ADDIS ABABA UNIVERSITY
COLLEGE OF NATURAL SCIENCE
SCHOOL OF INFORMATION SCIENCE

Classifying Insider Threat from Electronic Mail Communication

A Thesis Submitted to the School of Information Science of Addis Ababa University for the Partial Fulfillment of the Requirements for the Degree of Master of Science in Information Science

By

Firesenbet Adela

June 22, 2016

Addis Ababa, Ethiopia

ADDIS ABABA UNIVERSITY
COLLEGE OF NATURAL SCIENCE
SCHOOL OF INFORMATION SCIENCE

Classifying Insider Threat from Electronic Mail Communication

A Thesis Submitted to the School of Information Science of Addis Ababa University for the Partial Fulfillment of the Requirements for the Degree of Master of Science in Information Science

By

Firesenbet Adela

Advisor

Wondwossen Mulugeta (PhD)

ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL SCIENCE

SCHOOL OF INFORMATION SCIENCE

Classifying Insider Threat from Electronic Mail Communication

A Thesis Submitted to the School of Information Science of Addis Ababa University for the Partial Fulfillment of the Requirements for the Degree of Master of Science in Information Science

By

Firesenbet Adela

Name and signature of members of the examining Board

Name	Title	Date	Signature
Wondwossen Mulugeta (PhD)	Advisor	June 22, 2016	
	Chair Person	June 22, 2016	
Million Meshesha (PhD)	Examiner	June 22, 2016	
Dereje Teferi (PhD)	Examiner	June 22, 2016	

DECLARATION

I, the undersigned, declare that this thesis is my original work and has not been presented as a whole a degree in any other university and that all sources of materials used for the thesis have been duly acknowledged.

Firesenbet Adela
June 22, 2016

The thesis has been submitted for examination with my approval as university Advisor

---

Wondwossen Mulugeta (PhD)

June 22, 2016

ACKNOWLEDGMENT

I would like to express my gratitude and heartfelt thanks to my advisor, Dr. Wondwossen Mulugeta for his keen insight, guidance, and unreserved advising. I am really grateful for his constructive comments and critical readings of the study. I am very grateful to Ato Solomon Mekonnen (AAU) and Yalemsew A (JU), for constant support in accessing the data and providing appropriate comment, explanation and other supportive document for my research work. My special thanks also goes to Addis Ababa University Library staffs.

I am immensely thank you to my beloved family especially my father and mother (Eba and Eta) Emu, Getch, Sheks, My frinds Tseganeh, Mesi (Shu), Abresh (Debian), Solomon Aklilu, and all, but not mentioned here thank you for giving me unconditional care, love, time, patient and support throughout my life.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ANN…….....Artificial Neural Network

CALO……...Cognitive Assistance that learns and Organize

CERT……...Computer Emergency Readiness Team

CSI……..….Computer Security Institute

IT……..Information Technology

KDD…….Knowledge Discovery in Database

KKT…….….Karush Kuhn Tucker

MAE…….Mean Absolute Error

MCC…….Matthews Correlation Coefficient

MIT……...Massachusetts Institute of Technology

OCEAN……Openness, Consciousness, Extraversion, Agreeableness, and Neurotic

OLAP……..…Online Analytical Programming

PWC……...Processed Word Count

RAE……...Relative Absolute Error

RAND……..…. Research and Development

RMSE…….Root Mean Squared Error

ROC……...Receiver Operation Characteristics

RRSE…….…Root Relative Squared Error

SMO……..…Sequential Minimal Optimization

SVM…….….Support Vector Machine

USB……...Universal Serial Bus

TABLE OF CONTENTS

ABSTRACT

In a current interwoven global world the means of communication has been diversified. Electronic mail is one of the popular, simple and user-friendly for communication. The implication of this means of communication is reflected in various corners of the day to day activities of the modern world. Currently, email communication is set as a standard procedure for office communication in many organizations. Having the good face of such a communication approach, on the contrary unwanted distracting messages could bring institutional instability and even collapse.

The objective of this research work is to classify the level of being insider threat using email text classification techniques from the electronic communication. In order to meet the stated objective, data mining algorithms in Weka 7.8 software has been used to classify the email texts. The experiment was conducted using 9808 negative and positive dictionary words identified by psychologists for training. For testing individual email files are used. The Enron higher officials email text was investigated after extensive text preprocessing techniques. The text preprocessing technique includes removal of email header, signature, alphanumeric character etc.SMO Classifiers are employed to manage the experiment. Therefore, the text email analyzed was categorized into negative and positive word counts then the negative word count was further classified into five stages of threat levels. Among twenty eight higher officials investigated at Enron Company, 22 of the employees were found at the exploration stage, one on exploitation stage, two on execution stage and three of them classified under escape stage.

The evaluation of the classifier is acceptable and suitable for threat classification. Moreover, a court which was designated to investigate wire fraud, conspiracy and false audit report, convicted 3 of the officials spend in prison from 1.5 – 24.3 years. These individuals were classified under the escape stage of this study. Eventually, the output of this study indicates the promising use of text classification technique to trace and classify insider threats from email communication. Hence, further study and standardization of such a work could bring better result in organizational security and institutional functioning.

Key words: Insider threat, text classification, email classification, threat classification, organizational security

CHAPTER ONE
INTRODUCTION

1.1 Background

Information and communication (ICT) has been a dynamic scientific domain where its application falls almost in every sector of human progressive development and interaction. Particularly at this era, the role of ICT is more than any organizational elements which had been given priority in the older time. One of its typical ultimate effects is enhancing and promoting the connection of people to people, nation to nation, organization to organization etc. simply and rapidly. Among these, electronic mail is the prior and relatively the oldest to be mentioned (Janus, 2010). Electronic mail becomes one of the good mechanisms for intra-organization communication, scientific communication, business transaction, marketing designs, consultancy services and many more. This may be attributed because of its low cost, time saving, high speed, user-friendly and not bounded with geographic barriers. These features have made electronic email to be chosen in the wider population of the internet world where a great deal of important and sensitive data shared.

Currently in every corner of our planet the use of email communication for sensitive information and data is common among organization. However, organizations nowadays recognize the value of sensitive data concerning it as the asset and livelihood of their day to day activity. With the advent of ICT technologies and uncontrolled standard, organizations asset are easily available by smart phone, desktop and laptop users. In this regard employees' usage of Internet increases significantly. Due to this organizations become open to a range of threats. These lead to increase the opportunity of external and internal (Insider) threats and attacks against the organizations. There are several different definitions of the terms external and insider threat. Among these, one common definition is that an insider is defined as an a person who have privileged access right, represent, use, or decide about one or more assets of the organizations (Probst, 2008). External attack or threat on the other hand is defined as, a threat that executed by non-employee who have no authorized or privileged access, he/she exploit the organization resources by passing network firewalls and security devices illegally (Osman, et al, 2015). Different circumstances suggest that, there are many different types of insiders. From the many definitions of the term insider one can observe that these definitions depend on the interpretation and emphasis placed on, such as

access to the system is one of the main mechanisms to this insiders, including freely joining the network system through the wired and wireless network of the organization. Indeed, organizations' important information becomes uncovered to hackers attacks. These threats come from various sources, such as employees' (insiders) activity or external attack.

In order to protect these external (which comes from outside the organization) threat organizations put into practice a broad defense such as configuring strong firewall and latest hardware devices but it cannot address the risk from inside (within) the organization. In fact research works done by Jouini, etal. (2014), specify 70% of fraud is committed by insiders rather than by external hackers or attackers, but 90% of security management or control focused on external threat. As the result indicates, even though outsider threats are more outshine, insiders able to pose a significant risk to information security if they are agitated by different psychosocial factors such as stress, hostility, seeking money, depression, hopelessness, and job control (Sarkar, 2010).According to Vectra, (2014), 61% of Information Technology professionals states they could never prevent insider attacks and 59% mentioned they were unable to even identify. According to CERT (2015) statistical report, 85% of fraud problem committed by insider threat. As stated by Pennebaker (2001) individual emotion related to the psychosocial factor. The psychosocial factor affects individuals' personality to be good or bad. Moreover, these bad psychosocial factors pretense to risk.

The term psychosocial factor is the combination of psychological and social factors. Psychological factors embrace individual level of progression and meanings that dictate or influence the mental states of the person. Social factors consist of wide-range of factors at the level of human society concerned with social order and social processes that impose on the individual. Psychosocial factor also implies that the effects of social processes that are sometimes mediated through psychological understanding. The relationship between psychological factors and the physical body can be influenced by social factors. Psychosocial factors include social support, being alone, marriage status, social disruption, sadness, work environment, social status, and social incorporation (Stansfeld & Rasul, 2007). The terms or words individuals use in their day to day activities reflect the important characteristic of their social as well as psychological environment and social relationship they are in.

Language is one of the ways peoples reveal their internal thoughts and psychological emotion into a structure that others can understand. Languages are the standard by which, personality, cognitive, clinical, and social psychologists effort to understand human beings. ICT development with high speed computers, fast Internet speed and well-designed software's has helped us to bring a new era of psychological study of language. By studying the variety and number of words, researchers relate everyday language usage with psychological emotion, self-reported measure of social behavior, personality and cognitive styles. The smallest words in our vocabulary often reveal the most about our psychological emotion (Tausczik, et al. 2010). Recently the role of psychological emotion towards understanding organizations has made significant progress. Individuals develop different emotional states based on their living and working environment. These include positive and negative emotion. Positive emotions are associated with increased creativity, spontaneous and responsive to stimuli, easy involvement in professional and constructive approach to meet the requirements of professional activity. On the other hand most experimental studies implies negative emotion shows harmful effects, Such as the tendency to process negative information; this can be expressed by dissatisfying with work, terrorism, corruption, bribery, sabotage, nepotism and deliberate acts of espionage (Andrieș, 2011). Due to these negative emotions, organizations face numerous security threats. A number of these security threats may instigate from the "trusted" inside of an organization (Kandias et al, 2010). By using the Internet infrastructure, the employees communicate one to the other via email to come up with their common goal in line with the organization objective by using natural language as a medium of communication. Brown, et al (2013), has found that there is a strong correlation between word use and behavior of employee.

This research work investigates insider threat by using text classification technique from employees email content based on their word use during e-mail communication and associated psychological emotion using statistical analysis. That measure differences in the common words found in electronic communication may provide clues about potential insider threat and related psychological emotion. For this research purpose the Enron company employees' email data set is used. This company was an American commodity and energy service company founded in Huston, Texas in 1985 by Kenneth Lay. Enron was one of the biggest American companies cited as having prevalent audit failure in history of American (Bratton, 2001). The main reason Enron email archive were selected rather than Short Message Service (SMS), twitter and facebook is

that, the first reason is that because of email archive availability of online for research purpose. The second reason is that, the data set is simple to validate with employees committed wrong doing with court verdict that lead the organization to collapse.

In this study the word usage of the employee through their email communication was examined which would be important to classify the likelihood of being insider threat or not based on the fives stage model (Exploration, Experimentation, Exploitation, Execution, Escape) of insider threat(Young, et al,, 2013;Vectra,2015; Ted et al, 2013; Smith, 2014).

Exploration (Recruitment/Tipping Point) stage is characterized by involvement of new circumstances that may lead to irreversible action deliberately or unknowingly. All those investigated insiders not necessarily mean they are threat to the organization. But this stage is the first initial step where by individuals comes in to the next upper ladder (Experimentation) of the threat level.

Experimentation (Search/Reconnaissance) is characterized by having a target to obtain relevant information about the activity and resource held by the target through visual observation or other detection mechanism discovered through at the exploration phase.

The third stage is the exploitation (Utilizing the Weakness) whichis characterized by the insider unfairly treat the target resources or utilizing resources wrongly by finding weak points from policy documents, guideline or procedures identified at the exploration phase and tested on the experimentation phase.

Execution (Collection/Exfiltration) is characterized by unauthorized transfer, retrieval, coping of data from database or information sources. The insider perform the threat by using the above three steps (i.e. exploration, experimentation, exploitation)

Escape/Envision (Obfuscation) stage is characterized by making some unlawful materials presented as a legal justification, untruthful, shift blame or preparing cook book etc.

Words used by individuals classified into positive and negative classes by using Support vector machine (SVM). After the words are classified into positive and negative, negative words are used to classify into the five stages. Based on the percentage of negative word usage from the

total words used during a year, employees are classified into five cut of point. The first cut of point aligned with the first stage and the second cut of point aligned with the second stage etc.

1.2 Statement of the problem

Nowadays organization or institutions are a blend of employees, counselor, partners and multifaceted infrastructure. Due to this organizational nature, handling insider threat makes unapproachable or difficult. Those insiders can be former or current employee or contractor or business partner who had or has privileged access to an organization's data, Information systems and network are misused by the given privilege in a way that negatively affected the organization confidentiality and integrity (Vectra, 2014). An insider threat possibly obtain several forms, including dissatisfied workers, oppression, individuals under financial stress, data breach, terrorism, IP theft, corruption, bribery, sabotage, nepotism and deliberate acts of espionage (Brown, 2013). To do these forms of threat they use different communication methods such as telephone, face-to-face, social media and electronic mail. Knowing the content and the words used by employee (insider) email communication, help organizations to protect themselves from such risky insider threat. Due to insider do violence, organizations face potential damage through loss of revenue, loss of reputation, loss of intellectual property or even loss of human life.

A survey done by the Computer Security Institute in 2008 indicated about 44% of all organizations undergo abuse of computer systems in 2008 which dropped to 30% in 2009 and 42% reported loss of laptops both in 2008 as well as 2009, and 17% reported theft of customer data. The 2009 survey also indicate 25% of the respondents felt that 60% of the financial losses were due to insiders. The insider threat problem is more intangible and confusing than any other threat. Assessing the insider threat is the first step to conclude the probability of any insider attack. Therefore, employees' psychological emotion assessment is essential in facilitating the classification of insider threats. According to Christopher et al., (2013), there exist a strong correlation between word use and behavior of employee. The same research demonstrates that measurable differences in the frequency of common words found in electronic communication may offer clues about prospective insider threat risks.

According to the literature review made, the gap in the previous research works is that the earlier researchers are not able to classify insider threat according to its severity. Moreover, the aim of

this research is to fill the above gap by classifying insiders, into the five stages of insider threat: Exploration, Experimentation, Exploitation, Execution and Escape stage.

To fill the above mentioned gap, this research addressed the following research questions:

1. Can insiders' intention be characterized by analyzing their email text using text classification technique?
2. How is word or text usage related with insider threat?
3. What are the special features of insider threat associated with psychological emotion?
4. To what level/intensity/extent insiders can disrupted the services given to the others?

1.3 Objective

The following general and specific objectives are outlined in order to solve the problems that initiate this study.

1.3.1 General objective

The general objective of this research is to classify the level of being insider threat using email text classification techniques from an electronic communication using statistical analysis.

1.3.2 Specific Objectives

To achieve the aforementioned general objective, the following specific objectives are drawn.

- ☞ To undertake detailed literature review so as to understand related research work
- ☞ To prepare training dictionary and test dataset;
- ☞ To train the model using word dictionary;
- ☞ To analyze the association between insiders word usage and their psychological emotion as a threat;
- ☞ To classify the negatively agitated employees into the five stage model
- ☞ To affirm the degree of disruption by the insiders threat;
- ☞ To evaluate the performance of the classification technique.

1.4Scope and Limitation of the study

The scope of this study is to investigate insider threat from their email word usage distribution class (i.e. positive and negative) and associate with psychological emotion based on the five

stage model(Young, et al 2013; Vectra, 2015 ; Ted et al, 2013) to look into who are risk in the organization by using text classification and statistical techniques. For text classification we use algorithms on Weka software to decide the class of each word. After the class is known by employing five cutoff points manually, we classify employees into the five stages. The frequency of the word occurrences do not considered rather representatives of each word are used.

The problem coverage of this research work is to classify employee (insider) into five stages of insider threat. Insiders are classified based on the total number of negative words count. The data coverage of this research work pervade from January 2001 to May 2002 Gregorian calendar. The classification had been done based the conversation they made on this range of date. The approach we cover in this study all email files are merged to know the psychological emotion of the employee during the tenure since January 20001 to May 2002. Because merged files are more expressive about the employee emotion than single file at the possession within the organization.

One of the limitations of this research work is that, it classifies manually employees using simple statistical technique (i.e. cutoff points), into the five stages model after the words are classified using the classification algorithm. The other limitation of this study is that it doesn't consider the relationship of the term or the sentiment analysis because of the technical difficulty and the time given to complete this research.

The testing data set only work for those who worked in the Enron Company. Since the study basically trains the classifier using data from different sources, the classification result for other people with different work culture other than Enron Company may not be robust and valid. The attachment files with e-mail are not considered, because there is no attachment files within the email archive.

1.5 Significance of the Study

The greatest security threat comes from the person with authorized access (Jouini,et al. 2014). People design, develop and use as well as misuse information systems. It is, therefore, necessary to understand the psychology of people involved in both malicious and non-malicious insider activity. Protecting your organization from insider threats is necessary part of critical information security best-practices. This research has the following significance:

- ☞ Helps the organizations to know the psychological status of their employee. Knowing their status assist the organizations to coach and mitigate negatively agitated employees and to reward positively motivated.
- ☞ Helps to synthesis the association between psychological emotions with insider threat.
- ☞ Helps to implement effective work planning and control to reduce job pressure and maintain employee preference.
- ☞ Helps to improve awareness of employees to use proper words during conversation.
- ☞ The approach and technique can be used as a base line for those who want to further develop this research area.

1.6 Methodology

1.6.1. Research design

This study is an experimental desk based research where retrospective data were used to train the classification model. The dataset was collected and organized by the CALO Project (A cognitive Assistance that learns and organizes).

1.6.2. Data preparation

In order to classify insider threat from their word usage employees email data set was used. The training dictionary is already prepared by the psychologist. The dictionary totally contains 9808 positive and negative words. From these 4904 are positive words and 4904 are negative words. The classifier algorithm trained with 9808 words to classify the newly coming instances. The test set data were collected from the former American energy service company i.e. Enron. Enron Company was selected due to its freely available online for research purpose by the federal Energy regulatory commission during its investigation. Moreover, some of the employees were found guilty in the court. The data set contains 150 employees email archive. From these 150 employees only 28 Enron higher officials emails are classified into the five stage model. The main reason these higher officials (CEO, President and vice president) are selected for investigation is that, most of there were accountable for the biggest fraud committed in the annals of USA. These higher officials sentence from 1.5 -23.5 years to serve in prison. For each employee 340 email files are selected from individual email archive. These emails are

preprocessed (removing headers and signatures, numbers, abbreviations etc.) for further classification.

1.6.3. Experimentation tools and algorithms

The tools used to classify texts into positive and negative class were Weka 3.7.

The experiment was conducted by first training the model algorithm with the dictionary word and then the test words are classified into positive and negative classes by employing Support vector machine (Simple Minimal Optimization).

The main reason we select the above algorithm, for ease of understanding and interpretation of the result of the model. The test data are classified into positive and negative word classes. The main reason the we classify the data set in to two is that negative words are indicators for threat and positive words are indicators of protection, safety, and encouragement.

The percentage of the negative words as compared with the total number of words utilized by an individual helps to assign the individual into the five stage model. Because negative words have the potential to describe negatively agitated person. The five stage model is a model that indicates how much an individual is savior when move from stage one to stage five. The percentage of negative words from each individual are put in ascending order and adjusted into five cut off point to classify the tendency of employee risk level.

1.6.4. Evaluation procedure

In order to evaluate the performance of the classification technique employed the federal investigation office result was used. The federal investigation in collaboration with the court sentenced between 1.5- 23.4 years based on the crime Enron higher officials committed. The classification algorithm result found was evaluated by the court verdict.

CHAPTER TWO

LITERATURE REVIEW

In this chapter, an effort has been through different extensive review of literature on security, text and document classification, supervised and unsupervised machine learning approaches and classification algorithms, data mining, Insiders psychological characteristics in common, big five personality traits, how human psychology and emotion is related with the words they used; aimed to give background about the Insider classification.

Insider threat nowadays' known with it as information security problem. It particularly indicates the present employee or the former employee who have the chance to exploiting the information system by using their valid access in order to perform malicious activity. Threats which arise from the inside of an organization's yard becomes a considerable attention than from outside the, since it is difficult to distinguish them from compassionate activity of day to day worker. Due to insiders know information and potential about the organization, their consequence much more serious than the external hackers or attackers. According to Hamin, (2000) Insiders due to their affiliation with the organizations information system and system holes in security are in profitable position to misuse organization information system. The insider threat could have negative impact on the revenue of an organization or due to disclosure of classified information, negative perception of the community on the organization image, (Colwill, 2009). The following section define the term information security, insider, insider threat and the concept and different perspectives that represent insider threat.

2.1 Defining Insider threat, language and psychological emotion

2.1.1 Asset

Asset can include People, information and property. People possibly will include customers and member of staff along with other invited persons such as contractors or guests. Property assets include tangible and intangible thing that have a value. Intangible assets comprise character and proprietary information. Information perhaps embraces databases, software source code, significant organization records, and other intangible substance(Independent Security Consultants, 2010).

2.1.2 Insider and Insider threat

Insider is a person who is an affiliate or employee of a group or organization. The definition of Insider threat definitely depends on what an Insider is, Insider threat can obtain numerous forms, but the main part is the attack is initiated from inside your organization network. They can be compromised insider those negotiated by an outside threat in black market to sell confidential information; neglect insider those who expose the data carelessly or accidentally; and Malicious insiders those who steal data or destroy organization network intentionally(Bishop, 2008)..

2.1.3 Personality

Personality is a structured set of uniqueness held by a person that influences his or her cognitions, inspiration, and behaviors in diverse circumstances (Pennebaker, 2001).

2.1.4 Risk

It is potential for loss, destruction or damage of an asset as a result of a threat utilizing weakness in the strategies and management. Due to failure of organizations to implement new strategies and best data management technology as well as to employee monitor malicious behavior of an employee seeking financial gain or revenge (Independent Security Consultants, 2010).

2.1.5 Trait

Traitis the exceptional quality such as jealousy and envy, that makes one person unique from the other (Pennebaker, 2001).

2.1.6 Positive Words

Human beings in their day to day communication use positive words to express thinking about the good qualities of an object, person, process and understanding etc. positive words are characterized by exhibiting acceptance, very confident, constructive, optimistic, practical, useful, helpful, effective, beneficial, pragmatist, with no possibility of doubt in nature (Pennebaker, 2003).

2.1.7 Negative Words

Negative words are characterized by self-defeating talk that reduces their internal potential and confidence, they compare themselves adversely with others this leads them to greater stress, depression and anxiety, blaming others responsibilities for their bad luck, fear of failure and

making mistakes, over generalize and labeling peoples, discounting positive events and struggle to forgive themselves (Aspinwall, & Taylor, 1993; Collins, 1996). Negative words are on the contrary side of positive words. These words are used by peoples to state about the bad qualities of an object, person, understanding and process etc. It is characterized by the absence of positive words such as a word or statement that state denial, disagreement or refusal, refuse to accept, reject, dark, depressing, anti, uncooperative, and obstructive in nature (Pennebaker, 2003).

2.1.8 Emotional stability and instability

According to Barrick, & Mount, (2000) emotional stability is associated with being calm, steady, self - confident, and secure. From this a person who is emotionally stable can have psychologically consistent mood and affect despite forces that threaten to disturb it. This stability can be an indicator of peaceful and non-threaten behavior in the work environment. Emotionally unstable person can cause of a lot of damage and even breakups and can increase the risk of reacting with violent or harmful behavior. It is characterized by rage, sorrow, shame, panic, terror and long term emptiness and loneliness. Those people who show emotional instability can cause of threat in their work area and considered as insider threat.

2.1.9 Positive Thinking

The traits of positive thinker include optimism, belief, enthusiasm, courage, integrity, confidence, determination, calmness, and focus. They are characterized by expecting the positive outcomes as well in the difficulty, personal motivation that add value, trust themselves and others, have personal commitment, have interest to take risks and fears, assuring personal abilities and potential, they strive tirelessly for their objective, they are willing to wait for chance result from the others and oneself, and they have directed goals and vision.

According to Judge, et al., (1998) most managers would probably agree that positivity is something they value in employees. A broad personality trait, labeled positive self-concept or core self-evaluations, is a potentially important personality trait in the classification of job performance. Positive self-concept consists of four specific traits self-esteem, generalized self-efficacy, locus of control, and (low) neuroticism or emotional stability. Drawing from four motivation theories, Judge, et al., (1998) argue that the principal reason positive self-concept is

linked to job performance is because positive employees are more motivated to perform their jobs.

Baumeister, (1997) define "The term self-concept refers to the totality of inferences that a person has made about himself or herself' (p. 681). Positive self-concept is the favorability of these self-inferences. Thus, individuals with a positive self-concept evaluate themselves positively, and are likely to make favorable inferences about themselves and be accepting of their identity. Children from the basic elements of their self-concept very early in life and, although changes in self-concept do occur, the initial formation of self-concept probably has lasting consequences for the individual. Core self-evaluation is the fundamental premises that individuals hold about themselves and their functioning in the world. Self-esteem is the central element underlying on a positive self-concept or it is the overall value that one places on oneself as a person. Self-efficiency is defined as one's judgments of "how well one execute course of action required to deal with prospective situations" (Bandura, 1982, p. 122). The locus of control represents the perceived degree of control in life. Individuals with an internal locus of control believe their behavior controls their lives, whereas individuals with an external locus of control believe that their life is controlled by luck, chance, fate, or powerful others (Rotter, 1966).

2.2 Overview of Enron Corporation

In the annals of the United States of America, Enron Corporation represented as the largest fraud scandal. The government made a financial investigation on the alleged fraud and the company was forced to file for bankruptcy in December 2001 (Lucian & Cristina, 2007).

Enron Company founded by Kenneth Lay in 1985 and rooted at Omaha, Nebraska (US). Enron was an American largest energy company and engaged on providing natural gas, electricity and communication to wholesale and retail to customers. In 1985 the Huston natural Gas fused with Inter North Energy Company to form a big energy company that based in Huston, Texas. When Kenneth Lay the former chief executive officer of Huston natural Gas is assigned as chief executive officer in 1986 at this young energy company and he choose the name Enron Corporation. The company integrated several fuel pipeline systems and able to build the first nationwide natural gas pipeline system in United states of America. In 1987, after discovering and identifying fuel and oil traders in New York the company overextend its account by almost

$1 billion and strive its loss down to $142 million. This was the time that leads Enron Corporation to develop variety of services in order to reduce the risk of price fluctuation (Lucian, C., & Cristina, D, 2007).

After a year the company opened its first office in England out of USA. On the meeting “Come to Jesus” the company announced its new business strategy to the executives. The strategy was to pursue unregulated markets over its regulated fuel and oil pipeline. In 1989 Jeffrey Skilling joined Enron Corporation and commence Gas Bank program, which is the buyers of natural gas agreed long-term supplies at fixed prices. During this time the Company begins cooperation to offer financing for oil and gas producers. After a year Teesside power plant began operations in England. It would take as an attest to Enron company international strategy success. In 1992 the company expands and pushes its territory of service to South America by integrating Transporta dora de Gas Del Sur. In 1995 Enron joined the European wholesaler market with the establishment of gas company and trade center. These all efforts made the company to collect the biggest profit ever.

The profit collected from Europe helps the Enron to build Dabhol power plant in India. The company continued its policy of acquiring companies and acquired Wessex Water in 1999 from England which formed the foundation of its water subsidiary Azurix (Lucian, C., & Cristina, D, 2007).

2.2.1Causes of the downfall of Enron

Enron's hazy financial statements and complicated business model did not show clearly facet of its process and finances in relation with Shareholders. After a year later Azurix one-third its asset sold to public, the Enron problems become clear as the shareholder become fell sharply compared with after the early raise. In the same year Enron Energy services turned its first profit. Even if in 2000 the profit reached $100 billion that proves the growing importance of trading, the problems with Azurix continued. According to the 2000 report of Energy financial group Enron was ranked as the sixth profitable and capitalized market Energy Company in the world(Lucian, & Cristina, D, 2007).

In October 2001 Enron announced its $570 million bankrupt to the public by the California utility Pacific Gas and Electric Corporation. The critical dates of the company’s Scandal are

October 16, 2001 and November 8, 2001. This disclosed information initiates the US security exchange commission to conduct investigation against Enron high-ranking executives, investment banking partners and company's accounting firm. After the investigation Enron's accounts for the previous four years had not shown the true state of its huge indebtedness. Enron became collapsed (Lucian & Cristina, 2007).

2.2.2 The Enron Verdict

Many higher, middle and low level Enron executives worked for the company have been in the investigation and trail of wrongdoing at the company since its collapse. Most of them found guilty and some of them are still not receive their punishment.

From the top executives of the Enron Company Kenneth Lay chairman and chief executive convicted and charged with conspiracy, security fraud, wire fraud and banking fraud, but vacated after his death. Jeffrey K. Skilling chief executive convicted and charged conspiracy, securities fraud, insider trading, Perjury/lying to investigators/auditors and sentenced twenty four year and six months imprison he is still in prison. David W. Delainey charged with insider trading and sentenced two and half year in prison and he is released now. Andrew S.Fastow chief financial officer sentenced six years in prison for his charge of conspiracy. Paula H.Rieker the Board Secretary and manager of investor relation sentenced 2 years with the charge of insider trading. Richared A. Causey Chief accounting officer sentenced five years and six months with the charge of security fraud. (A. B. (2006, October 23). Enron Scorecard. Retrieved May 11, 2016, from http://www.nytimes.com/ref/us/20061023_ENRON_GRAPHIC.html).

From middle level executives Lea Fastow was an assistant treasurer and Lawrence M.Lawyer the financial executive charged with filing false tax returns and sentenced one year and two years respectively. Michael J.kopper vice president charged with conspiracy, Money laundering and sentenced thirteen months. (www.nytimes.com)

Daniel O.Boyle the financial executive deal for Nigerian Barage charged with Conspiracy, Wire fraud, perjury/lying to investigators/ auditors and sentenced three years and ten months. And many others are charged related with Nigerian Barage such as Daniel H.Bayly chief investment banking Merrill Lynch, James A.Brown Chief of asset leases Merrill Lynch, William R.Fushs

vice president Merrill Lynch, Robert S.Furst investment banker Merrill Lynch. (www.nytimes.com)

The Southampton, England deal of Enron, Gary Mulgrew a managing director of Natwest bank, Giles Darby a managing director of Natwest Bank, and David Bermingham finance specialist at Natwest bank, all charged wire fraud and waiting their sentence. (www.nytimes.com)

As stated on NewYork Times A. B. (2006, October 23). Ex-Enron Chief is sentenced to 24.3 Years [Enron Scorecard]. Retrieved May 11, 2016, from http://www.nytimes.com/ref/us/20061023_ENRON_GRAPHIC.html?ref=enron ; A. B. (2006, October 23). Enron Scorecard. Retrieved May 11, 2016, from http://www.nytimes.com/ref/us/20061023_ENRON_GRAPHIC.html) after the collapse of the huge energy company Enron, many top executives, presidents and vice presidents are charged and sentenced from 24.3 years to 1 year. Among the charged executives:

Jeffrey K. Skilling:-the former chief executive of Enron charged with conspiracy, Securities fraud, Insider trading, Purjury/laying to investigator/auditors sentenced 24.3 years and he is still in prison.

Kenneth L.Lay:- chairman and chief executive of Enron charged with Conspiracy security fraud, wire fraud, bank fraud; convicted, but vacated after he died and deceased.

David W.Delainey:- chief executive of energy division at Enron company, charged with Insider trading and sentenced 2.5 years.

Andrew S.Fastow:-chief financial officer at Enron and charged with conspiracy and sentenced 6 years in prison.

Richard A.Causey:- chief accounting officer at Enron charged with security fraud and sentenced 5.5 years, but he is free on bond.

Paula H.Rieker:- Board secretary, manager of investor relation charged with Insider trading and sentenced 2 years, but she is now on probation

2.3 Information System Security

Information security is defined as ‘protecting information and information systems from unauthorized access, use, disclosure, disruption, modification, perusal, inspection, recording, or destruction’ (Allen, 2001).

The central attention of security is protection of assets from threats, these threat can arise from inside or outside for misuse of protected assets. Nowadays organizations use of information systems to store, organize, preserve, process and disseminate important information assets. According to Daft, 2000, Information is defined as data that have been transformed into a meaningful and valuable context for the receiver. How much the information is important and valuable depends on the organization but intellectual property and organizations strategic information are important elements to take competitive advantage against its rivalry.

It is understandable that organizations having these information system are faced a plethora of threat arise from both inside and outside the organization. Protecting the assets of the organization is the role of owners who put worth on the assets. The threat agents also worth on the assets and inquire to misuse the asset in a different manner than the owner. At this time owners recognize the harm on the assets such that the worth of the asset is minimized. Security characteristics include confidentiality, integrity and availability. The owners of the assets will investigate the threats appropriate to their assets and their environment, formulating the risks connected with their organization. This investigation helps organizations to reduce the risk to information system security forced to require law of these countermeasures that is taken as most important to them. Countermeasures are obligatory to minimize vulnerabilities and to get together security policies of the owners of the assets. Remaining vulnerabilities may stay after the imposition of countermeasures. Threat agents take the chance of exploiting loop holes in information system or security measures. Owners will look for to reduce that risk given to other controls (Daft, 2000). The following figure summarizes the general context of information security.

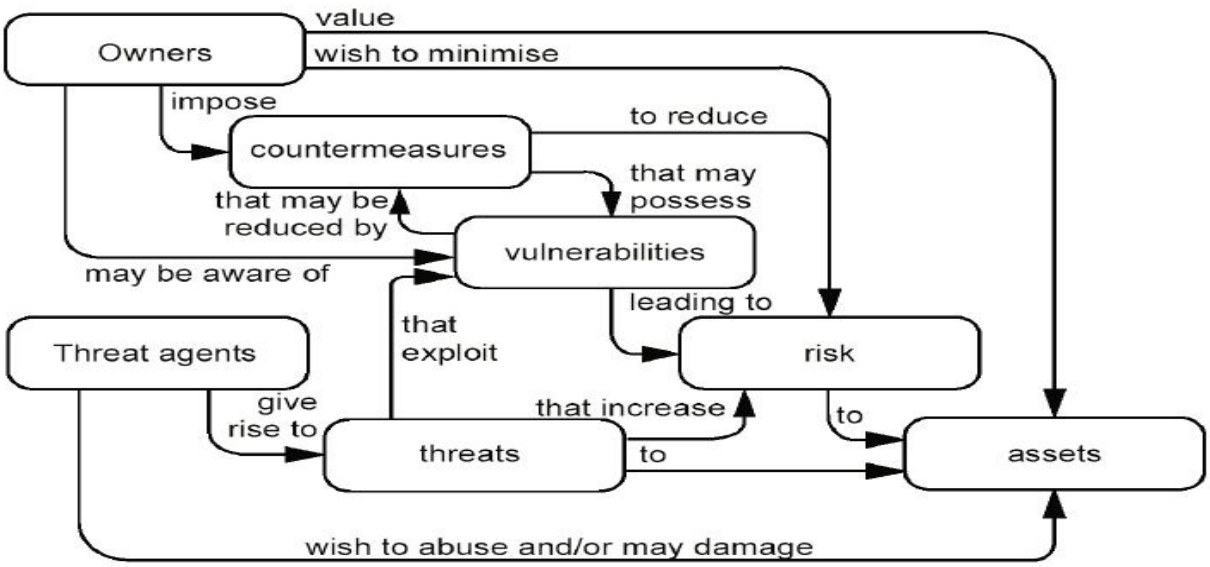

Figure 2.1 security concepts adapted from (ISO/IEC 15408-1, 2005(E))

Systems security failure could solicit unauthorized access, interruption or modification, unauthorized disclosure of information. According to ISO/IEC 15408-1, (2005(E)) these security failures are characterized by confidentiality, integrity and availability respectively.

Confidentiality can be defined as information is accessed when we need, it require cryptography and protect from disclosure. Integrity means ensuring that the data is correct information could not be modified, changed or ruined by the insider knowingly or accidently. Availability defined as information should be at hand when required by the organization to manage and achieve its objective and mission (Ezingeard et al, 2005). Alleviating these three security characteristics can lead to conflict of interest among different parties. For example commercial banks process large amount of customers' data including credits, deposit and withdrawal. Wrong integrity of customers' data could create more negative image on the bank and loss of its profits due to migration of these customers to other competitors. In medical institutions unavailability of information could directly leads to the patient to death. In courts or legal institution the non-confidentiality of both parties the innocent and guilty information could lead to wrong decision or penal complex.

We can generalize that organizations expected to safeguard their information from the threat agents that guide them loss of intellectual property, money, employee loss of lives and customers' as well.

2.3.1 Insider

Different interested research group individually, in group or in the form of organization including United States of secret service, Computer Emergency Readiness Team (CERT), Research and Development (RAND), university researchers and industries have participated in various insider threat research. But there is no one consistent definition for insider. In fact, we agree not having consistent definition for an insider delay research in differentiating or detecting threats from insiders (Bishop, 2008). Defining an insider helps us to conjure and detect different classification and detection approaches threats of the preferred type. In this thesis work I propose a definition of an insider from extensive peer reviewed literature that can span from corner to corner for various domains. Some definitions are acquired from the literature and described as follows.

I enlarge the definition of an insider from Bishop, (2008): *"a trusted entity that is given the power to violate one or more rules in a given security policy... the insider threat occurs when a trusted entity abuses that power*." According to this definition, to understand the set of rules and regulations embodied in the security policy of the organization are determining factor to recognize what an insider is. TBishop(2008) used two primitive actions in order to define an insider. These are breaching of security policy by utilizing genuine access right and breaching of an access control security policy by gaining illegal or unauthorized access right.

In the first action, those insiders are categorized under *compromised insiders*; these types of insiders have access with credential to access resources and have been negotiated by outsider threat to leak data or strategic intellectual and/or financial information. In the second action, the insiders are categorized under *malicious insiders*: these types of insiders can be the former employee who put malware on company network on his/her last day at work. These insiders are trusted employee contractor or third party partners. What makes these insiders from outsider is they are considered as part of the organization and imagined they attempt to achieve the

organization objective. In order to show the degree of insiderness the security person and insider threat researchers give attention to the most damaging cause to an organization greatest asset.

The other definition for insider given by Schultz,(2002) "*insiders would usually be contractors, employee, and temporary, consultants helpers and even personnel from third-party business partners and their contractors, consultants and so forth have the knowledge of the information system"*. From this definition of insider we can understand that it is hard to categorize insiders from outsiders due to the chain length of employment. When an attack is committed by a former employee or contractor who inject malicious program on the system, should the attack taken as from the inside or outside? The so- called "insider jobs" makes the definition more complex to distinguish insider from outsider. In addition to this insiders have the knowledge of how the system operates, security and procedural measures and where valuable asset of the organization is stored. These makes the process of insider threat detection complicated and even undetected. To some extent the definition works true, but relating these definitions to the actual world seems to be questionable.

Kandias et al., (2010) define insider as *"a human entity that has/had access to the information system of an organization and does not comply with the security policy of the organization"*.Kandias definition does not define the clearly the level of access, downloading or sharing or the level of skills required to achieve its objective. In these respect we have to understand that legitimate access is different from authorized access. Legitimate access can be given as an access but that is not authorized such as janitor to access desk and looking into bins for snooping sensitive information. While authorized access is the level grantee to system and network or database administrator with password and username. But simply if the human entity employed as a worker and had/has a privilege to do the given activity and bypass the given responsibility for misusing; and if this misusing contrast with the employer or organization policy of given privilege and responsibility, the human entity or employee, contractor, or consultant considered as insider.

In this thesis work the definition given by Cross disciplinary workshop on "countering Insider threat" Probst,et al, (2009) is considered accordingly they defined insider as a person that has been officially empowered to access and decide about one or more assets of the organizations'. This definition tells us insiders are the ultimate decision makers based on their credential, access

control or any given level of access on the organization assets. Their decision making capability depends on their psychological motives, these motive can be expressed by such as disgruntled employee, workers with financial problem, for espionage, corruption, terrorism etc (Christopher, et al., 2013; Kuheli, 2010; Greitzer et al., 2011).

Insiders are human being know or recognize the organization system structure and the loop hole in the system and they are in advantageous position to exploitation organizations information system (Hamin, 2000). Hamins' definition of insider helps us to understand intentional insiders make social engineering or by utilizing their position study weak spot on the system and harm the organization. Outsiders are protected by using firewall and other techniques to deny the access. Insiders are a trusted person in position and know the system better than outsiders that makes detection of insider threat detection very difficult.

According to Tagg,(2014) an insider is someone who has been given a role within an organization and has access to premises and/or internal systems and information. There are many types of roles; some are core to the business and performed by employees, while others are non-core and contracted out to service providers' such as cleaning, maintenance, or information technology (IT). The categories of insider may be classified as (Tagg,(2014)):

Current staff: work*s* directly for and under the control of the organization's management. This category includes employees as well as temporary staff, contractors, and consultants. Most of these people are located on the organization's premises and are connected to the internal network with access to internal information.

Departing staff represent one of the highest risks to an organization. These peoples consist of employees who have resigned or are planning to do so, temporary staff, contractors or consultants coming to the end of their contract, as well as all the people whose employment or services are being ended by the organization. These people still have access to internal information, and may be motivated to take that information with them they leave, or to commit sabotage in revenge for perceived wrongs.

Former staff: are those who are no longer employed by or providing services to the organization. This group still has insider has insider knowledge and without mitigating controls,

they can do substantial damage long after they have left the organization. Former staff can be highly motivated to attack their former employers.

Service-Providers organizations have many roles that are necessary for the smooth running of business but that are not core to the company mission. Examples of these roles are cleaning services, maintenance, and IT. Over the last 20 years, services providers have steadily moved up the service attach to perform core functions as well. It is now common for service providers to perform business operations and even to be the first point of contact with customers. Although not all service providers will need access to premises or internal systems, there are many roles that do. The key distinction between service providers and staff is that staff are working under the control of the organizations management and subject to its policies and procedures; in contrast, service providers are providing a specified service and all personnel are the responsibility of the service provider. This situation increases risk to the organization, as it has little control over service-provider staff, but the threats are the same as from internal staff.

Partners organizations are partners when they work together on a business venture. Access to information goes up to senior business leaders within the partners. Partnerships may be short or long term and partners may be competitors at the same time as being partners. This situation creates risk for organizations, which has to share information relevant to the partnership but not other internal information.

This study gives prefer for use: *a human entity that has/had access to the information system of an organization and does not comply with the security policy of the organization.* The access level of the information includes using organizations internet service/ information system for the purpose of their day to day activity including their electronic mail communication. Providing employee to use the internet is the level of access grantee to within perimeter of the organization either using wired or wireless devices.

2.3.2 Threat

There are numerous perceptions and understanding on the definition of insider threat. Before we are going to discuss insider threat let us look at some dictionary and scholar definition of “threat”:

The etymology of the word threat has Germany origin *verdrieten* 'grieve' means distress, unhappy, suffer etc. As written on (http://www.etymonline.com) Old English *thrēat* "crowd, troop," also "oppression, coercion, menace," In the old English language the term has given the sense of hostile or unfriendly intention.

According to Longman advanced American dictionary define the term "threat" as "*a statement that you will cause someone pain, unhappiness, or trouble, especially if they do not do something you want.*" This definition notify us if your desire or need you expect from your organization are not meet or if you became disgruntled with financial stress or other cases you may do something that create harm or trouble to the organization. With this action you are defiantly considered a threat to the organization and taken as insider threat.

The definition on (merriam-webster.com) *"a statement saying you will be harmed if you do not do what someone wants you to do" or "someone or something that could cause trouble, harm, etc." or "the possibility that something bad or harmful could happen"*. From these three definitions we can understand threat is related with two entities; the first entity is the person who commits harmful or bad things against the second entity for the response of loose of what they expect from the second one.

Bishop, (2005) in his book introduction to computer security define threat: "*is the potential violence of security."* The violation off course need not to be expected to happen but violation might take place, those leading actions for the violence to be happen considered as a threat.

Businessdictionary.com define "threat" as the action of breaching information system by using the loophole; this may be caused by opening of false information to deceive the users.

Brown, et al., (2013) psychological motivation are the main factors to trigger this violation. These motivations are expressed in terms of big five personality trait OCEAN (openness, consciousness, Extraversion, Agreeableness and Neurotic).

2.3.3 Insider threat

Threats to valuable asset are posed by threat agents initiate from inside or outside. Insider threats poses a greater risk level than even if outsiders gain publicity. The current research work use insider threat interchangeably with insider attack. In this thesis work also I used insider threat.

Among those definition according to Pfleeger, et al., (2010) define an insider threat as the way or techniques of an insider disclose the organizations' valuable asset, policy, process and resources at risk intentionally (for personal gain) or unintentionally.

According to Schultz, (2002), "*an insider attack can be defined as the intentional misuse of computer systems by users who are authorized to access those systems and networks*." From Schultz the term "intentional" misuse implies there is a psychological motivation to be intentional and violate or misuse the system. Motivations activate trigger, direct, process and sustain objective oriented activity and behavior of the employee. It is what reasons us to act intentionally; this can be being paid money by the third party to reduce the financial stress or may be simply to be tech savvy.

An insider threat is the action of breaching the organization security policy and disclosing resources in a manner that negatively influence the organization asset availability, confidentiality and integrity to the competitors or in collaboration with competitors again and infringed by former or current employee, business partner etc. who had or has authorized or legitimate right to the system. The threat can include sabotage, theft, fraud, corruption etc. (Smith, G, 2014).

Carroll, (2006) define "*insider threat can be either intentional or unintentional"*. Defining and discussing the word Unintentional is very important. This means insiders are the result of deficient in policy, failure to follow the policy, improper training, wrong assumption and lack of experience. In both cases whether in deficiency or failure of policy can lead to catastrophic problem.

Both Schultz, (2002) and Carroll, (2006) use the term "intentional" in their definition of insider threat. This inferred to us insider threat can be done intentionally by the insiders.

National counterintelligence and Security center of United States (www.ncsc.gov) define insider threat regarding the USA stand for the action. It arises when the employee has authorized access to U.S. government asset including equipment, networks, facilities and information; utilize that privilege to destruct the security of the government. This report marked more information are steal or carried out the perimeter by the flash disk or removable media in contrary with what the competitors or enemies given in hard copy in the annals of U.S. history. Over the century

malicious insider committed incalculable harm by planting logic bombs with secret motive. At that time the compromised insiders exhibited identifiable sign, but the sign were not reported due to lack of willingness of employee to believe them as disloyal to the U.S. government.

2.3.4 Threats from insiders

As Tagg, (2014) stated successful organizations need to have people in many different roles, with the primary distinction between business roles (business management, sales, customer services, and product design) support/infrastructure roles(IT, finance, logistics, human resources and the like). The people in each role need access to information and system to perform their role, and the key point to make at this stage is that the impact an insider attack makes on the organization is related to the insider role. A salesman leaving an organization to join a competitor will have had access to customer and product related information, whereas someone in product design is likely to have access to valuable intellectual property on current and future products. Customers' service center staff are likely to have access to customers' personality identifiable information that can be used to commit identity fraud.

2.3.5 Types of insider Threat

There are three main classifications of insider threats: accidental, malicious and non-malicious (Tagg, 2014).

i. Accidental Threats

Accidental threats are generally caused by mistakes; for example, staff may not follow operating procedures due to carelessness, disregard for policies, or a lack of training and awareness of the right thing to do. An example is a customer service representative who accidentally breaches client confidentiality by emailing client information to the wrong email address. Such errors may be caused by the use of email clients that have an auto-complete feature on the email address; staff under pressure to keep up with the volume of work may not notice the error before they send the data. This error is a particularly high risk for financial services organizations where in some jurisdictions a client confidentiality breach is a criminal offense (Tagg, 2014).

ii. Malicious Threats

Malicious threats deliberately try to damage the organization or to benefit the attack. Disgruntled IT administrators can sabotage IT systems, bringing an organization to a halt. There have been many incidents where both current and former administrators have deliberately caused system issues for various motives: enjoying the life style of traveling around the world in luxury to fix the problems they created, extorting money from the organization, or simply causing as much damage as possible. Some company information is highly valuable and specifically targeted by attackers. Personally identifiable information is one category that is sometimes illegally copied by staff to conduct identity theft and fraud, or to sell it criminals. Another category is intellectual property (IP) such as trade secrets. Staff may take this information to help them with their next job or to sell to competing companies. This crime is thought to be common with IT developers who often seek to take their source code with them. Industrial espionage sponsored by rival companies or foreign government is another common threat to IP. Information can leave an organization by being copied to removable storage such as USB flash or hard drives and CD/DVD writers. Portable 3TB drivers are now readily available, which means entire databases can be copied to a drive measuring less than 7 x 5 inches in size. With gigabit Ethernet becoming standard, the time required to copy this data is rapidly decreasing. Other common channels include emailing attachments to external email address, uploading files to external email services and to internet websites, and using cloud backup and cloud storage tools. To address these data-leakage channels, products known as data loss prevention system are increasingly being installed in organizations. There are also common physical threats, such as taking printed information from people's desk or from the office printers. Where logical access controls are strong, staff intent on stealing information can take photographs of documents or information on screen with the high-resolution cameras in today's mobile phones. There are also some incidents where the motive may be conscience based as well as having a desire to damage an organization. There have been a number of publicized incidents where insiders have provided list of offshore clients and accounts to country tax authorities (Tagg, 2014).

iii. Non-malicious Threats

Non-malicious threats are some actions takes deliberately by people without intent to damage the organization. Often, the motive is to increase productivity, and the mistakes occur due to a lack training or awareness of policies, procedures, and the risk. There have been many incidents in

which staff loaded internal information onto Internet-based systems, some of which have no access controls. This error makes the information available to anyone who uses the site and is often found and indexed by search engines. For instance one common example is the loss of personally identifiable information when staff copy information to laptop computer or to removable storage devices such as USB drives or CDs/DVDs, which are then lost or stolen (Tagg,G. L. 2014).

One common insider threat that can happen for both malicious and non-malicious reason is to email internal information to their home email address. Once on the staff member's own computer, the information is vulnerable to theft, successful attack on the computer or email account, or recycling of the machine by donating it to charity, or giving it a way to family or friends. There have been many media reports of people finding sensitive information on secondhand computer hard disk. The non-malicious motive for this practice is often to enable the employee to work from home, or the information is needed for a business trip. The malicious motive is to take the information with them, for reasons covered earlier (Tagg, 2014).

2.4Text classification and Machine learning techniques for Information security

According to Frawley, et al, (1991) every twenty months the world data estimated double the amount of the present data. In order to collect, store, update and modify these big data we use a database. Text classification play significant role in discovering the relation among this huge amount of data in the database. A database is a formally and logically associated or integrated one data and structured for easy data and information management as well as retrieve. Data mining process and techniques are applied to extract the hidden association or pattern from the database and summarizing the result as helpful information. The techniques applied in data mining are integrated from various disciplines such as machine learning, image processing pattern recognition, statistics, and database system.

The idea of finding important pattern in data has several names such as data mining, knowledge extraction, data pattern processing, and information discovery. In history *Knowledge discovery in database (KDD)* came into the first use and gain popularity after 1989 KDD workshop. KDD refers to the comprehensive process of discovering important or hidden patterns from data and

data mining specifically refers to the application of algorithms to find these hidden patters. The name KDD emphasize knowledge is the ultimate end product of knowledge spectrum or data-driven discovery. The connection between KDD and machine learning is extracting valuable high level knowledge from low level large data set. Data mining is one level of process in KDD and highly relies on machine learning techniques (Fayyad et. al, 1996).

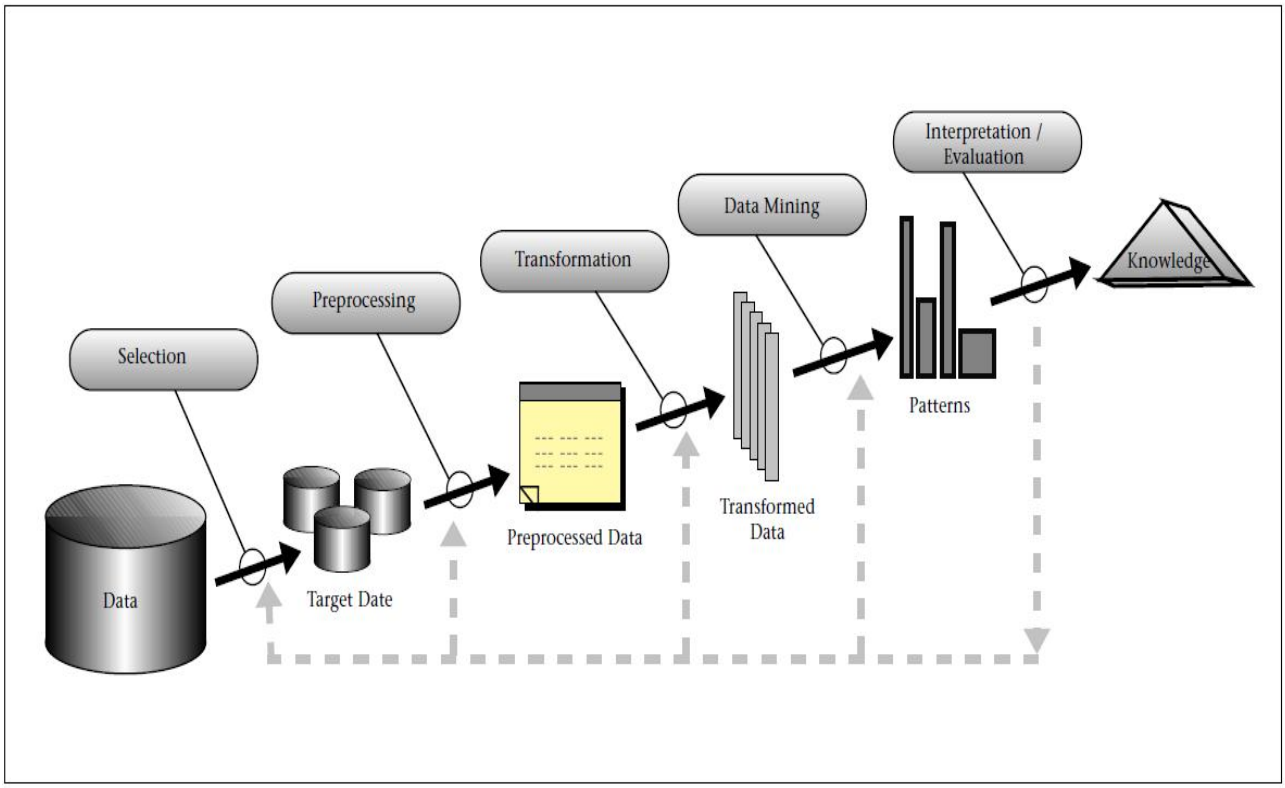

Figure 2.2 The general overview of KDD process adapted from Fayyad. et al, 1996

KDD is very important process to recognize valid, potential, novel and useful patterns in data. Data mining is one of the very important steps within KDD (Fayyad. et al, 1996). Due to the advent of computer and internet email communication create large amount of data that needs a strong security control due to its sensitive nature. There are various data mining techniques proposed on the literature (Verykios et al., 2004;Bertino et al., 2005). Among these KDD process data mining play significant role.

The above figure 2.2 shows the generic KDD process. The *data* are the raw facts and pattern is the result or output of the data after applying a model and an algorithm. *Preprocess* convert the

target data into processed data by applying data preprocessing techniques ranging from data understanding including, data cleaning, data integration up to normalization. *Transformation* change processed email data set into *transformed data*. This transformed data can be for the list of words used by the insider. *Data mining* techniques and algorithms are applied to identify patterns or association among verbal writing or word usage techniques of insiders on the transformed data. Finally *Interpretation and/or evaluation* of the transformed word usage of insiders are changed into high level knowledge based on their interestingness. However KDD use the above sequential steps, data mining process are generalized into three steps for successful insider threat classification. The first step is the previous historical data are processed to train patterns and models to classify the future behavior. These can include frequent word usage among the insider. The second step is the above patterns or models are used to attain a new email communication to determine their likelihood to exhibit the modeled insider word usage behavior. These attained results are used to act upon optimizing a insider threat identification process(Bertino et al., 2005).

Insider threat classification is a core area in information security and can lower the risk the organizations faces. Text classification can be used for insider threat classification for taking preventive and corrective measurement for the survival of the organization against the insiders. Classification techniques we employed can be used for building security intelligence and decision support solution.

2.4.1 Supervised Text classification

Supervised machine learning also known as directed or targeted modeling. Building a model that have the capability of making classifications or estimates the value attribute based on the given data set (classifier) is one of the goal of supervised machine learning. Patterns are recognized through adaptive algorithms and then the computer learns from the data. The algorithm obtains known data and response to the output and the model trained to generate fair or logical classification for the reply of new data. When a machine is uncovered to more data or observation the classification performance of the computer is improved. Supervised models are used to find word usage association among the writers or insiders. This model categorized into two, estimation and classification model.

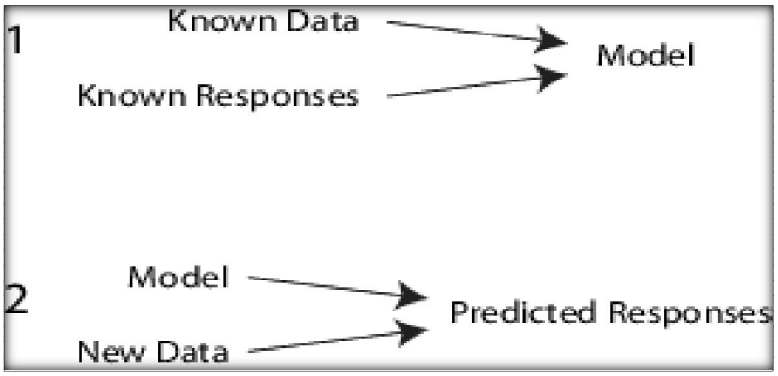

Figure 2.3 generic overview of supervised learning

In Classification models the directed or targeted groups are known from the beginning. It is the task of mapping input values into output class. Classifying these predefined groups to classify the incident is the aim of this model. The new produced model used as a measurement standard for new assigned cases or objects to earlier predefined classes. For example detecting spam email message and categorizing based on their header and content is a good example of classification. The model estimates the tendency of the measurement for each new object. The tendency of measurement signifies the most probable likelihood of occurrence of the directed group.

2.4.2 Text Classification

According to Kamruzzaman. et al, (2010) text classification is the process of classifying documents into correct class label based on their content. The set of labels are defined in advance. This method is an automated task to classify documents on their class label. Nowadays' the number of electronic documents increase from time to time. Such document represent massive amount of information that seeks easy accessibility. Most of the works are done using automated data mining.

Each document in the training corpus is labeled by their correct grouping. These correctly labeled documents features are extracted and given to the classifier algorithms. Based on the mined feature of the document a classifier model is built. This classifier model assists us for classify the correct class label of the new input data. The document we want to know the class, its feature extracted in the same manner as the training data through feature extractor.

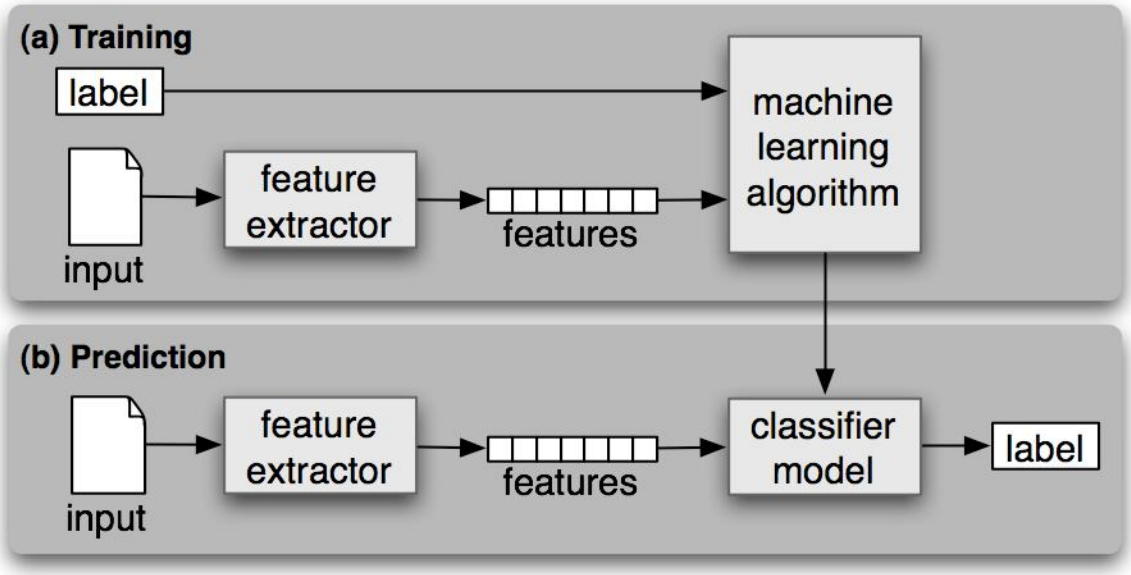

Figure 2.4 supervised document classification diagram

According to Radovanović,& Ivanović, (2008) movie review documents are classified based on their sentiment content whether it is negative or positive expressed opinion. The training input used for the classification was selected manually from positive and negative review. The feature extractor or sentiment analyzer select the sentiment that makes the opinion negative or positive. The words are identified as which have positive and negative sentiment. Based on these they built a classifier model that can able to classify the new movie review content sentimentally either negative or positive. The newly given movie opinion features are extracted and given to the classifier model and then the classifier model based on the previous learned sentiment words which have positive or negative sentiment the new document classified into one of the category of the review that have positive or negative opinion.

In estimation model are similar to the above model but what makes them unique is that, estimation model uses observed input attributes to classify the value of continuous field.

The following classification techniques such as decision trees, decision rules, logistic regression, neural networks, support vector machine, and Naïve Bayes are used for both estimation and classification. These classification techniques also known as classifier are a methodical approach to build classification models from an input objects or data set.

Support Vector Machine (SVM): is one of the algorithms that are used as the representative for the development of machine learning algorithm. It is the set of supervised machine learning methods for high dimensional spaces, target categories or nonlinear data set that can easily be separated by linear function or hyper plane or it is a discriminative classifier that can be applied for regression and classification. SVM is base at statistical learning theory and utilize a subset of training data to make the decision function the so called support vectors. The support vector is those data points that the hyper plane pushes up against. This decision is based on the decision plane that separate or classify objects having different class association. SVM is exercised for *classification, outliers' detection and regression.* The support vectors are the data points that are closest to the separating hyper plane; these points are on the boundary of the slab. An SVM classifies data by finding the best hyper plane that separates all data points of one class from those of the other class. The best hyper plane for an SVM means the one with the largest margin between the two classes. Margin means the maximal width of the slab parallel to the hyper plane that has no interior data points (Platt, 1998).

Sequential minimal optimization: Sequential minimal optimization (SMO) is an algorithm for efficiently solving the optimization problem which arises during the training of support vector machines (SVM), faster and has better scaling properties for difficult SVM problems than the standard SVM training algorithm. It was invented by Platt in 1998 at Microsoft Research (Platt, 1998). SMO is an iterative algorithm for solving the optimization problem by breaking the problem into a series of smallest possible sub-problems, which are then solved analytic quadratic programming rather than numerical quadratic programming as inner loop as in SVM. Because of the linear equality constraint involving the Lagrange multipliers $\alpha_{i\ \text{the}}$ smallest possible problem involves two such multipliers. Then, for any two multipliers α1and α_2 the constraints are reduced to:

$$0 \leq \alpha_1, \alpha_2 \leq C \quad (2.1)$$

$$y_1\alpha_1 + y_2\alpha_2 = k \quad (2.2)$$

Where C is an SVM hyper parameter and K(xi, xj) is the kernel function, both supplied by the user; and the variables are Lagrange multipliers and this reduced problem can be solved analytically.

The algorithm proceeds as follows:

1. Find a Lagrange multiplier that violates the Karush–Kuhn–Tucker (KKT) conditions for the optimization problem.
2. Pick a second multiplier and optimize the pair.
3. Repeat steps 1 and 2 until convergence.

When all the Lagrange multipliers satisfy the KKT conditions (within a user-defined tolerance), the problem has been solved. Although this algorithm is guaranteed to converge, heuristics are used to choose the pair of multipliers so as to accelerate the rate of convergence.

In general, According to Girma A,(2012) point out that the following are the general pseudo code for SVM algorithms are depicted as follows.

Introduce positive Lagrange multipliers, one for each of the inequality constraints. This gives Lagrangian:

$Lp= ½ ||W||^2 – \sum_{i=1}^{n} \alpha i yi(xi.w - b) + \sum_{i=1}^{n} \alpha i$…… (2.3)

Minimize Lp with respect to w, b. This is a convex quadratic programming problem.

In the solution, those points for which α i are called “support vectors” > 0

Even though the maximum margin allows the SVM to select among multiple candidate hyper planes, for many datasets, the SVM may not be able to find any separating hyper plane at all because the data contains misclassified instances. The problem can be addressed by using a soft margin that accepts some misclassifications of the training instances (Veropoulos et al. 1999). This can be done by introducing positive slack variable Si, i =1,..., N in the constraints, which then become:

W.Xi –b≥ +1- ϵ f or yi =+1……..……..……. (2.4)

W.Xi –b≥ -1 + ϵ f or yi = -1……..……..……... (2.5)

ε ≥ 0, Thus, for an error to occur the corresponding $\in i$ must exceed unity, so $\sum \mu_i \varepsilon_i$is an upper bound on the number of training errors. In this case the Lagrangian is:

$lp=1/2 ||w||^2 + C\sum_i \varepsilon i - \sum_i \alpha i\{yi(xi.w - b) -1 +e1\} - \sum_i \mu_i \varepsilon_i$…….. (2.6)

Where the μi are the Lagrange multipliers introduced to enforce positivity of the ε_i. Nevertheless, most real-world problems involve non- separable data for which no hyper plane exists that successfully separates the positive from negative instances in the training set. One solution to the inseparability problem is to map the data onto a higher- dimensional space and define a separating hyper plane there. This higher-dimensional space is called the transformed feature space, as opposed to the input space occupied by the training instances.

In this research work again the researcher use SMO algorithms to test the performance of the classification technique of SVM against the Enron higher official level of insiderness.

Naïve Bayes Classifier: it is based on supervised learning and also known as probabilistic machine learning classifier method. This method is based on the Bayes theorem and "naïve" postulate or assumption of independence between every pair of attribute. The aim of this algorithm is to classify the new cases classes based on the given training data. This classifier can hold an arbitrary number of independent variable whether continuous or categorical.

The Naive Bayes Classifier technique is based on the so-called Bayesian theorem its and methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of independence between every pair of features. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods. For instance in our case there is a training set which contain list of words indicating positive or negative emotion of the person. As shown on the next diagram the to the right bold line shows list of positive words and to the left of bold line shows list of negative words.

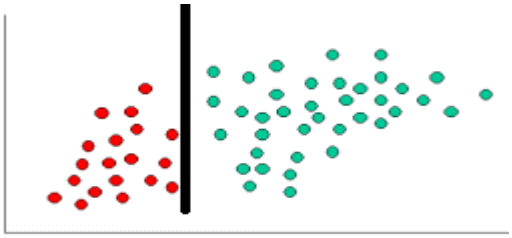

Figure 2.5 Naïve Bayes prior probability

To demonstrate the concept of Naïve Bayes Classification, consider the example displayed in the illustration above figure 2.5. As indicated, the objects can be classified as either RIGHT/POSITIVE or LEFT/NEGATIVE. Our task is to classify new word as they arrive, i.e., decide to which class label they belong either positive or negative, based on the currently exiting words.

Since there are twice as many POSITIVE words as NEGATIVE, it is reasonable to believe that a new case (which hasn't been observed yet) is twice as likely to have membership POSITIVE rather than NEGATIVE. In the Bayesian analysis, this belief is known as the prior probability. Prior probabilities are based on previous experience; in this case the percentage of POSITIVE and NEGATIVE words, and often used to classify outcomes before they actually happen.

$$\text{Prior probability for RIGHT/POSITIVE} = \frac{\text{No of POSITIVE Words}}{\text{Total number of words}} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2.7)$$

$$\text{Prior probability for LEFT/NEGATIVE} = \frac{\text{No of NEGATIVE Words}}{\text{Total number of words}} \quad \ldots\ldots\ldots\ldots\ldots \quad (2.8)$$

For instance there is a total of 60 objects, 40 of which are POSITIVE and 20 NEGATIVE, our prior probabilities for class membership are:

Priority for positive= $\frac{40}{60}=\frac{2}{3}$ and Priority for negative= $\frac{20}{60}=\frac{1}{3}$

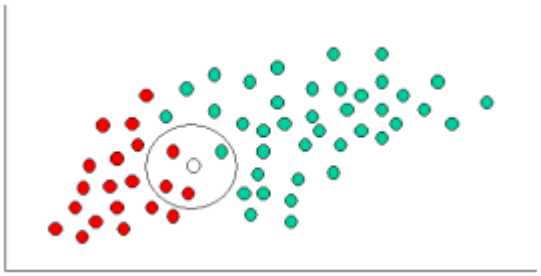

Figure 2.6 Naïve Bayes Likelihood Probability

Since the words are well clustered, it is reasonable to assume that the more POSITIVE (or NEGATIVE) words in the vicinity of X, the more likely that the new cases belong to that particular class. To measure this likelihood, we draw a circle around X which encompasses a

number (to be chosen a priori) of points irrespective of their class labels. Then we calculate the number of points in the circle belonging to each class label. From this we calculate the likelihood positive and negative:

Likelihood of X given RIGHT/POSITIVE=$\frac{NoofPOSITIVEWordsinvicinityofX}{TotalnumberofPOSITIVEwords}$……. (2.9)

Likelihood of X given LEFT /NEGATIVE =$\frac{NoofNEGATIVEWordsinvicinityofX}{TotalnumberofNEGATIVEwords}$……. (2.10)

Likelihood of X given RIGHT/POSITIVE=$\frac{1}{40}$

Likelihood of X given LEFT/NEGATIVE=$\frac{3}{20}$

Although the prior probabilities indicate that X may belong to POSITVE (given that there are twice as many POSITIVE compared to NEGATIVE) the likelihood indicates otherwise; that the class membership of X is NEGATIVE (given that there are more NEGATIVE words in the vicinity of X than POSITIVE). In the Bayesian analysis, the final classification is produced by combining both sources of information, i.e., the prior and the likelihood, to form a posterior probability using the so-called Bayes' rule. Equation (1) and (2) are combined with (3) and (4).

Posterior probability X being POSITIVE=prior probability of POSITIVE X Likelihood of given POSITIVE

Posterior probability X being POSITIVE=4/6 X1/40=1/60

Posterior probability X being NEGATIVE= prior probability of NEGATIVE X Likelihood of given NEGATIVE

Posterior probability X being NEGATIVE=2/6X3/20=1/20

Finally, we classify X as NEGATIVE since its class membership achieves the largest posterior probability. This example is intuitive to understand how the naïve Bayes is applied on this research. But Naïve Bayes classifiers can handle an arbitrary number of independent variables whether continuous or categorical.

Text classification is used to give helpful information from the big amount of data. It is one of the important research issues in the field of data mining. Text classification is the task of classifying words by their composition: that is, by the letters of which they are comprised. Text classifiers often don't use any kind of deep representation about language: often a word is represented as a bag of letters. Consider a word D, whose class is given by C. In the case of emotional words there are two classes C = P (positive) and C = N (negative). We classify D as the class which has the highest posterior probability P(C|D), which can be re-expressed using Bayes' Theorem:

$$P(C|D) = \frac{P(D|C)P(C)}{P(D)} \propto P(D|C)P(C) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.11)$$

In this research work the researcher naïve Bayes classifier to test the classification capability of insider threat by using equal number of positive and negative words from the employee email content to avoid the superiority or domination of posterior probability of one class over the other.

2.5 Potential Application of text mining

The application of data mining spans various industries. Telecommunications and insurance industries make use of data mining techniques to detect fraudulent activities. In medicine, data mining is used to classification the effectiveness of surgical procedures and medical tests.

Companies in the financial sector use data mining to determine market and industry characteristics as well as to classify individual company and stock performance, classify which customers buy new policies; Identify behavior patterns of risky customers (Srivatsa, 2013).

According to Crows, T, (1999) the two important keys for successful text mining are: to come up with a precise formulation of the problem you are trying to solve and using the right data. After getting the right data we need data transformation and combining it in significant way. The more the model builder can play with the data, build models, evaluate results, and work with the data some more (in a given unit of time), the better the resulting model will be.

Consequently, the degree to which a data mining tool supports this interactive data exploration is more important than the algorithms it uses. Ideally, the data exploration tools (graphics/visualization, query/OLAP) are well-integrated with the analytics or algorithms that build the models.

2.6 Psychology of Language

Natural language is used as money for most human social day to day activities and processes. We use words to communicate our emotions and opinions, feelings, to tell stories, and to understand the world. They disclose vast information about our social relation and personality. According to Macmillan dictionary psychology is "*the ability to understand the way that people think, that helps you to make people do what you want*" These definition affirm to us human have the ability to understand how other people think or perform and they can communicate with one another through their common understanding. These regular understanding makes people to share their knowledge and carry out what they want for each other. One of the mechanism human beings communicate nowadays is by writing an electronic mail. In order to write email we need to have words that express our intention. These words we use emotionally either negative or positive are related with our mode, emotion state and how we feel. Emotion is a way of expressing oneself in life or the quality of how one relates to life.

According to Scherer, (2005) in their work define emotion defined as an event of interconnected, synchronized alters in the states of all or most of the five organismic subsystems in response to the evaluation of an external or internal stimulus episodes as pertinent to major concerns of the organism. From this definition human emotion can be changed through internal or external stimulus. A variety of circumstances on the workplace can influence employees' emotion in positive ways, such as seeking to get money, authority, and getting promotion. On the contrary negative emotions such as being demoted or passed promotion can create envy. These emotions both positive and negative are important part of organizational life. Emotion influence and influenced by functioning in an organization.

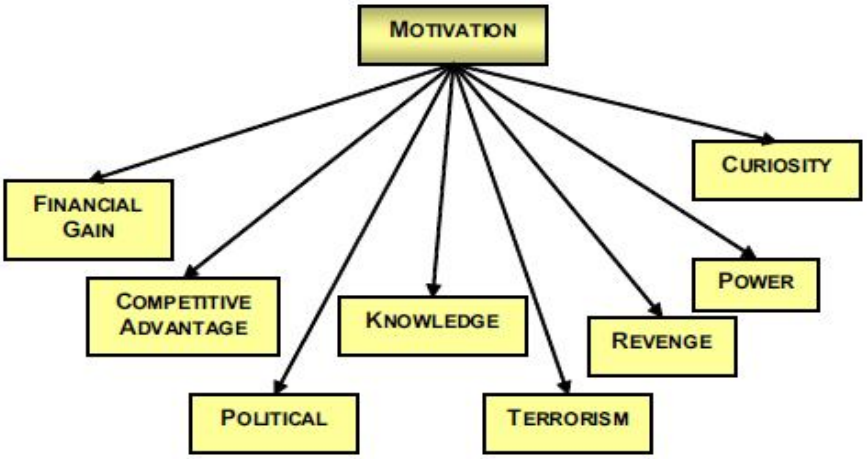

Figure 2.7 The components of motivation adapted from (Scherer, 2005)

Research in organizational psychology have shown that positive emotions are associated with increased creativity, cognitive flexibility, labor productivity and professional satisfactions, the availability of communication and negotiation skills, etc. Positive emotions, by organizing and constructive effects, optimize the quality of work. Employees in situations experiencing pleasant emotions tend to set higher goals and to engage in constructive activity and generative ways. With regard to negative emotions, most experimental studies show harmful effects, such as restricting the repertoire of thinking, the tendency to process negative information and maintaining the dysfunctional cognitive schemes, work dissatisfaction, low emotional engagement tend to leave the organization (employees wish to change their work), etc. (Maria Andries, 2011).

For the past 15 years many of duplicated publications verified that being either emotionally positive or negative during writing can influence the health condition of the patient which is related with immune function, blood pressure, stress hormones, and a host of social, academic, and cognitive variables. These effects manifested across cultures, ages, and diverse socio demographic variables (Pennebaker & Graybeal, 2001; Smyth, 1998).

Most language narrative researchers argue that language by definition is contextual and phrases or the whole texts that taken into consideration in relation to the aim of the speaker and audience (Pennebaker, & Mehl, 2003). Due to the complexity of this nature of communication, this study explores the word usage distribution that lead to positive or negative emotion.

According to Pennebaker, et al.(1997); Pennebaker & Mehl (2003) when we use more positive words than negative words our health getting better and we have a better life. From these finding we can assert that our words we write or speak are determinate to our health. The words people use is associated with the social, mental, mental health and physical state.

With the advent of modern computers and software that can analyze texts enables researchers confidently and quickly study individual word use and their linguistic styles (how it is being said). The words we use in our daily lives can express significant aspect of our social and psychological environments. These words we use related with our personality, social, situational and psychological as well as psychosocial involvement. The way in which these word are used in carry out relevant information about their speaker, audience and the situation they are in. The word usage of individual person is also associated with their motives, social status, sex and age. The word utterances help us to sense the speakers' deeper motives such as open to new experience, emotion, sadness, fear, happiness etc. The present language analysis categorized into three methodologies (Pennebaker, et al.1997). The *judge-based thematic content analysis* focused on identifying important text samples from the developed system. The second methodology is *thematic content* gives an attention on studying phenomena such as cognitive complexity, explanatory styles and level of thinking. The third new methodical *word pattern analysis* approach focused on the word pattern coverage across the whole content of the document (Pennebaker, J. W., Mehl, 2003).

As Krauss, & Chiu,(1998) describe in the hand book of social psychology language is every part of our social life. It is the primary means to transmit cultural knowledge and we gain access to the contents of others' minds. Language is inflicted in most of the situation that happen at the nucleus of social psychology such as attitude change, social insight, personal identity, social interaction, intergroup bias and stereotyping, attribution, and so on. Furthermore for social and behavioral psychologist language ideal means of medium by which subjects' responses are draw out. In addition to these for social psychologist language plays significant role in stimulus and response. As language interpenetrate every part of social life, the elements of social life comprise an inherent part of the way language is used. Linguists look upon language as an abstract structure that live independently of specific instances of usage, but any communicative exchange is situated in social circumstance that limit the linguistic forms participants use. How these

participants define the social situation, their perception of what others know, think and believe, and the claims they make about their own and others, identities will affect the form and content of their act of speaking and writing.

2.6.1 Word use and Individual Difference

That people differ from one another is evident. Different theoretical methodological techniques and approaches are able to recognize how peoples prefer or select words in writing their feelings and deeper motives. To define what psychometric is the first and foremost step before we are dealing with the link between word use and individual difference. Psychometrics deals with the frequency and word usage pattern across the time and the context. The present studies on psychometrics of word use assert that individuals word choice is enough and permanent over a period of time and stable across the topic to use language as an individual difference measurement for basic grammatical categories and psychologically based language (Pennebaker, 2003).

Demographic variables also have their own influence and makes difference on individuals language use such as age and sex. Women's and men's language are differ in a variety of scope. But there is no clear picture that has yet grownshows exactly their difference. Women's lack of authoritative power in society leads using less assertive speech that express, high emotional in language use, high degree of politeness, less searing, more frequent tags. In contrary men's are more precise, less emotional in their language use and directives. The other demographic factor is age. When age is changed over through the life span and language use also changed. When age became increasing peoples use more positive emotion words than negative words and first person singular pronoun "I" for self references, more future tense and less past tense verbs. The age of the individual is highly positively correlated with increase long words, causation words and insight word usage. In addition to these stereotype on aging can serve as main linguistic age marker (Pennebaker, J. W., Mehl, 2003).

2.6.2 Words reflect the situation and social process

The way how we say depends on the situation we live in. young children change the way they interact depending on the context. The voice characteristics and other non-verbal sign shift depend on the formality of the environment or situation, the nature of audience and the degree of

the speakers' connection with the actors. Social situations dimensions are related or associated with the language and word usage. These taxonomic dimensional structure helps to identify how and when the language is shifted and the psychological dimension of situation associated with language and communication including cooperatives and involvement. Formal and informal settings of the situation also determine the language use. Individuals based on their situation they make code switching, which is the way we change language, accent and dialect occur often among the group or speakers (Pennebaker, 2003).

Campbell in his book the psychology of computer criminals stated that although situational factors can account for some of the behaviors of some computer criminals, one must not discount the impact of personality factors on their illegal activities. National institute of justice define a computer criminal as any individual who uses computer or network technology to plan or perpetrate a violation of the law. Attitudes and behaviors are often the product of both situational influences and individual personality traits. It should be noted that there are few empirical studies that engage in a scientific examination of the personality traits of computer criminals, so without concrete evidence, the unreliable claims regarding pathological traits of cybercriminals should be interpreted with caution. In addition, simply having traits that are consistent with psychological disorder does not mean that one actually has disorder (Campbell, & Kennedy, 2009).

2.6.3 Psychology of Email writer

Nowadays' Email is an everywhere communication tool and comprise major interaction among the internet user and seems to be a normal habit for office worker. Understanding the psychology of email writer verbal or written communication is an inevitable task at this age of information overload. Knowing how the writer use words and the type of words help us to classify the personality of the writer. According to Shen, Brdiczka, & Liu,(2013); Brown, Watkins, & Greitzer, (2013) it is possible to infer the personality of the email writer from the content of the email, which helps us for better personalization, recommendation, advertising, for better insider threat mitigation and to identify individuals having the motivation and capability to damage the work environment, harm co-workers, or commit suicide in the workplace. Both researchers use the big five factor personality model to classify the personality of the employee based on their email content.

2.6.4 Big five factor model of personality

Personality has been conceptualized from a variety of theoretical perspectives, and at various levels of abstraction or breadth. Each of these levels has made unique contributions to our understanding of individual differences in behavior and experience. However, the number of personality traits, and scales designed to measure them, escalated without an end in sight. What personality psychology needed was a descriptive model, or taxonomy, of its subject matter. The Big Five personality dimensions do not represent a particular theoretical perspective but were derived from analyses of the natural-language terms people use to describe themselves (John, & Srivastava,1999).

The big five-factor model of personality has broad dimension and a hierarchical sense of personality traits that can provide significant personality classifier in terms of five basic dimensions: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to Experience Christopheret al (2013).

Neuroticism

Freud originally used the term “neurosis to describe a condition marked by mental distress, emotional suffering, and an inability to cope effectively with the normal demands of life. He suggested that everyone shows some signs of neurosis, but we differ in our degree of suffering and out specific symptoms of distress. Today neuroticism refers to the tendency to experience negative feelings”.

Neuroticism is a dimension of normal personality indicating the general tendency to experience negative effects such as fear, sadness, embarrassment, anger, guilt, disgust, vulnerability and depression. High scorers may be at risk of some kinds of psychiatric problems. A high Neuroticism score indicates that a person is prone to having irrational ideas, being less able to control impulses, and coping poorly with stress. A low Neuroticism score is indicative of emotional stability. These people are usually calm, even-tempered, relaxed and able to face stressful situations without becoming upset found that Neuroticism is a classifier of performance in various occupations (Rothmann, S., & Coetzer, 2003). Anxiety, anger, depression, self-consciousness, immoderation and vulnerability are the facets of neuroticism.

Extraversion

Extraversion is manifested by definite commitment with the external world. These include traits such as enjoy being with people, sociability, full of energy, assertiveness, enthusiastic, action-oriented, activity and talkativeness. They prefer to say "great" or "Let's go". Introverts lack energy, quite, deliberate, and disengaged from social work. This not having social engagement should not be interpreted as shyness or depression. Extraversion is characterized by positive feelings and experiences and is therefore seen as a positive effect (Rothmann,& Coetzer, 2003). Friendliness, gregariousness, assertiveness, activity level, excitement-seeking, and cheerfulness are the facets of extraversion.

Openness to Experience

Openness to Experience includes active imagination, aesthetic sensitivity, and attentiveness to inner feelings, a preference for variety, intellectual curiosity and independence of judgement. People scoring low on Openness tend to be conventional in behavior and conservative in outlook. They prefer the familiar to the novel, and their emotional responses are somewhat muted. People scoring high on Openness tend to be unconventional, willing to question authority and prepared to entertain new ethical, social and political ideas. Open individuals are curious about both inner and outer worlds, and their lives are experientially richer. They are willing to entertain novel ideas and unconventional values, and they experience both positive and negative emotions more keenly than do closed individuals (Rothmann, & Coetzer, 2003). Imagination, artistic interests, emotionality adventuress, intellect, and liberalism are the facets of this dimension.

Agreeableness

This dimension reflects personal difference in teamwork and social harmony. An agreeable individual is basically kind, unselfish, generous, honesty, trustworthy, decent helpful, and in return believes that others will be equally helpful. These people had optimistic view. In contrary disagreeable individual put self-interest above obtaining along with others including such as selfish, unconvinced of others' intentions, and competitive rather than co-operative (Rothmann, & Coetzer, 2003). The facets are trust, morality, altruism, cooperation, modesty, and sympathy.

Conscientiousness

Conscientiousness focused on the way in which we plan, organize, doing tasks, control, regulate, and direct our impulses that need a snap decision. The conscientious individual is purposeful, strong-willed and determined. Conscientiousness is characterized by hardworking, persistent, responsible, careful, orderliness, plan full and organized. On the negative side, high Conscientiousness may lead to annoying meticulousness, compulsive neatness or workaholic behavior. In contrast to conscientiousness unconscientious person have criticized for their lack of moral principles, reliability, ambition and they experience ephemeral pleasure and they are also less exacting in applying the manifestation of conscientiousness (Rothmann,& Coetzer, 2003). Facets of this dimension are self-efficiency, orderliness, dutifulness, achievement-striving, self-discipline, and cautiousness.

2.7 Evaluation Metrics

In this sub-chapter the researcher discuss the basic evaluation metrics or criteria to evaluate the performance of classification algorithm.

The Kappa statistic (or value) is a metric that compares an observed accuracy with expectedaccuracy (random chance). The kappa statistic is used not only to evaluate a single classifier, but also to evaluate classifiers amongst themselves. In addition, it takes into account random chance (agreement with a random classifier), which generally means it is less misleading than simply using accuracy as a metric.

$$kappa = \frac{(observed\_accuracy) - (Expected\_accuracy)}{1 - (Expected\_accuracy)} \quad \text{..... (2.12)}$$

According to (Landis, & Koch, 1977) in order to maintain consistent nomenclature when describing the relative strength of agreement associated with kappa statistics, the following labels will be assigned to the corresponding range of kappa.

Table 2.1 Kappa statistics (Landis, & Koch, 1977)

Kappa Statistics	Strength of Agreement
<0.00	Poor
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost Perfect

Let's denote the true value of interest as θ and the value estimated using some algorithm as $\hat{\theta}$. The Correlation tells us how much θ and $\hat{\theta}$ are related. It gives values between −1 and 1, where 0 is no relation, 1 is very strong, linear relation and −1 is an inverse linear relation (i.e. bigger values of θ indicate smaller values of $\hat{\theta}$, or vice versa).

Mean Absolute Error (MAE)

The MAE measures the average magnitude of the errors in a set of interest, without considering their direction. The MAE is the average over the verification sample of the absolute values of the differences between interest and the corresponding estimated value. The MAE is a linear score which means that all the individual differences are weighted equally in the average. If the MAE value is 0 means there is no error; 1 means strong error. When the MAE value increase from 0 to 1, the average magnitude of the error in a set of interest also increases.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{\theta}_i - \theta_i\right| \quad \text{..... (2.13)}$$

Where *N* is the number of measurements.

Root Mean Square Error (RMSE)

The RMSE is a quadratic scoring rule which measures the average magnitude of the error. The difference between interest and corresponding observed values are each squared and then averaged over the sample. Finally, the square root of the average is taken. Since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors. This means the RMSE is most useful when large errors are particularly undesirable.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\theta}_i - \theta\right)^2} \quad \text{..... (2.14)}$$

Relative absolute error (RAE)

This RAE equation tells you how far off you were in comparison to the overall measurement. A low relative error is, of course, desirable.

$$RAE = \frac{\sum_{i=1}^{N}\left|\hat{\theta}_i - \theta_i\right|}{\sum_{i=1}^{N}\left|\overline{\theta} - \theta_i\right|} \quad \text{..... (2.15)}$$

Where $\overline{\theta}$ is a mean value of θ.

Root Relative Squared Error (RRSE)

The root relative squared error is relative to what it would have been if a simple classifier had been used. More specifically, this simple classifier is just the average of the actual values. Thus, the relative squared error takes the total squared error and normalizes it by dividing by the total squared error of the simple classifier. By taking the square root of the relative squared error one reduces the error to the same dimensions as the quantity being classified.

$$RRSE = \sqrt{\frac{\sum_{i=1}^{N}\left(\hat{\theta}_i - \theta_i\right)^2}{\sum_{i=1}^{N}\left(\overline{\theta} - \theta_i\right)^2}} \quad \text{..... (2.16)}$$

Precision:- is a measure of result relevancy. Precision (P) is defined as the number of true positives (Tp) over the number of true positives (Tp) plus the number of false positives (Fp).

$$P = \frac{T_p}{T_p + F_p} \quad \text{..... (2.17)}$$

Statistical Precision

Cohen (1988, p. 6) defines the statistical precision of a sample statistic as "the closeness with which it can be expected to approximate the relevant population value. It is necessarily an estimated value in practice, since the population value is generally unknown" (Cohen, 1988, p. 6). This precision is usually estimated using a standard error, that is, the amount of chance fluctuation (or lack of precision) we can expect in sample estimates. We can use the standard error as an estimate of the precision of a statistic in two ways: descriptively or inferentially (Thompson, 2006, pp. 154-155).

Descriptively, when precision is estimated using a standard error, it is thought of as the amount of fluctuation from the population parameter that we can expect by chance alone in sample estimates. For example, for a sample mean (M), we can calculate the standard error of the mean (SE_M), which provides an estimate of how much fluctuation from the population parameter that we can expect in sample estimates of M. Since standard errors are distributed normally, we can expect sample means to vary by chance ±1 SE_M 68% of the time, ±2 SE_M 95% of the time, and ±3 SE_M 98% of the time (for a review of how these percentages work, see Brown, 1988, pp. 80-85; or 2005, pp. 116-123). For example, if the mean for a sample turned out to be 78 with a conveniently round SE_M of 2, we would expect such sample means to vary by chance between 76 and 80 (68% of the time), between 74 and 82 (95% of the time), and between 72 and 84 (98% of the time). The following equation can be used to calculate the standard error of the mean (SE_M):

$$SE_M = \sqrt{\frac{S^2}{n}}$$

Where: SE_M = standard error of the mean S = standard deviation n = group size

Inferentially, the standard error is also commonly used in estimating the statistical significance of differences between or among parameter estimates. For example, a t-test can be used to estimate the probability that an observed difference between two means (say between treatment-group and control-group means) is statistically significant (i.e., that the difference is due to other than

chance factors. One formula for the t-test (where the two samples are independent and are the same size) is as follows:

$$t = \frac{M_T - M_C}{\sqrt{\frac{S_T^2}{n_T} + \frac{S_C^2}{n_C}}}$$

Notice that the numerator represents the difference between the treatment-group mean and the control-group mean (M_T - M_C), and that the denominator contains the two standard errors for the treatment and control groups. Thus the t-test is simply a ratio of the mean difference to the square root of the sum of their standard errors, or put another way, the t-test is the mean difference in relation to the precision with which the two means were estimate than chance factors.

Though several factors can affect the precision of a parameter estimate, sample size is always a factor. As Cohen (1988, p. 6), put it, "depending upon the statistics in question, and the specific statistical model on which the test is based, reliability [i.e., precision] may or may not be directly dependent upon the unit of measurement, the population value, and the shape of the population distribution. However, it is always dependent upon the size of the sample"

Look at any of the equations above for various permutations of standard error, and notice that all of them have the *n*-size in the denominator. Hence, with all other factors held steady, as sample size increases, the standard error decreases, or gets more precise. Put another way, as the sample size increases so does the statistical precision of the parameter estimate. This has ramifications for both the descriptive and inferential uses of the standard error. Descriptively, as sample size goes up, parameter estimates become more precise. Inferentially, as sample sizes go up, parameter estimates are more precise, so differences between or among parameter estimates can be smaller and still turn out to be statistically significant.

Recall:- is a measure of how many truly relevant results are returned. Recall (R) is defined as the number of true positives (Tp) over the number of true positives (Tp) plus the number of false negatives (Fn).

$$R = \frac{T_p}{T_p + F_n} \quad \text{..... (2.18)}$$

The F-measure can be interpreted as a weighted average of the precision and recall, where an F-measure reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F-measure are equal. The formula for the F-measure is:

$$F - measure = 2 * \frac{(\Pr ecision * \Re call)}{(\Pr ecision + \Re call)} \quad \text{..... (2.19)}$$

Matthews correlation coefficient (MCC):- The MCC is in essence a correlation coefficient between the observed and predicted binary classifications; it returns a value between −1 and +1. A coefficient of +1 represents a perfect classification, 0 no better than random classification and −1 indicates total disagreement between classification and observation. While there is no perfect way of describing the confusion matrix of true and false positives and negatives by a single number, the Matthews correlation coefficient is generally regarded as being one of the best such measures.

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} \quad \text{..... (2.20)}$$

Receiver Operation characteristics (ROC) Area:- are a fairly standard way to evaluate model fit with binary outcomes. ROC curves only give us an idea of how the classifiers are performing in general. The curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. A stream of classification that as probabilities range between 0 and 1, 0 meaning that poor classification and 1 mean strong classification.

Confusion Matrix

The confusion matrix is more commonly named *contingency table*. In our case we have two classes, and therefore a 2x2 confusion matrix. The number of correctly classified instances is the sum of diagonals in the matrix; all others are incorrectly classified.

TP = true positives: number of words classified positive that are actually positive

FP = false positives: number of words classified positive that are actually negative

TN = true negatives: number of words classified negative that are actually negative

FN = false negatives: number of words classified negative that are actually positive

Based on the understanding and the knowledge of the researcher, this research problems has not yet attempted before and novel because different methodology and techniques are used for training and testing the classification model. The variables used such as negative and positive emotional words to classification insider threat from email communication word distribution using supervised machine learning approach algorithms are not used yet by other researchers. In addition to this weka latest version are used for classification of the text and taking this fact in mind, this research has initiated and conducted, and would be expected to contribute a lot to managers of the organization to make prevention control among their employee harmful behavior.

2.8. Review of Related Works

There have been done various attempts to deliver a robust insider threat classification and detection system. Insider threat classification and detection exploit ideas from intrusion detection or outsider detection (Parveen, et al., 2011). In this portion recent and related peer reviewed works with insider threat classification and detection would be discussed with their limitation and strength. Finally the researcher shows how this study differs from the work of others.

2.8.1 Classifying or Detecting Insider Threat

Nowadays' organizations are challenged with their employees activity. According to Christopher et al. (2013) many organization are challenged by the mounting risk from malicious or careless insiders and these may lead organizations open to liability, hurt morale and in extreme cases including the physical harm to the employees. The objective of this research was to investigate the potential for active monitoring of electronic communications as a method that may identify problems early, allowing for proactive mitigation through coaching, assistance program and where warranted, termination. To test this theoretical objective the researchers develop a prototype system that utilize a large body of real text samples and augmented with positive controls. The very fundamental argument to do this research was that word use (even if no single word can provide conclusive evidence on the insider threat) frequency (cumulative text

collection) reveals an individual's basic personality and that those personality factors that are used to deduce psychological gauge of insider misuse or abuse. The big five personality trait factor such as agreeableness, conscientiousness, neuroticism, extraversion, and openness are used to gauge the level of risk organizations faces by their employee. The text source include email messages, chat session and social media post such as face book and twitter. The authors of this article select the three personality trait factor from the five; these are agreeableness, conscientiousness, and neuroticism. The Linguistics word inquire count developed by Pennebaker are used with a special purpose text analysis software developed using C#. The data set used for this research is taken from former Enron company 150 senior level executives email text archive when Enron executives were actively committing a variety of financial crime. Three positive controls were added to the dataset. These controls were individuals who would represent a risk to any organization. After the message body is extracted from the email text file it was scored against 27 dictionaries that represent significant psychological factors. The analysis approach they follow to reach the stated objective they combine the frequency scores from a number of word categories into one of the three personality factor score of agreeableness, conscientiousness, and neuroticism. Only those word categories they give in a statistically important correlation to one of the three personality factor were used. Means and standard deviations were calculated first each word category using the function:

$$\sigma = \sqrt{\frac{\sum(x-\bar{x})}{(n-1)}}^{2}$$

Each email sender's word category Z-scores were then calculated the function:

$$z = \frac{x-\mu}{\sigma}.$$

The authors use Chebyshev's inequality to identify individual outlier in the distributions, the inequality which states that the probability of a random variable (I) exceeding any real value T>0 is:

$$p\left(\left|\frac{\tau-\mu_{\tau}}{\alpha_{\tau}}\right| \geq T\right) \leq \frac{1}{T^{2}}$$

The authors in order to use the Chebyshev theorem, they make an assumption that the population variance is finite and not zero. In addition to this they were confident in their assumption and the distribution is unimodal, which allows the use of a refinement proven by Vyochanskij and Petunin that allows Chebyshev's original inequality to be tightened by multiplying the right hand side by (4/9). Using the standard score in the expression, this yields the following variant of Chebyshev's inequality:

$$p\left(\left|Z_{\tau}\right| \geq T\right) \leq \frac{4}{9T^{2}}$$

According to the result attain from this Chebyshev's inequality the system correctly identified the positive controls as individual who may pose a risk to an organization. After extensive experiment they found there is a correlation between word use and behavior. In addition to this there is a slight but measurable differences in the frequency of common words found in email communication they may provide hint about the potential insider threat risk. As the result shows Agreeableness is positively correlated with eleven word categories. Individuals who score high in the agreeableness category are more willing to defer to others, comply with policy and avoid conflict. These individuals forgiving attitudes use inoffensive language and may be viewed by others as a pushover. However, for insider threat classification, the concern is not with individuals who are highly agreeable; rather than the goal is to identify those individuals who have unusually low scores for agreeableness. Individuals who score low in agreeableness have an increased frequency of conflict, and may be likely to engage in physical and relational aggression as well as likely to exhibit narcissist personalities.

On the other hand Conscientiousness is negatively correlated with twelve word categories. A person who produces higher frequencies of words in these categories may lack conscientiousness. Research on the trait of conscientiousness has revealed that highly conscientious individuals have a strong sense of purpose, high aspiration levels, tend to be well organized, diligent, through achievement oriented, reliable and self-determined; they may aspire to leadership, make long-term plan and have an organized support network. Individual low scores have a harder time coping with stress (Christopher, et al.2013).

The third personality factor was neuroticism which is positively correlated with ten word categories, therefore individuals whose messages contain a higher than normal percentage of words from these categories may be more likely to demonstrate neurotic personality traits. Individuals who score high for neuroticism are prone to depression, sadness, hopelessness and guilty, they are also more likely to engage in relational aggression. They may exhibit low self-esteem and pessimistic attitudes. They also tend to look at things more negative life events. Due to their negative attitudes and the frequency of negative attitudes and the frequency of negative life events, they may become hostile (Christopher, et al.2013).

The limitation of (Christopher, et al., (2013) insider threat prediction through linguistic analysis of electronic communication is that; the methodology the used seems to be automated predictive mechanism but that is not automated and they used manual data processing and predictive techniques for predicting the insiders behavior. Hence, the researcher understands the above limitation and develops an automated supervised machine learning predictive methods by combining the emotion of the insider either positive or negative and associated behavioral description with the class of emotion.

A Bayesian Network Model for predicting insider threat is the other significant research work written by Axelrad, et al (2013). This paper introduces the importance of a Bayesian network model for understanding the motivation and psychology of the harmful or malicious insiders. This study incorporates the five personality psychological variable agreeableness, conscientiousness, neuroticism, extraversion, and openness are used with the assumption of they are the indictor of a potential malicious insider. They begin their study by identifying psychological variables hypothesized to characterize a malicious insider and develop an initial model based on the result in the research literature to highlight critical variables for the prediction of malicious insider threat. The researchers then conducted a survey to measure these predictive variables in a common sample of normal participant to validate the association among measurable variable in the literature and the normal participant. Followed by these they develop a structural equation model based on half split of empirical literature and half split from the survey dataset to validate against each other. The study consists of 486 subjects that responded to a 112 item survey measuring the identified psychological variables. The Bayesian network was adjusted in light of the results of the empirical analysis. Finally they validate the model by

comparing predicted relationships between variables from surveys to the initial predictions derived from the literature.

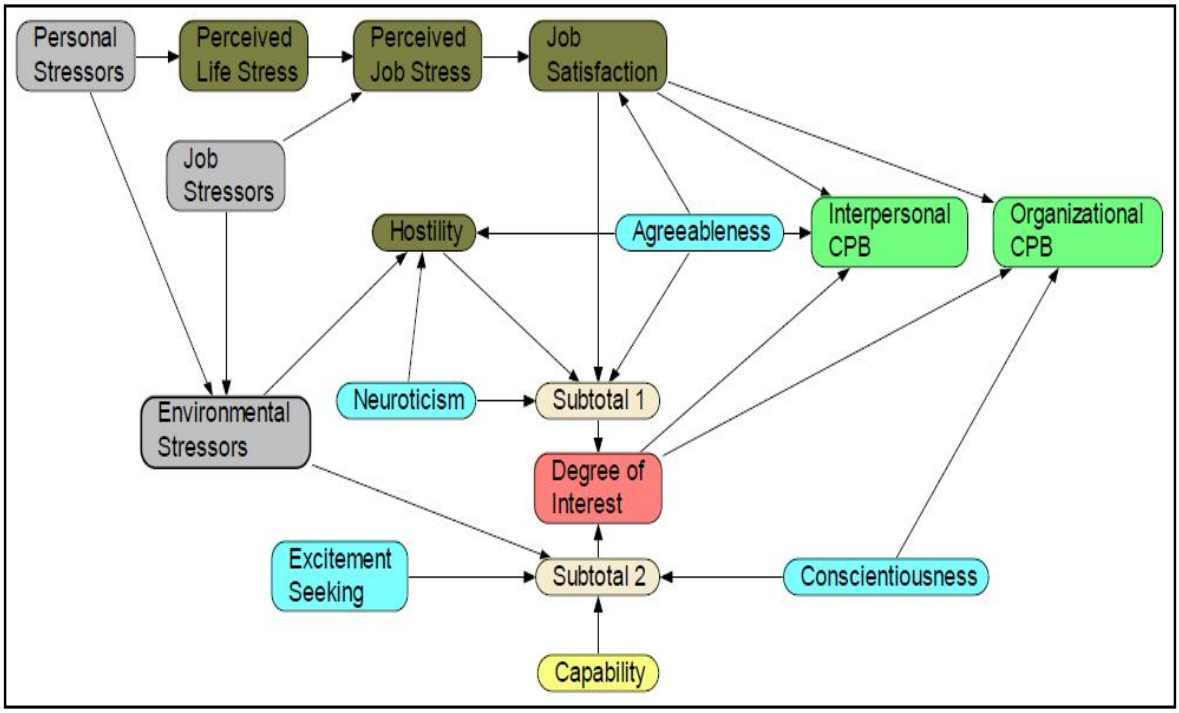

Fig 2.8Bayesian network variables and structure

The result suggested some changes to the original model that would improve model fit. Based on the result of that fit, they implemented those adjustments and tested the adjusted model on a holdout sample. The fit of the revised model was reasonable and they adjusted the original Bayesian network in the same manner. The result indicates that the derived relationships are valid, but identify several additional relationships that should be considered.

The main difference between the work of this paper and Axelrad, et al.(2013) research work they used a literature survey to predict the malicious or harmful insiders based on the big five personality trait. Based on the model they develop from the literature and employees observable behavior they develop a model using Bayesian network model to validate the relationship between the literature survey and the data collected by observation from employees. Henceforth, by understanding the gap to predict insiders' behavior the researcher build a model that able to classify insider behavior from their email communication.

According to Parveen, & Weger,et al., (2011) in their article supervised learning for insider threat detection using stream mining with the objective of testing an ensemble-based stream mining algorithm based on supervised learning and how stream mining can be effectively used to detect insider threats that can address the challenge of rare anomalies by maintaining an evolving collection of multiple models to classify dynamic data streams of unbounded length. An ensemble base classifier algorithm induces from incoming chunks of the data stream. Each chunk consists of prototypes and can be updated using instance selection technique when a new data have arrived. When a new data chunk is formed, ensemble model is also updated on the basis of weights assigned to each one-class classifier (Czarnowski, &Jędrzejowicz, 2014). The test data consisting of real-time recorded data of raw system calls is used to demonstrate the practicality of the approach. During the testing phase, test data is classified as normal or anomalous based on geometric deviations from the model. However, the approach is only applicable to bounded-length, static data streams. In contrast insider threat- related data is typically continuous and threat patterns evolve over time. Hence effective classification models must be adaptive and highly efficient in order to build the model from large amount of evolving data. The result is a classifier that exhibits substantially increased classification accuracy for real insider threat streams relative to traditional supervised learning and other single-model approach. The supervised learning achieves much higher accuracy 71% than the unsupervised learning 56% accuracy, 54% false positive rate and 42 % false negative rate.

As stated by Gavai,& Sricharan, et al., (2015) in their article supervised and unsupervised methods to detect insider threat from enterprise social and online activity data with the objective to discover insider threat by analyzing enterprise social and online activity data of employees. For this research social data including email communication patterns and content, and online activity data such as web browsing patterns, email frequency, and file and machine access patterns. The researchers take two approaches to detect insider threat by using supervised and unsupervised approach using the state of art anomaly detection methods. The result shows that receiver operating characteristic (ROC) score of 0.77 for the unsupervised approach, and a classification accuracy of 73.4% for the supervised approach. As the result shows that the proposed approaches are fairly successful in identifying insider threat events.

In general the aforementioned literature highlight that much has not been done and left out recommendation to try by other researcher in a way to improve the performance of classification of the insider threat behavior and get the actual problem solved. In this aspect, this study has been attempted to classify insider threat on the five stages and shows who more risky employee for the organization than others. In addition to this, the study indicates who are coming to on savior stage. Moreover this mitigate being insider threat by identifying employees in different stages.

CHAPTER THREE

METHODS AND APPROACHS

3.1 Study design

This study is an experimental desk based research where retrospective data were used to train the classification model. The dataset was collected and organized by the CALO Project (A cognitive Assistance that learns and organizes). It contains data from about 150 users, including senior management of Enron. The archive organized into folders and sub-folders with files. The corpus contains a total of about 500,000 messages. This data was originally made public, and posted to the web, by the federal Energy regulatory commission during its investigation. The dataset here does not include attachments, and some messages have been deleted "as part of a reduction effort due to requests from affected employees" (www.cs.cmu.edu. 2015/).

3.2 Architecture of the system

The difficult nature of classifying insider threat, initiate researchers to design various architecture based on the test and training data set. In order to address the insider threat classification in easy and robust way, better system architecture with all possible limitations should be designed.Therefore, in this study the design of the system architecture is depicts as observed in figure 3.1 below. The conceptual model that holds ideas comprising broader concepts depicted on the figure including text preprocessing, machine training, and text classification to classify the insiderness. Text preprocessing task involves selecting relevant section from large set of email documents. This includes removing unnecessary directory and email header, merging files and removing numbers and punctuation. Training machine aspect takes the role of learn or train from the word dictionary which are identified from different sources. These training words are composed of negative and positive words identified by psychologists. The final component of the conceptual model is classifying texts into positive and negative class based on the machine learning capacity for newly coming instances that passes the previous two stages.

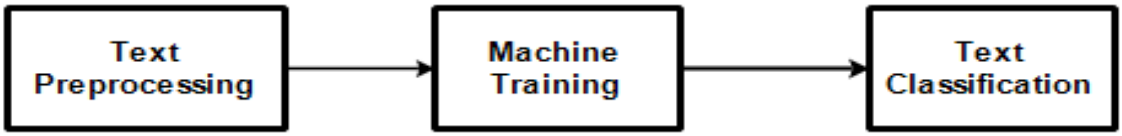

Figure 3.1: Conceptual model for insider threat words classification

Figure 3.2 shows how an insider threat words can be classified from the electronic mail communication intelligently. The system takes an identified list of positive and negative words for training purpose. These training words are proposed by researchers such as Hu, & Liu, 2004; Liu, & Cheng, 2005, are used in sentiment analysis of tweeter opinion. Both the negative and positive words indicate the persons' emotion. The content of the email communication archive should be preprocessed by using a preprocessing module that makes it suitable for further processing. For each individual 340 emails are taken for further processing. The number of emails for each individual selected based on the content of the email. The email content that has mote that one line is selected for further processing. Email files that embodied "well received", "Thanks", "My pleasure" etc. are removed by taking each words as a representative for the email. In addition to this the main reason 340 email files are selected most employees email files found on the range of 340 -349.Totally 9520 emails are preprocessed for the selected 28 Enron higher officials. The archived email communication of one employee email files pass a sequence of steps ranging from remove irrelevant directories, removing email header, merging email file, eliminate punctuation and umbers, data size reduction and move towards to the last stage whereby list of words are sorted. These list of words extracted from the email file are utilized by the employee to communicate their email partner. The composition of the word is alphabetic only and alphanumeric characters such as street address, Identification number are removed from these lists of words. The machine would be trained with list of positive and negative words that have psychological positive and negative emotion tendency. These two classes (positive and negative) of words learn the model through SMO algorithms. The test data are applied on the learned model. The models based on the training data set classify the test data set into positive class and negative class. The total words found from the email content classified into positive and negative classes. The number of negative words found prom the classification model helps us to classify employees into the five stages of insider (Exploration, Experimentation,

Exploitation, Execution and Escape). The people who use more positive words are more emotionally stable, have positive thinking and non-insider threat.

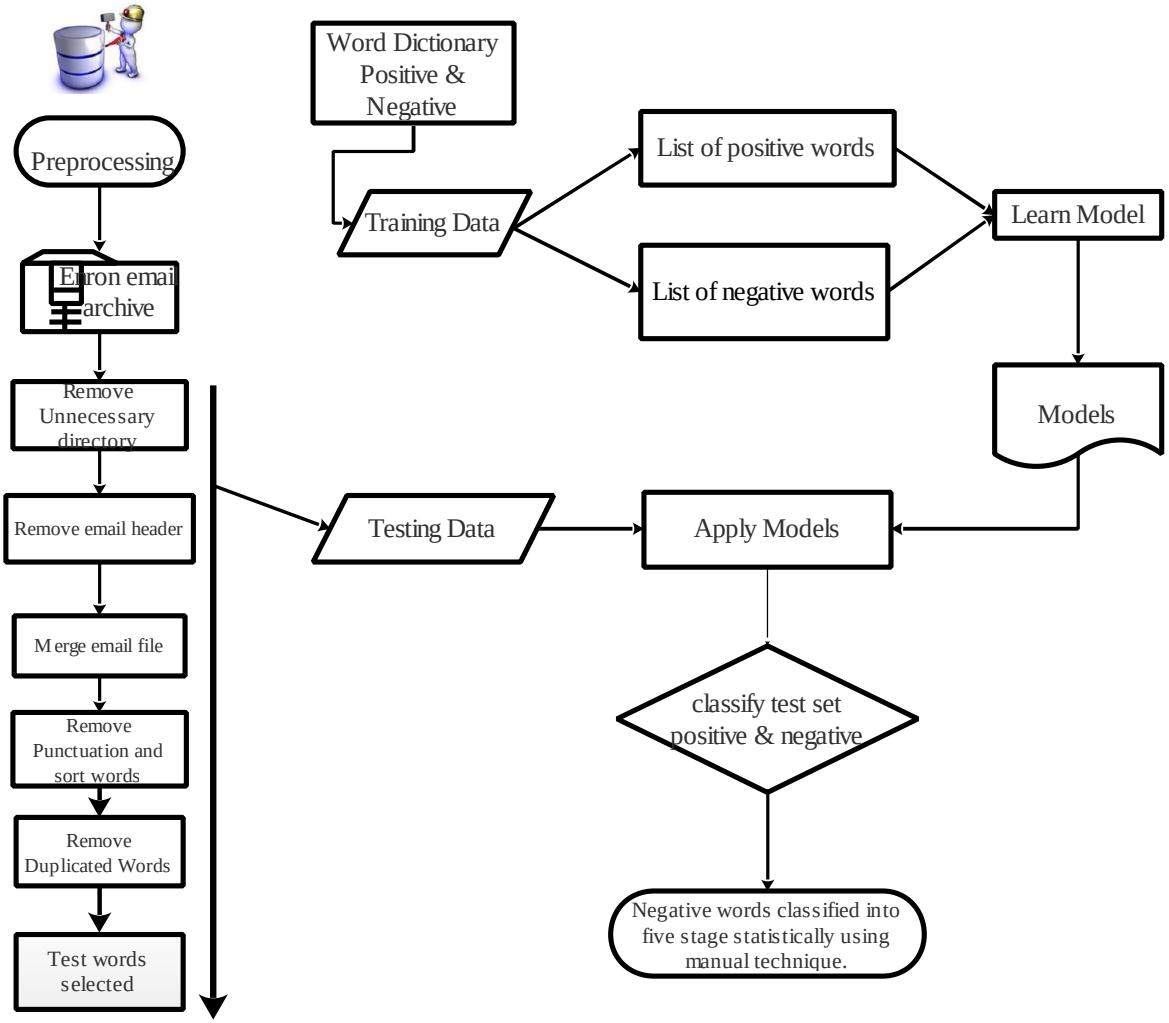

Figure 3.2 System Architecture

The architecture shows that the data mining task used for generating the required number of words used during the email communication from large collection of email archive for the person to be classified or suspected.

3.3. Test Data set

Data collection and preprocessing are the primary step in text classification. Henceforth**,** it is necessary to run the data trough preprocessing steps. Different companies collect and organize their data for various purposes. Enron corporation data set was collected and prepared by the cognitive assistant that learns and Organizes (CALO) project. The email archive contains including email of senior management of Enron Corporation. The email archive is organized into folders which are about 150 employees of the corporation. The corpus contains a total of 500,000 messages. This data was originally unclassified to the public and posted on the web, by the federal Energy regulatory commission during its investigation. The email data set does not include attachments and some messages have been deleted as part of a reduction effort due to the request from affected employees after the dataset was bought and posted by Leslie Kaleling at MIT (www.cs.cmu.edu/~/enron/). The dataset was used to classify insider threat personality by Brown, et al.(2013) by integrating with LIWC Dictionary (Linguistic Inquiry and Word Count).

For the testing purpose the Enron corporation Chief executive officer (CEO), President and vice president are selected and investigated to classify their psychological emotion. The main reason the we focus on this group is that including founder of the corporation Kenneth Lay, Jeffrey K. Skilling chief executive, David W. Delainey chief executive, Andrew S. FastowChief financial officer and others are charged related with Conspiracy, Securities fraud, Insider trading, Perjury/lying to investigators/ auditors, wire fraud, bank fraud etc. (Retrieved from http://www.nytimes.com/ref/us/20061023_ENRON_GRAPHIC.html_GRAPHIC.html?ref=enron on January 29,2016 by Pavel et. al, 2002, The Enron scandal).

Table 3.1 Description of data source and Number of records

Source of data center	Data coverage	Number of records	Number of attribute	Size of the data	Data type
Carnegie Mellon University	Year 2000-2001 G.C	517,402 files	17	1.32 GB	text/plain; charset=us-ascii

The classification performance of textclassification highly depends on the quality of given data. Understanding and knowing the data and its quality helps the researchers to go through the data preprocessing steps. The Enron organizational email corpus has been organized in the form directory and sub directory. The main mail corpus directory contains 150 individual mail directory labeled with employee last name. Moreover, each individual mail directory contain multiple subdirectory including self-defined and computer generated directories such as “sent_items”, “deleted_items”, “inbox”, “drafts”, “all documents”, “discussion thread” and different personal folder. Each email file has an attribute as a standard by the mail server such as header and another non-standard attribute such as content/body and signature block. From the above directories “discussion thread” and “all documents” are generated by the system not written by the sender.

As the email corpus shows the data has lack of incompleteness, missing directory, personal defined directories difficult to understand about their content, and empty mail box. Since the data quality affect the outcome of the data being analyzed. Basically the email files have three parts email header, body/text content and closing signature block. The first part of the email that is header has a standard structure, but on the content/body anything the sender wants to write can be written and on the signature block part there is no standard way of writing among the writers. Due to this the researcher give attention to the data preprocessing steps clearly. The following subsections deal with the data preprocessing steps.

3.3.1 Test dataset attribute Selection

Attribute selection is the process of selecting significant classifier for model construction and use. For this research the selected Enron Corporation email data needs cleaning and irrelevant attributes are discarded and relevant attribute were selected.

The description of selected attribute helps us to understand what the data looks like in terms of structure and content. The folder name is given as the employee’s last name, followed by a dash, followed by initial letter of employee’s first name. For instance “sturm-f” is named after Enron employee Sturm, Fletcher. The detail of employs name, contact, order in the corpus, and status are selected from the corpus and prepared for mapping folders to employee. Apendix A shows how employee name mapped with folder name. Each individual folder contain many types of

subfolders such as “inbox”, “sent”, “_sent_mail”, “sent_item”, “discussion_threads”, “all_documents”, “deleted_items”, and personal subfolders created by the employee work.

The corpus has a huge number of reproduced or duplicated emails in those folders. The folders “sent”, “_sent_mail” and “sent_item” from the corpus, that contain email files generated by email sender and each email file has different attribute as shown in section 3.3.2. When the employee sends an email some files are goes to “sent” folder and some others are to “sent_item” and others are “sent-mails” or in the three folders at the same time. Due to this the actual sender content found in these three folders are mingled with other email partner sent as a “forward” message. Moreover, these forwarded messages need cleaning because it doesn’t represent the actual sender.

From the email file attribute the body attribute of the email is selected for preprocessing by the assumption of the body of the email written by the sender and can classify the psychological emotion of the worker either positive or negative emotion as stated (Tausczik & Pennebaker, 2010). The header attributes generated by the system are not significant for this research including the signature block. For this research the selected attribute from the email message is that the body of the email produced by the employee and doesn’t include forwarded messages because there is an individual characteristics difference depending on the word usage.

In this research all the Enron higher official Chief executive officers, presidents and vice presidents going to be evaluated and classified on the five stage model. From the given data set the higher officials on the revealed status are 28 in number, from these all 28 or 100% the selected population is going to be investigated.

3.3.2 Descriptive properties of test data set

Email data has various encouraging features for research on insider threat classification. In the first place emails have defined format. An email message starts with the header, followed by body and signature block. An electronic mail message contains three parts: the header, body, and signature block. The header block embraces structured information, including receiver(s), sender, subject, date and time, carbon copy (Cc) etc. The body of the email part contains unstructured text. Information about who exchange information with who, with what subject they talk, at what time they talk can be retrieved from the header and the body. The signature block gives the

contact information of the sender and his/her position in the corporation. These components enable the researchers to trace past events. In this section the researcher discuss the 17 parts of the email attribute.

Email header: The email header section is a block of codes that hold information about who send e-mail or from where it comes and the methods how the message gets in touch at the destination. Headers contain the sender email address and/or the receiver. Each header field begins with a field name, followed by a "**:**", then a space and then the field value. Reviewing the email headers help us to determine the source of the message, to analyze the timestamp in line with delivery route and if there is delaying etc. In general it embraces control information and data regarding the message. In this email from line 1 to 15 are header, and the body line is begin after the header. Here is what the sample email taken from Enron corpus specifically Jeffrey K. Skilling Chief executive of the Corporation.

Table 3.3 sample Email file with its header

Message-ID: <23167422.1075840163215.JavaMail.evans@thyme> Date: Mon, 14 May 2001 19:46:00 -0700 (PDT) From: jeff.skilling@enron.com To: gmarmol@gmarmol.com Subject: Re: Luminant Mime-Version: 1.0 Content-Type: text/plain; charset=us-ascii Content-Transfer-Encoding: 7bit X-From: Jeff Skilling X-To: Gil Marmol<gmarmol@gmarmol.com> X-cc: X-bcc: Philippe A Bibi <Philippe A Bibi/HOU/ECT@ENRON>, Joannie Williamson <Joannie Williamson/Corp/Enron@ENRON> X-Folder: \jskillin\Sent Items X-Origin: SKILLING-J X-FileName: jskillin.pst
Gil, Jeff asked me to let you know that he really appreciates this opportunity, but his time in the office is at such a premium right now due to his heavy travel schedule, it would be difficult at best to work them in. He has turned down several invitations to join boards because he just can't commit the time. So, while he is flattered by their interest, if and when Mr. Quackenbush calls we will decline. It was nice talking to you. Glad to hear all is well.
Yours faithfully, Sherri ;-)

Message-ID: the Message-ID is a distinctive string generated by the mail server system that sends your message out in charge of the client, for a particular version of a digital message when the email message is first produced by the sender.

Date: This show the date and time as well as the time zone of email message send/received by the sender/receiver respectively.

From: This indicates the email address from who the message sent.

To: This part shows the address of who is the receiver of the message.

Subject: this part of the header shows the subject of the email message.

Mime-Version: Multipurpose Internet Mail Extensions (MIME) defined by RFCs (Request for comment) standards. MIME used as a means to improve the incomplete capabilities of email. It add the following services such as able to send many attachment via single message, able to write unlimited message length, able to use character set rather than ASCII code and utilization of rich layouts, fonts and color etc. The MIME-Version is the standard used in the email message and the current version is 1.0.

Content-Type: Content-Type describes the data type contained in the message body or the format of the message (plaintext or html) and the character set separated by semi-colon.

Content-Transfer-Encoding: Content transfer encoding describe the encoding used by the message body.

X-From: This section of header reminds us the name of the original sender of the message.

X-To: This section tells us the name of the original receiver of the message.

X-cc: The term Cc stands for Carbon copy of the email. In this field recipients are others peoples whom the sender of the email wishes to publicly inform of the message.

X-bcc: The bcc stands for blind carbon copy. Those who are BCCed are not visible to anyone who receives the email. The main deference between Cc and BCc is in case of Cc recipients are visible to all other recipients.

X-Folder: This section shows the original directory of the location of the message.

X-Origin: This section of header shows tells to us the name of who is the original sender of the message.

X-FileName: The x-filename shows the original file name of the email.

Body of the email content/ message: Message body is the actual or the main part or content of the email message, written by the sender. It can contain message's text, links, and image and attachments.

Signature block: This section is the last line below the body or content of the message, the senders write about his/her reminding, greeting and address. This part of the email has no general standard among the email users what must be included in this section.

3.4 Training Data set

In this section the researcher clearly describe the training dataset source and what they are. Splitting the data set into training and test set is an important for evaluation of the model. As shown on the system architecture the training set is implemented to build up a model and used to train the classifier. The training data set is used to build a model with the proposed algorithm. In this research work the training data set comprises two set of word classes that have negative and positive emotion when we used during communication. The training data set is composed of Action verbs, adjectives, and noun words that signify positive and negative within training set.

3.4.1 Training dataset attribute Selection

Attribute selection also known with its name variable selection or feature selection. These attribute selections help us to create an accurate classification model and decrease the training time. In order to come up with accurate classification model removing irrelevant and redundant instances and attributes from the data that does not contribute anything for the accuracy of the model. Small number of attributes are preferable than large number of attributes because less attributes diminish the complexity of the classification model and easy to understand as well as to explain.

For this research, the researcher selected two training attribute. The positive and negative words, those have positive and negative meaning. The detail of how these words are selected, their

composition, and number of words discussed on the next paragraph. In addition to this, to know what positive and negative words look the literature review part.

The lists of words are identified by Hu, & Liu,2004 and Liu, B., Hu, & Cheng, 2005, are used in sentiment analysis of tweeter opinion. Another research web site (http://positivewordsresearch.com) uses list of positive and negative words to describe the behavior of people. The University of North Caroline identified list of positive and negative (http://www.unc.edu/~ncaren/haphazard/positive.txt retrieved on March, 11, 2016) (http://www.unc.edu/~ncaren/haphazard/negative.tx retrieved on March 11, 2016) words respectively. In this research for the training purpose from the above sources unique words are selected for positive and negative word category to increase the performance of the model.

a) **List of positive Words:-**List of positive words were collected from three different source (http://www.unc.edu/~ncaren/haphazard/positive.txt; Bing Liu, Minqing Hu and Junsheng Cheng, 2005; http://positivewordsresearch.com/list-of-positive-words/). What makes different these three sources is that the number of words they contain. The first source (http://www.unc.edu/~ncaren/haphazard/positive.txt) contains 575 positive words. The second source (Bing Liu, Minqing Hu and Junsheng Cheng, 2005) contain 2006 positive words. The third source (http://positivewordsresearch.com/list-of-positive-words) contains 1197 positive words. From (http://systemagicmotives.com/Positive%20Noun%20Glossary.htm retrieved on

   March 11, 2016.) 1262 noun words collected. Totally from the above four sources we found that 5040 words including the duplicated words. These duplicated words refined by a special purpose python program that can identify unique words only. Finally we have 4904 unique positive words are organized to train the model.

b) **List of Negative words:-**List of negative words were gathered from the three different source (http://www.unc.edu/~ncaren/haphazard/negative.txt; Bing Liu, Minqing Hu and Junsheng Cheng, 2005; http://positivewordsresearch.com/list-of-negative-words/). Again like list of positive words these three sources contain different number of negative words. The first source (http://www.unc.edu/~ncaren/haphazard/negative.txt) contains 575 negative words. The second source (Bing Liu, Minqing Hu and Junsheng Cheng, 2005) contain 4783 negative words. The third source (http://positivewordsresearch.com/list-of-positive-words) contains 4542 negative

words. From the total 9900 negative words list of negative words are refined by a special type of python program that can able to identify list of unique negative words. Finally have around 4904 unique negative words.

3.5 Ethical consideration

Disclosing individual email files to the public is very unethical, but organizationl mails are created and used by the employees to do their office tasks. According to the owner (CALO project) of this email archive, the data set can be used by any researcher who wants to do further research on email data. Organization need to take too much care or even the government tahe care in collaboration with courts and ministry of justice before unclassified the email documents to the public.

CHAPTER FOUR

DATA PRE-PROCESSING AND EXPERMENATATION

Data preprocessing play a significant role in text classification. Text classification is used to give helpful information from the big amount of data. It is one of the important research issues in the field of data mining. Text classification is the task of classifying words by their composition. In order to build a better classification model we employee the preprocessing techniques. In this chapter we conduct a serious of data preprocessing and an extensive experimentation over the training dictionary and the test data set. This chapter has been organized into two-subchapters. The first part clearly describe data preprocessing phase including removing irrelevant directories, data size reduction, data cleaning, tokenization, data transformation etc. The second part shows how the experimentation is done.

4.1 Data Preprocessing and Experimentation

In this section, we discussed all the techniques that have been used in developing a model to classify whether insider threat exist or not using the word usage distribution categorized as positive and negative words. Moreover, the experimentation integrates the typical stages that mark a text classification process and insider threat.

Data preprocessing technique plays important role in text classification. This technique can often have a significant impact on classification performance of a learning algorithm (Kotsiantis, et al, 2006). Real world data, most of the times are unfinished, inconsistent and lacking measurable trend and prone to many error for analysis. Data preprocessing involves how these real world data are transformed into understandable format either by human or machine. In this research the data preprocessing steps include removing unnecessary directories, removing email header and signature, merging email, removing punctuation and segmenting sentences to word level. Lastly the content of the email data extracted for further processing.

4.1.1 Training data cleaning

The training data cleaning step involves removing irrelevant or duplicated words and punctuation from positive and negative word class. After the words are collected from different sources the

next step was cleaning of the data. During the data collection words from some sources were coma separated. Moreover, words are difficult to know wheatear they are repeated or not. In order to reach to these numbers of words different tasks have been done such as removing duplicated words, numbers and punctuation from each category by writing appropriate program module. In order to perform the above program python 2.7 is used. The python program automates the removal of duplicated words and removal of punctuation marks and numbers such as (!',()-[]{};:'"\,<>./?@#$%^&*_~=1234567890). Appendix B (program 1) is developed to remove punctuation from the training dataset.

The appendix B program 1, removing punctuation, it sorts words alphabetically. This makes easy for human being to know whether words are repeated or not. At the time of removing punctuation and sorting words, the researcher faces heap number of duplicated words from different source. Removing these duplicated words manually was very difficult and tedious. Removing duplicated words helps to normalize the length of the instances. Due to this reason a special purpose program module is developed that remove these duplicated words.

The appendix B, program 2acceptsthe output from program 1 and remove duplicated words from the list.

The output of the above program (program 2) is a list of words without repetition. Finally the words are arranged into two classes i.e. positive and negative. The words are researched and identified by the psychologists that have positive and negative classes. For each class 4904unique words are selected for training purpose.

4.1.2 Test data cleaning

In order to clean the test data set found from Enron email archive the following steps are followed.

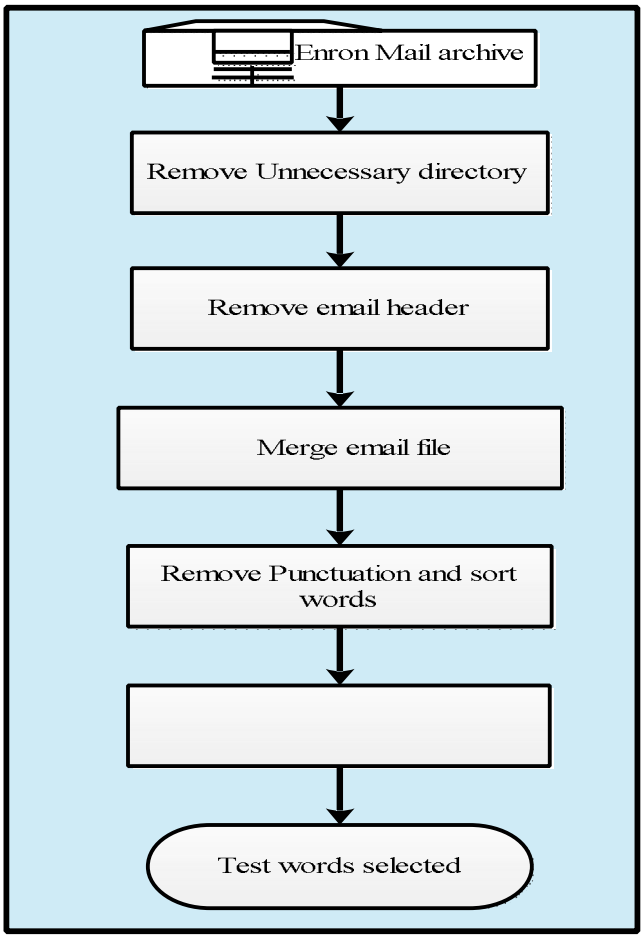

Figure 3.3 Test data cleaning phase

Mapping Folders to Employee: mapping folders to employee is the process of assigning corpus folders name with employee detail. Moreover, this is used for filtering employees that interest us to know about his/her status. Enron email corpus contains 150 mostly senior managers and whose emails are selected for investigation by the federal investigation office of US. The corpus embraces 500,000 messages. In the corpus individual employee's emails are organized by folders. The files in the folder are saved in .pst file format and need conversion to text or .txt file format for further preprocessing.

The first step in this research is to map folders and its ownership of file with the corresponding employee. The mapping task is put into operation by finding the name of the employee from "sent" folder files and the header section which contain "From" tells us who are the owner of the file and the corresponding email. The cases of employee-folder matching existed in the dataset.

This research was conducted with the assumption of one folder to one employee. For instance from user "allen-p" subfolders "_sent_mail", "sent_item" and "sent" 1509 emails are identified. From these 1500 email are sent by "Phillip K Allen", 8 emails are sent by "Ian Rangel" and 1 email sent by "Pam Butler" are found. From this we can say "Ian Rangel" seems to be secretary or assistant of Allen k Philip, but there is no evidence. In this case by considering the majority of the email are sent by Allen K Philip and asserted to be he is the only owner of the folder "allen-p".

Appendix A shows that 28 senior higher officials of Enron are identified from the data set including their first name, last name, email one, email two, order name in the folder, and their Responsibility in the Corporation.

a) Removing unnecessary directory

Data cleansing is one of the major step in text classification. In this data cleaning phase removing or deleting unnecessary directories are found very crucial. Removing these unwanted directories help as to avoid any confusion among the directories to be investigated or analyzed. From the individual mail archive "inbox", "deleted items", "personal folders" and many more directories are deleted except "sent", "sent_item" and "sent_mail" directories. Because these three directories file are in need to be investigated. In order to remove unnecessary directories from the individual mail archive a special purpose of program module is developed by the researcher. This program code automatically removes unwanted directories from the archive based on the given directories by the researcher. The Apendex B, program 3 is used to delete unnecessary directories. Python 2.7 is used to run this program.

b) Removing email header and signature

An email header as indicated in chapter three table 3.1 sample email file is generated by the system. The system automatically assigns message id, sender, receiver, and many other 14

attributes. Due to this, the system generated part of the email found to be irrelevant for this research and deleted from the content. The other part that is not significant for this research is the closing remark or salutation or address or held position in the organization. In order to remove the header the researcher developed a python program. In this regard the closing remarks are deleted manually because there is no standard way of writing closing remark among the email partner. The researcher tried to remove the last 3 or 4 lines from the email but some writers end their email without closing. In this case the last 3 or 4 lines of messages are deleted from the body of the email, this makes lose of important terms from the body of the email. For some other email files removing the last 3 or 4 lines work perfectly but in order to make the data to be consistent the researcher prefer to remove the closing remarks manually.

Email communication is like a thread. Each single file contains the message exchanged among many email partners. For example from "delainey-d/sent" directories we found that an email file labeled as "1_" in this file the email is exchanged among David W.Delainey and Brett R. Wiggs. In order to investigate the word usage only words written by David W.Delainey are selected because the directory belongs to David. The other words in this email file written for David email response are deleted. The main reason the word written by respondent (Brett R) threads are deleted "our words are ourselves" (Pennebaker, et al., 2003). According to Pennebaker in his annual review of psychology our words we use in our day to day life express ourselves. The words people use is associated with the social, mental, mental health and physical state. From this sentence and other context of the review the word use is directly related with the social health of the employee. This social health is expressed through pleasing interpersonal relationship with our families, friends, work colleagues etc. Each of this relationship expressed through string communication skill among those peoples. The communication skills include writing, listening, speaking and reading. Appendix B, program 4 removes the header from the email file.

c) Merging email files

In text classification, there are various methods of data reduction. Merging email files is one of the essential method of reducing large number of data into a single file. For Instance, from the sample set of data "Presto_k" has 956 files in the directory of presto_k/sent_item and 7 files in presto_k/sent directory. After applying a program that remove the header and manually deleting

the closing remarks and forwarded email on these directories the content of email files are merged into a single file. The files need to be combined into a single file. In order to combine these files the researcher developed a program that can merge these 956 files into a single file. The 7 files in the directory of “presto_k/sent” into single file. Sent directory files are merged as a single file and sent_item directory files are merged as a single file. The first merged file from “sent” directory and the second merged file from “Sent-mail”directory are merged together as a representative of “Presto-K”. Finally “Presto_k” have single file that contain all his sentences during the email communication. Appendix B, program 5 combines multiple files into a single file.

d) Punctuation and Tokenization

Punctuation and tokenization process involves removing special characters, symbols, personal name, place name and breaking up a sequence of strings into words known as tokens. Tokens can become the input for another subsequent process including text classification. The electronic mail text is a linear sequence of phrases, words, character or symbols. In English language, words or tokens are separated by a white space character. The texts used during the communication needs to be segmented and listed into word level. The segmentation of low-level tokenization can be conducted such as separating hyphenated words. At this stage, punctuation, numbers and alphanumeric characters are removed by a special purpose program that remove these special characters and only words are leaved for further preprocessing.

The next step after merging the files is that removing symbols and numbers such as !',()-[]{};:'"\,<>./?@ #$%^&*_~=1234567890 and splitting sentences into word level. Appendix B, program 6 remove symbols, punctuation mark and numbers and split sentences into word level.

Program 6 provides list of words used during the email communication including abbreviations, human and corporation name. List of words are selected from the output of the above program. When we run the above program there are many duplicated words in the sentences. The following table 3.3 shows number of words count by executing the above program step by step.

It is obvious that it is unethical to disclose the name of the email owner, but the data set prepared by the CALO project declare that anyone who wants to do research on email can utilize the data set freely.

Table 4.1: List of Enron corporation higher officials

No	LastName	FirstName	Folder Name	Total email	No_email Selected	Total word count
1	Arora	Harry	Arora-h	379	340	594
2	Clair	Carol	stclair-c	1328	340	46,220
3	Corman	Shelley	corman-s	630	340	29,358
4	Davis	Dana	davis-d	363	340	11,600
5	Delainey	David	delainey-d	1814	340	27,243
6	Fossum	Drew	fossum-d	2067	340	18,780
7	Hayslett	Rod	hayslett-r	639	340	6,390
8	Horton	Stanley	horton-s	482	340	8,149
9	Kean	Steven	kean-s	1773	340	42,495
10	Kitchen	Louise	kitchen-l	1132	340	19,450
11	Lavorato	John	lavorato-j	1878	340	13,460
12	Lay	Kenneth	lay-k	527	340	29,570
13	Martin	Thomas	martin-t	683	340	11,237
14	McCarty	Danny	Mcarty-d	348	340	5,400
15	Neal	Scott	neal-s	1102	340	8,237
16	Presto	Kevin	presto-k	962	340	21,417
17	Shankman	Jeffrey	shankman-j	1084	340	10,652
18	Shapiro	Richard	Shapiro-r	344	340	600
19	Shively	Hunter	shively-h	683	340	6,072
20	Skilling	Jeffrey	skilling-j	605	340	10,942
21	Steffes	James	steffes-j	1325	340	18,140
22	Steponovich	Joe	Steponovich-j	341	340	871
23	Sturm	Fletcher	sturm-f	417	340	6,003
24	Tholt	Jane	tholt-j	684	340	11,965
25	Whally	Greg	Whally-g	344	340	658
26	William	Jason	William-j	341	340	900
27	Zipper	Andy	zipper-a	352	340	7,193
28	Zufferli	John	zufferli-j	340	340	6,000

As shown in the table 4.1 the column “Folder name” is the caption name of individual named from first and last name, “total email” the number of emails found from “Sent”,”Sent_item”, “Sent mail” directories. For each individual 340 email files are processed. The last column Word Count (WC) indicates the total number of words used by the respective officials. The total number of words are computed from the files found in each folder. For instance “Arora Harry” 594 words are found by counting the total WC after the merging files found in the “sent” and “sent item”.

4.1.3 Data Size Reduction and Selecting Test Data Set

At this stage the data size is reduced and transformed or consolidated into forms that are applicable to the data mining process. The content or body of the message size reduced and transformed. In this research work the words that are tokenized into word level are sorted and duplicated words are removed and the data size is reduced into unique words only by a python program. The main reason for removing these duplicated words is that, only the word distribution of investigates are studied rather than frequency.

In order to remove these duplicated words the researcher developed appendix B (programs 7) that identify unique words only. Program 7 gives the list of words used by the employee throughout the email communication. The output contains irrelevant letters (A, B, C, D etc), abbreviation (PPL, LLC, etc) and employee name (Allen, James, Jeff,Prestoetc). These irrelevant items do not have psychologically negative or positive implication by themselves. Due to this reason the researcher deleted these items from the list of words from each individual employee utilized through the communication process. The final lists of words identified for each employee are used for testing purpose.

Table 4.2 The unique word count for Enron higher official

No	LastName	FirstName	Folder Name	No_email Selected	Total word count	Processed word count
1	Arora	Harry	Arora-h	340	594	350
2	Clair	Carol	stclair-c	340	46,220	2,100
3	Corman	Shelley	corman-s	340	29,358	3,111
4	Davis	Dana	davis-d	340	11,600	1,467

5	Delainey	David	delainey-d	340	27,243	3,643
6	Fossum	Drew	fossum-d	340	18,780	2,900
7	Hayslett	Rod	hayslett-r	340	6,390	1,120
8	Horton	Stanley	horton-s	340	8,149	1,542
9	Kean	Steven	kean-s	340	42,495	299
10	Kitchen	Louise	kitchen-l	340	19,450	1,749
11	Lavorato	John	lavorato-j	340	13,460	1,219
12	Lay	Kenneth	lay-k	340	29,570	1098
13	Martin	Thomas	martin-t	340	11,237	1337
14	McCarty	Danny	Mcarty-d	340	5,400	1199
15	Neal	Scott	neal-s	340	8,237	1058
16	Presto	Kevin	presto-k	340	21,417	1,965
17	Shankman	Jeffrey	shankman-j	340	10,652	2,179
18	Shapiro	Richard	Shapiro-r	340	600	299
19	Shively	Hunter	shively-h	340	6,072	576
20	Skilling	Jeffrey	skilling-j	340	10,942	1192
21	Steffes	James	steffes-j	340	18,140	2399
22	Steponovich	Joe	Steponovich-j	340	871	449
23	Sturm	Fletcher	sturm-f	340	6,003	969
24	Tholt	Jane	tholt-j	340	11,965	1,290
25	Whally	Greg	Whally-g	340	658	509
26	William	Jason	William-j	340	900	309
27	Zipper	Andy	zipper-a	340	7,193	9,77
28	Zufferli	John	zufferli-j	340	6,000	840

From table 4.2 the “processed word count” meaning that the list of words found unique from the total email conversation. Moreover, this column words are used to classify the employee as a threat.

4.1.4 Challenges of Cleaning Enron Corporation Email Archive

From the Enron Corporation email data set the researcher in general identified four main problems. The first problem is that some of Enron higher official are still in prison but there is no dataset of these individuals rather they send their information through their office assistances. From the data set Kenneth Lay was one of the CEO but the message was disseminated by his office Assistant Rosalee Fleming due to this it is difficult to investigate Kenneth Lay in this

research, but some email files are found in Rosalees' folder written by him and these files are investigated. Second, multiple email address and personal name are used by the same person. Appendix A shows employees multiple email address and their name as organized in the archive. Even if the email address and name of the sender are different the target receivers obtain their email. Off course these alternatives will not affect the delivery because the destination address is supplied in the protocol, but the main difficulty is here mapping these names correctly greatly increased. In addition to this one employee can have additional domain email address such as hotmail, gmail, yahoo etc. so mislabeling employee may lead us to confusion and wrong conclusion. The third and the most challenging task is the fact that duplicate email are found in the archive. For Instance when X sends an email to Y, that email would be found in the "Sent" of X and "Inbox" of B simultaneously. If there are multiple recipients, each one of them would have a copy of that email. These duplicate emails should be removes for the research purpose. Fourth, the content of the email should be extracted. Extracting the email content written by X for Y, Z,W and again Y,Z,W replay for X and again X replies for them .Extracting Mr. X email content from the thread is found to be very cumbersome and challenging task in this research. The email content is normally blended with the closing remarks, quotation etc.

The researcher used list of words from the email content that can classify emotional status of the employee as negative and positive emotion and the corresponding stages to be with negative number of words classified.

4.1.5 Decision processes

In this phase, classification model is used to build a tree structure. It breaks down a data set into positive and negative word distribution class. The decision process phase compares the word distribution between the training dataset and testing data set. After the models are learn by the algorithm and training data, test data are supplied for testing. The learned models classify the number of instances given as a test classify into positive and negative class. Henceforth, the total number of words used by one employee during email communication is classified into these two attributes. The total number of negative words as compared with the total number of words in terms of percentage will help the investigator to pass the decision. In order to pass the decision the researcher uses the five stage model of insider threat classification will be used as suggested by (Young, et al,2013;Vectra,2015; Ted et al, 2013; Smith, 2014). From the five stages model as

clearly discussed in literature review part individuals who fall from stage two up to stage fivewill be classified as insider threat. In order to categorize employees into these five stages, the total number of negative words used by each individual is separated by five cutoff points. He cutoff points are determined by the negative words used by the email owner. The first cutoff point class into the first stage, the second cutoff point assigned into second stage, the third cutoff point assigned into third stage and so forth.

4.2 Data Conversion

Data conversion is one of the basic methods we apply to make the data suitable for further preprocessing and classification. The list of words is extracted for 28 individuals by applying the above stated methods and techniques such as removing irrelevant directory, deleting email header, merging files, identifying unique words etc.

The researcher used Weka 3.7 version for this research and found that the attribute relation file format (arff) file type is a good data type for this software and other such as RapidMiner, KNIME to make a better classification over the newly arrived instances. An ARFF (Attribute-Relation File Format) file is an ASCII text file that describes a list of instances sharing a set of attributes. After the files are merged as a representative for each individual listed in the table 4.2in text files format need to be changed to arff file format. The Weka simple command line interface allows the users to convert (.txt) files to (.arff) file format by applying a single line of code using weka "TextDirectoryLoader". "TextDirectoryLoader" is wekas special program that helps us to convert the text files to arff. For this research purpose the researcher use the following command to create the attribute relation file format. For instance in order to covert Jeffery skilling text files to the arff file type put the J.skilling text file in the directory wekaclass which is located on my desktop *" C:\Users\Vostro 2520\Desktop\wekaclass.*

The files converted to arff file type are saved in other directory on mydesktop *"C:\Users\Vostro 2520\Desktop\weka\jskilling.arff"*.Bothe the training file and test files need to be similar file format.

```
// "-dir" shows the beginning of directory
// the symbol ">" means change the directory the new arff file saved.
java weka.core.converters.TextDirectoryLoader -dir "C:\Users\Vostro 2520\Desktop\wekaclass" >
"C:\Users\Vostro 2520\Desktop\weka\skilling.arff"
```

The above Weka command (the “textdirectoryloader”) program gives the following output.

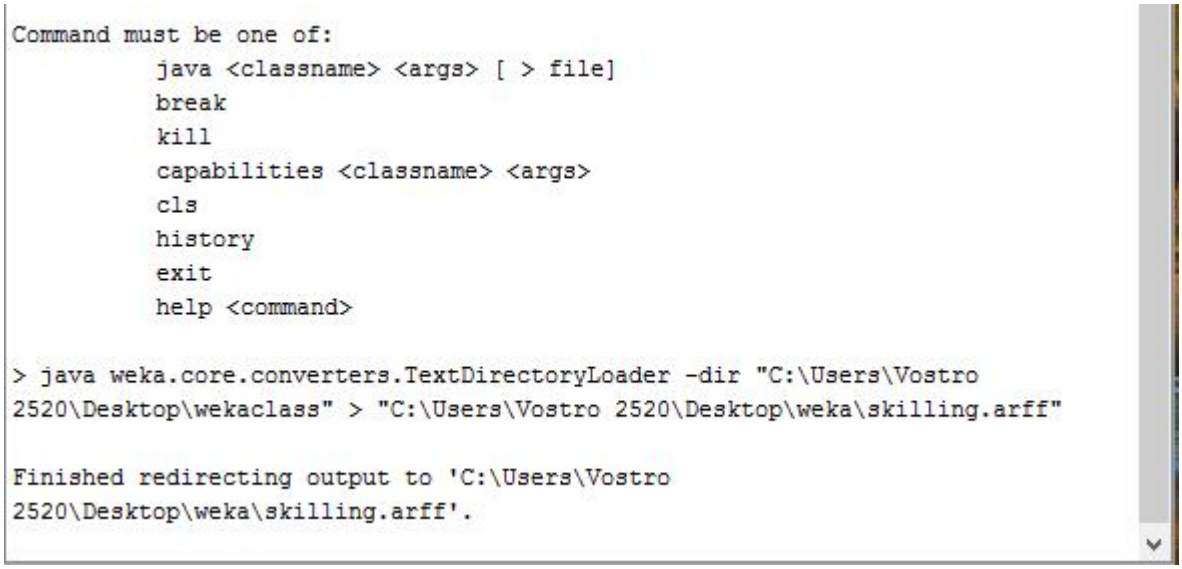

Figure 4.1 Weka file conversion from txt to arff

As can be seen from the above screenshot of Weka simple command line interface output our “textdirectoryloader” run without any problem and“Finished redirecting output to C:\Users\Vostro 2520\Desktop\weka\skilling.arff” meaning that our program works correctly and the .txt files are changed to .arff file format.

One of the biggest challenges for this research were to create a compatible attribute file format for both training and test set. Unless we make similar the header attributes of the training and test set Weka could not run the dataset.

In order to alleviate the above challenge the researcher followed batch filtering techniques to create a compatible header attribute file format. *Batch filtering techniques* is utilized if a test set, needs to be processed with the same statistics as training set.

Before running the batch filtering process first of all change your training and tesing files on weka*openfile>selctfilter>unsupervisedatt>attribute>numerictonomianl>edit>select_attribute> attribute as a class>ok>* and finally save the model.

The above steps give us the following attribute arrangement for training and test. As depicted in the following screenshot figure 4.2 before a batch filtering process applied on test set and training set.

```
@relation 'training-weka.filters.unsupervised.attribute.Reorder-
R1,2-weka.filters.unsupervised.attribute.NumericToNominal-
Rfirst-last'

@attribute text string
@attribute @@class@@ {pos,neg}

@data

able, pos
access, pos
added, pos
rankle, neg
rollercoaster, neg
selfserving, neg
fright, neg
subpar, neg
stigma, neg
accuse, neg
fear, neg
dispiriting, neg
```

Figure 4.2 Wekalists of attributes

ARFF files have two distinct sections. The first section is the **Header** information, which is followed by the **Data** information.

The researcher applied string to word vector batch filtering techniques on both training and test set. String to word vector conversion converts String attributes into a set of attributes representing word occurrence information from the text contained in the strings. The set of words (attributes) is determined by the first batch filtered (typically training data). Up to the researcher understanding the researcher prefer to run the batch filtering techniques for each individual test file with one training file is excellent.

The following command line are used to make a batch filtering process

```
java weka.filters.unsupervised.attribute.StringToWordVector -b -itrain.arff -o output_train1.arff -c last
-r test.arff -s output_test.arff -R ,1,2 -O -C -T -I -N 0 -M 1
```

Here, the first input/output pair (**-i**/**-o**) initializes the filter's statistics and the second input/output pair (**-r**/**-s**) gets processed according to those statistics. To enable batch filtering, one has to provide the additional parameter **-b** on the command line.

The data type after applying batch filtering looks like the following

```
@relation 'training-weka.filters.unsupervised.attribute.Reorder-
R1,2-weka.filters.unsupervised.attribute.NumericToNominal-
Rfirst-last-
weka.filters.unsupervised.attribute.StringToWordVector-R1-W1000-
prune-rate-1.0-C-T-I-N0-stemmerweka.core.stemmers.NullStemmer-
M1-O-tokenizerweka.core.tokenizers.WordTokenizer -delimiters \"
\\r\\n\\t.,;:\\\'\\\"()?!\"-
weka.filters.unsupervised.attribute.Reorder-
R2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25
,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,

9797,9798,9799,9800,9801,9802,9803,9804,9805,9806,9807,9808,9809
,9810,1-weka.filters.unsupervised.attribute.NumericToNominal-
Rfirst-last-weka.filters.supervised.attribute.AddClassification-
Wweka.classifiers.rules.ZeroR-
weka.filters.unsupervised.attribute.ReplaceMissingValues-
weka.filters.unsupervised.attribute.ReplaceMissingValues'

@attribute abandoned {0,6.370825}
@attribute abandonment {0,6.370825}
@attribute aberration {0,6.370825}
@attribute abhorred {0,6.370825}
@attribute abhorrence {0,6.370825}
@attribute abhorrent {0,6.370825}
```

Figure 4.3 Weka batch filtering result

As table 4.1processed Word Count (PWC) columns shows the total number of unique words found for each individual. For instance “Sturm Fletcher” used 969 unique words and these words are classified into positive and negative word classes based on the training data and the classification model.

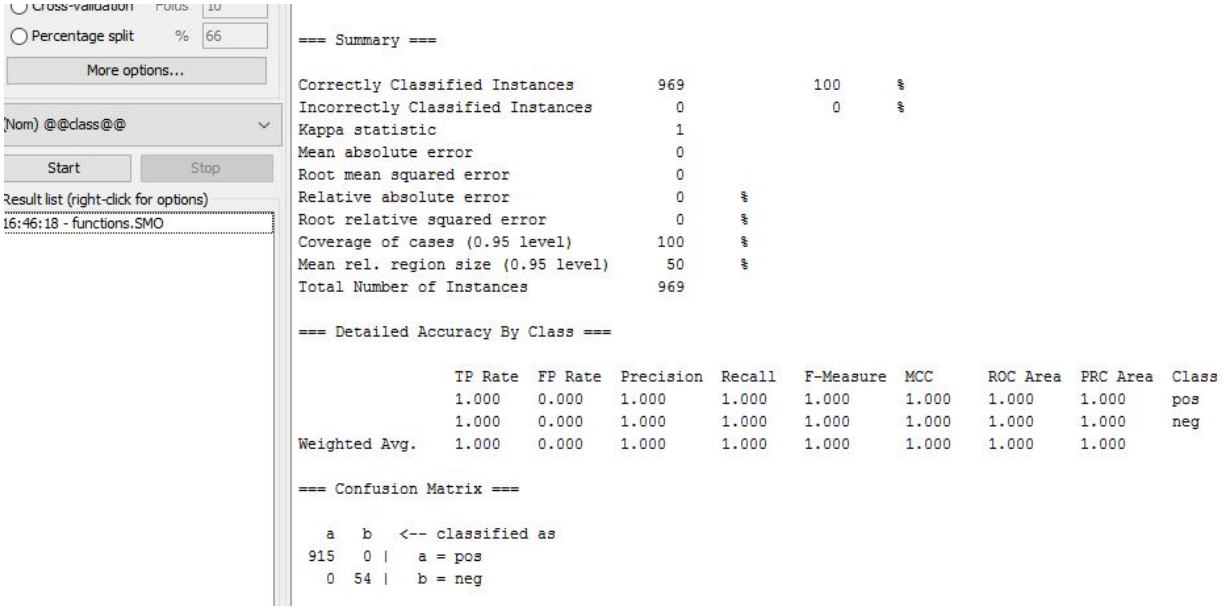

Figure 4.4 Sturm Fletchers' words are classified into positive and negative classes

4.3 Model Building

In text classification, building a model is one of the vital tasks which is carried for text classification because text classification play significant role to deal with very large amount of text documents. In this phase several text classification techniques are applied and their parameters are adjusted to optimal values. Typically, different techniques can be utilized for similar data mining problems. Some of the tasks include selecting the modeling technique, experimental setup or design, building a model and evaluating the model.

Selecting appropriate model depends on goal of text classification. Consequently, to attain the objectives of these research two classification techniques has been selected for model building. The analysis was performed using Weka environment. Among the different available classification algorithms in Weka, SMO (Sequential minimal optimization) for Support vector machine are used for experimentation of this study.

Experimental Setup

In all research including text classification, experimental set up introduces how the experiment is conducted. The testing procedure follows classification technique. In classification, precision and error rates encountered were used for measuring the performance of the model. The researcher

studies two different scenarios for the selected two algorithms i.e. naïve Bayes and Sequential minimal optimization (SMO) as stated below. The training and test set are prepared separately and there is no percentage split and k-fold cross validation method for training and test set. The training data contain 4904 positive and 4904 negative words when we used them that have emotionally positive and negative meaning respectively. The test set are contain different number of words for each individual as stated on table 4.1;i.e. processed word count that could be classified as positive and negative depending on the performance of the classification model.

4.4Experimental result

In this sub topic the researcher run the experiments, interpret the result in one scenario for the selected classification algorithm; i.e.SMO. For this research work 9808 words are used in two categories i.e positive and negative word classes. For SMO the training classes are correctly classified as 4904 words for each positive and negative class.

The SMO classifier employs a very complex hypothesis function. If the model is able to fit the training dataset very well it would have a low bias. There is no error resulting from inaccuracies in its hypothesis class.

The confusion matrix on figure 4.4 shows the result of true positive is 915 meaning that all positive words classified in positive classes as a positive word. Moreover, this matrix also shows 54 negative words are classified as threat word class.

SMO model

After the model is trained by training dictionary the test data set are reevaluated against the model. The percentage of true negative is used to classify insider threat. The percentage of true negative is calculated as follow:

$$\text{Percentage of Threat Words} = \frac{\text{Wordcount} * 100\%}{\text{Threat words}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (4.1)$$

Table 4.3 the classification result of SMO

No	LastName	FirstName	Status	Word Count (WC)	Non-threat words	Threat words	% of threat words
1	Arora	Harry	V.President	350	322	28	8.0
2	Clair	Carol	V.President	2100	1944	156	7.43
3	Corman	Shelley	V.President	3111	2194	917	29.47
4	Davis	Dana	V.President	1467	1375	92	6.27
5	Delainey	David	CEO	3643	2054	1589	43.62
6	Fossum	Drew	V.President	2899	2253	646	22.28
7	Hayslett	Rod	V.President	1120	1074	46	4.11
8	Horton	Stanley	President	1542	1427	115	7.46
9	Kean	Steven	V.President	299	288	11	3.68
10	Kitchen	Louise	President	1749	1609	140	8.0
11	Lavorato	John	CEO	1219	1133	86	7.05
12	Lay	Kenneth	CEO	1098	728	370	33.7
13	Martin	Thomas	V.President	1337	1248	89	6.6
14	McCarty	Danny	V.President	1199	1120	79	6.59
15	Neal	Scott	V.President	1058	1028	30	2.83
16	Presto	Kevin	V.President	1964	1837	128	6.51
17	Shankman	Jeffrey	President	2179	2046	133	6.1
18	Shapiro	Richard	V.President	299	288	11	3.68
19	Shively	Hunter	V.President	576	554	22	3.82
20	Skilling	Jeffrey	CEO	1192	791	401	33.64

21	Steffes	James	V.President	2399	1710	689	28.72
22	Stepenovitch	Joe	V.President	449	431	18	4.0
23	Sturm	Fletcher	V.President	969	915	54	5.57
24	Tholt	Jane	V.President	1290	1209	81	6.28
25	Whalley	Greg	President	509	477	32	6.28
26	Williams	Jason	V.President	309	293	16	3.14
27	Zipper	Andy	V.President	977	898	79	8.08
28	Zufferli	John	V.President	840	783	57	6.79

From Table 4.3 the percentage of true negative word is calculated for each individual. After the percentage true negative words are calculated, the numbers are sorted with ascending order and cut into five cutoff point to know that who classified on stage one, two, three, four and five of insider threat classification stage. The first cutoff point used for the first stage, the second cutoff point for the second stage, the third cutoff point for the third stage etc. The result of cutoff point is indicated in chapter five, table 5.1.

The above table, table 4.3 shows the Enron higher officials word usage. The table contains both threat and non threat words. As shown on the first row of the table Arora Harry has the vice president status and from the total 350 words used 322 words are non-threat words; 28 of the words found are threat words, meaning that from the total words (350) 8.0% of words are threat indicator. On the other hand 92 % of words are non threat indicator.

Table 4.3 shows Delainey Davied the chief executive officer of the Enron Company, score highest amount of threat words relative to his colleagues; which is rated as 43.62 % of threat words. Moreover, Delainey was sentenced and goes to through to jail.

Table 4.3 shows that Nail scote was more non-threat employee as compared to his colleagues. The number of threat words computed were 2.83 %; meaning that 97.17% were non-threaten employee.

CHAPTER FIVE

RESULT AND DISCUSSION

This chapter describes experimental study of the text classification algorithms and the results found which are described in the previous chapter, on insider threat classification from the distribution of words found in email communications.

5.1 Analysis of the result

In this section the interpretation of results found in the previous chapter is discussed. As stated on the literature review chapter. The five stage model (Exploration, Experimentation, Exploitation, Execution, and Escape) is a commonly used model to categorize or classify an insider threat.

The dataset are not full for all higher officials such as Andrew S. Fastwow, Richard A.Causey and many others. All the data of Enron higher officials CEO, President, Vice President are found from the archive are investigated in terms of their word usage in their email communication. The words found in the email are classified as true negative and true positive.

In this research the true negative result found from Weka 3.7 used to classify the insider threat problem in line with category of Young, et al, (2013); Vectra, (2015); Ted et al, (2013) five stages and the psychological meaning of using more negative words in their communication.

In order to classify the level of insider threat from the analyzed data, the data set is divided into five cutoff points. The cutoff points are calculated as follows:

$$CutoffPo\text{int} = \frac{(Maximum - Minimum)}{Number\_of\_Class} \quad \text{..... (4.1)}$$

For this research we look the data with from the above five stages of insider threat. Due to this the numbers of classes or cutoff points are determined to five categories.

$Cutoffpo\text{int} = \frac{(43-2)}{5} = 8.2$ is the cutoff point for the class.

After sorting the result by ascending order the first cutoff point from 0 to 8.2. In order to calculate the second cutoff point adds 8.2 on the first cutoff point i.e. 8.2 to 16.4. For the third cutoff point add 8.2 on the second cutoff point i.e. 16.4 to 24.6 etc.

Table 5.1: Category of employee on the stages of insider threat

Class No	Class by cutoff Point by percentage (%)	No. employee	Stages of threat
2-8.2	2.83, 3.14, 3.68, 3.68, 3.82, 4, 4.11, 5.57, 6.1, 6.27, 6.28, 6.28,6.51, 6.59, 6.6, 6.79, 7.05, 7.43, 7.46, 8, 8, 80.8	22	Exploration
8.3-16.5		0	Experimentation
16.6-24.8	22.28	1	Exploitation
24.9-33.1	28.72, 29.47	2	Execution
33.2-	33.64, 33.7, 43.62	3	Escape/Envision

Table 5.1 shows that the 28 Enron higher official CEO, President and Vice presidents are categorized into five stages of insider threat. As the table 5.1 shows from the total of twenty eight (28) officials: twenty two (22) of them are categorized under exploration stage, none of them are at stages of experimentation, one (1) at the exploitation, two (2) at the Execution and three (3) of them on the stage of Escape/Envision. The level of risk the organization encounters increase from exploration to escape/envision.

Table 5.2 Employee classification on the stages of insider threat

Class No	Class by cutoffPoint for Threat words by percentage (%)	Stages of threat	Enron Officials
1	2.83, 3.14, 3.68, 3.68, 3.82, 4, 4.11, 5.57, 6.1, 6.27, 6.28, 6.28,6.51, 6.59, 6.6, 6.79, 7.05, 7.43, 7.46, 8, 8, 80.8	Exploration	Arora; Clair; Davis; Hayslett; Horton;Kitchen;Zipper;Zufferli;Williams; Whalley; Tholt; Sturm; Stepenovitch; Shively; Shapiro; Shankman;Presto; Neal; McCarty; Martin; Lavorato, kean
2		Experimentation	

3	22.28	Exploitation	Fossum.D
4	28.72, 29.47	Execution	Steffes Jams,Corman
5	33.64, 33.7, 43.62	Escape/Envision	D. Delainey, J.Skilling, Kenneth-Lay

5.2 Interpretation of the Result

In this section the interpretation of the result found from the analysis given in detail by supporting different literatures from human psychology, neurology, hormonal science and the risk level meet the organizations due to the presence of positive or negative emotion among the employee.

According to Talarovicova, Krskova, et al. (2007), Hariri, et al, (2002) and Broeren, et al, (2011), study seeing negative words through the video flash such as “NO” in a huge Functional magnetic resonance image (fMRI) machine within a microsecond point out unexpected or sudden discharge of many stress producing hormones and neurotransmitters. This discharged hormones and neurotransmitter chemicals right away disrupt the normal functioning of human brain, impairing logic, reason, language and communication processing, negative emotion such as depressed person feel worse fear, sadness, embarrassment, anger, guilt, disgust, and vulnerability, even can damage the vital structures that control human memory, feelings and emotion. This interrupts human long-term joy and pleasure. On the contrary the word “YES” helps to generate hormones and neurotransmitters that makes human happy, having positive emotion such as develop lifelong satisfaction, improves communication process, hardworking, persistent, responsible, careful, orderliness, plan full and organized.

As stated by Pennebaker, Mayne,et al.(1997) and Pennebaker, Mehl, (2003), when we use more positive words than negative words our health getting better and we have a better life. From these finding we can assert that our words we write or speak are determinate to our health and negative words clearly affect our health and mental condition due to suddenly produces hormones and neurotransmitter chemicals. Moreover, these chemicals are generated in response to internal and external stimuli’s. Both the internal and external stimulus causes can be their work environment such as get money, authority, knowledge, terrorism and getting promotion etc.

Research in organizational psychology (Maria Andries, 2011; Scherer, 2005) have shown that positive emotions are associated with increased creativity, cognitive flexibility, responsibility, labor productivity, carefulness, professional satisfactions, the availability of communication and negotiation skills, etc. Positive emotions, by organizing and constructive effects, improve the quality of work. Employees in situations experiencing pleasant emotions tend to set higher goals and to engage in constructive activity and generative ways. With regard to negative emotions, most experimental studies show harmful effects, such as restricting the repertoire of thinking, the tendency to process negative emotion and maintaining the debilitated cognitive schemes, work dissatisfaction, low emotional engagement tend to leave the organization.

As stated by Pennebaker, Mayne, et al.(1997) and Pennebaker, Mehl, (2003), negative words are indicators of negative emotion meaning that depressed person feel worse fear, sadness, embarrassment, anger, guilt, disgust, and vulnerability etc. As stated by Maria Andries,(20110, negative emotion incapacitate the cognitive schema, work dissatisfaction and increases the tendency of employees to leave or change their work.

Emotion in organization plays significant role how the organization communicates itself and the outside community. As stated by Ashkanasy& Ashton-James(2007), positively motivated employee favor their organization to obtain a better outcome such as achievement, profit, high quality social acceptance and job improvement. On the opposite negatively motivated employees' shows deviance mode such as fear, sadness, anger etc lead the employees to the deliberate desire to cause harm an organization and the workplace.

From the analysis section Table 5.2 shows that 28 Enron higher official CEO, President and Vice presidents are categorized into five stages of insider threat.

As the table 5.2 shows from the total of twenty eight (28) officials: twenty one 22 (78.57%) of them are categorized under exploration stage, none of them are at stages of Experimentation, one 1(3.57%) at the exploitation, two (7.14%) at the Execution and three 3 (10.71%) of them on the stage of Escape/Envision. The level of risk the organization encounters increase from exploration to escape/envision.

From table 5.2 the majority of the higher officials that accounts for 78.57% are classified at the *Exploration stage*. Of the total number of population on this stage none of them are charged with any official breaking rules and regulation.

As shown on table 5.2 none of them are classified into Experimentation stage. The reason this stage has became vacant in terms of employee is that there is big number difference among the employees being evaluated.

As shown on table *5.2* 3.57% employee are classified on Exploitation stage. Up to the researcher information there is no legal documents that shows they commit wrong doings and charged in the prison.

As shown on table 5.2 shows 7.14% employee was classified on Execution insider threat stage, the researcher not yet found legal documents that shows they are charged.

As shown on table 5.2 10.71% employees from the total population are classified on *Escape/Envision* level of insider threat.

From the total population of Enron higher official they are 28 in number including CEO, president and vice president. Off the total population 28(100%) classified as insider for the organization is that 21.42% of them are found on different stages of insider threat range from stage (*exploitation)* up to stage (*escape*). As per the classification technique we utilize off the total of 21.42%, classified 10.7% of them are charged and accepted their verdict in the court. The rest 3(10.7%) are not accused in the court.

In general from the above interpretation we can conclude that although the level of threat stages increases from the first stage (i.e. exploration) to the final level (i.e escape). It is difficult to make 100% sure that all of them are a threat to the organization. The classification result of the system can be used as an input to strengthen the year to be sentenced on the suspected employee. In addition to this other physical documents and extra investigation are used to verdict the employee. The result found from the classification is very promising because the three employees classified under stage five found guilty and sentenced from 1.5 up to 24.3 years in prison. The result of the classification indicates that those employees using more negative words

found on the savior stages of being insider threat and we can conclude that negative words are more indicative to negatively agitated employees.

5.3 Finding of the Study

In this study an extensive experiment had been done to classify insider threat into the five stages model. The classification has been done based on the negative word usage. The number of negative words found in the email conversation indicates the extent of being insider threat. From the result found most of the employees among the higher officials classified into exploration stage. Moreover, employees found on the escape stage are got their verdict in the court. The classification technique used the classify insider threat found significant. Though, from the result found I redefine insider threat according to their email word distribution behavior. An insider threat is a person who works for an organization, negatively agitated employee manifested through more negative words in their conversation.

5.4 Evaluation of the Classification Technique

There are a multiple number of patterns that can be discovered from the result of text classification task. In this research, the patterns discovered from the email texts found applicable to the classification of insider threat. As stated by (Young, et al,2013;Vectra ,2015; Ted et al, 2013; Smith, 2014) the stages from exploration to escape increases the chance of being an insider threat. As the result of this interpretation indicates that from the classified six higher official employees three of them are sentenced and serving in prison as classified on stage five. From the result the researcher found that, six of them are categorized from stage three up to five. Moreover, the result shows that there is clear boundary among the identified four stages (experimentation, exploitation, execution, escape/envision). The result shows that the three employees classified under stage five are charged and served in prison. This result is very promising and crucial to classify insider threat. The two employees classified under stage four and the one employee classified under stage two need strong follow up and control to avoid the risks against the organization.

As stated on literature review the cause for decease or fall of Enron Corporation is that the higher officials hide or cooked financial papers to the US financial intelligent agent. After the intelligent agent investigates the company was forced to file bankruptcy (Lucian, & Cristina, 2007). The

intelligent agent announces in the annals of USA, Enron Corporation executed the biggest fraud scandal. Due to these illegal tasks may of innocent employees lose their job and the corporation come closed down.

As stated on the specific objective of this research one of the specific objective is that to affirm the degree of disruption created by insider threat. The degree of disruption formed by these insiders was incalculable in terms of the tax and salary to be paid for the state and the employee respectively. The financial audit made on Arthurs Anderson office shows that $63 million were found fraudulent by the US intelligent office. According to Gillan, Stuart and Martin, John D. (November 2002), Using Enron's January 2001 stock price of $83.13 and the directors' beneficial ownership reported in the 2001 proxy, the value of director stock ownership was $659 million for Lay, and $174 million for Skilling.

The other specific objective is to look into the association between insider word usage and being threat or not. From the result we found that those higher officials use more negative words bring into being an insider threat.

From the result found the evaluation of the classification technique we utilized very crucial and promising. As discussed on the literature review section 2.2.2 the verdict given by the court as compared the result found highly positively correlated results. The level/intensity/extent insiders' disruption the services given to the others clearly indicated; meaning that these insiders accused for the collapse of the company.

CHAPTER SIX

CONCLUSION AND RECOMMENDATION

6.1 Conclusion

Electronic mail communication has been becoming the central element of the modern world. Daily millions of email communications have been made for several issues which range from simple personal greeting to high level international businesses. It is well known that after the advent of computer and internet several institutions and organization use internet as a means of

their daily channel to discuss about their job progress, report as well as personal issues. However, the emergences of such technologies have brought unprecedented counter technologies that are causing personal and institutional damage, i.e. hackers. On the other side of the face, such technologies have been playing a paramount role for the dynamic growth of the information technology world while efforts are exerted in response to the possible damage encountered due to hackers. Such hackers or intruders may reside in different corner of our planet to pose the possible attack. From the outside, network intruders or crackers becoming a growing problem. Within a given organization, insider threat poses a great deal of fear than external attack as stated by Jouini, et al.(2014); Kuheli Roy Sarkar, (2010); Solutions, (2014). In order to study the insider threat attack different methods are proposed. Among these one of the states of the art in psychometrics computing that scrutinize word usage using computer in the field of psychology. Many researchers Tausczik, et al (2010); Christopher et al., (2013); Liu,(2010) studied the correlation between the word usage and psychological emotion. The individuals word used for communication via electronically or paper based platform can be used to classify the psychological emotion of the person.

Accordingly organization can use suitable computer-based system that specifically employ text classification technology which bases on word type and uses in day to day email communication among workers. Henceforth, the email content must be analyzed by the organizations using the text classification technique. Consequently, this will help organizations to contrive possible strategies that can possibly mitigate the impact from insiders. In addition, it would further help to assess the organization spirit among workers.

In this research, an attempt has been made to apply text classification techniquein line with human language psychology in support of identifying and classify insider threat problem using the email data from careenage melon University data archive particularly Enron corporation (former American energy service corporation) organizational email archive. In text classification, basically we follow the following three steps: "Preprocessing" the text to refine the documents into a structured format; reducing the results into a more practical size and finally mining the reduced dataset. Lastly classification methods help us to interpret the newly discovered information context with human language psychology which in turn informs as to take a measure to protect our organization from insider threat.

Accordingly, this research uses the six step data cleaning (remove irrelevant directory, remove email header and salute, merge email content, remove punctuation, remove duplicated words, test words selected) and the supervised text classification method. The test data used in this research has been downloaded from carnage melon University data archive. The test data set was email archive of former American energy Service Corporation (i.e Enron). The dataset contain 150 employees of the corporation and from these 28 higher official used as a test set. The email archive has been taken for investigation with the reason of these officials was accountable for the fall of the company and they sentenced different years in the prison depending on their charge. The training set contains 9808 words which express positive and negative emotion. From 9808 words 4904 are negative and 4904 words are emotionally positive. The training data set is collected from different research sites. The data has been preprocessed and make suitable for text classification. The data preprocessing and technical challenges take 75% of the study period.

In the study the text classification done using SMO algorithms achieves better result. The number of words used by each individual during the email communication is classified as true positive and true negative class. Classification for being insider threat or not was made based on the number of true negative percentage over the total words used. Of course there are neutral words in our word usage, but the main goal of this research was to classify insider threat. To be regarded as a threat to an organization a number of test set of negative word distribution on the training data set were examined. As stated by different psychologist negatively agitated employees' shows deviation in terms of mode, such as fear, sadness, and anger etc. These may lead the employee to affect the organization performance deliberately. On the contrary positive words are indicative for the presence of better institutional environment where workers can excel their contribution in the organization. Moreover, the finding of this research indicates that not only low level employees harm their organization, but the higher official can also be a weed to the organization.

The study find out that out of the total 28 Enron higher official, six of them are classified as an insider threat. From the classified six employees as a threat three of them sentenced ranging from 1.5 up to 23 years in prison with the charge of conspiracy, wire fraud, insider trading, lying to auditor etc. It is obvious that making corruption by lying auditing report, conspiracy and many

other related problems harm the organization. The main source of motivation to commit the above crime is that internal and external psychological motivation or rewards.

In general the use of text classification technique and human language psychology bring us vital knowledge to classify insider threat. The results found from the experiment shows that it can be used as a supportive means to accuse a suspected employee with different charges in the court. In order to strength the study email data cleaner is needed highly. Moreover, increasing the training dictionary is very crucial to classify texts into the two statistical analysis classes.

6.2 Recommendation

In this research work an effort has been made to find out the potential applicability of email text classification techniques and human language psychology to examine the occurrence of insider threat within the organization. Although this research has been prepared for the academic purpose, its results disclosed for the potential application to be functional in addressing and mitigating insider threat problem.

Based on the findings of this research, the following recommendations are suggested or forwarded:

- ☞ This research shows that the email text classification technique in line with language psychology approves the classification of insider threat from word distribution. However,

this research uses a step-by-step data cleaning technique to make the email data suitable for classification. Due to the thread nature of email, extracting the required content makes the classification task difficult and time consuming. Hence, We recommend that developing integrated software for email cleaning which considers all the six step of data cleaning used in this research and the thread character of the email is vital to make the cumbersome task easy and lesser error in classification.

- Psychological emotions used in this research are positive and negative emotion. However the degree of negativity can be high, medium and low based on the word we use. We recommend that for language psychology researchers, preparing list of words categories that gives more classification on high, medium and low negative emotion.

- After insider threats are identified, investigating the association of the insider threat with other person in the organization through mail contact analysis (who send to whom) makes the classification more accurate by identifying the social circle of the insider threat.

- This research is done by using word level segmentation to classify insider threat. To improve the classification of the system, we recommend that using English phrasal verb (blow up, break down, do away with etc.) advance the classification.

- In this research the training dictionary is collected and prepared from different source. The number of positive and negative words class are totally 9808 and 4904 for each class. To increase the performance of classification increasing the number of training words enhances the potential applicability of the text classification technique.

- The training word set has limited number of synonym words. The researcher found that increasing the number of words similar to a single word could improve the classification performance.

- ☞ Email partners use informal words during their communication and abbreviated words. Henceforth, the researcher recommends that using more informal words in the dictionary enhances the performance of the system.

Reference

1. Andre. B. (2006, October 23). Enron Scorecard. Retrieved May 11, 2016, from http://www.nytimes.com/ref/us/20061023_ENRON_GRAPHIC.html
2. Chapman ,P., Clinton, J., Kerber ,R., Khabaza, T., Daimlerchrysler ,T. R., and Shearer ,C(.2000) "CRISP-DM 1.0: Step-by-step data mining guide." CRISP-DM consortium. SPSS.USA
3. Allen, J. H. (2001, May). Cert system and network security practices. In *Proceedings of the Fifth National Colloquium for Information Systems Security Education (NCISSE'01), George Mason University, Fairfax, VA USA* (pp. 22-24).
4. Andrieș, A. M. (2011). Positive and negative emotions within the organizational context. *Global Journal of Human Social Science, 11*(9).

5. Ashkanasy, N. M., & Ashton-James, C. E. (2007). Positive emotion in organizations: A multi-level framework. *Positive organizational behavior*, 57-73.
6. Axelrad, E. T., Sticha, P. J., Brdiczka, O., & Shen, J. (2013, May). A Bayesian network model for predicting insider threats. In *Security and Privacy Workshops (SPW), 2013 IEEE* (pp. 82-89).
7. Bertino, E., Fovino, I. N., & Provenza, L. P. (2005). A framework for evaluating privacy preserving data mining algorithms*. *Data Mining and Knowledge Discovery*, *11*(2), 121-154.
8. Bishop, M. (2005). *Introduction to computer security*. Boston, MA: Addison-Wesley.
9. Bishop, M., & Gates, C. (2008, May). Defining the insider threat. In *Proceedings of the 4th annual workshop on Cyber security and information intelligence research: developing strategies to meet the cyber security and information intelligence challenges ahead* (p. 15). ACM.
10. Bratton, W. W. (2001). Enron and the dark side of shareholder value. *Tul. L. Rev.*, *76*, 1275.
11. Broeren, S., Muris, P., Bouwmeester, S., van der Heijden, K. B., &Abee, A. (2011). The role of repetitive negative thoughts in the vulnerability for emotional problems in non-clinical children. *Journal of child and family studies*, *20*(2), 135-148.
12. Brown, C. R., Watkins, A., &Greitzer, F. L. (2013, January). Predicting insider threat risks through linguistic analysis of electronic communication. In *System Sciences (HICSS), 2013 46th Hawaii International Conference on* (pp. 1849-1858). IEEE.
13. C. W. Probst, J. Hunker, M. Bishop, and D. Gollmann, 2008. "Countering insider threats," Dagstuhl Seminar Proceedings.
14. Campbell, Q., & Kennedy, D. M. (2009). The Psychology of Computer Criminals. *Computer security handbook*.
15. Carroll, M. D. (2006, September). Information security: examining and managing the insider threat. In *Proceedings of the 3rd annual conference on Information security curriculum development* (pp. 156-158). ACM.
16. Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nded.). Hillsdale, NJ: Lawrence Erlbaum Associates.
17. Colwill, C. (2009). Human factors in information security: The insider threat – Who can you trust these days? *Information Security Technical Report, 14*(4), 186-196. doi:10.1016/j.istr.2010.04.004

18. Czarnowski, I., &Jędrzejowicz, P. (2014). Ensemble classifier for mining data streams. *Procedia Computer Science*, *35*, 397-406.
19. Daft, R. L. (2000). Management, 5 de druk., pp. 670.
20. Ezingeard, J. N., McFadzean, E., &Birchall, D. (2005). A model of information assurance benefits. *Information Systems Management*, *22*(2), 20-29. pp. 20-29.
21. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37.
22. Frawley, W. J., &Matheus, C. J. (1991). *Knowledge discovery in databases* (pp. 1-27). G. Piatetsky-Shapiro (Ed.). Menlo Park, CA: AAAI Press.
23. Freud, S. (1923/1949). The ego and the id. London, England: Hogarth Press. (Original work published 1923)
24. Girma, Aweke. (2012). *predicting hiv infection risk factor using volantory counseling and testing data: a case of african aids initiative international (aaii)* (Unpublished Master's Thesis). Addis Ababa University.
25. Gavai, G., Sricharan, K., Gunning, D., Hanley, J., Singhal, M., & Rolleston, R. (2015). Supervised and Unsupervised methods to detect Insider Threat from Enterprise Social and Online Activity Data. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA)*, *6*(4), 47-63.
26. Gillan, Stuart and Martin, John D. (November 2002)., Financial Engineering, Corporate Governance, and the Collapse of Enron
27. Greitzer, Frank L., Ph.D. and Hohimer, Ryan E.. (2011): "Modeling Human Behavior to Anticipate Insider Attacks." *Journal of Strategic Security* 4, no. 2 (25-48).
28. Hamin, Z. (2000). Insider cyber-threats: Problems and perspectives. *International Review of Law, Com-puters& Technology*, *14*(1), 105-113. doi:10.1080/13600860054944
29. Han, J. (2006). Kamber, Data mining: concepts and techniques.
30. Hariri, A. R., Tessitore, A., Mattay, V. S., Fera, F., & Weinberger, D. R. (2002). The amygdala response to emotional stimuli: a comparison of faces and scenes. *Neuroimage*, *17*(1), 317-323
31. http://google-dictionary.so8848.com/ retrieved on February 10,2016
32. http://positivewordsresearch.com /list-of-negative-words/ retrieved on march 14,2016

33. http://positivewordsresearch.com/list-of-positive-words retrieved on March 14,2016
34. http://systemagicmotives.com/Positive%20Noun%20Glossary.htm retrived on March 11, 2016.
35. http://www.businessdictionary.com/definition/threat.html retrieved on February 4,2016
36. http://www.etymonline.com/index.php?term=threat retrieved on February 4,2016
37. http://www.macmillandictionary.com/thesaurus-category/british/psychology-and-psychoanalysisretrieved on February 5,2016
38. *http://www.merriam-webster.com/dictionary/threat* retrieved on February 5,2016
39. http://www.ncsc.gov/issues/ithreat/retrieved on February 4,2016
40. Hu, M., & Liu, B. (2004, August). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 168-177). ACM.
41. ISO/IEC 15408 (2005(E)). Information technology Security techniques Evaluation criteria for IT security. ISO/IEC Switzerland.
42. Janus, J. (2010). Email communication of Jeffrey Janus (ISCO with the author.
43. John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, *2*(1999), 102-138.
44. Jouini, M., Rabai, L. B. A., & Aissa, A. B. (2014). Classification of security threats in information systems. *Procedia Computer Science*, *32*, 489-496.
45. Kamruzzaman, S. M., Haider, F., & Hasan, A. R. (2010). Text classification using data mining. *arXiv preprint arXiv:1009.4987*.
46. Kotsiantis, S. B., Kanellopoulos, D., & Pintelas, P. E. (2006). Data preprocessing for supervised leaning. *International Journal of Computer Science*, *1*(2), 111-117.
47. Krauss, R. M., & Chiu, C. Y. (1998). Language and social behavior. *The handbook of social psychology*, *2*, 41-88.
48. Kuheli Roy Sarkar (2010), Assessing insider threats to information security using technical, behavioral and organizational measures. Information security technical report. doi:10.1016/j.istr.2010.11.002

49. Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159-174.
50. Liu, B. (2010). Sentiment Analysis and Subjectivity. *Handbook of natural language processing*, *2*, 627-666.
51. Liu, B., Hu, M., & Cheng, J. (2005, May). Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web* (pp. 342-351). ACM.
52. Lucian, C., & Cristina, D. (2007). Fraud Case Analysis: Enron Corporation.
53. Maria Andries, A. (2011). Positive and Negative Emotions within the Organizational Context. *Global Journal of Human-Social Science Research*, *11*(9).
54. Ophoff, J., Jensen, A., Sanderson-Smith, J., Porter, M., & Johnston, K. (2014). A descriptive literature review and classification of insider threat research. *Proceedings of Informing Science & IT Education Conference (InSITE) 2014* (pp. 211-223). Retrieved fromhttp://Proceedings.InformingScience.org/InSITE2014/InSITE14p211-223Ophoff0543.pdf
55. Osman, T. T., & A/Nabi Mustafa, D. B. (mar-april 2015). Internal & External Attacks in cloud computing Environment from confidentiality, integrity and availability points of view. *IOSR Journal of Computer Engineering (IOSR-JCE),17*(2), 93-96. doi:10.9790/0661-17259396
56. Parveen, P., Weger, Z. R., Thuraisingham, B., Hamlen, K., & Khan, L. (2011, November). Supervised learning for insider threat detection using stream mining. In *Tools with Artificial Intelligence (ICTAI), 2011 23rd IEEE International Conference on* (pp. 1032-1039). IEEE.
57. Pavel, T., Encontro, M., &Mve, F. R. The Enron scandal.
58. Pennebaker, J. W., &Graybeal, A. (2001). Patterns of natural language use: Disclosure, personality, and social integration. *Current Directions in Psychological Science*, *10*(3), 90-93.
59. Pennebaker, J. W., Mayne, T. J., & Francis, M. E. (1997). Linguistic predictors of adaptive bereavement. *Journal of personality and social psychology*, *72*(4), 863.
60. Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, *54*(1), 547-577.

61. Pfleeger, S. L., Predd, J. B., Hunker, J., & Bulford, C. (2010). Insiders behaving badly: Addressing bad actors and their actions. *Information Forensics and Security, IEEE Transactions on*, *5*(1), 169-179.
62. Platt, John (1998).Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. MIT Press.
63. Post, J. M., Ruby, K. G., & Shaw, E. D. (2000). From car bombs to logic bombs: The growing threat from information terrorism. *Terrorism and Political Violence*, *12*(2), 97-122.
64. Radovanović, M., &Ivanović, M. (2008). Text mining: Approaches and applications. *Novi Sad J. Math*, *38*(3), 227-234.
65. Retrived from: https://www.cs.cmu.edu/~./enron/ November 10-2015.
66. Rothmann, S., & Coetzer, E. P. (2003). The big five personality dimensions and job performance. *SA Journal of Industrial Psychology*, *29*(1), p-68.
67. Scherer, K. R. (2005). What are emotions? And how can they be measured?.*Social science information*, *44*(4), 695-729.
68. Schultz, E. E. (2002). A framework for understanding and predicting insider attacks. *Computers & Security*, *21*(6), 526-531.
69. Shen, J., Brdiczka, O., & Liu, J. (2013). Understanding email writers: Personality prediction from email messages. In *User Modeling, Adaptation, and Personalization* (pp. 318-330). Springer Berlin Heidelberg.
70. Smith, G. (2014). Combating Insider Threat. *Coast Guard Journal of Safety & Security at Sea, Proceedings of the Marine Safety & Security Council*, *71*(4).
71. Solutions, V. E. (2014). Verizon 2014 data breach investigations report. *verizon. com*.
72. Srivatsa, s. (2013). Segmentation of customers for prediction analysis using supervised and unsupervised data mining techniques.
73. Stansfeld, S., &Rasul, F. (2007). Psychosocial factors, depression and illness. In A. Steptoe (Ed.), *Depression and physical illness* (pp. 19–52). Cambridge: Cambridge University Press.
74. Tagg, G. L. (2014)The Insider Threat. *Computer Security Handbook, Sixth Edition*, 13-1.
75. Talarovicova, A., Krskova, L., & Kiss, A. (2007). Some assessments of the amygdala role in suprahypothalamic neuroendocrine regulation: a minireview. *Endocrine regulations*, *41*(4), 155-162.

76. Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, *29*(1), 24-54.

77. Threat, vulnerability, risk - commonly mixed up terms - INDEPENDENT SECURITY CONSULTANTS. (2010). Retrieved January 01, 2016, from http://www.threatanalysis.com/2010/05/03/threat-vulnerability-risk-commonly-mixed-up-terms/

78. Young ,Ted, E., Goldberg, H. G., Memory, A., W. T., Rees, B., Pierce, R., ...&Essa, I. (2013, August). Detecting insider threats in a real corporate database of computer usage activity. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1393-1401). ACM.

79. Thompson, B. (2006). *Foundations of behavioral statistics: An insight-based approach.* New York: Guilford.

80. Vectra. (2015). Detect Insider Attacks in Real Time. Retrieved May 10, 2016, from www.action-one.ch/fileadmin/.../Vectra_White-Paper_Insider-Threat-Detection.pdf.

81. Veropoulos, K., Campbell, C., &Cristianini, N. (1999, July). Controlling the sensitivity of support vector machines. In *Proceedings of the international joint conference on AI* (pp. 55-60).

82. Verykios, V. S., Bertino, E., Fovino, I. N., Provenza, L. P., Saygin, Y., &Theodoridis, Y. (2004). State-of-the-art in privacy preserving data mining. *ACM Sigmod Record*, *33*(1), 50-57

83. Young, W. T., Goldberg, H. G., Memory, A., Sartain, J. F., & Ted, E. (2013, May). Use of domain knowledge to detect insider threats in computer activities. In *Security and Privacy Workshops (SPW), 2013 IEEE* (pp. 60-67). IEEE

Appendix A: List of Enron Higher officials Extracted from the Archive

No	LastName	FirstName	E-Mail1	E-Mail2	Archive name	status
1	Arora	Harry	harry.arora@enron.com	harora@enron.com	arora-h	Vice President
2	Clair	Carol	carol.clair@enron.com	carol.st.clair@enron.comom	stclair-c	Vice President
3	Corman	Shelley	shelley.corman@enron.com	scorman@enron.com	corman-s	Vice President
4	Davis	Dana	dana.davis@enron.com	dana_davis@enron.com	davis-d	Vice President
5	Delainey	David	david.w.delainey@enron.com	david.delainey@enron.com	delainey-d	CEO
6	Fossum	Drew	drew.fossum@enron.com	drew_fossum@enron.com	fossum-d	Vice President
7	Hayslett	Rod	rod.hayslett@enron.com	rod_hayslett@enron.com	hayslett-r	Vice President
8	Horton	Stanley	stanley.horton@enron.com	stanley_horton@enron.com	horton-s	President
9	Kean	Steven	steven.j.kean@enron.com	steven.kean@enron.com	kean-s	Vice President
10	Kitchen	Louise	louise.kitchen@enron.com	lkitchen@enron.com	kitchen-l	President
11	Lavorato	John	lavorato@enron.com	john.j.lavorato@enron.com	lavorato-j	CEO
12	Lay	Kenneth	kenneth.lay@enron.com	kenneth_lay@enron.com	lay-k	CEO
13	Martin	Thomas	thomas.a.martin@enron.com	thomas.martin@enron.com	martin-t	Vice President
14	McCarty	Danny	danny.mccarty@enron.com		mccarty-d	Vice President
15	Neal	Scott	scott.neal@enron.com		neal-s	Vice President
16	Presto	Kevin	kevin.m.presto@enron.com	m.presto@enron.com	presto-k	Vice President
17	Shankman	Jeffrey	jeffrey.a.shankman@enron.com	a.shankman@enron.com	shankman-j	President
18	Shapiro	Richard	richard.shapiro@enron.com		shapiro-r	Vice President
19	Shively	Hunter	hunter.s.shively@enron.com	s.shively@enron.com	shively-h	Vice President
20	Skilling	Jeffrey	jeff.skilling@enron.com	jeffrey.skilling@enron.com	skilling-j	CEO
21	Steffes	James	james.d.steffes@enron.com	d..steffes@enron.com	steffes-j	Vice President
22	Stepenovitch	Joe	joe.stepenovitch@enron.com	joe_stepenovitch@enron.com	stepenovitch-j	Vice President
23	Sturm	Fletcher	fletcher.j.sturm@enron.com	fletcher.sturm@enron.com	sturm-f	Vice President
24	Tholt	Jane	jane.tholt@enron.com	jane.m.tholt@enron.com	tholt-j	Vice President
25	Whalley	Greg	greg.whalley@enron.com	gwhalley@enron.com	whalley-g	President

Appendix A: List of Enron Higher officials Extracted from The Archive						
26	Williams	Jason	jason.williams@enron.com	jason.r.williams@enron.com	williams-j	Vice President
27	Zipper	Andy	andy.zipper@enron.com		zipper-a	Vice President
28	Zufferli	John	john.zufferli@enron.com		zufferli-j	Vice President

Appendix B:Program Module

Program 1: A program that removes punctuation and sort words

```
'''A python program that remove/erase punctuation marks and sort words alphabetically for looking the words are duplicated or not. '''
#Developed by Firesenbet Adela
punctuations = '''!',()-[]{};:'"\,<>./?@#$%^&*_~=1234567890'''
#The following line of code take a string form the user
listofwords_str = input("Enter a string: ")
no_punct = ""
for char in listofwords:
   if char not in punctuations:
no_punct = no_punct + char
''' print no_punct Program to sort alphabetically the words form a string provided by the user. '''
''' In this program, we take a string form the user. Using the split() method the string is converted into a list of words. The split() method splits the string at whitespaces. The list of words is then sorted using the sort() method and all the words are displayed. '''
#take input from the user breakdown the string into a list of words

email = no_punct.split()
# sort the list .lower() returns a version with all upper case #characters #replaced with lower case characters.
email.sort(key = lambda k : k.lower())
# display the sorted words
for word in email:
   print(word)
```

Program 2: A program that removes duplicated words

```
'''The list of words must be arranged in alphabet with new line'''
#Developed by Firesenbet Adela
''' The following (Import re) is a regular expression(RE) specifies a set of strings that matches it; the functions in this module let you check if a particular string matches a given regular expression'''
import re
data =""" list of words to be cheeked taken from Python program 1"""
# use f.readlines() instead if reading from file
lines = data.splitlines()
# split the words and only take ones that are all alpha
words = filter(lambda x: re.match('^[^\W\d]+$', x), lines)
# remove duplicates and print out
print ('\n'.join(set(words)))
```

Python Program 3 : A program that removes selected directories

```
``` This program code is written by Firesenbet Adela. This program deletes selected directories. Please cite the source when you use this program for your academic purpose.```

import os
emptyDirs = []
#the path must be double slash in order to solve“Unicode Error ”unicodeescape" codec can't decode bytes…
#Cannot open text files in Python 3
path="C:\\Users\\Firesenbet\\Desktop\\enron_mail_20150507~\\maildir\\arnold-j\\inbox"

defdeleteFiles(dirList, dirPath):
    for file in dirList:
        print ("Deleting " + file)
```
```

```
os.remove(dirPath + "/" + file)


defremoveDirectory(dirEntry):
    print ("Deleting files in " + dirEntry[0])
deleteFiles(dirEntry[2], dirEntry[0])
emptyDirs.insert(0, dirEntry[0])

tree = os.walk(path)
for directory in tree:
removeDirectory(directory)

for dir in emptyDirs:
    print ("Removing " + dir)
os.rmdir(dir)
```

Python program 4: A program that remove header from the email

```
import sys, os, re, StringIO
import email, mimetypes


invalid_chars_in_filename='<>:"/\\|?*\%\"+reduce(lambda x,y:x+chr(y), range(32), '')
invalid_windows_name='CON PRN AUX NUL COM1 COM2 COM3 COM4 COM5 COM6 COM7 COM8 COM9 LPT1 LPT2 LPT3 LPT4 LPT5 LPT6 LPT7 LPT8 LPT9'.split()


atom_rfc2822=r"[a-zA-Z0-9_!#\$\%&'*+/=?\^`{}~|\-]+"
atom_posfix_restricted=r"[a-zA-Z0-9_#\$&'*+/=?\^`{}~|\-]+" # without '!' and '%'
atom=atom_rfc2822
dot_atom=atom + r"(?:\." + atom + ")*"
quoted=r'"(?:\\[^\r\n]|[^\\"])*"'
local="(?:" + dot_atom + "|" + quoted + ")"
domain_lit=r"\[(?:\\\S|[\x21-\x5a\x5e-\x7e])*\]"
```

```
domain="(?:" + dot_atom + "|" + domain_lit + ")"
addr_spec=local + "\@" + domain
email_address_re=re.compile('^'+addr_spec+'$')
class Attachment:
def __init__(self, part, filename=None, type=None, payload=None, charset=None, content_id=None,
description=None, disposition=None, sanitized_filename=None, is_body=None):
self.part=part          # original python part
self.filename=filename  # filename in unicode (if any)
self.type=type          # the mime-type
self.payload=payload    # the MIME decoded content
self.charset=charset    # the charset (if any)
self.description=description    # if any
self.disposition=disposition    # 'inline', 'attachment' or None
self.sanitized_filename=sanitized_filename# cleanup your filename here (TODO)
self.is_body=is_body        # usually in (None, 'text/plain' or 'text/html')
self.content_id=content_id# if any
        if self.content_id:
# strip '<>' to ease searche and replace in "root" content (TODO)
            if self.content_id.startswith('<') and self.content_id.endswith('>'):
self.content_id=self.content_id[1:-1]
defgetmailheader(header_text, default="ascii"):
    """Decode header_text if needed"""
    try:
        headers=email.Header.decode_header(header_text)
    except email.Errors.HeaderParseError:


        return header_text.encode('ascii', 'replace').decode('ascii')
    else:
        for i, (text, charset) in enumerate(headers):
            try:
                headers[i]=unicode(text, charset or default, errors='replace')
```

```
            except LookupError:
# if the charset is unknown, force default
                headers[i]=unicode(text, default, errors='replace')
        return u"".join(headers)
defgetmailaddresses(msg, name):
    """retrieve addresses from header, 'name' supposed to be from, to,  ..."""
addrs=email.utils.getaddresses(msg.get_all(name, []))
    for i, (name, addr) in enumerate(addrs):
        if not name and addr:
# only one string! Is it the address or is it the name ?
# use the same for both and see later
            name=addr
        try:
# address must be ascii only
addr=addr.encode('ascii')
        except UnicodeError:
addr=''
        else:
# address must match address regex
            if not email_address_re.match(addr):
addr=''
addrs[i]=(getmailheader(name), addr)
    return addrs
defget_filename(part):
"""Many mail user agents send attachments with the filename in  the 'name' parameter of the 'content-type' header instead of in the 'filename' parameter of the 'content-disposition' header."""
    filename=part.get_param('filename', None, 'content-disposition')
    if not filename:
filename=part.get_param('name', None)
# default is 'content-type'
```

```
    if filename:
# RFC 2231 must be used to encode parameters inside MIME header
        filename=email.Utils.collapse_rfc2231_value(filename).strip()

    if filename and isinstance(filename, str):
# But a lot of MUA erroneously use RFC 2047 instead of RFC 2231
# in fact anybody miss use RFC2047 here !!!
        filename=getmailheader(filename)
    return filename
def _search_message_bodies(bodies, part):
    """recursive search of the multiple version of the 'message' inside
    the the message structure of the email, used by search_message_bodies()"""

type=part.get_content_type()
if type.startswith('multipart/'):
# explore only True 'multipart/*'
# because 'messages/rfc822' are also python 'multipart'
        if type=='multipart/related':
# the first part or the one pointed by start
            start=part.get_param('start', None)
related_type=part.get_param('type', None)
            for i, subpart in enumerate(part.get_payload()):
                if (not start and i==0) or (start and start==subpart.get('Content-Id')):
                    _search_message_bodies(bodies, subpart)
                    return
elif type=='multipart/alternative':
# all parts are candidates and latest is best
            for subpart in part.get_payload():
                _search_message_bodies(bodies, subpart)
elif type in ('multipart/report',  'multipart/signed'):
# only the first part is candidate
```

```
            try:
                subpart=part.get_payload()[0]
            except IndexError:
                return
            else:
                _search_message_bodies(bodies, subpart)
                return
elif type=='multipart/signed':
# cannot handle this
            return
        else:
# unknown types must be handled as 'multipart/mixed'
# This is the peace of code could probably be improved, I use a heuristic :
# - if not already found, use first valid non 'attachment' parts found
            for subpart in part.get_payload():
tmp_bodies=dict()
                _search_message_bodies(tmp_bodies, subpart)
                for k, v in tmp_bodies.iteritems():
                    if not subpart.get_param('attachment', None, 'content-disposition')==":
# if not an attachment, initiate value if not already found
bodies.setdefault(k, v)
            return
    else:
        bodies[part.get_content_type().lower()]=part
        return

    return

defsearch_message_bodies(mail):
    """search message content into a mail"""
    bodies=dict()
```

```
    _search_message_bodies(bodies, mail)
    return bodies


defget_mail_contents(msg):
    """split an email in a list of attachments"""
    attachments=[]
# retrieve messages of the email
    bodies=search_message_bodies(msg)
# reverse bodies dict
    parts=dict((v,k) for k, v in bodies.iteritems())
# organize the stack to handle deep first search
    stack=[ msg, ]
    while stack:
        part=stack.pop(0)
type=part.get_content_type()
if type.startswith('message/'):
# ('message/delivery-status', 'message/rfc822', 'message/disposition-notification'):
# I don't want to explore the tree deeper her and just save source using msg.as_string()
# but I don't use msg.as_string() because I want to use mangle_from_=False
            from email.Generator import Generator
fp = StringIO.StringIO()
            g = Generator(fp, mangle_from_=False)
g.flatten(part, unixfrom=False)
            payload=fp.getvalue()
            filename='mail.eml'
attachments.append(Attachment(part, filename=filename, type=type, payload=payload,
charset=part.get_param('charset'), description=part.get('Content-Description')))
elifpart.is_multipart():
# insert new parts at the beginning of the stack (deep first search)
            stack[:0]=part.get_payload()
        else:
```

```
            payload=part.get_payload(decode=True)
            charset=part.get_param('charset')
            filename=get_filename(part)
            disposition=None
            if part.get_param('inline', None, 'content-disposition')==":
                disposition='inline'
elifpart.get_param('attachment', None, 'content-disposition')==":
                disposition='attachment'
attachments.append(Attachment(part, filename=filename, type=type, payload=payload, charset=charset,
content_id=part.get('Content-Id'),description=part.get('Content-Description'),disposition=disposition,
is_body=parts.get(part)))
    return attachments
defdecode_text(payload, charset, default_charset):
    if charset:
        try:
            return payload.decode(charset), charset
        except UnicodeError:
            pass
    if default_charset and default_charset!='auto':
        try:
            return payload.decode(default_charset), default_charset
        except UnicodeError:
            pass
    for chset in [ 'ascii', 'utf-8', 'utf-16', 'windows-1252', 'cp850' ]:
        try:
            return payload.decode(chset), chset
        except UnicodeError:
            pass
    return payload, None
if __name__ == "__main__":
    raw="""content here"""
```

```
    if len(sys.argv)>1:
        raw=open(sys.argv[1]).read()


msg=email.message_from_string(raw)
    attachments=get_mail_contents(msg)
    for attach in attachments:
# dont forget to be careful to sanitize 'filename' and be carefull
# for filename collision, to before to save :
#print '\tfilename=%r is_body=%s type=%s charset=%s desc=%s size=%d' % (attach.filename, attach.is_body, attach.type, attach.charset, attach.description, 0 if attach.payload==None else len(attach.payload))
        if attach.is_body=='text/plain':
# print first k lines
            payload, used_charset=decode_text(attach.payload, attach.charset, 'auto')
            for line in payload.split('\n')[k ]:
# be careful console can be unable to display unicode characters
                if line:
                    print '\t\t', line
```

Python program 5: A program that combines multiple email files

```
''' This python program is developed by the researcher in order to merge multiple number of email files into a single file '''
import glob
read_files=glob.glob("C:\Users\Firesenbet\Desktop\enron_mail_20150507~\maildir\prestok\sent_items .txt")
efilenames = ['C:\Users\Firesenbet\Desktop\enron_mail_20150507~\maildir\presto-k\sent_items\1',
            'C:\Users\Firesenbet\Desktop\enron_mail_20150507~\maildir\presto-k\sent_items\2',
            'C:\Users\Firesenbet\Desktop\enron_mail_20150507~\maildir\presto-k\sent_items\3' ]
```

```
with open(' C:\Users\Firesenbet\Desktop\enron_mail_20150507~\maildir\presto-k\presto.txt', 'w') as outfile:
    for fname in efilenames:
        with open(fname) as infile:
            for line in infile:
outfile.write(line)
```

Python program 6: A program that erases punctuation and splitting sentences into word level

```
#the following program erase punctuation and by splitting sentences sort the words alphabetically
punctuations = '''!',()-[]{};:'"\,<>./?@#$%^&*_~=1234567890'''
my_str = input("Enter a string: ")
no_punct = ""
for char in my_str:
   if char not in punctuations:
no_punct = no_punct + char
  print no_punct
''' Program to sort alphabetically the words form a string provided by the user In this program, we take a string form the user. Using the split() method the string is converted into a list of words. The split() method splits the string at whitespaces.The list of words is then sorted using the sort() method and
all the words are displayed. take input from the user breakdown the string into a list of words'''
email = no_punct.split()

# sort the list
## .lower() returns a version with all upper
#case characters replaced with lower case characters.
email.sort(key = lambda k : k.lower())
```

```
# display the sorted words
for word in email:
   print(word)
```

Python program 7: A program that identify unique words only

```
import re

data ="""" list of words""""
# use f.readlines() instead if reading from file
lines = data.splitlines()

# split the words and only take ones that are all alpha
words = filter(lambda x: re.match('^[^\W\d]+$', x), lines)
# remove duplicates and print out
print ('\n'.join(set(words)))
```