for contracted form in email body
8. **if** contracted form found then
9. Replace it with long form
10. Search for stop words in data
11. **if** stop words found then
12. Remove the stop words
13. **End**

iii. Process the data

After cleaning data from the garbage data then the email data is ready to be processed by vectorization. Here we will use the TFIDF vectorization process to split and calculate the importance of any word in the dataset.

Algorithm 3.2: Process the cleaned email

Input: cleaned email

Require: process the cleaned email

1. Begin
2. Remove stop words from raw tweets
3. Convert the raw email into lowercase form
4. Tokenize the tweets
5. Calculate the TFIDF matrix

13. **End**

iv. Train the Model

To predict the classes of the email we need a mathematical model that can specify the class of the email based on their features. We have used Logistic Regression classifier algorithm. There are some more algorithm for classification. We use Logistic Regression as it is a probabilistic method and we have a small amount of training dataset

Logistic Regression

Algorithm 3.3: Logistic Regression learning algorithm

Inputs: Training data, x

Require: Train model to classify

1. Begin
2. Initialize w
3. **for** i=1 to n **do**
4. z(i)=∑w(i)*x(i)
5. **end for**
6. **for** j = 0 to d **do**
8. **for** i = 1 to n **do**
9. θ(j) = SOFT-MAX(z(i))
10. **end for**
11. **end for**
12. **End**

v. Collecting the Mail Data in Real Time

In the real time classification process we use a crawling algorithm and store it in a database. We crawl the data and the communicating employee's name. Then it stores in another database for further analyzation.

Algorithm 3.4: Collect the email data from the profile

Inputs: Automated process

Require: Take the data to another database

1. Begin
2. Access the storing database of the email
3. **if** the email is not still taken
4. take the email
5. **End**

vi. Generating Output

By using the model that we have built in the previous steps, we can classify email body. The classification result is 0 or 1. According to this result we can show the type of the email.

Algorithm 3.5: Classification of real-time email

Inputs: model file

Require: Classification of the email

1. **Begin**
2. classifier = load(model)
3. **for** each email in the **emaildata** table in database **test**
4. take the **email body**

5. type = classifier.predict(email body)

6. **if** type = 0 **then**

7. *result* = “It is not infected”

8. **else** if type = 1 **then**

9. *result* = “Infected”

10. store the email with the sender and receiver name in another database

11. ***show the result***

11. **End**

vii. Complexity Analysis

Our system has three parts keeping the issue of time in concern. The time complexity of our system may be described as follows:

viii. Complexity of Email Cleaning

Let, n is the number of letter in a email, m is the number of email in the dataset So time require to clean all the email are O (nm).

ix. Complexity of storing the database of Email

Let we have n number of email in any email database. So it will take to crawl the data from the dataset to another dataset. It will take the complexity of O (n).

x. Complexity of Predicting an Email

Let, n is the number of word in a email, C is the number of words in the feature vector. So time require to predict an email is O (nC).

3.6 Implementation Requirements

After the proper analysis on all necessary statistical or theoretical concepts and methods, a list of requirement has been generated that must be required for such a work of Email Conspiracy Classification. The probable necessary things are:

Hardware/Software Requirements

- Operating System (Windows 7 or above)
- Hard Disk (minimum 4 GB)
- Ram(more than 1 GB)
- Web Browser(preferably chrome)

Developing Tools

- Python Environment
- Jupyter (Anaconda3)
- Notepad++
- Bootstrap

Chapter 4:

Experimental Results and Discussion

4.1 Introduction

In this chapter of implementation of conspiracy detection module, we will discuss about the overall implementation procedure of the project. It is a challenging task to implement this module. In the first chapter 4.1 we will describe our experimental setup. In the next section 4.2 shows the system that can be used to exchange the email data among the employee of the company. Section 4.3 shows the detection and exchanging email procedure. And in the last one section 4.4 we will conclude the chapter of implementation. We have tested our system with extensive experiment. In this section, we first introduce how data are collected for our Conspiracy Detection Model. Then we will present the performance of the system and compare it with existing systems.

4.2 Experimental Setup

An Email Conspiracy Detection Framework has been developed on a machine having the Windows 10, core i5 processor with 8GB RAM. The system has been developed in Python and Php in the backend and javascript is used in the front end. Mysql is used for storing related data in this framework.

For coding in python, we have used the latest version of PyCharm which is 2018.2.4 with python version 3.6. For coding in php, we have used the latest version of phpstorm which is 2018.2.2 with php version 7. The system architecture following illustrates the internal and external structure of system modules integrated together in one package to form one system. The following subsections provide a brief background overview of the tools used and the implementation details of the different modules of the developed system starting from the back-end to the front-end. The whole system was developed on Windows Operating System and using pycharm and phpstorm IDE.

4.3 Experimental Results

After completing the classification of Bangla news, User Interface shown in figure bellow:

i. Email Exchanging System

In this section we will discuss about the mail server that can exchange the mail between the employees of the company. Here we build a system that can show us an interface to login to the mail server and check the inbox, outbox, sent box and logout. In this system we build the process to compose the mail to any other individual of the company. In this system anyone can exchange the mail to anyone of this company. This system can be used from anywhere of this world using internet. This system is hosted with a Public IP. Thus anyone in anywhere can exchange the mail with having the Id and Password of this system.

We have named the system as “DIU Mail”. And hosted it in a public IP with classified user id and password.

Fig 4.1 shows the snapshot of the system of exchanging the email. In here we will give the snapshot of the login Page of the mail system interface:

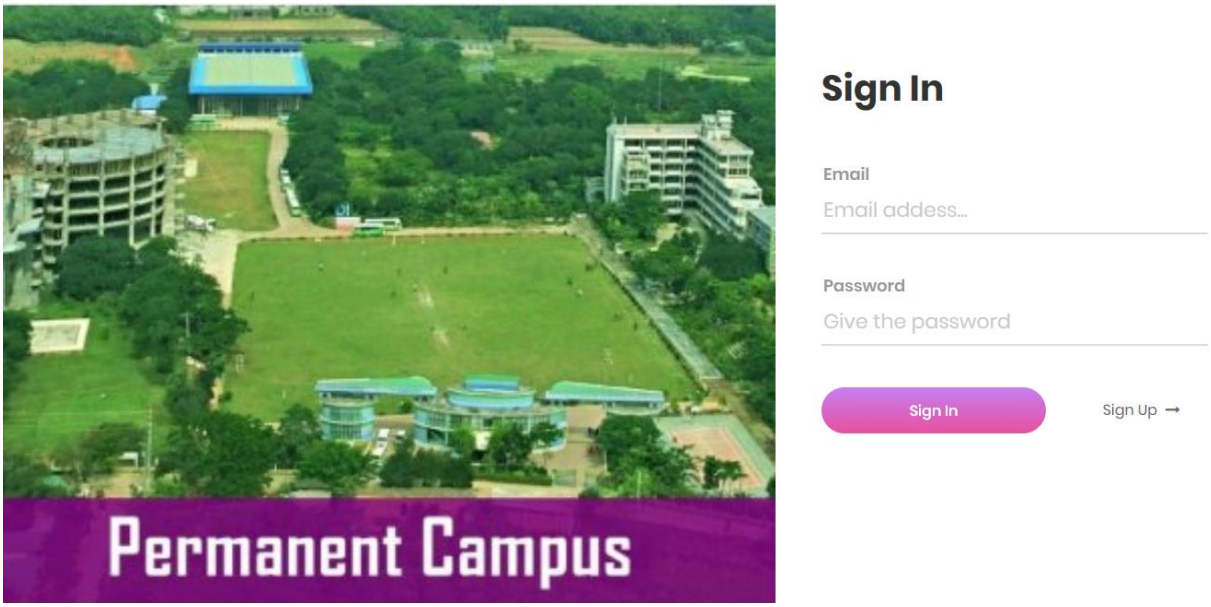


Figure 4.1: Login interface of DIU Mail.

Now we will give the inbox interface of our system that will show the mails that are sent to the individual of this account. This page will be seen after the authorized login

Fig 4.2 will show the figure below:

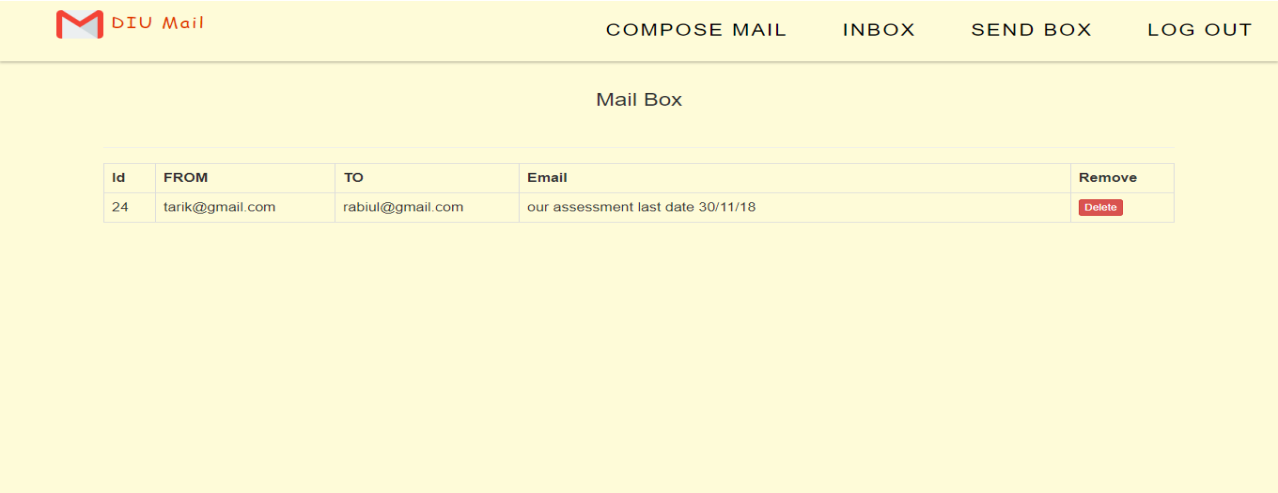


Figure 4.2: Inbox of the System

Here we will give the necessary interface of the Send box of the system.

Fig 4.3 will show the interface:

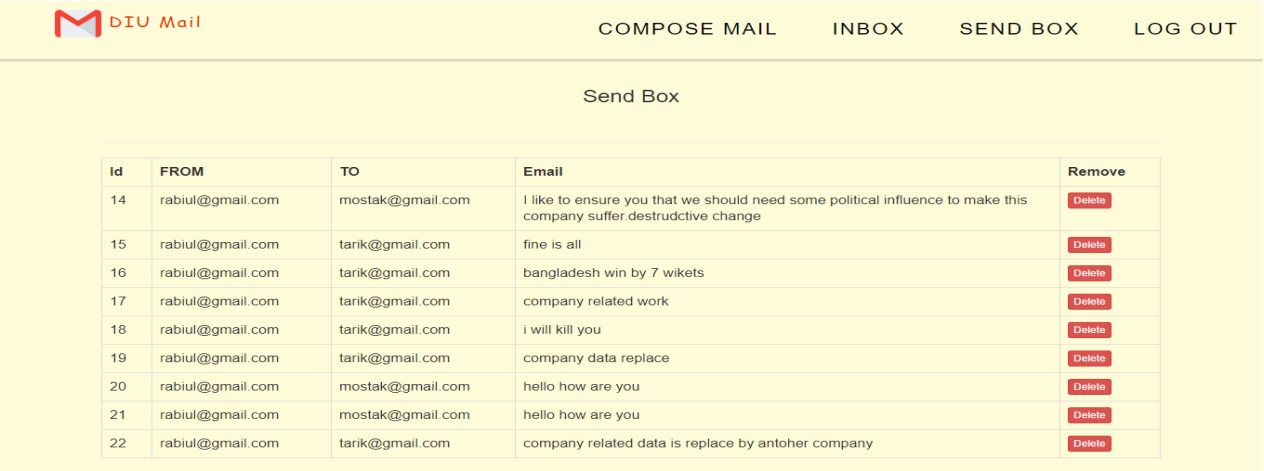


Figure 4.3: The Interface

Now the last one is the composing mail interface, in where we can compose the new mail to another end user.

Fig 4.4 will show the figure below:

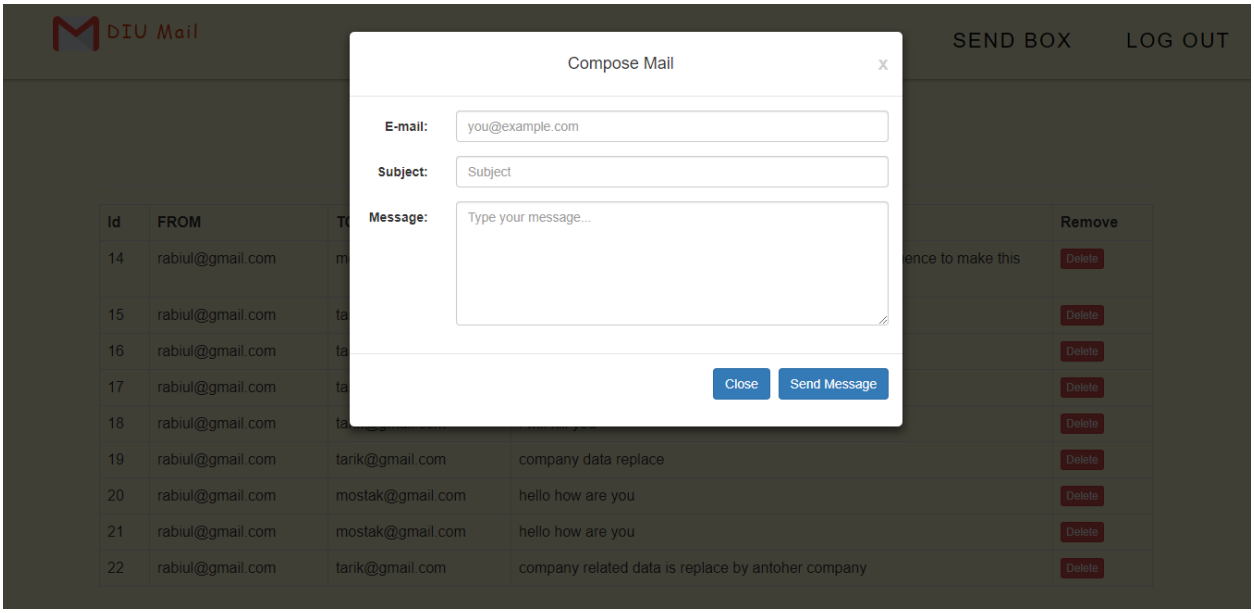


Figure 4.4: Compose Mail Interface

In this every section we use three database table to store the email data and for analyzing the data. The two database looks like:

Fig 4.5 shows the email_data database table for analyzing the email data:

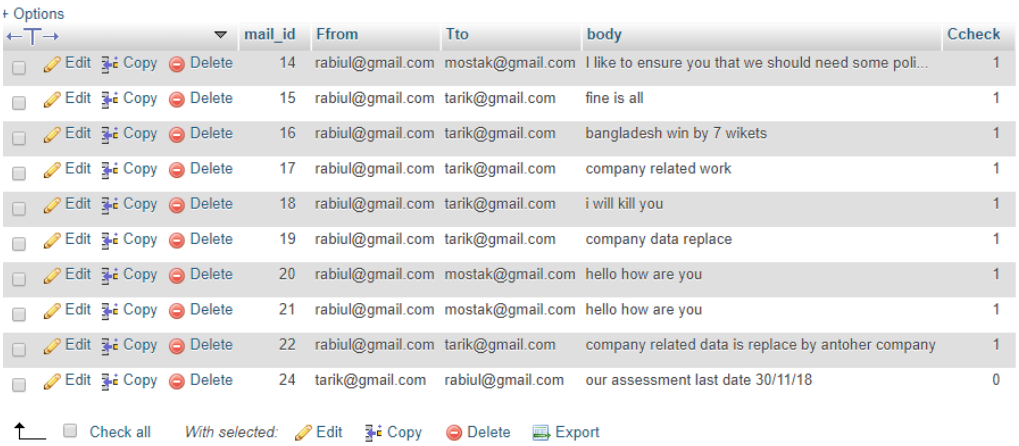


Figure 4.5: 'Email Data' Table Interface

Then the next figure will show the interface for the infected email that are analyzed to show the governing body of the company

mail_id	Ffrom	Tto	body	verdict
14	rabiul@gmail.com	mostak@gmail.com	I like to ensure you that we should need some poli...	infected
17	rabiul@gmail.com	tarik@gmail.com	company related work	infected
19	rabiul@gmail.com	tarik@gmail.com	company data replace	infected
22	rabiul@gmail.com	tarik@gmail.com	company related data is replace by antoher company	infected
23	rabiul@gmail.com	mostak@gmail.com	replace	infected

Figure 4.6: Infected 'Result' Database Table

Now we will show the user of the company database. It means the employee who has the authorized id and password to enter into the mail server:

Md.Rabiul Hasan	rabiul@gmail.com	0
Tarikul Islam	tarik@gmail.com	0
hasan	hasan@gmail.com	1234
rubel	rubel12@gmail.com	123456
jannatul Ferdous	sopna@diumail.com	12345

Figure 4.7: User Verification Table Interface

4.4 Descriptive Analysis

In this section we will discuss about the interface of the detection model. Here the company management can have the alert and got the verdict about the mail had exchanged between the employees in real time. If any employee delete the data from his inbox or send box, still the system can monitor the email in real time. It stores the email after analyzing in a new database and will give the proof of the email.

In this section we will show the system as a whole in the figure 4.8:

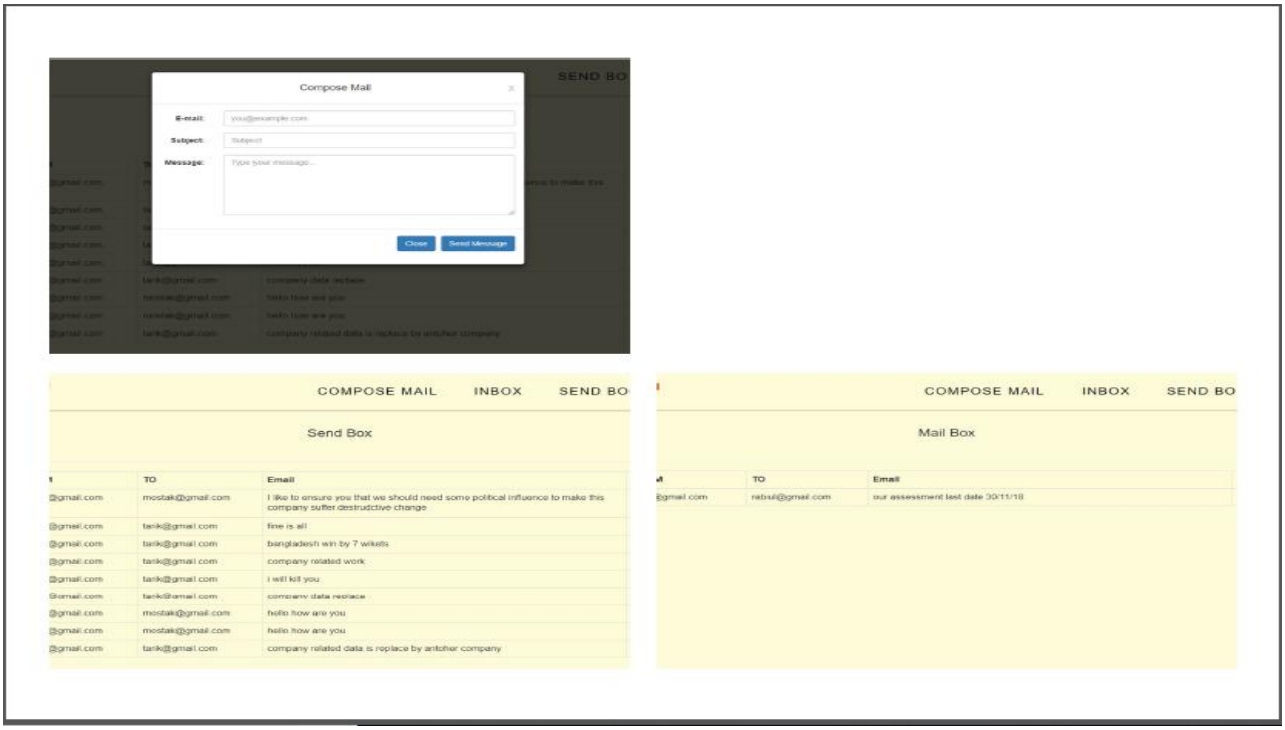


Figure 4.8: Email Sending System

In the next one we will show detection in the verdict page for any mail exchanged through this server.

Fig 4.9 will show the interface of this affair:

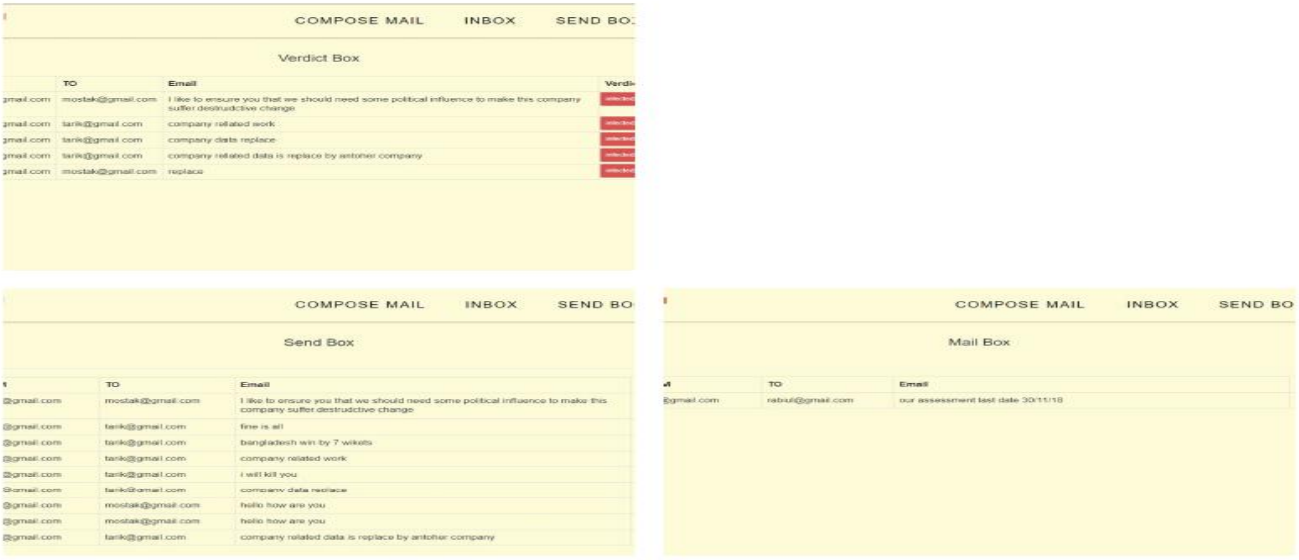


Figure 4.9: Detection Illustration

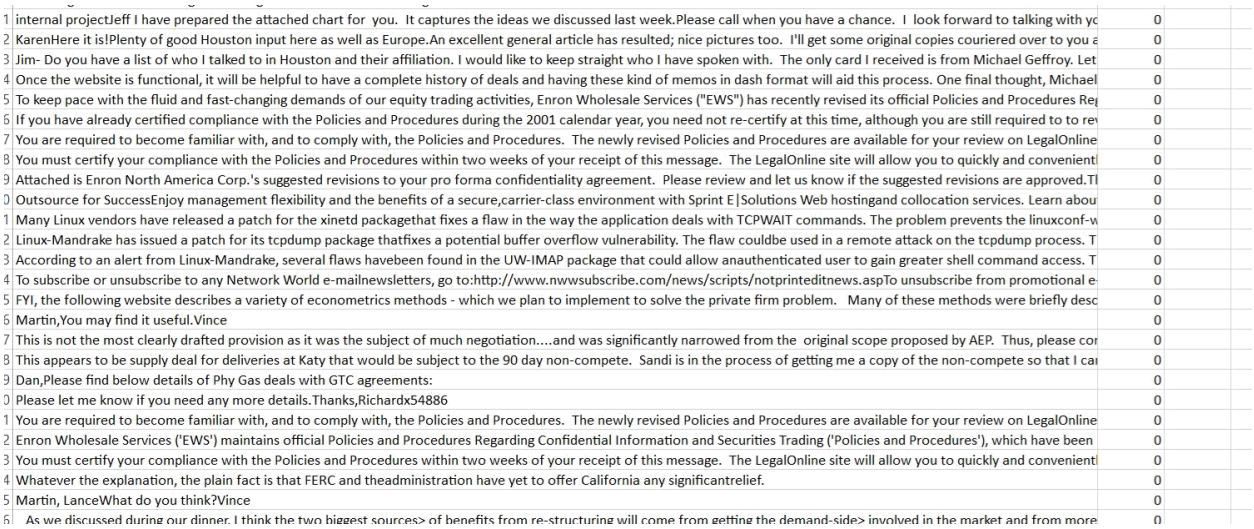
4.5 Data Collection

For Conspiracy detector we collect the data from the real environment. First of all finding the email data is not so easy. There is only some of real email dataset available in this world that are free for general research. We uses Enron Dataset. So many researches has been done successfully with this set of data.

i. Green Data Collection

As we are going to use data of two classes. So far our plan was to collect the official email like office affair, work related email, deal related, client related, gratitude related, personal email, internal component operation, legal advice, humor, friendship affection related, jokes, forwarding email, Logistic arrangement etc. We have uses the data with perfectly classified with these sort of classes. These sort of data or email are frequently exchanged between the employees of the company. So we have to classify these data as a green data.

We collect the email from that dataset and labeled it with class 0(zero). These are our Green dataset. We have stored these data in a csv file with two rows. Row one is for the body of the email. And the second row is for the sentiment or class. Class row holds the zero as the sentiment.



1	internal projectJeff I have prepared the attached chart for you. It captures the ideas we discussed last week.Please call when you have a chance. I look forward to talking with yc	0
2	KarenHere it is!Plenty of good Houston input here as well as Europe.An excellent general article has resulted; nice pictures too. I'll get some original copies couriered over to you a	0
3	Jim- Do you have a list of who I talked to in Houston and their affiliation. I would like to keep straight who I have spoken with. The only card I received is from Michael Geffroy. Let	0
4	Once the website is functional, it will be helpful to have a complete history of deals and having these kind of memos in dash format will aid this process. One final thought, Michael	0
5	To keep pace with the fluid and fast-changing demands of our equity trading activities, Enron Wholesale Services ("EWS") has recently revised its official Policies and Procedures Re	0
6	If you have already certified compliance with the Policies and Procedures during the 2001 calendar year, you need not re-certify at this time, although you are still required to to re	0
7	You are required to become familiar with, and to comply with, the Policies and Procedures. The newly revised Policies and Procedures are available for your review on LegalOnline	0
8	You must certify your compliance with the Policies and Procedures within two weeks of your receipt of this message. The LegalOnline site will allow you to quickly and convenientl	0
9	Attached is Enron North America Corp.'s suggested revisions to your pro forma confidentiality agreement. Please review and let us know if the suggested revisions are approved.TI	0
0	Outsource for SuccessEnjoy management flexibility and the benefits of a secure,carrier-class environment with Sprint E|Solutions Web hostingand collocation services. Learn abou	0
1	Many Linux vendors have released a patch for the xinetd packagethat fixes a flaw in the way the application deals with TCPWAIT commands. The problem prevents the linuxconf-w	0
2	Linux-Mandrake has issued a patch for its tcpdump package thatfixes a potential buffer overflow vulnerability. The flaw couldbe used in a remote attack on the tcpdump process. T	0
3	According to an alert from Linux-Mandrake, several flaws havebeen found in the UW-IMAP package that could allow anauthenticated user to gain greater shell command access. T	0
4	To subscribe or unsubscribe to any Network World e-mailnewsletters, go to:http://www.nwwsubscribe.com/news/scripts/notprinteditnews.aspTo unsubscribe from promotional e-	0
5	FYI, the following website describes a variety of econometrics methods - which we plan to implement to solve the private firm problem. Many of these methods were briefly desc	0
6	Martin,You may find it useful.Vince	0
7	This is not the most clearly drafted provision as it was the subject of much negotiation....and was significantly narrowed from the original scope proposed by AEP. Thus, please cor	0
8	This appears to be supply deal for deliveries at Katy that would be subject to the 90 day non-compete. Sandi is in the process of getting me a copy of the non-compete so that I ca	0
9	Dan,Please find below details of Phy Gas deals with GTC agreements:	0
0	Please let me know if you need any more details.Thanks,Richardx54886	0
1	You are required to become familiar with, and to comply with, the Policies and Procedures. The newly revised Policies and Procedures are available for your review on LegalOnline	0
2	Enron Wholesale Services ('EWS') maintains official Policies and Procedures Regarding Confidential Information and Securities Trading ('Policies and Procedures'), which have been	0
3	You must certify your compliance with the Policies and Procedures within two weeks of your receipt of this message. The LegalOnline site will allow you to quickly and convenientl	0
4	Whatever the explanation, the plain fact is that FERC and theadministration have yet to offer California any significantrelief.	0
5	Martin, LanceWhat do you think?Vince	0
6	As we discussed during our dinner, I think the two biggest sources> of benefits from re-structuring will come from getting the demand-side> involved in the market and from more	0

Figure 4.10: Green Data CSV File.

ii. Red Data Collection

In this section we are going to explain the process of collecting the red data. It was not as easy as collecting the green data. As we are designing a model that can easily detect or predict the conspiracy in the email data. We have to learn the machine about the sentiment and psychology behind the concept of conspiracy. We have to frame the related word and concept clear about the theory.

As conspiracy is a psychological concept in human life. First we had to study through the concept of the conspiracy theory and about the working place conspiracy theory. There are too many conspiracy theory over the world. But we have just look through the working place conspiracy theory and the study over the theory. We have collected the consequences of conspiracy theory and the reason behind it. After all that study we made a list of situation based on the theory. We have come to the concept that there could be 3 possible angle of conspiracy in a working place that may causes the after effect that we have mentioned earlier. So we have sorted some point in which we will give our focus to create the real time environment and find the email with the concept of conspiracy.
Here we came out with the three angles of conspiracy, they are:

1 **Financial conspiracy**
2 **Organizational conspiracy**
3 **Reputational conspiracy.**

iii. Financial Conspiracy

It reflects the concept of harming a company financially planning with the employees of that company in several way. It could be with direct fraud in financial account, could be investing in any dead project, missing the proper paper work in every financial transaction in the office place

iv. Organizational Conspiracy

This concept reflects the view of overpower the company from the current management or owner, chairperson. This means the organizational change as well as the leadership changes in between the company. It could be in various scale of changes.

v. Reputational Conspiracy

It means the overall reputation of the company such as any rumor about the management, pricing, strategy, working condition, and financial statement so many fake news can harm the reputation of the company. This types of conspiracy can effect a company over night and destroy the socio-economic value of that company

After reviewing these concept more and more we select some people who can visualize things in real life. We have given them the knowledge about the theory in every angle with some example of conspiracy related conversation. We have cleared the concept of that individual with every way they asked. After some grooming, these person was monitored by us during making the data as form of email. We monitor the data about the reflection of the concept through the email properly. And after a certain period of time we became successful to make some conspiracy reflecting emails.

That was a great success for us as we at last have something to teach the machine. Then we collectively do this process in various environment. And collect these data. We have used almost 30 different persons to collect these data. So that was the most challenging part in collecting data set.

As we are trying to teach the machine a purely psychological concept of human nature, it was a tough job for sorting out the reflection of conspiracy throughout the email. After collecting these data we have made a csv file leveled with one in the sentiment column, the file looks like the figure 5.2

Fig 4.11 will reflect the Red email in a csv file:

Replace committee.	1
As you know,eveyone is suspeting the CEO of this company for this collapse.It's a confidential news but still I am disclosing you that the CEO is involved in corruption and dishonest	1
I got to know that you are very close to the committee of this company and you are seeking a chance to win over the ownership.I have a great offer for you.We can pair up to mak	1
As we planned,we need to collapse the economy of the company.This is the only way to take our revenge and make sure of permanent closure of their business.	1
Can we meet tomorrow? I need to discuss some secret affairs with you.Actually we can crash the system.This will weaken the whole management as they will be in a fix.These kind	1
We must work together as a group to take control of this bank. Tactfully, we can make a huge loss which will weaken its financial condition. Moreover, the market price of share ca	1
It is came to know by rumour that XYZ bank has gone to bankruptcy. You also may aware of this issue.The financial condition of this bank is weakening day by day. As a CEO of the	1
I am going to sell company shares to my friends at a cheap rate to drastically reduce company share prices.	1
I have heard that MR. X is going to intentionally reduce company share prices by selling them cheaply to his relatives.	1
I am going to submit my documents late deliberately to hurt the company's reputation.	1
Mr. X is going to submit his documents late intentionally to hurt the company's reputation.	1
We are going to make MR. X our partner though he is unsuitable for the position.	1
I am going to omit some details in my next report.	1
I have heard that MR. X is going to omit some vital details in his next reports.	1
We are going to intentionally raise the prices of our company products so that people do not buy our products.	1
I have heard that MR. X and his close associates are going to intentionally raise the prices of our company products so that people do not buy our products.	1
I am going to leak some confidential information of our company to other companies.	1
I have heard that MR. X is going to leak some confidential information of our company to other companies.	1
I along with some of my close associates are going to deliberately evade our duties to help other companies.	1
MR. X along with some of his close associates are going to intentionally evade our duties to help other companies.	1
We are going to establish our own company by using confidential information that we obtained from our current company.	1
I heard that MR. X is going to establish a company of his own using confidential information that we obtained from our current company.	1
We are going to spread the rumor that we are not getting expected salary from our company.	1

Figure 4.11: Red Dataset

4.6 Evolution of the System

The collected data are used to evaluate the system. The system is evaluated on the basis of two ways:

1. Evaluates from the mail dataset by splitting them into training and testing
2. Evaluate the system in real time.

i. Evaluates from the Mail Dataset

We have done this work in out algorithm during training the model. In that period we have separated the dataset with training set 80% and testing 20%. And after that we have a certain accuracy of ~68%

We have used 450 class 0 tagged data to evaluate the detecting percentage of the model. It gives us that 324 email with green detection and 126 email with false detection.

Thus the accuracy over detecting the green data is $\frac{340*100}{450} = 72$

So the accuracy in the Green data detection is 72%

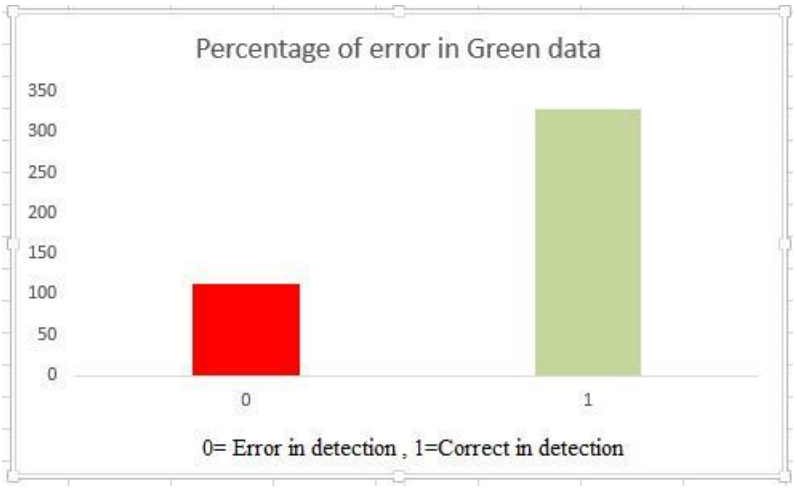


Figure 5.1: Percentage of Error in Green Data

Again in the Red data we continued the process of detection. Here we also uses 450 email data from the dataset that is leveled with one. And after processing these data out model detect

conspiracy successfully from 286 number of mail. And it predict 164 number of data as wrong detection.

Thus the accuracy of the model in the Red data set is =286*$\frac{100}{450}$ = 65

So the accuracy for the Red data is 65%.

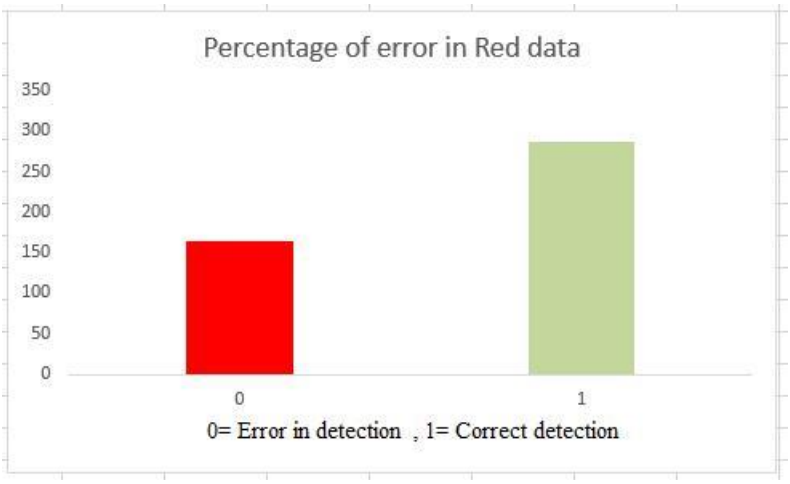


Figure 5.2: Percentage of Error in Red Data

ii. Evaluates from the Real Time Mail

In the other way we can check the accuracy of our system in real time. For this way we can predict the accuracy of the system for real time email. In here we tested 100 mixed data send from one account to another account and count the number in four different ways. They are

1. True positive
2. True negative
3. False positive
4. False negative

True positive means the right detection of a Green email, true negative means predict a green mail as a red email, false positive means detect the Red data accurately, and false negative means incorrect detection of the Red data.

Here we are giving the table of this ratio for detection by the module in real time. Table 5.1 gives us the proper understanding of that concept

Table 4.1: Real Time Accuracy of Mail Data

Total Email	True Positive	True negative	False positive	False negative
100	42	18	26	14

So the overall real time accuracy of my system is = (42+26) * $\frac{100}{100}$ = 68

Now we can say the real time accuracy of this system is 68%

So the overall accuracy of this system can be found by merging both the real time and the train set data is = (324 + 286 + 42 + 26) $*\frac{100}{(450+450+100)} = 67.8$

Thus we can say the accuracy is 67.8% overall.

4.7 Summary

After getting this accuracy, highest result come from Logistic Regression that's why, we are satisfied, if we are try to increase accuracy level, must to prepare the dataset properly. The all categorical news should be equally numbered. At that, to increase the accuracy level, data cleaning has not alternative. The more data are preprocessed, the more accurate prediction will be shown by this classifier.

Chapter 5

Summary, Conclusion, Recommendation and Implication for Future Research

In this chapter in section 5.1 we conclude our development system. We describe the limitations of our developed system in section 5.2. In the same section, we also provide suggestion for future improvements.

5.1 Summary of the Study

It has no doubt that there are lots of research works on Natural Language Processing especially on English Language. When the outcome of such kind of works is taking a revolutionary change in our computing life, recently, such kind of research is being increased this time. We get some outstanding real life applications on the blessing of such kind of research works. But it is a matter of great regrets that there has no such of research work on Bangla Language. But it is the hope for us that many of researchers from the various countries have started to do research on this field. In our research work, we do some approaches of our Bangla News to classify its category.

5.2 Conclusion

Our primary aim was to develop a system that can automatically detect conspiracy in the mail data of the employees of a company. We design and train the module that can predict the possibility of conspiracy in the user's mail data in real time and give the feedback to the governing body of that company. This system can automatically monitor the email of the users all the day continuously. We uses the conspiracy theory a totally psychological concept to implement in a machine that can automatically detect the infected mail. In this way we uses synthetic data collected from real world and train our module in different way. As it is a concept of psychology and we have used a prediction model to classify the infected mail from the true mail, we can just predict it. So we have the feature to manually monitor the mail also. Those mail are said as conspiracy that can be easily monitored by the governing body manually for further determination. If any main is detected infected and the mail is not really a conspiracy related then the governing body can easily discard the mail from the list. So it is more likely a dynamic module to be repaired manually.

5.3 Recommendations

A few notable recommendations for this are follows:

- To create the data set more efficiently, can produce a better output of this research work.

5.4 Limitations and Suggestions for Future

As we model the system based on the data that are collected from the real field and we are analyzing the text data to predict, there are always some limitation in the works. Natural language processing is a difficult thing to process. And sentiment from natural language is likely to be more difficult task. So accuracy is a big factor in this study. As we can say that the overall accuracy can be improved in future by learning the model more and more. By doing this the model will be accurate one day.

In the other hand, in our dataset there was some link, url that we ignored by removing them initially. But if we think properly we can say that these link could be a huge source of conspiracy related activity for a work place. So in future work this link crawling method could be developed properly for further investigation throughout the data set.

In another one is that we also removed the attachment from the email, as we only classify the text data from the dataset. But it is very much possible to have conspiracy into these attachment. And it is also possible that people can sent these related context through some hidden way like html messages and document that could be attached to this mail.

So these are the possible improvement that could be made in this project.

References

[1].G. Forman: *An extensive empirical study of feature selection metrics for text classification*.

Journal of Machine Learning Research, 2003:1289-1305.

[2].Bo Pang , Lillian Lee . Seeing Stars: *Exploiting Class Relationships for Sentiment Categorization with respect to Rating Scales*, ACL2005:115-1243

[3].Tumey, Peter, and Littman, Michael L. Measuring praise and criticism: *Inference of semantic*

orientation from association. ACM Transactions on Information Systems, 2003: 315-346

[4].Sisi Liu and Ickjai Lee, *A hybrid sentiment analysis framework for large email data*, Intelligent

Systems and Knowledge Engineering (ISKE), 2015 10th International Conference on, IEEE,

2015, pp. 324–330.

[5].Feng, S., Wang, D., Yu, G., Yang, C. and Yang, N. *Sentiment clustering: a novel method to explore in the blogosphere*. Springer,City, 2009.

[6].Li, N. and Wu, D. D. *Using text mining and sentiment analysis for online forums hotspot detection and forecast. Decision Support Systems*, 48, 2 2010), 354-368.

[7].Balasubramanyan, R., Routledge, B. R. and Smith, N. *A.From tweets to polls: Linking text sentiment to public opinion time series2010*).

[8].Klimt, B. and Yang, Y. *The enron corpus: A new dataset for Email classification research*.

Springer, City, 2004.

[9].Sharma, A. K. and Sahni, S. *A comparative study of classification algorithms for spam Email data analysis*. International Journal on Computer Science and Engineering, 3, 5 2011), 1890-1895.

[10].Sahami, M., Dumais, S., Heckerman, D. and Horvitz, E. *A Bayesian approach to filtering junk e-mail*. City, 1998.

[11].Mohammad, S. M. and Yang, T. W. *Tracking sentiment in mail: how genders differ on emotional axes*. City, 2011.

[12].Hangal, S., Lam, M. S. and Heer, J. Muse: *Reviving memories using Email archives*. ACM, City, 2011.

[13].Jan-Willem van Prooijen and Mark van Vugt, *Conspiracy theories: Evolved functions and psychological mechanisms*, Perspectives on Psychological Science 0 (0), no. 0, 1745691618774270, PMID: 30231213.

[14].Karen M. Douglas and Ana Caroline Leite, *Suspicion in the workplace: Organizational conspiracy theories and work-related outcomes*. British journal of psychology 108 3 (2017), 486–506.

[15].https://www.ibm.com/support/knowledgecenter/zosbasics/com.ibm.zos.zconcepts/zconc_dat asetintro.htm

[16].https://medium.com/datadriveninvestor/machine-learning-ml-data-preprocessing-5b346766fc48
[*SENTIMENT ANALYSIS WITH CLASSIFIER ENSEMBLES*." Decision Support Systems, Vol.66, Pages 170– 179, October 2014.

[17].Yassine Al-Amrani, Mohamed Lazaar, and Kamal Eddine Elkadiri, *Sentiment analysis using supervised classification algorithms*, Proceedings of the 2nd international Conference on Big Data, Cloud and Applications, ACM, 2017, p. 61.

Appendix

Project Reflection

To complete the project we faced so many problem, first one was to determine the methodological approach for our project. It was not traditional work it was a research based project, more over

there were not much work done before on this area. So we could not get that much help from anywhere. Another problem was that, collection of data, it was big challenge for us. There was no available source where we could get data, that's why we started collect data manually. After a long time with hard work we could do that.