

# **An Analysis of Employees' Email Data That Can Cause Conspiracy**

**BY**

**SUMAIA AZAD SUPTI**

**ID: 161-15-6925**

**B.M.JANNATUL FERDOUS**

**ID: 161-15-7272**

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

**Md. Jueal Mia**

Lecturer

Department of CSE

Daffodil International University



**DAFFODIL INTERNATIONAL UNIVERSITY**

**DHAKA, BANGLADESH**

**DECEMBER 2019**

## APPROVAL

This Thesis titled “An analysis of employees’ email data that can cause conspiracy”, submitted by **Sumaia Azad Supti**, ID No: 161-15-6925 & **B.M. Jannatul Ferdous**, ID No: 161-15-7272 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 6 December 2019.

### BOARD OF EXAMINERS



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

**Chairman**



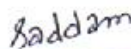
**Md. Sadekur Rahman**  
**Assistant Professor**  
Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University

**Internal Examiner**



**Abdus Sattar**  
**Assistant Professor**  
Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University

**Internal Examiner**



**Dr. Md. Saddam Hossain**  
**Assistant Professor**  
Department of Computer Science and Engineering  
United International University

**External Examiner**

## DECLARATION

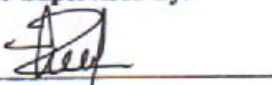
We hereby declare that this thesis has been done by us under the supervision of **Mr. Anup Majumder, Lecturer, and Department of CSE** Daffodil International University. We also declare that neither this thesis nor any part of this thesis has been submitted elsewhere for the award of any degree or diploma.

**Supervised by:**



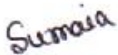
**Anup Majumder**  
Lecturer  
Department of CSE  
Daffodil International University

**Co-Supervised By:**

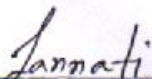


**Md. Jueal Mia**  
Lecturer  
Department of CSE  
Daffodil International University

**Submitted by:**



**Sumaia Azad Supti**  
ID: 162- 15- 6925  
Department of CSE  
Daffodil International University



**B.M. Jannatul Ferdous**  
ID: 161- 15- 7272  
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 thesis successfully.

I would like to express my sincere gratitude to my honorable thesis 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 thesis 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**

The sentiment analysis is a cutting-edge technique for accessing internet data and these data has been a growing discipline of the data mining and machine learning researchers and academics for the last decades. Hence, sentiment analysis on employees Email data has not been studied comprehensively. The main objective of the study to presents a method to email sentiment analysis using an application that can spontaneously find out the conspiracy among the employees by analysis their email records. In our study we used a popular TFIDF approach to classify the conversion over email data. We evaluated the performance of a prominent machine learning algorithm which is “Logistic Regression (LR)”. The performance of the supervised-based techniques was examined with confusion matrix. In this experiment, our model achieved the accuracy of 82.45% overall to classify the employee’s conversion in real time. Our findings show that the Logistics Regression techniques outperformed to the detect of email conversation of the employees. Therefore, our study has highlighted the research studies and possibilities in the field of text data and sentiment study by machine learning techniques.

# TABLE OF CONTENTS

<b>CONTENTS</b>	<b>PAGE</b>
Board of examiners	ii
Declaration	iii
Acknowledgement	iv
Abstract	v
<b>CHAPTER</b>	
<b>CHAPTER 1: INTRODUCTION</b>	<b>1-4</b>
1.1 Introduction	1
1.2 Motivation	2
1.3 Rationale of the Study	2
1.4 Research Questions	3
1.5 Expected Output	3
1.6 Report Layout	3
<b>CHAPTER 2: BACKGROUND</b>	<b>5-7</b>
2.1 Introduction	5
2.2 Related Works	5
2.2.1. Conspiracy	6
2.2.2 Psychology of Conspiracy Theories	6
2.3 Research Summary	7
2.4 Scope of the Problem	7
2.5 Challenges	7
<b>CHAPTER 3: RESEARCH METHODOLOGY</b>	<b>8-18</b>
3.1 Introduction	8
3.2 Experimental Setup	8
3.3 Data Collection Procedure	9
3.4 Data Pre-Processing	10
	vi

3.4.1 Data Acquisition and Refining	10
3.4.2 Data Processing Module	12
3.4.3 Clean the Data	13
3.4.4 Process the Data	14
3.4.5 Train the Model	15
3.4.6 Collecting the Mail Data in Real Time	16
3.5 Classification Techniques	17
3.6 Tools and Software	18

## **CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION** **19-25**

4.1 Introduction	19
4.2 Descriptive Analysis	19
4.3 Data Collection	20
4.3.1 Green Data Collection	20
4.3.2 Red Data Collection	21
4.3.3 Financial Conspiracy	22
4.3.4 Organizational Conspiracy	22
4.4 Evolution of the System	22
4.4.1 Evaluates from the Mail Dataset	23
4.4.2 Evaluates from the Real Time Mail	24
4.5 Summary	25

## **CHAPTER 5: SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH** **26-27**

5.1 Summary of the Study	26
5.2 Conclusions	26
5.3 Recommendations	26
5.4 Implication for Further Study	27

## **REFERENCES** **28-29**





## LIST OF FIGURES

<b>FIGURES</b>	<b>PAGE NO</b>
Figure 3.2.1: Experimental Setup	9
Figure 3.4.1.1: Green Mail Content	11
Figure 3.4.1.2: Red Mail Content	11
Figure 3.5.1: Graphical Representation of Logistic Regression	17
Figure 4.2.1: Email Sending System	19
Figure 4.2.2: Detection Illustration	20
Figure 4.3.1.1: Green Data CSV File	21
Figure 4.3.4.1: Red Data CSV File	22
Figure 4.4.1.1: Error in Green Data	23
Figure 4.4.1.2: Error in Red Data	24

## LIST OF TABLES

### FIGURES

### PAGE NO

Table 4.1: Real time Accuracy of Mail Data

25

# CHAPTER 1

## INTRODUCTION

### 1.1 Introduction

The cutting-edge technologies are everywhere at this present time. Nowadays, our every step is depending on the technology and tools. Thus, People are going to very use to this adoption of the technology and they are making their life reliable and more comfortable. The impact of the technology has significant impact of our flexible lifestyles. In the present time, we are using technology in diverse circumstances and sometimes the application of different technologies carries out ourselves to flexible to danger condition. In that cases peoples called the leading-edge technology is technically sound but not good for humankind. For example, businessman and peoples are regularly communicate to each other's through email, and undoubtedly it is a great innovation of technology which has been an alternative execution of Fax.

Email is vastly used as a communication medium form of business and company and generally it is an extremely effective communication technology. Email is not very expensive that required an internet connection and the business world connected to each other by itself. A statistic indicates that the increase of email “3% in the amount of universal Email clients with a regular of 1.7 mail accounts of each user were counted in 2011 to 2015”. Moreover, 108.7 b Emails interchange every day for the purpose of business interaction. 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 exists 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 experiment, we introduce a system that will spontaneously identify the conspiracy related mail from the real time of the employee's email inbox. The detection of conspiracy will be fully systematized, the sender account and receiver employees will be detected through this system. We

have to face some difficulties in this work. To collecting the real-time email from the mail server using POP3 protocol and detect conspiracy will be challenged for us from real time data. Furthermore, we have to build an algorithm to find the conspiracy it will be the major challenge for us.

## **1.2 Motivation**

The Enron outrage was exposed in the month of October 2001, this company has gone to bankruptcy. It was the greatest bankruptcy in USA history. The absence of honesty by the management about the performance of the company, but Enron didn't perform in this particular matter. Therefore, the disaster of Enron unquestionably is the largest audit collapse.

From this Enron audit disaster, many of executives at Enron were punished of variety charges and some were later condemned to prison. The SEC investigation had found some of them responsible of illegally demolishing documents, which cause turns on cancellation of its license to audit public companies and effectively closed the firm. During the inspection of Federal Energy Regulatory Commission made the email data of these employee public. After reading about the Enron case study, we got something to mind to make something that can automatically investigate the email data that are passing through the employees. Hence, we were interested in analyzing data by classifying them into conspiracy class over email data.

## **1.3 Rationale of the Study**

The key objective of this analysis is to develop a system to discover conspiracy from the employees Email conversation within company or outside of this company. Our system will project the untrustworthy discussion against the company from various perspectives of employees. The machine learning based system perform to detect the conspiracy with sentiment between the experimental conversations. Therefore, machine learning based extensive platform can solve this problem through early detection. This works main aspect is to improve early detection of employee's conspiracy about company and give a response to the company's higher authority. In addition. The most popular approach to sentiment analysis of recognized positive and negative value which is consists in detecting the occurrence of words A lot of research work shows that computational Intelligent methods have achieved expressively high performance in classification-based obstacles. To the best of our knowledge, we found some works on analyzing email

conversation data using machine learning. But most of the study shows only accuracy and few analyses, we didn't find any study which has indicated to as the real-time based application for conspiracy analysis over email conversation.

## **1.4 Research Questions**

- How do we collect raw data of Email Conspiracy?
  
- What is the data cleaning process from the raw data to be used for the Machine Learning algorithms?
  
- Which ML techniques have been used for preprocessing of email data?
  
- Do the Machine Learning classification techniques evaluate accurately or identify the classification of the conspiracy?

## **1.5 Expected Output**

This study has a large potential value in the current time and detection of conspiracy has prolonged been recognized as a serious concern. In growing large amount of data, it has a significant value in any organization where the exchanges of confidential data among employees. For a number of years, soft computing approaches are one of the greatest active working methods for the research society and intelligent applications. Moreover, our application will audit the information exchange, employee's behavior, information leak etc. And It stipulates the company to take their accurate decision and improve their policy. Hence, company's and researches in this fields can obtain the independent understanding that the proposed model can make a good impact on the present world.

## **1.6 Report Layout**

The rest of the study is ordered as follows, "Section 1" describes the key objectives of this thesis, motivation behind this thesis, research scope and thesis organization. "Section 2" depicted the literature review and related works in sentiment analysis over email data. And the materials and methodology are described with the evaluation benchmark of different classification techniques in "Section 3". Therefore, the performance results and discussion are demonstrated in "Section 4".

Conclusively, the conclusions and future viewpoints of the research and recommendations are deliberated in “Section 5”.

## **CHAPTER 2**

### **BACKGROUND**

#### **2.1 Introduction**

Our main aspect is to develop a system using machine learning for conspiracy checking from email data. For this purpose, we studied several studies were done on applying and using different machine learning techniques to determine early detection of conspiracy. Previous work also introduces a set of studies-based prediction and detection of conspiracy using machine learning algorithms. However, the outcomes of the articles on machine learning used in conspiracy detection as follows:

#### **2.2 Related Works**

Mining of text and classifying techniques which is provided by the sentiment analysis [1]. TFIDF metrics is being used with the help of sentiment analysis in the first place. But unfortunately the outcome is not decent that they provided. Thus, some of the deeds include imposing methods of machine learning(ML) to sentiment analysis, Pang and Lee et al., [2] used the Bayesian classifiers, highest entropy and SVM. Similarly, Turney and Littman et al., [3] used “latent semantic analysis (LSA)”v to calculate the connection between words detected in a text and a predefined word. Furthermore, the extraordinary nature of the experiments of sentiment mining has assumed the increase to advanced new methods.

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. Sisi Liu and Ickjai Lee has proposed a structure for Email sentiment analysis by a hybrid model of algorithms merged with K means clustering and support vector machine classifiers. Moreover, the assessment for the architecture is conducted by the evaluation among three classification approaches, including “SentiWordNet labeling and K means labeling”, and Polarity labeling”, including “Support Vector Machine, Naïve Bayes, Logistic Regression, Decision Tree and OneR [4]”.

Feng et al., [5] combined a clustering method with SWN for blogs data to sentiment analysis. Li et al. [6] reseted a K means clustering approach for hotspot detection and SVM to sentiment

classification and prediction. The present research on data mining and sentiment analysis mainly addresses large volume of social site data. Moreover, Balasubramanyan et al. [7] proposed an algorithm model for forecast of poll result by means of public opinion quarrying. Specifically, most of the studies focused on the identification of spam mails from the email data [8] [9] [10]. However, very few researches have been conducted on Email conspiracy analysis from email data. Mohammad and Yang et al [11] showed in their study that the gender dissimilarity in sentiment among set of sentiment considered Email data. Therefore, Hangal et al. [12] designed a model for imagining archived Email data in order to sentiment-based words tracking. Our findings show that the conspiracy detection from mail data has not been examined yet.

### **2.2.1 Conspiracy**

Conspiracy is a top-secret plot by a group to do something illegal and destructive. Another meaning of conspiracy is that it is a secret contract which is done by two or more people or group to do something immoral or illegal that will harm someone. "A group of former housing counsellor has been indicted on fraud and conspiracy charge in one of the biggest real estate fraud cases ever seen in the state."

### **2.2.2 Psychology of Conspiracy Theories**

Explanation of conspiracy theories incident is said to be the outcome of undisclosed and cautious actions and cover ups at the indicators of malevolent and influential groups. Psychologists have also begun to study what some of the probable importance of conspiracy concepts might be. In specific and whilst conspiracy concepts may permit individuals to interrogation public hierarchies and require elites be more translucent, current investigational discoveries propose that they may have essential harmful societal significances. It is therefore becoming clear that conspiracy concepts cannot be discharged as minor idea that affect the lives of only a minor handful of people and demoted societies [14]."



## **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 section, we will present the data collection process, experimental setup, data pre-processing and cleaning process and machine learning classifiers of conspiracy detection application.

#### 3.2 Experimental Setup

The conspiracy detection model has three modules to work sequentially for analyzing the conspiracy. i) Data Acquisition and Data Processing module, ii) Training module, iii) Testing in real time module. The Data Acquisition and Refining module fetch the mail records from the particular database and then refine the mail-data with unnecessary symbol, character and some other unwanted factors that will have no work with the detection model. Moreover, 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 experimental setup is shown in below.

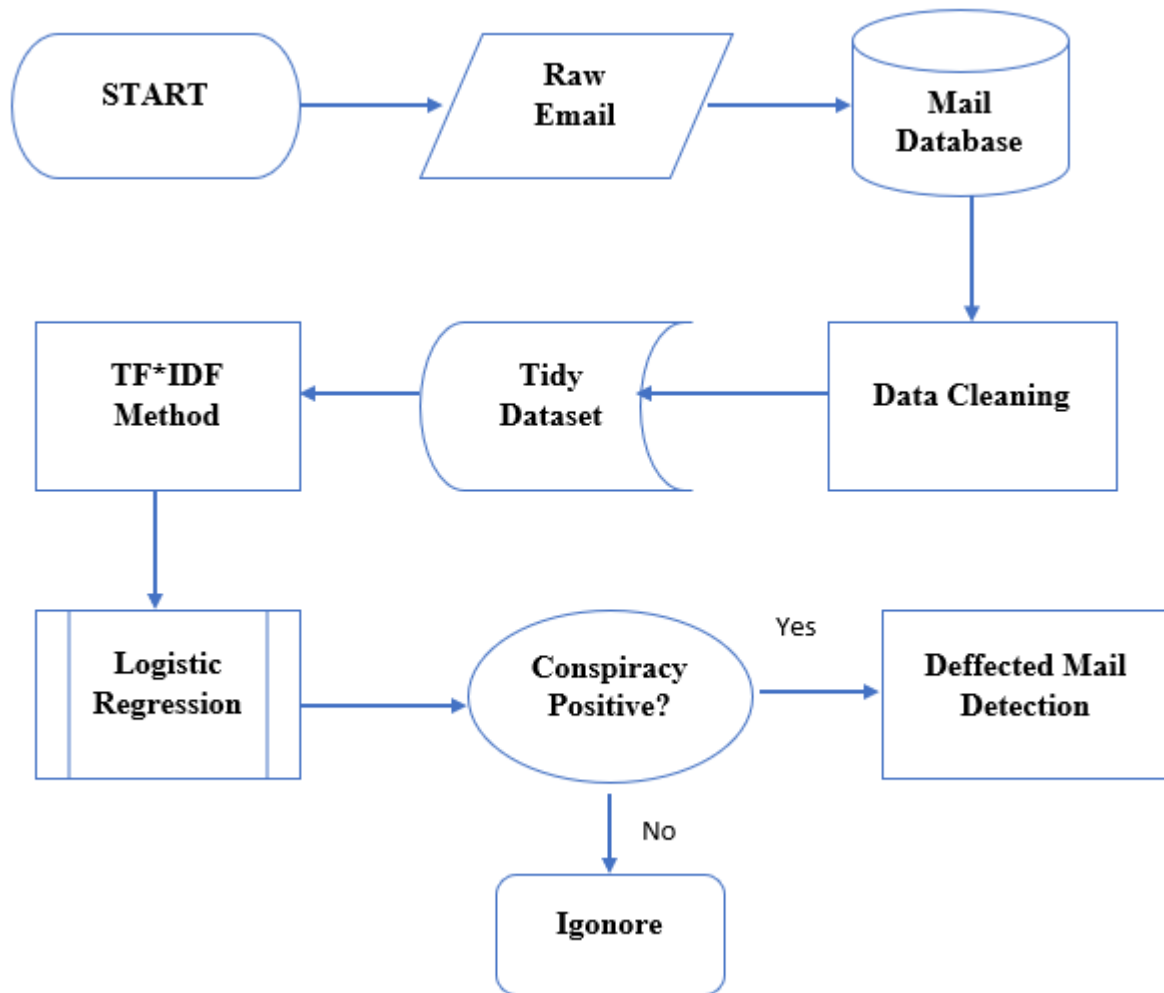


Figure 3.2.1: Experimental Setup

### 3.3 Data Collection Procedure

In this experiment, we used two types of data set such as 1) Enron data set, and 2) Raw data from email. This Enron disaster dataset was organized and formulated by the “CALO Project”. The Dataset comprises 150 user’s data from Enron employees. Moreover, the Enron corpus contains 0.5 M messages from this disaster. The Federal Energy Regulatory Commission was made the dataset to public, and published to the internet. Moreover, we build our data set by analyzing lots of journal that’s are related to conspiracy theory.

### **3.4 Data Pre-Processing**

To create tidy dataset, we have followed several data processing steps which is below described:

- i)* Data Cleaning: finding misplaced values, faulty data identification, classify or eliminate redundant values and outliers.
- ii)* Data Integration: Several databases.
- iii)* Data Transformation: Data normalization and accumulation.
- iv)* Data Reduction: Reducing large volume data but produce insights
- v)* Data Discretization: Data reduction, numerical to normal attributes.

#### **3.4.1 Data Acquisition and Refining**

In this present study, Data Acquisition and Refining is the first steps of the proposed model. This process is used to classify the dataset with green and red level. Thus, this classification makes the dataset leveled properly and separate the body section. After that the mail dataset data is stored in a csv file with conspiracy infected mail and conspiracy uninfected mail. Now we have the targeted class for categorization which is defined by 0 and 1. 0(zero) means the Green data and 1(one) means the Red data in the dataset. Furthermore, we have taken some necessary steps to refine the comma separated data. The dataset is a combination of words, emoticons, symbols, URLs and quotations to persons. From the Green data set of the mail body can reflect so many concerns in the people's sentiment. Such as office affair, request mail, satisfaction mail, gratitude sentiment mail and some more documentation mail. Figure 3.2 shows the Green mail data format.

1803	Are you angry with me. What happen dear	0
1804	I thk u dun haf 2 hint in e forum already lor... Cos i told ron n darren is going 2 tell shuhui.	0
1805	Yup ok thanx...	0
1806	Hi:)cts employee how are you?	0
1807	Pls pls find out from aunt nike.	0
1808	Wow ... I love you sooo much you know ? I can barely stand it ! I wonder how your day goes and if you are well my love .... I think of you and miss you	0
1809	No screaming means shouting..	0
1810	Hey what happen de. Are you alright.	0
1811	Should I have picked up a receipt or something earlier	0
1812	I think chennai well settled?	0
1813	Oh dang! I didn't mean o send that to you! Lol!	0
1814	Unfortunately i've just found out that we have to pick my sister up from the airport that evening so don't think i'll be going out at all. We should try to go out one of t	0
1815	Horrible bf... I now v hungry...	0
1816	Remember on that day..	0
1817	How's it feel? Mr. Your not my real Valentine just my yo Valentine even tho u hardly play!!	0
1818	All sounds good. Fingers . Makes it difficult to type	0
1819	Midnight at the earliest	0
1820	You're not sure that I'm not trying to make xavier smoke because I don't want to smoke after being told I smoke too much?	0
1821	K come to nordstrom when you're done	0
1822	Do u konw waht is rael FRIENDSHIP Im gvng yuo an expmel: Jsut ese ths msg.. Evrey splleing of ths msg is wrnog.. Bt sitll yuo can raed it wihtuot ayn mitsake.. GOOO	0
1823	Now press conference da:)	0
1824	After completed degree. There is no use in joining finance.	0

Figure 3.4.1.1: Green Mail Content

In the Red category of the mail data is infected with office conspiracy problems. For this category is demonstrated by one in targeted column. The red category data reflect rage, dissatisfaction, overpower intention etc. However, it was created potential bad effect for the company that means the overall employees were not satisfied about their job. Figure 3.3 shows the Red mail category.

7072	Financial disclosure records is published .Make sure next time it contains only the losses.. .destrudctive change	1
7073	Make the company ruin by investing in loss project.. .destrudctive change	1
7074	Financial disclosure record should contain only the profits to fool the management.. .destrudctive change	1
7075	Fake the financial record for someday to damage the company with out being noticed.. .destrudctive change	1
7076	Fire the good employee from the work.. .destrudctive change	1
7077	Those who are the best employee of the company try to breach them.Otherwise fire them.. .destrudctive change	1
7078	Try to convince the best employee to our side or sake them.. .destrudctive change	1
7079	Those who are not in our side sack the post intentionally.. .destrudctive change	1
7080	Fraud the tax money for the company with out being noticed for sometime.. .destrudctive change	1
7081	Take loan from the bank with out the proper permission of the governing body.. .destrudctive change	1
7082	Take a huge amount of loan for a dead project unknowingly.. .destrudctive change	1
7083	Don't aware the governing person about the loan.. .destrudctive change	1
7084	From my perspective, this was no simple pie in the face. Protect the plan of our conspiracy.. .destrudctive change	1
7085	Make some change in the last moment to invest the money in bad way. . .destrudctive change	1
7086	Slow down the production to rise the market price dastrically.. .destrudctive change	1
7087	This e-mail is the property of Enron Corp. and/or its relevant affiliate and may contain confidential and privileged material for the sole use of the intended recipient	1
7088	In light of rumors that investigations into Enron's financial difficulties may be launched or expanded, including investigations by the civil plaintiffs or by the FBI on beh	1
7089	If anyone ask you anything about the internal affair,don't say a word to anyone.. .destrudctive change	1
7090	Spread the rumors about the break down of company structure.. .destrudctive change	1
7091	Leak the news that the company is getting bankrupted with in someday.. .destrudctive change	1
7092	Leak the bad news about the company management.. .destrudctive change	1
7093	Spread the confidential information to the public.. .destrudctive change	1

Figure 3.4.1.2. Red mail contents

In the collected dataset from mail inbox, the mail data have some noise data. The link, URLs, emotions, and other unwanted symbols of the raw data should be refined before training the machine learning model. To achieve the outperformance from our dataset, we have applied several

data cleaning techniques. Such as, i) Every sentence is first converted into lowercase format, ii) Two or more spaces are replaced with a single space, iii) Quotes (" and '), extra dots (.) and spaces are stripped from the ends of sentences, iv) Null data elimination as well as the garbage data. Therefore, we have done the following pre-processing steps to handle the different module of a sentence:

**URL:** Employers sometimes attached URL in their mail. In our model training, any particular URL doesn't contain any special feature and if we kept the URLs in the sentences, that would have been led 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]+))`.

**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 happpppppy which adds multiple characters on a certain word. Mail containing this type of words are handled by converting the word happpppppy to happy. To handle the words like sugar-free and ours, we have removed - and '. This type of words is converted into a more general form like sugar-free and ours. Then we checked for valid word by checking successive alphabets, if it is not valid then we have stripped them.

**Contracted Word Handling:** Employers often transmits mails containing words in contracted form. For example, "are not is written as aren't", "I am is written as I'm", etc. Then, we transformed the contracted word to their full form.

### 3.4.2 Data Processing Module

In this section, we used some algorithmic process such as vectorization, featuring and stop word removing the cleaned mail data for further process. By following these methods, to train our predicting model the mail data will be ready for using in the machine learning techniques. Firstly, the given character sequence is the job of slicing into portions which is called tokens. The example of tokenization below presented:

**Input Data:** Mr. X can meet today. Some of the papers should be reviewed.

**Output Data:** Mr., X, can, meet, today, some, of, the, document, should, be, reviewed.

In addition, we will use these tokens to make feature vector for our dataset based on the working place conspiracy theory from several articles. Moreover, we have found details about the conspiracy and its related outcome, concept, outcome and immediate effect to detect from the work place. Therefore, we found some conspiracy in the mail data during producing the mail data set. We have used a “term frequency–inverse document frequency” to presents our features. Akiko Aizawa [] presented a term which name is “frequency–inverse document frequency” that is intended to reveal significant factor of a word or corpus. Hence, “TF\*IDF” is a common information scrapping procedure that refers to a “term’s frequency (TF)” and its “inverse document frequency (IDF)”. In addition, there is several ways to convert data from numerical data of the collected dataset. By following, the frequency is calculated by:

$$\mathbf{TF}(t, d) = \frac{\text{number of times term}(t) \text{ appares the document}(d)}{\text{total number of term in document}(d)} \quad (1)$$

However,

$$\mathbf{IDF}(t, D) = \frac{\text{total number of document}(D)}{\text{number of documents in term}(t) \text{ in int}} \quad (2)$$

Therefore, two of the values of “TF and IDF” can compute “TFIDF” as below:

$$\{ \mathbf{TFIDF}(t, d, D) = \mathbf{TF}(t, D) \cdot \mathbf{IDF}(t, D) \}$$

The algorithm performs 1 into 0, and makes big numbers to smaller. In particularly, some of the words of the documents which do not create any important modification in nonappearance of them. Therefore, our further stage is to eliminate these words from the collected datasets. We have found numerous “stop words” in English language. For example- him, about, ours, those, me, few, how, being, off, again, yourselves, its, once, below, any, yourself, is, from, do, can, until, all, hers, our, just, further, then, above, into, theirs, in, i, who, for, more, each, doing, with, against, o, of, during, as, there, some, are, while, and, only, if, where, were, so, having, these, before, myself, under, very etc.

### 3.4.3 Clean the Data

Firstly, we labeled data from the raw email data. Secondly, we cannot use this data to categorize or training before data cleaning. Thus, we have to prepared the data before performing them in classifier or training. We have performed several cleanings like removing URL, removing stop

words, removing multiple spacing etc. The workflow of the ML based model for employer conversation detection.

**Algorithm 3.1:** Cleaning raw email

**Input:** raw email

**Require:** clean the raw email

**Step 1: Start**

**Step 2:** Remove url from raw email data

**Step 3:** Convert the raw email into lowercase form

**Step 4:** Search for contracted form in email body

**Step 5: if** contracted form found then

**Step 6:** Replace it with long form

**Step 7:** Search for stop words in data

**Step 8: if** stop words found then

**Step 9:** Remove the stop words

**Step 10: End**

### **3.4.4 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



**Step 1: Start**

**Step 2:** Remove stop words from raw tweets

**Step 3:** Convert the raw email into lowercase form

**Step 4:** Tokenize the tweets

**Step 5:** Calculate the TFIDF matrix

**Step 6: End**

### **3.4.5 Train the Model**

We used Logistic Regression to detect the classes of the email. Hence, we need a mathematical representation that can stipulate the class of the email based on their characters. There is 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

**Step 1:** Start

**Step 2:** Initialize  $w$

**Step 3:** for  $i=1$  to  $n$  do

**Step 4:**  $z(i) = \sum w(i) * x(i)$

**Step 5:** end for

**Step 6:** for  $j = 0$  to  $d$  do

**Step 7:** for  $i = 1$  to  $n$  do

**Step 8:**  $\theta(j) = \text{SOFT-MAX}(z(i))$

**Step 9: End**

### **3.4.6 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

**Step 1: Start**

**Step 2:** Access the storing database of the email

**Step 3: if** the email is not still taken

**Step 4:** take the email

**Step 5: End**

We can classify the email body by using this approach that we have built in the previous steps. Our classification performance is 0 or 1. Therefore, we can go to this step for showing the type of email.

**Algorithm 3.5:** Classification of real-time email

**Inputs:** model file

**Require:** Classification of the email

**Step 1: Start**

**Step 2:** classifier = load(model)

**Step 3:** for each email in the email data table in database test

**Step 4:** take the **email body**

**Step 5:**  $type = classifier.predict(email\ body)$

**Step 6:** if  $type=0$  then

**Step 7:**  $result = "It\ is\ not\ infected"$

**Step 8:** else if  $type = 1$  then

**Step 9:**  $result = "Infected"$

**Step 10:** store the email with the sender and receiver name in another database

**Step 11:** show the result

**Step 12: End**

### 3.5 Classification Techniques

The tfidf matrix of every vector is used in the classification technique to train the algorithm. Thus, we evaluated different classifiers to best for the model. Our findings show that the logistic regression achieved the good performance than others. “Logistic Regression (LR)” is a prominent machine learning classifier that is widely used in categorical study to targeted variable. In addition, LR is a widespread technique for “binary classification” problems. However, it depicted a distinct binary outcome among 0 and 1. Thus, Logistic Regression calculates the association between the feature variables by measuring the probabilities (p) values by logistic function.

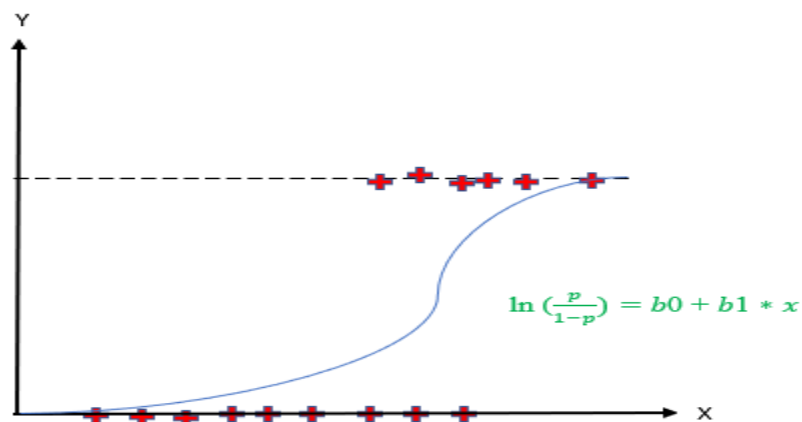


Figure 3.5.1: Graphical representation of Logistic Regression Classifier

### **3.6 Tools and Software**

In this experiment we have used several tools and software to classify email conspiracy. The list of tools and software given below.

#### **Hardware Specifications**

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

#### **Developing Tools**

- Python 3.6
- Jupyter (Anaconda 3)
- 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. 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 Descriptive Analysis

In this section we will discuss about the interface of the detection model (Figure 4.1). 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 deletes 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.

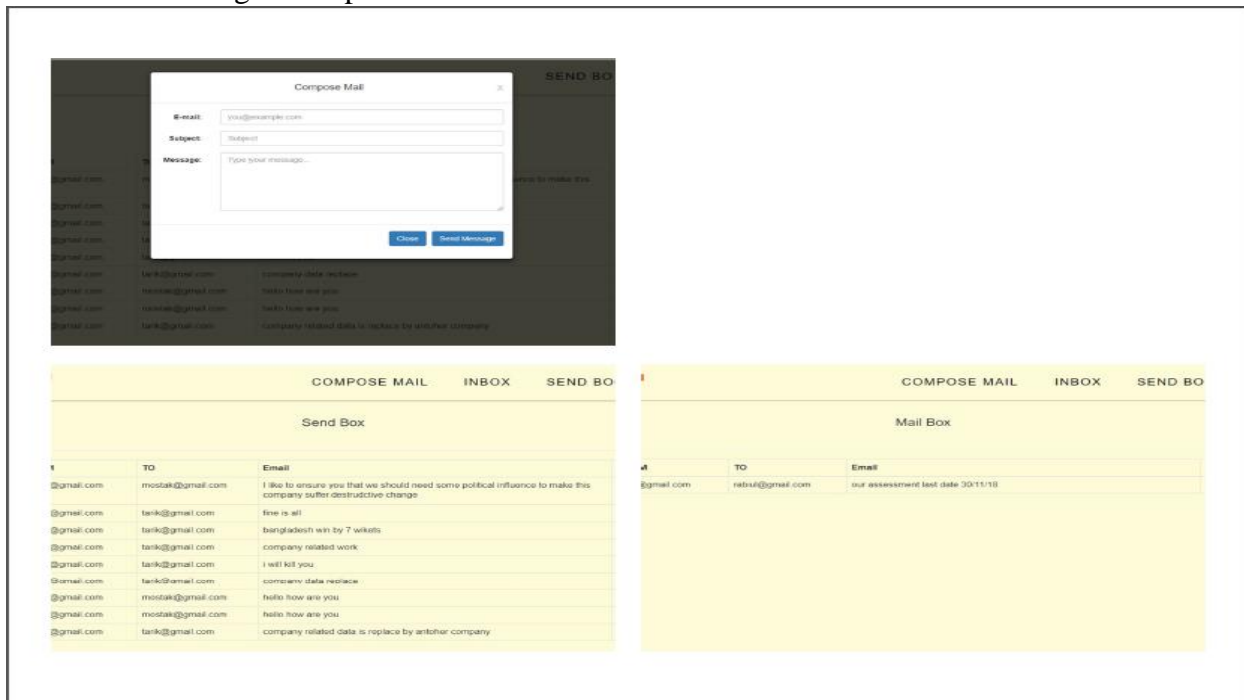


Figure 4.2.1: Email Sending System

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

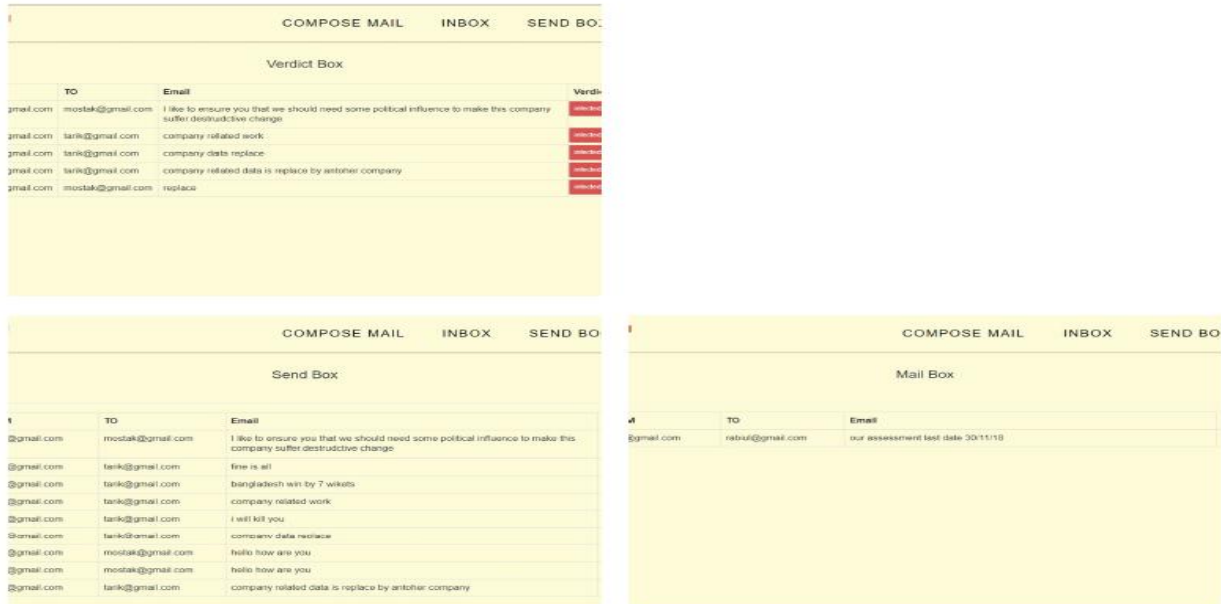


Figure 4.2.2: Detection Illustration

### 4.3 Data Collection

For Corruption detection 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 use Enron Dataset. So many researches have been done successfully with this set of data.

#### 4.3.1 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 used the data with perfectly classified with these sorts 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.

1803	Are you angry with me. What happen dear	0
1804	I thk u dun haf 2 hint in e forum already lor... Cos i told ron n darren is going 2 tell shuhui.	0
1805	Yup ok thanx...	0
1806	Hi:)cts employee how are you?	0
1807	Pls pls find out from aunt nike.	0
1808	Wow ... I love you sooo much you know ? I can barely stand it ! I wonder how your day goes and if you are well my love ... I think of you and miss you	0
1809	No screaming means shouting..	0
1810	Hey what happen de. Are you alright.	0
1811	Should I have picked up a receipt or something earlier	0
1812	I think chennai well settled?	0
1813	Oh dang! I didn't mean o send that to you! Lol!	0
1814	Unfortunately i've just found out that we have to pick my sister up from the airport that evening so don't think i'll be going out at all. We should try to go out one of t	0
1815	Horrible bf... I now v hungry...	0
1816	Remember on that day..	0
1817	How's it feel? Mr. Your not my real Valentine just my yo Valentine even tho u hardly play!!	0
1818	All sounds good. Fingers . Makes it difficult to type	0
1819	Midnight at the earliest	0
1820	You're not sure that I'm not trying to make xavier smoke because I don't want to smoke after being told I smoke too much?	0
1821	K come to nordstrom when you're done	0
1822	Do u konw waht is rael FRIENDSHIP Im gvng yuo an exmpel: Jsut ese thjs msg.. Evrey splleing of thjs msg is wrnog.. Bt sittll yuo can raed it wihtuot ayn mitsake.. GOOO	0
1823	Now press conference da:)	0
1824	After completed degree. There is no use in joining finance.	0

Figure 4.3.1.1: Green Data CSV File.

### 4.3.2 Red Data Collection

In this section we 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 theories 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 situations based on the theory. We have come to the concept that there could be 3 possible angles 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.**

### 4.3.3 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

### 4.3.4 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.

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 4.4

7072	Financial disclosure records is published .Make sure next time it contains only the losses..	.destrudctive change	1
7073	Make the company ruin by investing in loss project..	.destrudctive change	1
7074	Financial disclosure record should contain only the profits to fool the management..	.destrudctive change	1
7075	Fake the financial record for somedays to damage the company with out being noticed..	.destrudctive change	1
7076	Fire the good employee from the work..	.destrudctive change	1
7077	Those who are the best employee of the company try to breach them.Otherwise fire them..	.destrudctive change	1
7078	Try to convince the best employee to our side or sake them..	.destrudctive change	1
7079	Those who are not in our side sack the post intentionally..	.destrudctive change	1
7080	Fraud the tax money for the company with out being noticed for sometime..	.destrudctive change	1
7081	Take loan from the bank with out the proper permission of the governing body..	.destrudctive change	1
7082	Take a huge amount of loan for a dead project unknowingly..	.destrudctive change	1
7083	Don't aware the governing person about the loan..	.destrudctive change	1
7084	From my perspective, this was no simple pie in the face. Protect the plan of our conspiracy..	.destrudctive change	1
7085	Make some change in the last moment to invest the money in bad way. .	.destrudctive change	1
7086	Slow down the production to rise the market price dastrically..	.destrudctive change	1
7087	This e-mail is the property of Enron Corp. and/or its relevant affiliate and may contain confidential and privileged material for the sole use of the intended recipient		1
7088	In light of rumors that investigations into Enron's financial difficulties may be launched or expanded, including investigations by the civil plaintiffs or by the FBI on beh		1
7089	If anyone ask you anything about the internal affair,don't say a word to anyone..	.destrudctive change	1
7090	Spread the rumors about the break down of company structure..	.destrudctive change	1
7091	Leak the news that the company is getting bankrupted with in somedays..	.destrudctive change	1
7092	Leak the bad news about the company management..	.destrudctive change	1
7093	Spread the confidential information to the public..	.destrudctive change	1

Figure 4.3.4.1: Red Data CSV file

## 4.4 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.

#### 4.4.1 Evaluates from the Mail Dataset

We have done this work in our 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 ~83%

We have used 500 class 0 tagged data to evaluate the detecting percentage of the model. It gives us that 428 email with green detection and 72 email with false detection.

Thus, the accuracy over detecting the green data is  $\frac{428 \times 100}{500} = 85.6$

So, the accuracy in the Green data detection is 85.6%

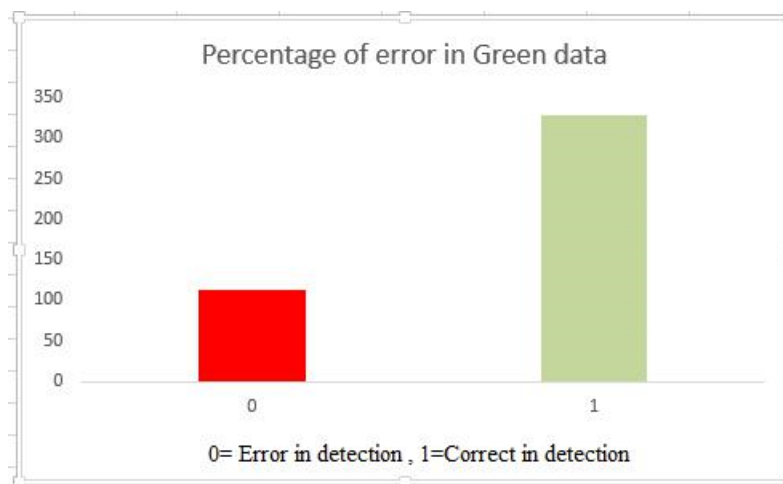


Figure 4.4.1.1: Error in Green Data

Again, in the Red data we continued the process of detection. Here we also use 500 email data from the dataset that is leveled with one. And after processing these data our model detected conspiracy successfully from 395 number of mails. And it predicts 105 number of data as wrong detection.

Thus, the accuracy of the model in the Red data set is  $= 395 \times \frac{100}{500} = 79\%$

So, the accuracy for the Red data is 79%.

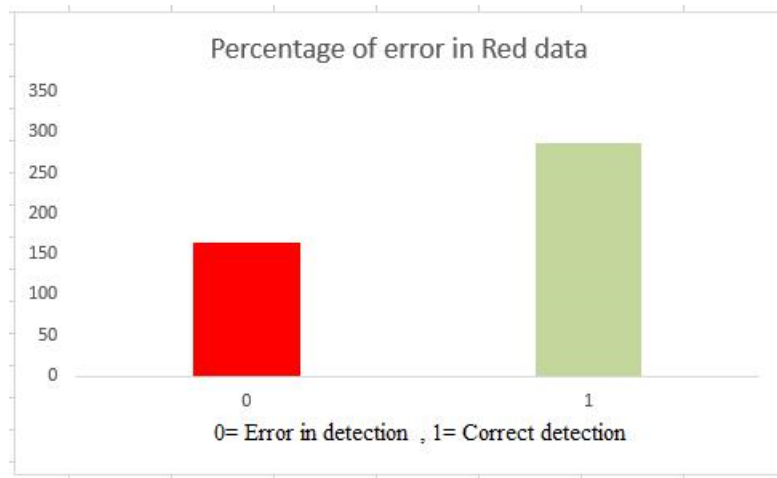


Figure 4.4.1.2: Error in Red Data

#### 4.4.2 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(100)	Positive	Negetive
Positive	True Positive(56)	True Negative(18)
Negetive	False Postive(10)	False Negative(6)

So, the total real time accuracy of my system is  $= (56+28) * \frac{100}{100} = 84$

Now we can say the real time accuracy of this system is 84%

So the total accuracy of this system can be found by merging both the real time and the train set

data is  $= (\text{Green Data} + \text{Red Data} + \text{True Positive} + \text{True Negative}) * \frac{100}{(\text{Total Dataset} + 100)}$

$$=(428+395+56+28) * \frac{100}{(500+500+100)} = 82.45$$

Thus, we can say the accuracy is 82.45% overall.

## 4.5 Summary

In the summary, with respect to accuracy Logistic Regression achieved the highest performance.

But the performance can be more improved. In our future work, we will take more data and classification techniques to classify the conspiracy problems on a company.

## **CHAPTER 5**

### **SUMMARY, CONCLUSION, RECOMEDATION AND IMPLECATION FOR FUTURE RESARCH**

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

There have been lots of research works on “Natural Language Processing” and “machine learning” areas. The outcome of like these technologies are taking a revolutionary transformation in our computing life. Recently, we have got some wonderful applications on this area. But very few of research works on Bangla Language are available. In our study, we have followed some approaches to classify its conspiracy category on Bengali text.

#### **5.2 Conclusion**

The main objective is this work was to develop a system that can automatically detect employee’s conversation in the mail data. We design and train the algorithms that can predict the enable to detect conspiracy in the user’s mail data in real time and give the response to the administration of this company. The detection system can automatically inspect email inbox of the users during all the day simultaneously. We used synthetic data collection process from real email inbox and train the datasets to our module in different way. However, the conspiracy is a concept of psychology and we have used a prediction model to classify the infected mail from the true mail. In particular, no we can just predict the affected mail. In future, we will enable to detect the conspiracy message also.

#### **5.3 Recommendations**

There are one potential recommendations for further study to create the data set more efficiently, can produce a better output of this research work.

#### **5.4 Limitations and Suggestions for Future**

We are collecting data from real time email conversation through users’ email, there are always some restriction in the works. “Natural language processing” is a difficult thing to procedure. And sentiment from natural language is likely to be more difficult task. This work has some limitation

such as accuracy can be improved, our datasets has some URL links, initially we removed the links. There is very close possibility we can say that these links could be a huge source of conspiracy related activity for a work place. Hence, we are working on this work to improve our system performance.

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

## REFERNCES

- [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 series 2010).
- [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](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**

### **Thesis Reflection**

To complete this thesis, we faced so many problems, first one was to determine the methodological approach for our project. Moreover, 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.



## Plagiarism Screenshot of Report

### Email

#### ORIGINALITY REPORT

<b>21</b> %	<b>15</b> %	<b>11</b> %	<b>16</b> %
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

#### PRIMARY SOURCES

<b>1</b>	Submitted to Daffodil International University Student Paper	<b>8</b> %
<b>2</b>	Sisi Liu, Ickjai Lee. "A Hybrid Sentiment Analysis Framework for Large Email Data", 2015 10th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), 2015 Publication	<b>3</b> %
<b>3</b>	Simon Fong, Yan Zhuang, Jinyan Li, Richard Khoury. "Sentiment Analysis of Online News Using MALLET", 2013 International Symposium on Computational and Business Intelligence, 2013 Publication	<b>2</b> %
<b>4</b>	Douglas, Karen M., and Ana C. Leite. "Suspicion in the workplace: Organizational ....."	<b>1</b> %