Conspiracy Detection by Real Time Email Analysis

Md. Ikramul Hoque

ID: 1304115

October, 2018

Conspiracy Detection by Real Time Email Analysis

This thesis is submitted in partial fulfillment of the requirement for the degree of Bachelor of Science in Computer Science and Engineering.

Md. Ikramul Hoque
ID: 1304115

Supervised by

Abu Hasnat Mohammad Ashfak Habib
Associate Professor,
Department of Computer Science and Engineering (CSE)
Chittagong University of Engineering and Technology (CUET)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
CHITTAGONG UNIVERSITY OF ENGINEERING AND TECHNOLOGY (CUET)
CHITTAGONG – 4349, BANGLADESH.

The thesis titled “**Conspiracy Detection by Real Time Email Analysis**” submitted by ID 1304115, Session 2016-2017 has been accepted as satisfactory in fulfillment of the requirement for the degree of Bachelor of Science in Computer Science and Engineering(CSE) as B.Sc. Engineering to be awarded by Chittagong University of Engineering and Technology (CUET).

Board of Examiners

1. ______________________________ Chairman

Abu Hasnat Mohammad Ashfak Habib
Associate Professor
Department of Computer Science and Engineering (CSE)
Chittagong University of Engineering and Technology (CUET)

2. ______________________________ Member
(Ex-officio)

Dr. Mohammad Shamsul Arefin
Professor and Head
Department of Computer Science and Engineering (CSE)
Chittagong University of Engineering and Technology (CUET)

3. ______________________________ Member
(External)

Dr. Asaduzzaman
Professor
Department of Computer Science and Engineering (CSE)
Chittagong University of Engineering and Technology (CUET)

Statement of Originality

It is hereby declared that the contents of this project is original and any part of it has not been submitted elsewhere for the award if any degree or diploma.

Signature of the Supervisor
Date:

Signature of the Candidate
Date:

Acknowledgement

Prima facie, I am grateful to the Almighty for giving me the strength for successful completion of this project. Then I would like to express my sincere gratitude to my honorable project supervisor Abu Hasnat Mohammad Ashfak Habib, Assistant Professor, Department of Computer Science and Engineering, Chittagong University of Engineering and Technology, for his valuable advices, constructive suggestions and sincere guidance with all the necessary facilities for assimilation, research and preparation for the project. I place on record, my sincere gratitude to Dr. Asaduzzaman, Professor, Department of Computer Science and Engineering, Chittagong University of Engineering and Technology, for his kind encouragement and cooperation. 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

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 text documents has been a growing field of text mining among researchers for the past few decades. Nevertheless, Email data sentiment analysis, a general means of social networking and communication, has been studied strongly.

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.

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 vectorization 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.

Table of Contents

Chapter 1

Chapter 2

Chapter 3

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

We live in the era 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. Today's technology is mostly an advancement of old technology, the influence of technology in modern life is immeasurable, we are using technology in various ways and sometimes the way we execute different technologies ends up offending our lives or the society. What we say modern technology is technically not that new in most cases. For instance, communication technology has developed with years, right now we use email which has been an improvement of Fax.

Email is broadly used as a way of business communication and it is a highly effective communication tool. Email is mostly inexpensive, it only requires an internet connection that is already present in the business. In the result as online social networking and communication is appealing to the public. Statistics indicates an improvement of 3% in the number of Email users with an average of 1.7 from 2011 to 2015. Email accounts per user counted in 2015. Moreover, business Email communication accounts for the majority of the total Email with over 108.7 billion Emails exchange daily, and Email remains the most common process of business pace 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 Previous Works

The most common process to text sentiment analysis consist of detecting the frequency of features (words) of given positive or negative semantic value. In that way, sentiment analysis has a lot in similarity with classical text classification and mining [1], and any one would be enticed to implement statistical keyword significance metrics, for example chi-square or TFIDF. Unfortunately, these do not provides expected result for sentiment classification. For better result some work has been done using standard machine learning techniques to sentiment analysis, such as Pang and Lee [2] who used Bayesian classifiers, maximum entropy and SVM. Similarly, Turney and Littman [3] used latent semantic analysis (LSA) to calculate the relationship between words observed in a text and a predefined praise word. But the unique character of the challenge of sentiment classification has given improve to innovative new technique as well.

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 framework for sentiment analysis in Email using a hybrid scheme of process combined with Kmeans clustering and a classifier , support vector machine. The measurement for the design is conducted through the correlation among three labeling methods, SentiWordNet labeling, Kmeans labeling, and Polarity labeling, and there were five classifiers, including Logistic Regression, Decision Tree, OneR and Support Vector Machine, Naïve Bayes, [4].Combination of clustering approach with SWN lexicon for sentiment analysis in blogs was proposed by Feng, Wang, Yu, Yang and Yang [5]; Li and Wu [6] proposed Kmeans clustering for hotspot indication and SVM sentiment classification and prediction. Current analyzation on text mining and sentiment analysis mostly addresses large scale social media data, such as Twitter corpus, blogs and Facebook data. For instance, Li and Wu [6] proposed a study on hotspot finding through online forum.

Moreover, Balasubramanyan, Routledge and Smith [7] propose an algorithm for the prediction of poll results by using public opinion. As for Email data classification, most studies focus on the identification of spam emails, discarded emails, the study of social networking among mails, and priority issues. [8] [9] [10]. However, few research has been conducted on mail conspiracy analysis. Mohammad and Yang [11] analyzes the gender classification of sentiment axis among set of sentiment labeled mail data, and Hangal, Lam and Heer [12] proposed a system for visualizing archived mail data with sentiment words. Other than these studies, a systematic and structured design for conspiracy detection from Email data has not been investigated yet.

1.3 Present State and Contribution

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

1.4 Motivation

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

Many executives at Enron were blamed for a variety of charges and many were later sentenced to prison. Many of them found guilty of unauthorized destruction of documents relevant to the SEC investigation, that voided its license to audit general 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.

1.5 Prospects

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. This study will help a company to monitor and predict any upcoming fall or loss in many ways. Any working place can analyze their entire system by implementing this project in their system. It also proposed a model and framework to convert a psychological theory into a machine. It is also an important aspect of research. So we can say that this thesis and proposed model of system can make a good impact on the digital automated world.

1.6 Organization of the Project

This report is arranged with six 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 and in chapter five, experimental results and evaluations are explained. Finally, chapter six concludes our overall work.

Chapter 2

Literature Review

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 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

Another definition of conspiracy is found in the Cambridge dictionary that a secret agreement made between two or more people or groups to do something bad or illegal that will harm someone else:

Conspiracy against sb It is my client's opinion that there has been a conspiracy against him.

conspiracy between sb (and sb) The group of optometrists denied there was any conspiracy between them and other industry associations to stop the sale of lenses to mail order houses.

conspiracy to do sth The four directors have denied conspiracy to defraud pensioners by misusing shares that belonged to pension funds.

A group of former housing counsellors has been indicted on fraud and conspiracy charges in one of the biggest real estate fraud cases ever seen in the state.

2.3 Conspiracy Theory

Conspiracy theories are omnipresent among members of modern and traditional societies. A common definition of conspiracy theory is the conviction that a group of actors meets in secret agreement with the purpose of attaining some malevolent goal. Contrary to the view that belief in such theories is pathological, large portions of the human population believe conspiracy theories. In 2004, 49% of New York City residents believed the U.S. government to be complicit in the 9/11 terrorist attacks. In addition, in a nationally representative sample of the U.S. population, 37% answered "agree" to the following statement: "the Food and Drug Administration is deliberately preventing the public from getting natural cures for cancer and other diseases because of pressure from drug companies." Another 31% answered "neither agree nor disagree," and only 32% disagreed with this statement. Belief in conspiracy theories is thus a widespread societal phenomenon and has increasingly drawn the research attention of social [28].

Although the definition provided above is rather general, here we explicate the specific underlying features of conspiracy theories. We argue that a conspiracy theory contains at least five critical ingredients [28].

First, conspiracy theories make an assumption of how people, objects, or events are causally interconnected. Put differently, a conspiracy theory always involves a hypothesized pattern [28].

Second, conspiracy theories stipulate that the plans of alleged conspirators are deliberate. Conspiracy theories thus ascribe intentionality to the actions of conspirators, implying agency [28].

Third, a conspiracy theory always involves a coalition, or group, of actors working in conjunction. An act of one individual, a lone wolf, does not fit the definition of a conspiracy theory [28].

Fourth, conspiracy theories always contain an element of threat such that the alleged goals of the conspirators are harmful or deceptive [28].

Fifth, and finally, a conspiracy theory always carries an element of secrecy and is therefore often difficult to invalidate. Conspiracy theories that turn out true—such as Watergate or the Iran-Contra scandal—are no longer conspiracy "theories." Hence, in judging the validity of conspiracy theories, there is always room for error [28].

People hold many beliefs that share some of the key elements of conspiracy theories, such as supernatural beliefs. Indeed, conspiracy theories and supernatural beliefs are positively correlated. What distinguishes conspiracy theories from supernatural beliefs is that they necessarily involve a coalition element of deceptive or potentially dangerous other human beings acting in unison. For conspiracy theories to occur, however, these nonhuman stimuli need, at the very least, to be connected to the real or suspected presence of a coordinated group of deliberate actors. Unlike other forms of beliefs, a hostile coalition is a prerequisite of any conspiracy theory [28].

One can find many lay theories that fit the key ingredients of a conspiracy theory (patterns, agency, coalitions, threats, secrecy). They usually involve powerful groups such as societal leaders, governmental institutions (e.g., secret services), influential branches of industry (e.g., oil companies, the pharmaceutical industry), or stigmatized minority groups (e.g., Muslims, Jews). Besides the context of citizens’ perception of geopolitical events, conspiracy theories emerge frequently in the micro level setting of organizations, as employees often suspect their managers of conspiring toward evil goals such as pursuing their self-interest at the expense of employees and the organization [28].

Furthermore, although the term conspiracy theory may sometimes be used to invalidate legitimate accusations of corruption [28].

2.4 Psychology of Conspiracy Theories

Conspiracy theories explain events as the result of secret, deliberate actions and cover ups at the hands of malicious and powerful groups [29].

Psychologists have also begun to consider what some of the potential consequences of conspiracy theories might be. In particular, whilst conspiracy theories may allow individuals to question social hierarchies and require elites be more transparent, recent empirical findings suggest that they may have important negative societal consequences. It is therefore becoming clear that conspiracy theories cannot be dismissed as trivial notions that affect the lives of only a small handful of individuals and marginalized communities [29].

2.5 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. We differentiate organizational conspiracy theories from various associated concepts. Specifically, rumor and gossip more often implicate individuals than groups and do not necessarily allege conspiracies between individuals. General mistrust, whilst correlated with conspiracy belief about societal events refers to broader negative feelings about individuals or groups rather than specific allegations of dishonesty and wrongdoing by groups. Like general conspiracy beliefs, organizational conspiracy beliefs may also thrive under conditions of powerlessness and uncertainty. Specifically, in situations where workers lack control (e.g., have little responsibility, or little control over their duties) or under conditions of uncertainty (e.g., new management, concern about the motives of managers), organizational conspiracy theories may prosper [29].

What effects are such organizational conspiracy theories likely to have on the workplace? Our investigation focuses on one of the most important outcomes for organizations – employee turnover. Turnover represents a significant challenge to organizations and can be very costly, resulting in financial losses associated with training employees who opt to leave, associated recruitment and other administrative costs as well as the potentially disastrous loss of valuable individuals. Turnover can also affect organizational performance outcomes, including customer service, profits, and revenues. Specifically, workers may be more likely to consider leaving their organization to the extent that they view it as a negative place where groups act secretly and maliciously in the pursuit of their own selfish interests. Indeed, some recent research shows support for this idea, demonstrating that belief in workplace-related conspiracies – as a result of despotic or laissez-faire leadership – is associated with turnover intentions. As yet, however, much is unknown about the relationship between conspiracy theorizing in organizations and workers' intentions to leave their workplace [29].

2.5.1 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 [29].

We argue that organizational conspiracy theories will decrease organizational identification. If an organization is riddled with perceptions of conspiracy, such as beliefs that managers are deliberately trying to harm employees, this is likely to weaken the importance of the organization to the individual and reduce the positive self-esteem they derive from it. Research has shown that self-esteem and identification are positively influenced by receiving procedural justice from a group [29].

2.5.2 Organizational Commitment

Similarly, organizational commitment can be understood as the psychological link people have to the organization. Research suggests that commitment is also a significant predictor of turnover intentions. Further, Van Prooijen and de Vries (2016) found that organizational commitment was associated with belief in workplace-related conspiracies and that it mediated the association between organizational conspiracy belief and turnover intentions. It remains to be seen whether organizational conspiracy belief exerts a causal influence on organizational commitment. Specifically, Jolley and Douglas (2014a) showed that conspiracy theories concerning the government weakened political engagement by generating feelings of powerlessness and uncertainty [29].

Like political engagement, organizational commitment entails willingness to act on behalf of the interests of the collective [29].

2.5.3 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 [29].

Like organizational identification and organizational commitment, we argue that organizational conspiracy theories will decrease job satisfaction. A perceived climate of conspiracy or specific beliefs that groups are conspiring against employees are likely to lead to disappointment and dissatisfaction in the job, or at least with specific aspects of the job [29].

2.5.4 Implications

More generally, our findings suggest that managers and employees may need to be mindful of the effects that conspiracy theories could have on the workplace. The current research suggests that, like broader societal conspiracy theories, organizational conspiracy theories may have clear and detrimental consequences for employees and the organization as a whole. Management especially should be mindful of workplace conspiracy theories that may not only damage their reputation, but force them to lose valuable employees, or even keep disengaged employees on their team.

It is plausible that interventions that would focus on improving specific aspects of the organizational climate could strengthen employees' commitment to the organization, their satisfaction with their job, and, consequently, decrease intentions to leave the organization [29].

2.6 Machine Learning

Evolved from the study of pattern recognition and computational learning theory in artificial intelligence, machine learning explores the study and construction of algorithms that can learn from and make predictions on data such algorithms overcome following strictly static program instructions by making predictions or decisions, through building a model from sample inputs. Machine learning is employed in a range of computing tasks where designing and programming explicit algorithms with good performance is difficult or infeasible; applications include email filtering, detection of network intruders or malicious insiders working towards a data breach, optical character recognition (OCR), learning to rank, and computer vision. Computer Vision and Machine Learning are two core branches of Computer Science that can function, and power very sophisticated systems that rely on CV and ML algorithms exclusively but when we combine the two, we can achieve even more. Machine learning tasks are typically classified into two categories, depending on the nature of the learning "signal" or "feedback" available to a learning system. These are –

2.6.1 Supervised Learning: The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs. The training data consists of a set of examples which is called training examples. In order to solve a given problem using supervised learning there are some steps:

- Deciding what kind of data will be in the training set.

- Gathering a set of input object and related output object in training set.

- Determining input feature representation of the learned function.

- Determining the structure of the learned function and corresponding learning algorithm.

- Running the learning algorithm in previously defined training set.

- Evaluating the accuracy.

2.6.2 Unsupervised Learning: No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end. For example cluster analysis is an unsupervised learning method. This explores data analysis and find hidden patterns or grouping in data.

We did out project using supervised learning.

2.7 Text Classifier –The Basic Building Blocks

Classification is the process of predicting the class of given data points. Classes are sometimes called as targets/ labels or categories. Classification predictive modeling is the task of approximating a mapping function (f) from input variables (X) to discrete output variables (y).

For example, spam detection in email service providers can be identified as a classification problem. This is s binary classification since there are only 2 classes as spam and not spam. A classifier utilizes some training data to understand how given input variables relate to the class. In this case, known spam and non-spam emails have to be used as the training data. When the classifier is trained accurately, it can be used to detect an unknown email.

Classification belongs to the category of supervised learning where the targets also provided with the input data. There are many applications in classification in many domains such as in credit approval, medical diagnosis, target marketing etc.

There are two types of learners in classification as lazy learners and eager learners.

1. Lazy learners

Lazy learners simply store the training data and wait until a testing data appear. When it does, classification is conducted based on the most related data in the stored training data. Compared to eager learners, lazy learners have less training time but more time in predicting.

Ex. k-nearest neighbor, Case-based reasoning

2. Eager learners

Eager learners construct a classification model based on the given training data before receiving data for classification. It must be able to commit to a single hypothesis that covers the entire instance space. Due to the model construction, eager learners take a long time for train and less time to predict.

Ex. Decision Tree, Naive Bayes, Artificial Neural Networks [15].

2.8 Sentiment Analysis

Sentiment Analysis is the most common text classification tool that analyses an incoming message and tells whether the underlying sentiment is positive, negative our neutral.

Sentiment analysis is a text mining challenge which deals with determining the opinion expressed by the author of a text. This task has become pressing in recent years, as the amount of opinionated texts online, such as blogs, editorials, and reviews, has skyrocketed. Being able to quickly and accurately measure people's opinions from the material the write and post online would allow governments and companies to better tailor and adapt their policies and products. Most research on sentiment analysis has been carried out on subjective texts such as blogs and product reviews. Authors of such text typically express their opinion freely, using clearly positive or negative language. The situation is different when dealing with news articles: in order to maintain an appearance of objectivity, journalists will often refrain from using clearly opinionated vocabulary [13]. The writing style is also different than in opinion pieces, and uses more complex sentences and specialized vocabulary.

Sentiment analysis is the automated process of understanding an opinion about a given subject from written or spoken language. In a world where we generate 2.5 quintillion bytes of data every day, sentiment analysis has become a key tool for making sense of that data. This has allowed companies to get key insights and automate all kind of processes.

Sentiment Analysis also known as *Opinion Mining* is a field within Natural Language Processing (NLP) that builds systems that try to identify and extract opinions within text. Usually, besides identifying the opinion, these systems extract attributes of the expression e.g.:

1. Polarity: if the speaker express a *positive* or *negative* opinion,

2. Subject: the thing that is being talked about,

3. Opinion holder: the person, or entity that expresses the opinion.

Currently, sentiment analysis is a topic of great interest and development since it has many practical applications. Since publicly and privately available information over Internet is constantly growing, a large number of texts expressing opinions are available in review sites, forums, blogs, and social media.

With the help of sentiment analysis systems, this unstructured information could be automatically transformed into structured data of public opinions about products, services, brands, politics, or any topic that people can express opinions about. This data can be very useful for commercial applications like marketing analysis, public relations, product reviews, net promoter scoring, product feedback, and customer service [14].

2.9 Dataset

The term data set 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.

Any named group of records is called a data set. Data sets can hold information such as medical records or insurance records, to be used by a program running on the system. Data sets are also used to store information needed by applications or the operating system itself, such as source programs, macro libraries, or system variables or parameters. For data sets that contain readable text, you can print them or display them on a console (many data sets contain load modules or other binary data that is not really printable). Data sets can be catalogued, which permits the data set to be referred to by name without specifying where it is stored.

In simplest terms, a record is a fixed number of bytes containing data. Often, a record collects related information that is treated as a unit, such as one item in a database or personnel data about one member of a department. The term field refers to a specific portion of a record used for a particular category of data, such as an employee's name or department.

The records in a data set can be organized in various ways, depending on how we plan to access the information. If you write an application program that processes things like personnel data, for example, your program can define a record format for each person's data [16].

2.10 Features

Features are the variables found in the given problem set that can strongly/sufficiently help us build an accurate predictive model.

Eg : To predict the sale price of a house, size of the house is a feature.

1. Features are a column of data given as the input. They are also called as attributes or might sometimes be referred as dimensions.

2. A particular problem data set can have several features tagging to them. It is important to select the features that are more relevant to our problem so that the accuracy of the model improves. It also reduces the complexity of the model as we avoid the least significant / unnecessary feature data. The simpler model is simpler to understand and explain.

3. This Process is called feature engineering / selection and is one of the crucial step of pre-processing. *Different algorithms* can be used to implement it.

4. The Features can be of different types.

5. Simple Supervised selection where they are simple values like numbers and characters.

6. Eg: Size of the house (number).

7. In unsupervised learning, the model is itself trained to recognize the features and work on it.

8. Eg: In character recognition, features may include histograms counting the number of black pixels along horizontal and vertical directions, number of internal holes, stroke detection and many others.

Eg. : Loan Granting Problem

Let us build a model that tells us if to give loan to a particular customer or not.

Now its data will have many features/attributes attached to it:

Loan id, Cust. Name, Cust. id, Cust.Addr, Employed (or) not, Age, Marital Status, Has already avail loan, Annual Income, no.of open accounts, tax liens, credit score, current balance and so on.

Using the feature selection it can be observed that for a particular Customer,

Employed (or) not, age, current credit score, annual income, already availed loan can significantly explain / contribute to the model accuracy better than the others.

Thus they become the features for building our model for this particular problem.

In Artificial Intelligence, features are observable or derived properties of instances in a model's domain. For instance, for an intelligent agent which learns to classify types of fish, features could be size, colour, shape, scale patterns, etc. of fish. The model then learns the correlation between these features and the class of fish, in time becoming able to determine which class of fish a given instance is only by looking at these features.

At last we can say that features are those properties of a problem based on which you would like to predict results.

2.11 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.

How is this done? Just like medical professionals getting a patient prepped for surgery so is data preprocessing, it prepares raw data for further processing. Below are the steps to be taken in data preprocessing

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. [17]

2.12 Python

Dating from 1991, the Python programming language was considered a gap-filler, a way to write scripts that “automate the boring stuff” (as one popular book on learning Python put it) or to rapidly prototype applications that will be implemented in other languages. However, over the past few years, Python has emerged as a first-class citizen in modern software development, infrastructure management, and data analysis. It is no longer a back-room utility language, but a major force in web application creation and systems management, and a key driver of the explosion in big data analytics and machine intelligence [18].

2.12.1 Advantages

Python is easy to learn and use: The number of features in the language itself is modest, requiring relatively little investment of time or effort to produce your first programs. The Python syntax is designed to be readable and straightforward. This simplicity makes Python an ideal teaching language, and it lets newcomers pick it up quickly. As a result, developers spend more time thinking about the problem they’re trying to solve and less time thinking about language complexities or deciphering code left by others [18].

Python is broadly adopted and supported: Python is both popular and widely used, as the high rankings in surveys like the Tiobe Index and the large number of GitHub projects using Python attest. Python runs on every major operating system and platform, and most minor ones too. Many major libraries and API-powered services have Python bindings or wrappers, letting Python interface freely with those services or directly use those libraries. Python may not be the *fastest* language, but what it lacks in speed, it makes up for in versatility [18].

2.12.2 Uses [18]

Python is used in many ways:

1) General application programming with Python

2) Data science and machine learning with Python

3) Web services and RESTful APIs in Python

4) Metaprogramming and code generation in Python

2.12.3 Matplotlib

Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shells, the Jupyter notebook, web application servers, and four graphical user interface toolkits. Matplotlib tries to make easy things easy and hard things possible. You can generate plots, histograms, power spectra, bar charts, errorcharts, scatterplots, etc., with just a few lines of code. For simple plotting the pyplot module provides a MATLAB-like interface, particularly when combined with IPython. For the power user, we have full control of line styles, font properties, axes properties, etc, via an object oriented interface or via a set of functions familiar to MATLAB users [19]

2.12.4 Numpy [20]

NumPy is the fundamental package for scientific computing with Python. It contains among other things:

1) A powerful N-dimensional array object

2) Sophisticated (broadcasting) functions

3) Tools for integrating C/C++ and Fortran code

4) Useful linear algebra, Fourier transform, and random number capabilities

Besides its obvious scientific uses, NumPy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined. This allows NumPy to seamlessly and speedily integrate with a wide variety of databases

2.12.5 Scikit-learn [21]

Scikit-learn is a library function that is used in python for machine learning.

1) Simple and efficient tools for data mining and data analysis

2) Accessible to everybody and reusable in various contexts

3) Built on NumPy, SciPy and matplotlib

4) Open source, commercially usable – BSD license

2.12.5.1 History [22]

Scikit-learn was initially developed by David Cournapeau as a Google summer of code project in 2007.

Later Matthieu Brucher joined the project and started to use it as apart of his thesis work. In 2010 INRIA got involved and the first public release (v0.1 beta) was published in late January 2010.

The project now has more than 30 active contributors and has had paid sponsorship from INRIA, Google, Tinyclues and the Python Software Foundation.

2.12.6 Pandas

Pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language. *Pandas* is a NumFOCUS sponsored project. This will help ensure the success of development of pandas as a world-class open-source project, and makes it possible to donate to the project [23].

2.12.6.1 Usage of Pandas

Python has long been great for data munging and preparation, but less so for data analysis and modeling. Pandas helps fill this gap, enabling you to carry out your entire data analysis workflow in Python without having to switch to a more domain specific language like R. Combined with the excellent IPython toolkit and other libraries, the environment for doing data analysis in Python excels in performance, productivity, and the ability to collaborate. *Pandas* does not implement significant modeling functionality outside of linear and panel regression; for this, look to statsmodels and scikit-learn. More work is still needed to make Python a first class statistical modelling environment, but we are well on our way toward that goal [23].

2.12.7 PyMySQL

This package contains a pure-Python MySQL client library, based on PEP 249.

Most public APIs are compatible with mysqlclient and MySQLdb.

NOTE: PyMySQL doesn't support low level APIs _mysql provides like data_seek, store_result, and use_result. You should use high level APIs defined in PEP 249. But some APIs like auto commit and ping are supported because PEP 249 doesn't cover their use case.[24]

Requirements

- Python – one of the following:
    - CPython : 2.7 and >= 3.4
    - PyPy : Latest version
- MySQL Server – one of the following:
    - MySQL >= 5.5
    - MariaDB >= 5.5

2.12.8 wordCloud

Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance. Significant textual data points can be highlighted using a word cloud. Word clouds are widely used for analyzing data from social network websites.

For generating word cloud in Python, modules needed are – matplotlib, pandas and wordcloud. To install these packages, run the following commands:

```
Pip install matplotlib

Pip install pandas

Pip install wordcloud
```

2.12.9 BeautifulSoup

Beautiful Soup is a Python library for pulling data out of HTML and XML files. It works with parser to provide idiomatic ways of navigating, searching, and modifying the parse tree. It commonly saves programmers hours or days of work.

2.12.10 CountVectorizer and LogisticRegression

CountVectorizer builds a count matrix where rows are occurrences counts of different words taking into account the high-dimensional sparsity.

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables

2.12.11 TF-IDF

TF-IDF, which stands for term frequency—inverse document frequency, is a scoring measure widely used in information retrieval (IR) or summarization. TF-IDF is intended to reflect how relevant a term is in a given document.

The intuition behind it is that if a word occurs *multiple times in a document*, we should boost its relevance as it should be more meaningful than other words that appear fewer times (TF). At the same time, if a word occurs many times in a document but also *along many other documents*, maybe it is because this word is just a frequent word; not because it was relevant or meaningful (IDF).

We can come up with a more or less subjective definition driven by our intuition: a word's relevance is proportional to the amount of information that it gives about its context (a sentence, a document or a full dataset). That is, the most relevant words are those that would help us, as humans, to better understand a whole document without reading it all.

As pointed out, relevant words are not necessarily the most frequent words since stopwords like “the”, “of” or “a” tend to occur very often in many documents.

There is another caveat: if we want to summarize a document compared to a whole dataset about a specific topic (let’s say, movie reviews), there will be words (other than stopwords, like *character* or *plot*), that could occur many times in the document as well as in many other documents. These words are not useful to summarize a document because they convey little discriminating power; they say very little about what the document contains compared to the other documents [25].

2.13 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.

Chapter 3:

Conspiracy Detection Methodology

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.

3.1 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

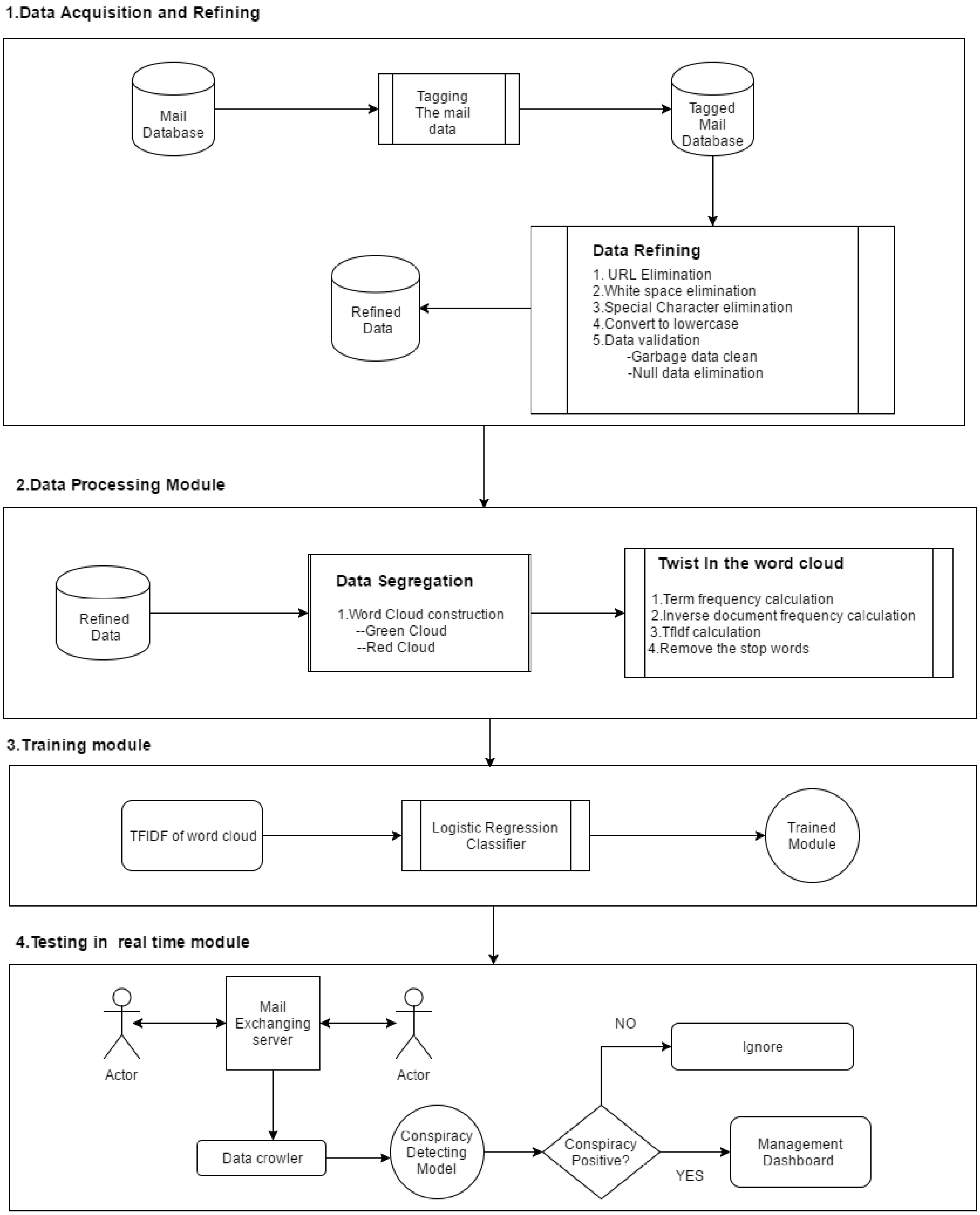


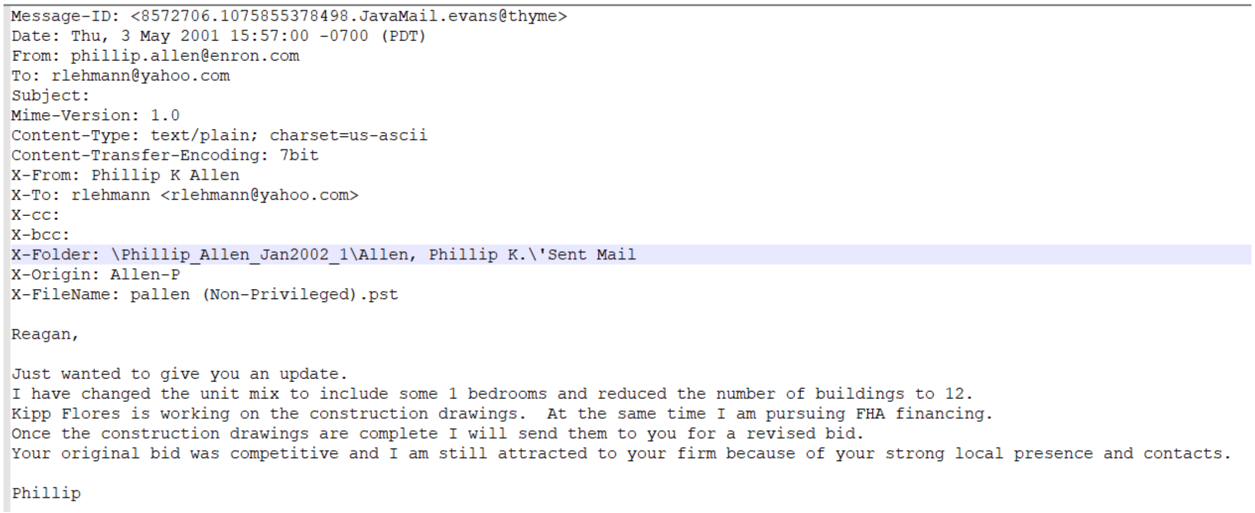
Fig 3.1: Conspiracy Predictor Model Architecture.

3.1.1 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.

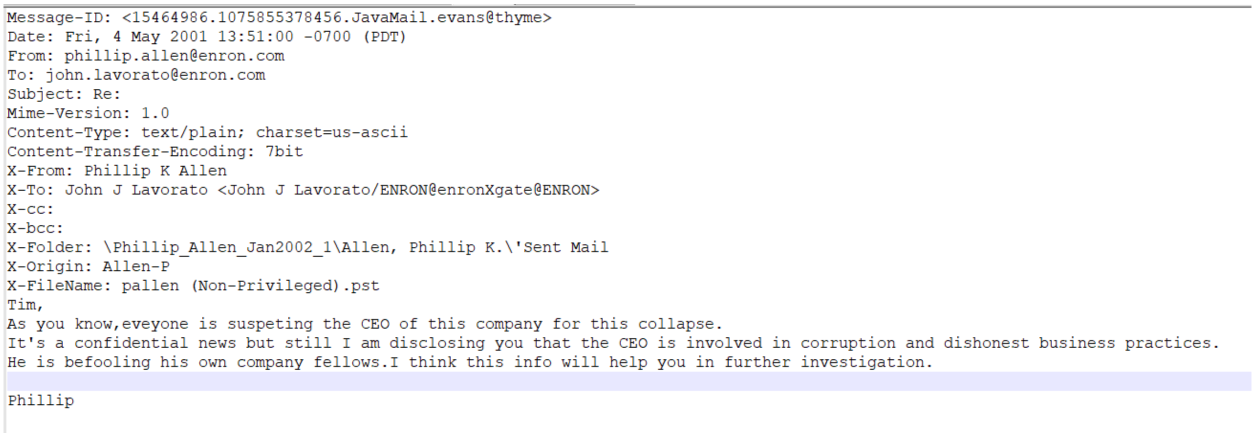
In the first category of Green set the mail body can reflect so many affair in the sentiment. Such as office affair, request mail, satisfaction mail, gratitude sentiment mail and some more documentation mail. Fig 2 shows the Green mail data in csv file



```
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
```

Fig 3.2: Green Mail Content

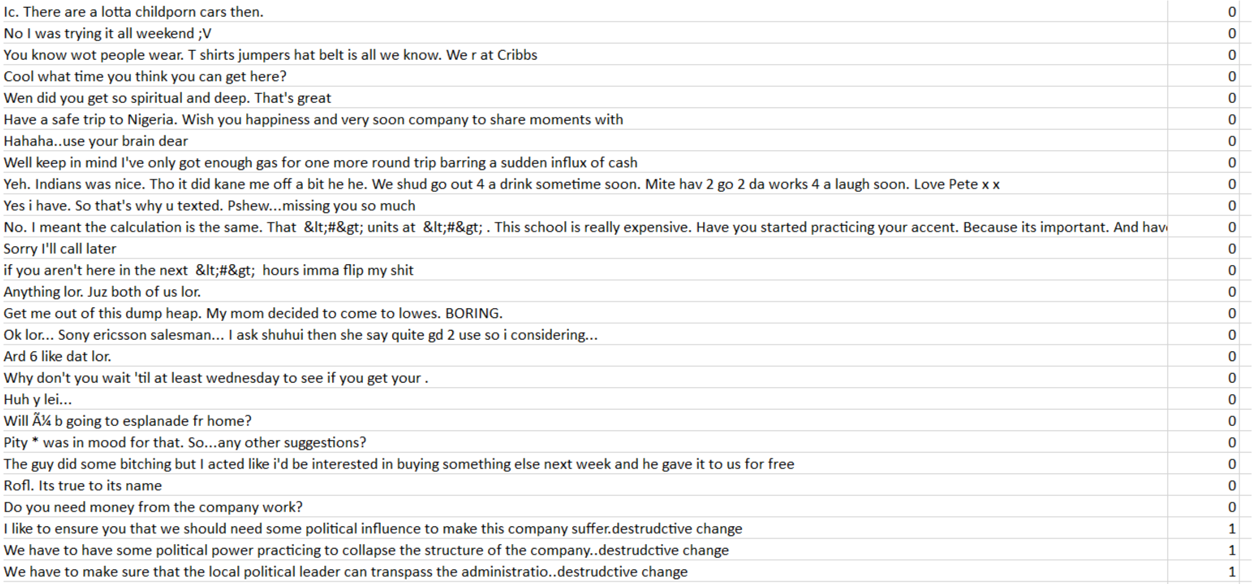
In the next category of the Red set of mail infected with office conspiracy. This category is leveled by one.This sort of mail data reflect rage, dissatisfaction, overpower intention etc that means the overall bad effect for the company. Fig 3 shows the Red mail category



```
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.
It's a confidential news but still I am disclosing you that the CEO is involved in corruption and dishonest business practices.
He is befooling his own company fellows.I think this info will help you in further investigation.

Phillip
```

Fig 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 &lt;#&gt; units at &lt;#&gt; . 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 &lt;#&gt; 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 Ã¼ 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

Fig 3.4: A Peek into the Dataset

3.1.1.1 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

3.1.2 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.

3.1.2.1 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.

These tokens are often loosely referred to as terms or words, but it is sometimes important to make a type/token distinction. A token is an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing. A type is the class of all tokens containing the same character sequence. A term is a (perhaps normalized) type that is included in the IR system's dictionary.iii

These tokens are further used to make feature vector.

3.1.2.2 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.

A common style based analysis method is called writer invariant or author invariant. It claims that all texts written by the same author are similar or invariant. In other words texts written by the same author will be more similar than those written by different ones. Even though we are not interested in the authors of tweets, style based features is an important analysis method, since the topic all employee write about is similar and the purpose of the mail is the same, mainly to make successful their propaganda, it is reasonable to believe that the style of writing they have might be similar. We used frequency of words, presence of words, count of words and use of punctuations as style based features.

We have used TF*IDF to represent our features. TF*IDF is an information retrieval technique that weighs a term's frequency (TF) and its inverse document frequency (IDF). Each word or term has its respective TF and IDF score. The product of the TF and IDF scores of a term is called the TF*IDF weight of that term. It shows the relative importance of a feature in a document or text.

TFIDF: 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:

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

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

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

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.

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. There are a lot of stop words in English language such as: 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.

3.1.3 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.

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 1. The function transforms a real number input x, into a number between 0 and 1. It has one set of parameters w, which represents the slope of the increase around the midpoint, 0.5. The intercept term b denotes the input value where the function output crosses the midpoint. A logistic classifier would predict the positive class if the sigmoid output is greater than 0.5 and the negative class otherwise. By varying w and b, one can control where that change in decision occurs m and how fast the decision should respond to changing input value around that point

Logistic regression is one of the most popular machine learning algorithm for binary classification. This is because it is a simple 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 [27]:

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

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:

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

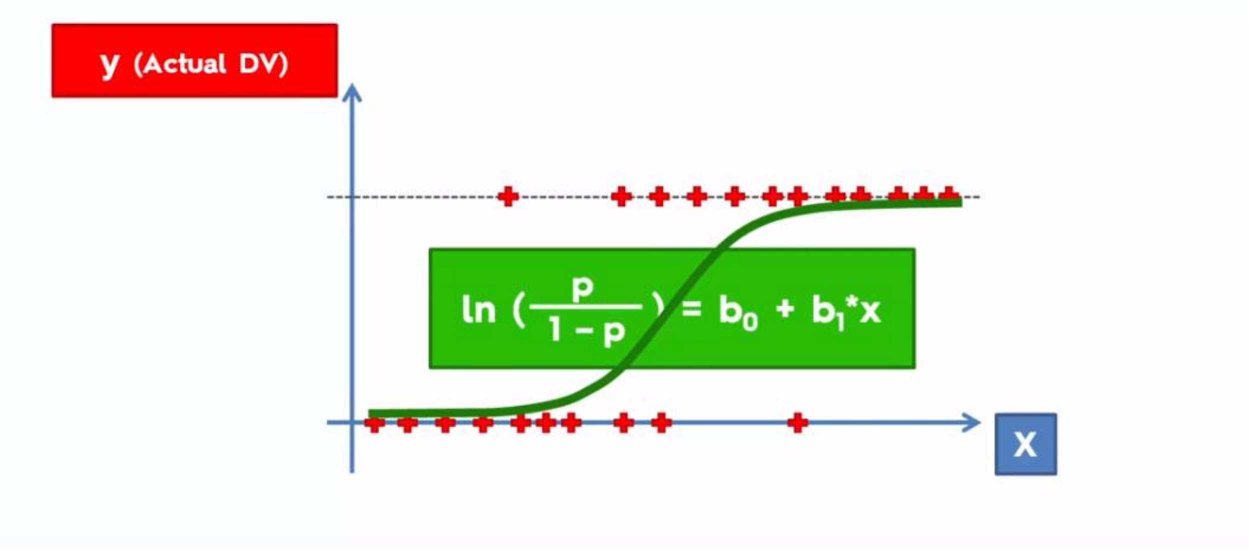


Fig 5: Graphical representation of Logistic Regression Classifier

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.

3.1.4 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.

3.2 Analytical Representation of the Architecture

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. 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 Windows operating system, PhpStorm IDE and PyCharm IDE platforms

3.2.1 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.

3.2.2 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
```

```
Search for contracted form in email body
if contracted form found then
    Replace it with long form
Search for stop words in data
if stop words found then
    Remove the stop words
End
```

3.2.3 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

```
Begin
Remove stop words from raw tweets
Convert the raw email into lowercase form
Tokenize the tweets
Calculate the TFIDF matrix
End
```

3.2.4 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

3.2.4.1 Logistic Regression

Algorithm 3.3: Logistic Regression learning algorithm

Inputs: Training data, x

Require: Train model to classify

```
Begin
Initialize w
for i=1 to n do
    $z(i)=\sum w(i)*x(i)$
end for
for j = 0 to d do
    for i = 1 to n do
        $\theta(j)$ = SOFT-MAX(z(i))
    end for
end for
End
```

3.2.5 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

```
Begin
Access the storing database of the email
if the email is not still taken
    take the email
End
```

3.2.6 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

```
Begin
classifier = load(model)
for each email in the emaildata table in database test
    take the email body
    type = classifier.predict(email body)
    if type = 0 then
        result = “It is not infected”
    else if type = 1 then
        result = “Infected”
        store the email with the sender and receiver name in another database
        show the result
End
```

3.3 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:

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).

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).

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).

Chapter 4:

Implementation of Conspiracy Detection Framework

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.

4.1 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.

4.2 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 “CUET 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:

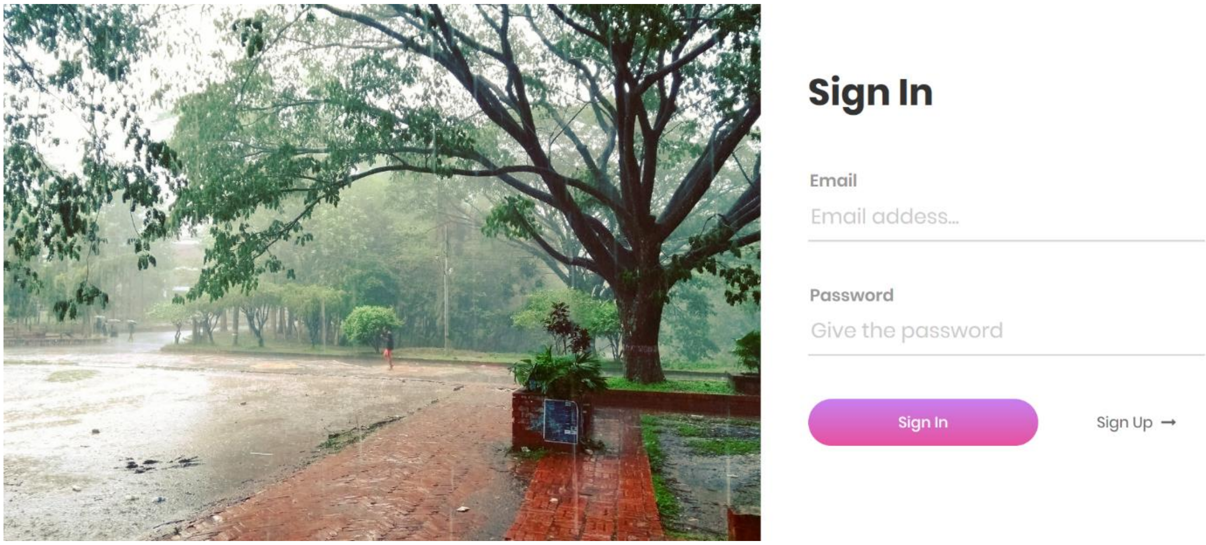
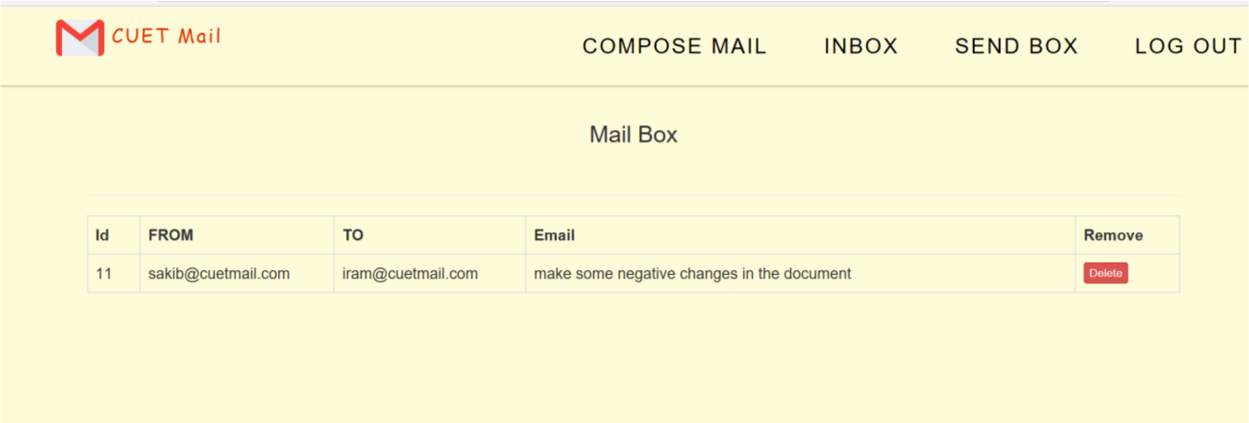


Fig 4.1: Login interface of CUET Mail.

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:



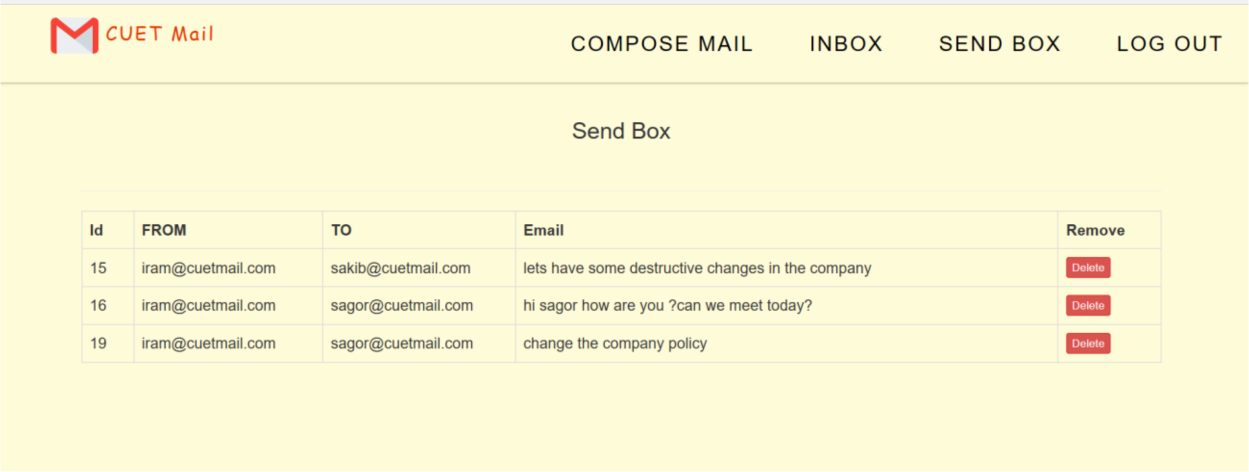
Mail Box

Id	FROM	TO	Email	Remove
11	sakib@cuetmail.com	iram@cuetmail.com	make some negative changes in the document	Delete

Fig 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:



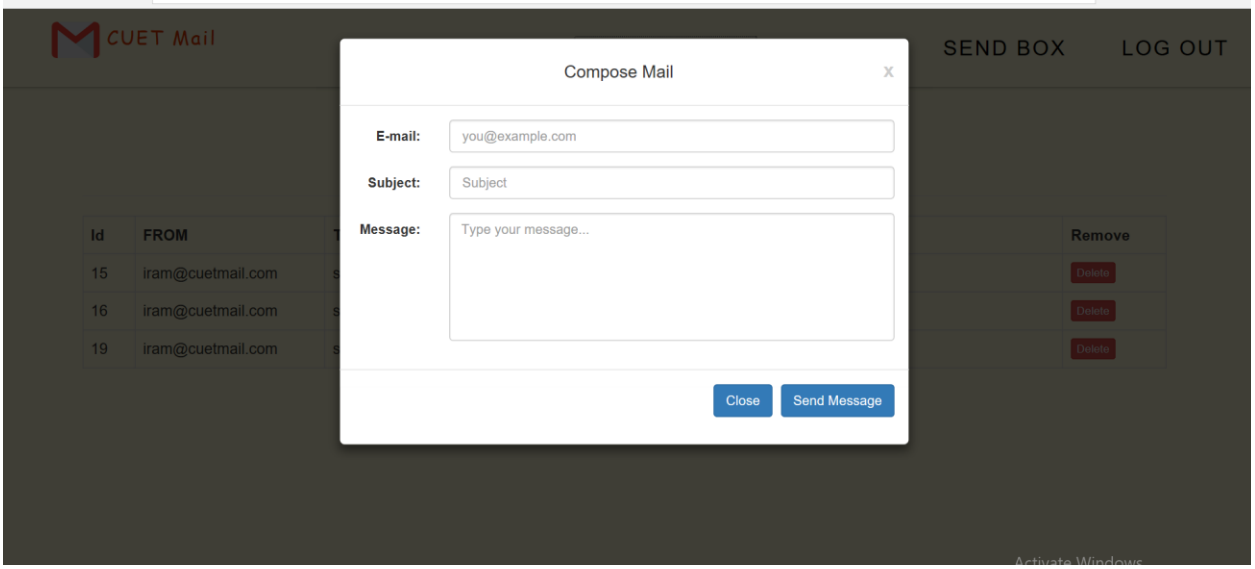
Send Box

Id	FROM	TO	Email	Remove
15	iram@cuetmail.com	sakib@cuetmail.com	lets have some destructive changes in the company	Delete
16	iram@cuetmail.com	sagor@cuetmail.com	hi sagor how are you ?can we meet today?	Delete
19	iram@cuetmail.com	sagor@cuetmail.com	change the company policy	Delete

Fig 4.3: Send Box 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:



E-mail: you@example.com

Subject: Subject

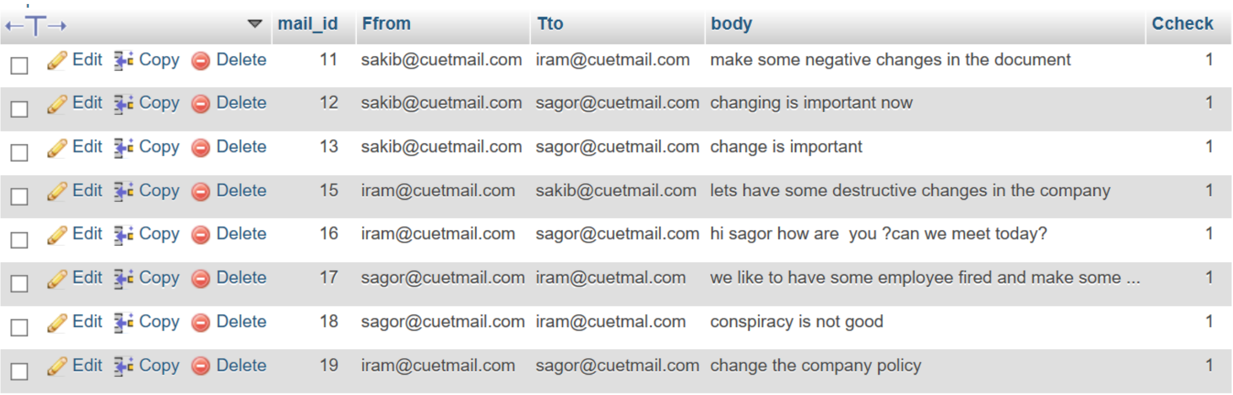
Message: Type your message...

Id	FROM	Remove
15	iram@cuetmail.com	Delete
16	iram@cuetmail.com	Delete
19	iram@cuetmail.com	Delete

Fig 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:

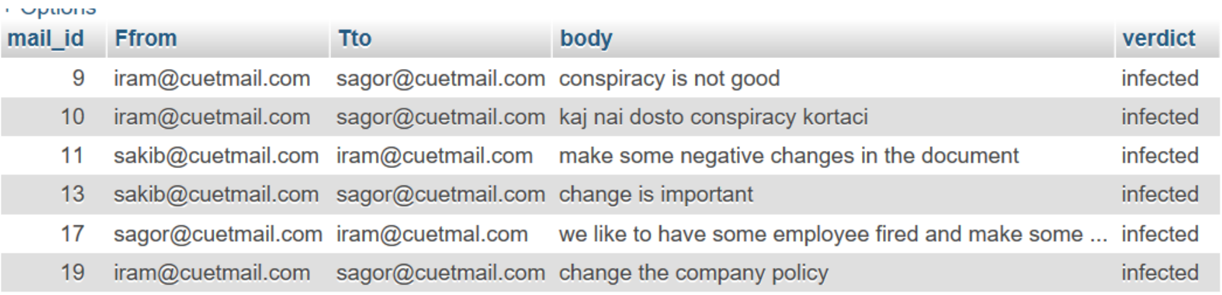
Fig 4.5 shows the email_data database table for analyzing the email data:



←T→ ▼	mail_id	Ffrom	Tto	body	Ccheck
□ Edit Copy Delete	11	sakib@cuetmail.com	iram@cuetmail.com	make some negative changes in the document	1
□ Edit Copy Delete	12	sakib@cuetmail.com	sagor@cuetmail.com	changing is important now	1
□ Edit Copy Delete	13	sakib@cuetmail.com	sagor@cuetmail.com	change is important	1
□ Edit Copy Delete	15	iram@cuetmail.com	sakib@cuetmail.com	lets have some destructive changes in the company	1
□ Edit Copy Delete	16	iram@cuetmail.com	sagor@cuetmail.com	hi sagor how are you ?can we meet today?	1
□ Edit Copy Delete	17	sagor@cuetmail.com	iram@cuetmal.com	we like to have some employee fired and make some ...	1
□ Edit Copy Delete	18	sagor@cuetmail.com	iram@cuetmal.com	conspiracy is not good	1
□ Edit Copy Delete	19	iram@cuetmail.com	sagor@cuetmail.com	change the company policy	1

Fig 4.5: ‘Email Data’ Table Interface

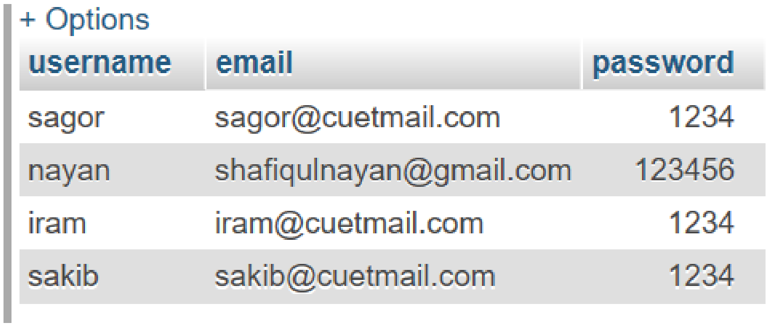
Then the next figure will show the interface for the infected email that are analyzed to show the governing body of the company



mail_id	Ffrom	Tto	body	verdict
9	iram@cuetmail.com	sagor@cuetmail.com	conspiracy is not good	infected
10	iram@cuetmail.com	sagor@cuetmail.com	kaj nai dosto conspiracy kortaci	infected
11	sakib@cuetmail.com	iram@cuetmail.com	make some negative changes in the document	infected
13	sakib@cuetmail.com	sagor@cuetmail.com	change is important	infected
17	sagor@cuetmail.com	iram@cuetmal.com	we like to have some employee fired and make some ...	infected
19	iram@cuetmail.com	sagor@cuetmail.com	change the company policy	infected

Fig 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:



+ Options

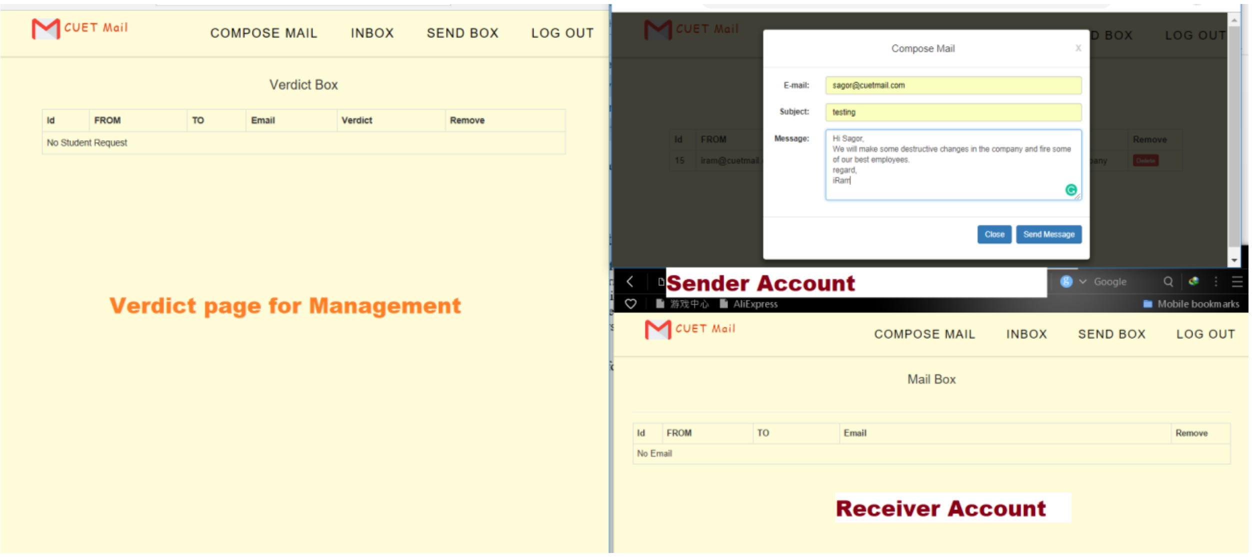
username	email	password
sagor	sagor@cuetmail.com	1234
nayan	shafiqulnayan@gmail.com	123456
iram	iram@cuetmail.com	1234
sakib	sakib@cuetmail.com	1234

Fig 4.7: User Verification Table Interface

4.3 Detection System Implementation

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:



Verdict Box

Id	FROM	TO	Email	Verdict	Remove
No Student Request					

Verdict page for Management

Compose Mail

E-mail: sagor@cuetmail.com

Subject: testing

Message:
Hi Sagor,
We will make some destructive changes in the company and fire some of our best employees.
regard,
iRam

Sender Account

Mail Box

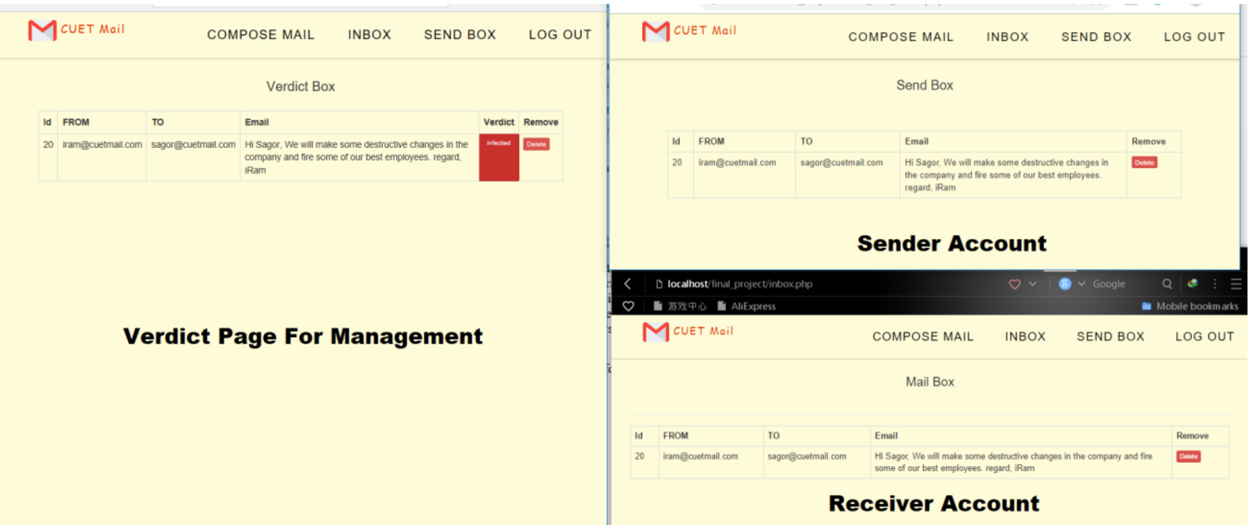
Id	FROM	TO	Email	Remove
No Email				

Receiver Account

Fig 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:



Verdict Box

Id	FROM	TO	Email	Verdict	Remove
20	iram@cuetmail.com	sagor@cuetmail.com	Hi Sagor, We will make some destructive changes in the company and fire some of our best employees. regard, iRam	infected	Delete

Verdict Page For Management

Send Box

Id	FROM	TO	Email	Remove
20	iram@cuetmail.com	sagor@cuetmail.com	Hi Sagor, We will make some destructive changes in the company and fire some of our best employees. regard, iRam	Delete

Sender Account

Mail Box

Id	FROM	TO	Email	Remove
20	iram@cuetmail.com	sagor@cuetmail.com	Hi Sagor, We will make some destructive changes in the company and fire some of our best employees. regard, iRam	Delete

Receiver Account

Fig 4.9: Detection Illustration

4.4 Conclusion

In this chapter we have tried to give an overview about our system’s implementation. Specifically, we have described the experimental set up, system layout and output generation.

Chapter 5

Experimental Results

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.

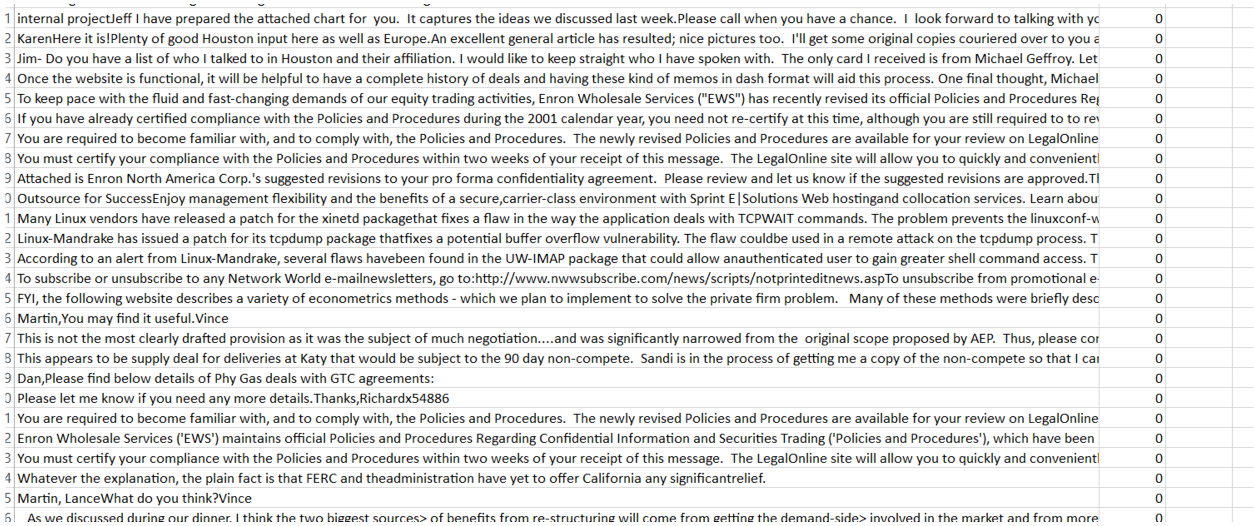
5.1 Data Collection

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.

5.1.1 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 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 rev	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.Tl	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 cai	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

Fig 5.1: Green Data CSV File.

5.1.2 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.**

5.1.2.1 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

5.1.2.2 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.

5.1.2.3 Reputational Conspiracy

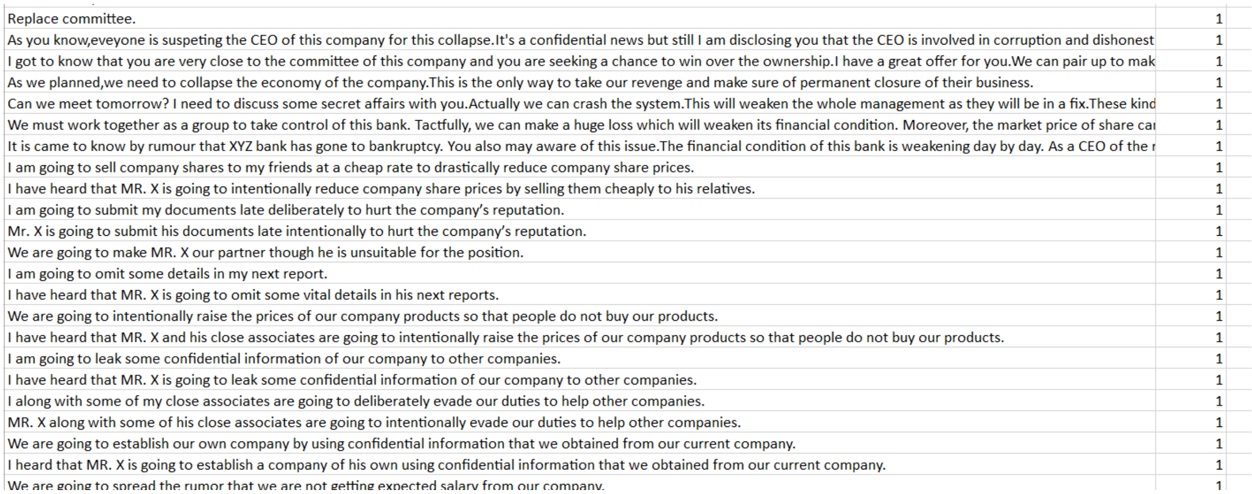
It means the overall reputation of the company such as any rumor about the management, pricing, strategy, working condition, and financial statement so many fake news can harm the reputation of the company. This types of conspiracy can effect a company over night and destroy the socio-economic value 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 5.2 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 r	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

Fig 5.2: Red Dataset

5.2 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.

5.2.1 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 $\frac{324*100}{450}$ =72

So the accuracy in the Green data detection is 72%

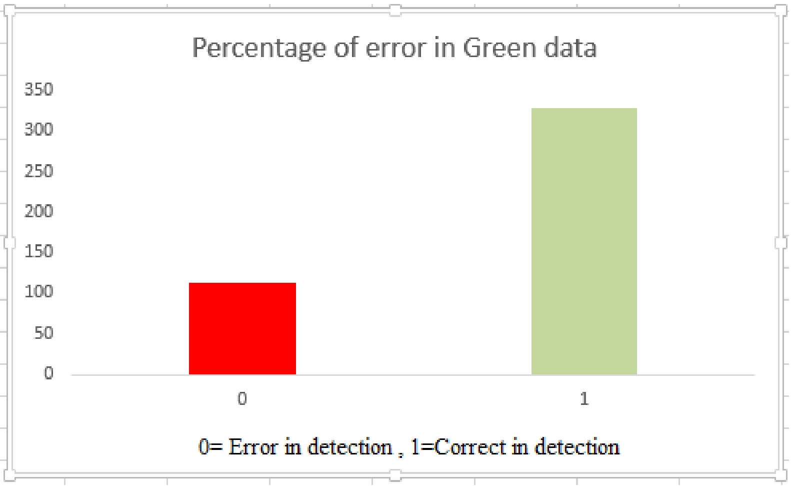


Fig 5.3: Percentage of Error in Green Data

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%.

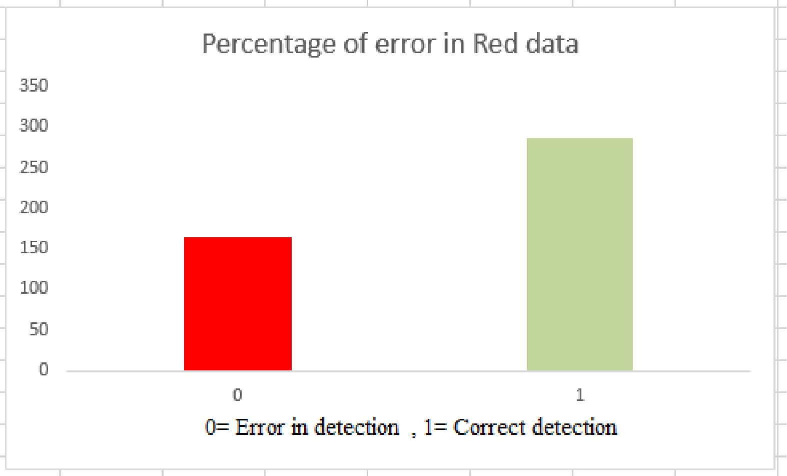


Fig 5.4: Percentage of Error in Red Data

5.2.2 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

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

Table 5.1: Real Time Accuracy of Mail Data

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.

Chapter 6

Conclusion and Future Recommendation

In this chapter in section 6.1 we conclude our development system. We describe the limitations of our developed system in section 6.2. In the same section, we also provide suggestion for future improvements.

6.1 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.

6.2 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.

Bibliography

[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].Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., van der Goot, E., Halkia, M., Pouliquen, B., Belyaeva, J.: *Sentiment Analysis in the News*. In: Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010), Valletta, Malta, May 19-21, 2010:2216–2220.

[14].https://monkeylearn.com/sentiment-analysis/

[15].https://towardsdatascience.com/machine-learning-classifiers-a5cc4e1b0623

[16].https://www.ibm.com/support/knowledgecenter/zosbasics/com.ibm.zos.zconcepts/zconc_datasetintro.htm

[17].https://medium.com/datadriveninvestor/machine-learning-ml-data-preprocessing-5b346766fc48

[18]. “What is the Python programming language? Everything you need to know | InfoWorld”

https://www.infoworld.com/article/3204016/python/what-is-python.html#toc-1

[19]. “Matplotlib: Python plotting — Matplotlib 3.0.0 documentation”

https://matplotlib.org/

[20]. “NumPy — NumPy” http://www.numpy.org/

[21]. “scikit-learn: machine learning in Python — scikit-learn 0.20.0 documentation”

http://scikit-learn.org/stable/

[22]. “A Gentle Introduction to Scikit-Learn”

https://machinelearningmastery.com/a-gentle-introduction-to-scikit-learn

[23]. “Python Data Analysis Library — pandas: Python Data Analysis Library”

http://pandas.pydata.org/

[24].PyMySQL · PyPI

https://pypi.org/project/PyMySQL/

[25].WTF is TF-IDF?

https://www.kdnuggets.com/2018/08/wtf-tf-idf.html

[26].Nádia F.F. da Silva, Eduardo R. Hruschka, Estevam R. Hruschka Jr. "*TWEET SENTIMENT ANALYSIS WITH CLASSIFIER ENSEMBLES*." Decision Support Systems, Vol.66, Pages 170–179, October 2014.

[27].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.

[28]. 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.

[29]. 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.