

A MACHINE LEARNING APPROACH TO DETECT INSIDER THREATS IN EMAILS CAUSED BY HUMAN BEHAVIOUR

Dissertation

by

ANTONIA MICHAEL

Submitted in fulfilment of the requirements for the degree

MASTER OF SCIENCE (COMPUTER SCIENCE)

in the

**FACULTY OF ENGINEERING, BUILT ENVIRONMENT AND INFORMATION
TECHNOLOGY**

at the

**UNIVERSITY OF PRETORIA
PRETORIA, SOUTH AFRICA**

SUPERVISOR: Prof. J.H.P. Eloff

November 2020

Acknowledgements

This dissertation would not have been possible without the Lord's grace, abundant blessings and inspiration. I would also like to express my gratitude to the following people who provided me with assistance and a great deal of support:

- Prof J.H.P. Eloff, my supervisor, for his patience, guidance, expertise, critical insight, unwavering support and for his belief in my ability. I have learnt such a lot through this process and I am ever so thankful to have had Prof as my supervisor. I am also grateful for the opportunity he afforded me to attend the HAISA conference in Cyprus and to publish in the proceedings, as well as in the Emerald journal, Information and Computer Security.
- Language editor, Ms Isabel Claassen, for her efficient and high-quality editing service, as well as her advice.
- Prof Steven Furnell and Prof Nathan Clarke from the HAISA Conference, for providing me with the opportunity to present my research at the conference and their assistance with the journal publication.
- The lecturers in the Department of Computer Science at UP, for laying the foundation of my academic career.
- Ms Stacey Baror and colleagues in the ICSA Research Group of UP who encouraged me to pursue my Master's degree.
- My family for their consistent love, care and support as well as their motivation and guidance to help me achieve this milestone.
- My friends and fellow students who supported and advised me during this process.

"Human behaviour flows from three main sources: desire, emotion and knowledge" - Plato

Abstract

In recent years, there has been a significant increase in insider threats within organisations and these have caused massive losses and damages. Due to the fact that email communications are a crucial part of the modern-day working environment, many insider threats exist within organisations' email infrastructure. It is a well-known fact that employees not only dispatch 'business-as-usual' emails, but also emails that are completely unrelated to company business, perhaps even involving malicious activity and unethical behaviour. Such insider threat activities are mostly caused by employees who have legitimate access to their organisation's resources, servers, and non-public data. However, these same employees abuse their privileges for personal gain or even to inflict malicious damage on the employer. The problem is that the high volume and velocity of email communication make it virtually impossible to minimise the risk of insider threat activities, by using techniques such as filtering and rule-based systems. The research presented in this dissertation suggests strategies to minimise the risk of insider threat via email systems by employing a machine-learning-based approach. This is done by studying and creating categories of malicious behaviours posed by insiders, and mapping these to phrases that would appear in email communications. Furthermore, a large email dataset is classified according to behavioural characteristics of employees. Machine learning algorithms are employed to identify commonly occurring insider threats and to group the occurrences according to insider threat classifications.

Keywords: cybersecurity, insider threats, insider threat detection, big data, emails, corporate, machine learning

DECLARATION

I herewith declare that I,

Isabel M Claassen (APSTrans (SATI)),

full-time freelance translator, editor and language consultant

of

1367 Lawson Avenue, Waverley, Pretoria

(cell 082 701 7922)

and

accredited member (No. 1000583) of the South African Translators' Institute (SATI)

completed the *language editing** of an MComp. dissertation entitled

A MACHINE LEARNING APPROACH TO DETECT INSIDER THREATS IN EMAILS CAUSED BY HUMAN BEHAVIOUR

which had been submitted to me by

Ms Antonia Michael

E-mail: **tonia.michael94@gmail.com**

Date completed: 06-11-2020

****Please note that no responsibility can be taken for the veracity of statements or arguments in the document concerned or for changes made subsequent to the completion of language editing. Also remember that content editing is not part of a language editor's task and is in fact unethical.***

Contents

Contents.....	5
Chapter 1.....	10
Introduction	10
1.1 Introduction	10
1.2 Dissertation statement	12
1.3 Problem statement	13
1.4 Objectives and research questions	13
1.5 Scope and context of the study	14
1.6 Methodology.....	15
1.6.1 Literature survey	15
1.6.2 Data gathering.....	15
1.6.3 Design of prototype	15
1.6.4 Prototype implementation, experimentation and results gathering	15
1.7 Terminology	16
1.7.1 Cybersecurity	16
1.7.2 Insider threat.....	16
1.7.3 Machine learning	16
1.8 Layout of the dissertation	17
Chapter 2.....	19
Use Case: The Enron Email Corpus	19
2.1 Introduction	19
2.2 The chosen dataset	19
2.3 Enron: Company profile and background	20
2.4 The Enron email dataset	21
2.5 Insider threats detectable in Enron email communications.....	25
2.5.1 Disgruntlement leading to insider threats.....	25
2.5.2 Fraudulent activities as insider threats.....	27
2.5.3 Insiders benefiting from confidential information whilst concealing the truth from investors.....	31
2.5.4 Insider threats accidentally constituted by non-malicious employees	32
2.6 Conclusion.....	33
Chapter 3.....	35
Defining Insider Threats.....	35

3.1	Introduction	35
3.2	Definition of an insider threat.....	35
3.2.1	An introduction to insiders	35
3.2.2	Defining insiders from a cybersecurity perspective.....	37
3.2.3	An introduction to threats	40
3.2.4	Defining threats from a cybersecurity perspective.....	42
3.2.5	Defining insider threat from a cybersecurity perspective	44
3.3	Preparing the context-specific insider threat definition.....	48
3.4	Proposed definition of insider threat for use in this research.....	48
3.5	Conclusion.....	52
Chapter 4.....		53
Insider Threat Behaviours		53
4.1	Introduction	53
4.2	Emails as a platform for attacks.....	53
4.3	Insider threats in organisations	54
4.4	Characteristics of insider threats.....	54
4.5	Insider threats driven by human behaviour	58
4.6	Conclusion.....	64
Chapter 5.....		65
Types of insider threat		65
5.1	Introduction	65
5.2	Categories of insider threats.....	65
5.2.1.	Insider Information Technology (IT) sabotage.....	65
5.2.2.	Insider intellectual property (IP) theft	68
5.2.3.	Insider fraud	71
5.2.4.	Negligence.....	73
5.3	Existing research on approaches to detect insider threats within emails	76
5.3.1.	Clustering approach	76
5.3.2.	Clustering and classification – a combined approach.....	77
5.3.3.	Human-behaviour-based approaches to insider threat discovery	79
5.3.4.	Key components of prototypes developed in past research	80
5.4	Conclusion.....	80
Chapter 6.....		82
Requirements for Prototype Development		82

6.1	Introduction	82
6.2	Overview	82
6.3	Functional requirements.....	83
6.4	Technical requirements	87
6.5	Conclusion.....	89
Chapter 7.....		90
Prototype to Detect Insider Threats in Corporate Emails.....		90
7.1	Introduction	90
7.2	Process overview	90
7.3	Data preparation.....	92
7.3.1.	Executable: Email data gathering	93
7.3.2.	Document: Email body.....	93
7.3.3.	Executable: Email body normalisation.....	93
7.3.4.	Executable: Email body cleaning.....	96
7.3.5.	Executable: Email data preparation.....	97
7.3.6.	Document: Pre-processed email body dataset.....	97
7.4	Data discovery.....	97
7.4.1.	Document: List of insider threat types	99
7.4.2.	Document: Phrases based on insider threat types.....	99
7.4.3.	Document: Centroid emails of phrases.....	99
7.4.4.	Document: Wordlists of phrases.....	101
7.4.5.	Executable: Supervised machine learning algorithms	104
7.4.6.	Executable: Performance metrics.....	106
7.5	Detection.....	106
7.5.1	Document: Machine learning model	107
7.5.2	Executable: Evaluate results	107
7.6	Conclusion.....	108
Chapter 8.....		109
Results Obtained from the Prototype Experimentation.....		109
8.1	Introduction	109
8.2	Research question.....	109
8.3	Experimentation	109
8.3.1	Experimentation environment.....	110
8.3.2	Experimentation data	110

8.3.3	Experimentation iterations	111
8.3.4	Data preparation results	112
8.3.4.1	<i>Attribute selection</i>	112
8.3.4.2	<i>Feature engineering</i>	112
8.3.4.3	<i>Data cleaning and normalisation</i>	112
8.3.5	Data discovery and detection results.....	114
8.3.5.1	<i>Centroids based on insider threat types</i>	114
8.3.5.2	<i>Results obtained from the execution of the K-means algorithm</i>	116
8.3.5.3	<i>Wordlists based on insider threat types</i>	120
8.3.5.4	<i>Results obtained from the execution of the Regular Expression Pattern Matching algorithm</i>	120
8.3.5.5	<i>Supervised machine learning algorithm results</i>	123
8.4	Observations	127
8.5	Verification.....	128
8.6	Conclusion.....	128
Chapter 9	130
Discussion and Conclusion	130
9.1	Introduction	130
9.2	Problem statement and main objective of the research	130
9.2.2	Summary and conclusion	135
9.3	Main contributions.....	135
9.4	Work flowing from this research	136
9.5	Future research.....	136
Appendix A	138
Ethics approval clearance	138
Appendix B	139
Research detail on the selection of the algorithms to detect insider threats within corporate email datasets.....		139
Appendix C	147
Python scripts in accordance with requirements to yield results in Chapter 8.....		147
Appendix D	169
Results from experimentation with the prototype to detect insider threats in a corporate email dataset		169
Appendix E	177
Larger subsets of datasets labelled by the unsupervised algorithms.....		177

The email in the dataset most similar to each centroid, as well as the similarity percentage.....	178
Appendix G	186
Wordlists for each insider threat type for Regular Expression Pattern Matching algorithm	186
BIBLIOGRAPHY	191

Chapter 1

Introduction

1.1 Introduction

Verizon's 2019 Data Breach Investigations Report shows that 34% of data breaches for the previous year were conducted by internal employees, and a further 2% of the attacks were conducted by partners (Verizon, 2019). IBM (2019) also reported that 82% of privilege misuse and compromises by insiders took months, if not years, to be detected. Insider incidents (also commonly referred to as insider threats) such as the above, caused enormous harm to large enterprises, as these were not detected and mitigated before they could cause damage. In 2016, a disgruntled employee, responsible for the Citibank IT systems brought 90% of the networks down due to a poor performance review obtained from management (Cluley, 2016).

From the aforementioned it is evident that dissatisfied employees can pose a major threat to their organisation. Research into employee-driven threats, resulting from resentful human behaviour, seems relevant and timely, and insider threats will therefore constitute the main theme of this research. Since Dasgupta and Dey (2016) argue that in most organisations the main risks and threats lurk inside employee, contractor and vendor email communications, the threats contained in employees' email communications will serve as a main theme of this research.

According to Chi et al. (2016), it is a challenge for organisations to detect and differentiate between normal business behaviours and malicious operations in email communications. Furthermore, due to the large volumes involved, as well as the lack of time, software or infrastructure, employees' email communications are usually not inspected by the employers. In addition, some organisations do not monitor the use of company email accounts, and limited or no controls exist to eliminate misuse. This points to inadequate awareness and understanding in organisations regarding the potential insider threats caused by employee emails.

A number of examples highlight what type of insider threats can be found in email communications and why it is necessary for organisations to start implementing detective and preventative controls to mitigate these risks. For instance, an insider threat in the publicly available Enron email dataset (Cukierski, 2015) was discovered in the email communication between top executives, John Lavoreto and Tim Belden (Tribolet, 2016). From the emails, it was clear that both parties were aware of Enron's active manipulation of the Canadian energy market in August 2000. It was found that the executives were both part of Project Stanley, which was the scam operation responsible for unscrupulous operations (Tribolet, 2016). Inspection of the emails revealed the fraudulent and unethical activities of the top executives, which posed a serious threat to Enron. An email sent by Jeffery Sherrick explicitly referred to the manipulation of balance sheet data (Sashikanth, 2015), while emails proving how two Enron employees bribed important individuals in Puerto Rico were located in this corpus.

Another example that demonstrates the richness of employee email communications is the Gupta leaks case in South Africa in 2017 (amaBhungane & Scorpio, 2017). The emails concerned show how executives from South African companies (also well-known audit firms) communicated with corrupt businessmen (the Gupta brothers), and indulged in fraudulent transactions or overlooked corrupt actions. Furthermore, the emails reflected evidence of the brothers' influence over political decisions made in the country, alleged money laundering and their stake in state contracts.

Malicious email use by employees often occurs when potential financial or personal gain is involved, or when revenge is planned due to feelings of disgruntlement or anger towards the organisation (Young et al., 2014) Such malicious employees or insiders are a great threat because they not only have a strong motivation to carry out the malicious activity, they also have widespread access to the organisation's non-public data and systems, and are often technically inclined and skilled (Chi et al., 2016).

There are, however, possible solutions to deal with the malicious use of company email accounts by employees. The large organisation Goldman Sachs uses surveillance methods to detect certain phrases in employee email communication so that action can be taken if suspicious behaviour is detected (Whitman, 2016). Phrases such as "I am not happy", "don't worry I'll take care of it", "split the difference", "where did my money go" and "embezzled

the account” are searched for in email messages. Scanning emails to detect threat-related phrases can be useful to obtain evidence of malicious behaviours of employees.

Goldman Sachs adhere to the requirement of the U.S. Financial Industry Regulatory Authority (FINRA), namely that firms must retain for at least three years all business communications such as email communications on all devices or websites used (Gural, 2013). FINRA’s 2017 Examination Priorities Letter stated that employee email communications are essential to business security (Robertson, 2017). Thus, FINRA dictated that US firms must maintain employee business email communications to be inspected for violation of business conduct (Robertson, 2017). FINRA also performs checks on how firms and financial advisors are monitoring and maintaining email communications, and it issues penalties to firms that violate these regulations (Robertson, 2017).

This research is relevant and necessary because of the widespread use of emails in organisations and the risks posed by employees, which can lead to financial and reputational losses. Weak cybersecurity measures allow staff to conduct malicious activities via email or expose them to becoming victim to threats via email, such as phishing scams. Thus, detecting whether employees use their email accounts to partake in insider trading, commit fraud, rant about the company, violate laws or company policies, or even to fall victim to phishing, malware or ransomware attacks, would be useful. This knowledge could guide the management of a company to understand what type of insider threat exists in their email environment. Furthermore, it will assist management to prioritise their resources and time so as to implement cybersecurity controls and governance to minimise any threats emerging from employee email communications.

1.2 Dissertation statement

The hypothesis or statement examined in this dissertation is as follows: *Employees potentially utilise their company email accounts for malicious and negligent activity and they often go undetected. Analysis of a corporate company’s email dataset can provide evidence of insider threats driven by human behaviour, such as employees conducting malicious activities and violating laws and company policies.*

Proving the above statement will be both the foundation and central focus of this research.

1.3 Problem statement

The main problem addressed by the research reported in this dissertation was formulated as follows:

Companies are unaware of malicious or negligent activities and behaviours of their staff due to weak or no controls being in place to govern the email content that is sent and received. These malicious activities could include fraud, leaking of confidential information and company data, or being the victim of phishing scams, ransomware and malware attacks – to name a few. These usually have harsh consequences for a company, such as financial and reputational damage. Companies are often negligent and ignorant about the fact that their email communication platforms are a major source of malicious activities. Due to the large size and structure of a company's email dataset, it is often considered unfeasible to inspect the content. Most companies do not possess the infrastructure or expertise required for such an operation. Therefore, the objective of this research is to prove that human-behaviour-driven insider threats can be accurately detected in a large corporate email communication dataset with the development of a comprehensive model.

1.4 Objectives and research questions

The main objective of this research was to prove that insider threats can be detected in corporate email communication data. The main research question relating to this objective was as follows:

How can insider threats caused by human behaviour be accurately detected within a large corporate email dataset with the development of a comprehensive model?

For this question to be answered, various research activities and experimentation were required to address the following sub-questions relating to the sub-objectives of this dissertation:

- i. *What are the main types of insider threat found in a corporate environment and what are the human behaviours that drive these insider threats?*

Answering this sub-question involved conducting a literature review to identify insider threat types and associated behaviour within a corporate environment.

- ii. *How can machine learning be used to detect insider threats with specific reference to corporate email systems?*

This sub-question links to the sub-objective which required the author to review various insider threat detection techniques and approaches from past research applied successfully within various domains. Another sub-objective that links to this sub-question involved determining whether machine learning email classification techniques applied in past research were relevant to the detection of insider threats.

- iii. *How can insider threats be identified in a given corporate email dataset, based on a set of phrases that link to different insider threat types and that are related to specific behaviours associated with insiders?*

To answer this sub-question, an insider threat classification prototype was developed, based on phrases identified and using machine learning techniques. This included the acquisition of a large email corpus, application of data-cleaning techniques and machine learning algorithms.

The next section will outline the scope and context of the research.

1.5 Scope and context of the study

The scope of this dissertation firstly covered the analysis of published research and literature in order to compile a literature review. Secondly, it included the application of various machine learning techniques to address the main research statement. The scope of the experimentation contained in this research was limited to email data, specifically to the publicly available Enron dataset. The entire dataset of 517 401 emails was used in the experimentation, but only selected fields of the email data and metadata were used in this work.

Note that the emails used for experimentation in this research were specific to the energy sector and to a large corporate organisation.

It must be acknowledged that there are privacy laws in some countries that protect employee email communications, despite the employer maintaining ownership and operation of the email platform. Thus, these organizations would be unable to utilise evidence pertaining to employee email communications for their risk management.

1.6 Methodology

This section outlines the methodology for the work conducted as part of this research. Specifically, the steps below address how the objectives and research questions in Section 1.4 were addressed.

1.6.1 Literature survey

The first step in conducting the literature survey involved obtaining an understanding of insider threats in general and human-behaviour-driven insider threats in particular. In addition, it involved an investigation into different insider threat detection techniques and approaches that were successfully applied to different domains. Furthermore, the focus was on email data and email classification techniques. A critical examination of the various insider threat detection and email classification techniques was conducted, in order to identify the most relevant approaches that could be applied by the prototype developed in this research.

1.6.2 Data gathering

An important part of this dissertation was to identify a suitable email dataset to be used for the proposed experimentation. The Enron Corporation email corpus, prepared by the CALO Project (a Cognitive Assistant that Learns and Organises), which was made publicly available in 2002, was retrieved from Kaggle.com (Cukierski, 2015). This email dataset was the only email dataset utilised in this research.

1.6.3 Design of prototype

In this step, a prototype was constructed to detect human-behaviour-based insider threats in corporate email communications, based on the work done by Van der Walt and Eloff (2018).

1.6.4 Prototype implementation, experimentation and results gathering

Once the prototype was constructed and implemented, several experiments were executed. These were done to test if the proposed solution was adequate to assist in addressing the problem statement. Various metrics such as accuracy and precision were tracked during the running of the different machine learning models used in the prototype to determine which technique was the most accurate.

Once this process was completed, a critical analysis was made to note areas of improvement in future work and to establish what could be gleaned from the results obtained.

1.7 Terminology

The following section presents brief explanations of the main terminology used so as to allow the reader to understand the context in which the different terms are used in this research dissertation.

1.7.1 Cybersecurity

Cybersecurity is defined as the “preservation of confidentiality, integrity and availability of information in the Cyberspace” (ISO, 2012). Specifically, cybersecurity involves the protection of assets that could be at risk when exposed to the Internet or cyberspace (Reid & Van Niekerk, 2014). Various controls can be put in place to eliminate or mitigate cybersecurity risks, and according to Pfleeger and Pfleeger (2012), preventative controls such as firewalls prevent cyberattacks from outside the organisation. However, attacks from within the organisation are caused by insiders who have been granted a wide range of system and network privileges to conduct their daily job, detective controls would offer a more suitable control of insider threats (Pfleeger & Pfleeger, 2012).

1.7.2 Insider threat

Kowalski et al. (2008) define insider threat as a threat initiated by a malicious current or former employee or by someone who has had some affiliation with the organisation. Such an insider has had legitimate access to the company’s network, system and non-public data and exploited this access to the extent that the confidentiality, integrity and availability of the organisation’s data or systems are compromised (Kowalski et al., 2008). Insider threats are driven by human behaviour and this behaviour is the focus of this research (Bell et al., 2019).

Kowalski et al. (2018) extend their definition to include threats caused by non-malicious negligent employees. These employees are also classified as insiders because they are careless about utilising security mechanisms or following proper security procedures.

1.7.3 Machine learning

The third component addressed in this study is machine learning. Machine learning refers to a type of artificial intelligence (AI) that aids systems to learn through various means, such as the observation of patterns in data, and not through explicit programmed instructions (Varone et al., 2019). Two main types of machine learning algorithms are supervised and unsupervised algorithms; the former require a labelled set of training data to learn from, such

that the algorithms can use this acquired knowledge to label an unclassified dataset (Cohen et al., 2018). The latter – unsupervised learning algorithms – are built to detect patterns in data without the use of pre-labelled training data (Mayhew et al., 2015).

1.8 Layout of the dissertation

The layout of the dissertation is as follows.

CHAPTER 1 contains the introduction for this topic and presents the layout of the rest of the document.

CHAPTER 2 presents a relevant case study that will be referenced throughout this dissertation, containing various examples of insider threats identified within the Enron email corpus.

CHAPTER 3 provides dictionary, as well as cybersecurity definitions for the terms: insider, threat, and insider threat. The aim of the chapter is to compile a definition for insider threat in the context of this research, based on the existing definitions, such that the reader is able to traverse through the dissertation with a clear understanding of the term.

CHAPTER 4 presents a discussion on the main components of this work. Specifically, it introduces the email platform as a means to facilitate insider threats. The chapter also elaborates on the severity of the insider threat problem that organisations face. The attributes of insider threat and the types of human behaviour conducted by insiders are also included.

CHAPTER 5 presents an in-depth study of the different types of insider threat, as well as the damage caused by these. The types of insider threat are mapped to human behaviours and various phrases linked to these, that could be evident in an employee's email communications.

CHAPTER 6 covers the established functional and technical requirements for the prototype developed in this work.

CHAPTER 7 discusses the conceptual design for the prototype to detect insider threats in a corporate email dataset, based on past research techniques and implementations.

CHAPTER 8 implements the prototype and presents the results of each step of the experiment that was conducted. The chapter also contains a discussion and validation of the results obtained during the experimentation phase.

CHAPTER 9 presents a summary of the conclusions, findings and contributions of this research. A discussion on the potential topics for future work concludes the chapter.

Chapter 2

Use Case: The Enron Email Corpus

2.1 Introduction

In the previous chapter, the main theme of this research was introduced as insider threats located within corporate email communications. It was noted that insider threats are an increasing problem in organisations, often detected only after the damage has been inflicted. Email communications were noted to be an important source of evidence of potential insider threats within organisations.

The purpose of Chapter 2 is to highlight real-world examples of insider threat within email datasets – specifically in the publicly available Enron email dataset. In the chapters that follow, references will be made to this chapter since concepts can be better explained through the use of the examples presented here. In addition, Chapter 2 shows that using various detection methods in large email datasets can aid with discovering insider threats.

2.2 The chosen dataset

Noever (2020) states that Google Scholar contains 20 200 published articles that referenced or used the Enron email dataset in some way, and 1360 of these are cybersecurity-specific papers. Agarwal et al. (2014) found that a large number of researchers focused their work on connecting people in the Enron network via their Enron email communications. Research also covered the language and sentiment contained in email content of the Enron dataset in attempts to analyse behaviours among colleagues (Agarwal et al., 2014). According to Zaki et al. (2017) the Enron dataset is useful in big data research that considers email communications.

It has been stated that a good way to determine an employee's interests is by studying their email traffic (Okolica et al., 2006). Nowadays, due to data privacy and personal information protection laws, it has become increasingly difficult to source a large organisation's email dataset for use in research. As such, the readily available Enron Corporation email corpus, applicable to various research domains, is commonly used within Computer Science research,

as can be seen from work done by authors such as Wang et al. (2010), Brown et al. (2013), Homoliak et al. (2018), Jiang et al. (2018) and Noever (2020). The Enron Corporation email corpus is also used in insider threat research, as it is easily accessible and includes relevant examples of such threats (Leber, 2013; Jiang et al., 2018).

Jiang et al. (2018) suggest that, before the Enron scandal broke, the emails of various insiders reflected negative emotions, which may have led to fraud and theft of confidential information. Chapter 2 will therefore refer to email examples from the Enron dataset that portray negative and suspicious human behaviour.

The following section introduces the case study (also referred to as use case) for this dissertation. It provides the background information of the Enron Corporation and the infamous scandal that transpired.

2.3 Enron: Company profile and background

According to Wilson and Banzhaf (2009), the Enron Corporation based in Houston, Texas, was one of the largest and most innovative energy and natural gas companies in the world. It employed more than 22 000 people (Noever, 2020; Wilson & Banzhaf, 2009). Towards the end of its existence, Enron lost a number of deals and its debts started multiplying. Amidst this, executives manipulated the oil prices and conducted various unethical activities. Investors started suspecting that the share prices of Enron were unrealistically inflated (Altman, 2003), and when analysts prompted the corporation to supply balance sheets and profit-related information, Enron was unable to do so. The CEO, Jeffrey Skilling, resigned around this time after only serving a short six-month term and did not provide the real reason for his sudden decision (Altman, 2003). The founder, Kenneth Lay, reclaimed his position as CEO, and attempted to raise the spirits of his employees, emphasising that Enron was in the best position it had ever been (Altman, 2003).

The CEO, however, aware of the dire situation Enron was in, was inconspicuously selling his stock, while the share price was plummeting. Enron slowly started collapsing and many employees, who possessed large Enron pension funds, witnessed this (Noever, 2020). The year prior to the bankruptcy, at the same time that the retirement funds lost a significant \$1 billion, 144 of the top executives at Enron received millions in performance packages (Noever, 2020).

On 2 December 2001, the corporation declared Chapter 11 bankruptcy due to corruption and fraudulent accounting activity to conceal debt and losses from shareholders (Wilson & Banzhaf, 2009). Enron's was the largest bankruptcy case in the United States and also one of the most severe audit crimes (Noever, 2020). More than 4000 of the company's employees were immediately retrenched and the share price plummeted to 1\$ (Noever, 2020). The auditor of the company, Arthur Anderson, was found to have assisted the fraudulent activity, which included providing sign offs on misleading financial statements containing overstated profits (Altman, 2003).

Once the scandal broke, in-depth investigations were conducted and amidst these, the Enron email corpus was released to the public. The next section covers the background, format and structure of this corpus and the individual emails.

2.4 The Enron email dataset

In 2003 when the Federal Energy Regulatory Commission conducted an investigation into the fraudulent activity that took place at Enron, various chunks of information such as audio transcripts, documentation as well as the Enron email dataset were made publicly available on the internet (Brown et al., 2013; Noever, 2020). The dataset consisted of more than 500 000 actual emails from 150 senior executives (Homoliak et al., 2018). It must be noted that some of the emails of senior executives such as Kenneth Lay (Chief Executive Officer) and Andrew Fastow (Chief Financial Officer) were actually sent by a personal assistant (Brown et al., 2013). The emails in the dataset included all Sent Items as well as Drafts created between 1998 and 2002 (Wilson & Banzhaf, 2009).

Although the dataset was cleared of most confidential and personally identifiable information (due to the risk of identity theft), it is still one of the largest publicly available real email datasets in the world (Noever, 2020). Investigation of the dataset revealed that the Enron employees sent not only business-related emails, but also emails containing personal information, spam, pornography and viruses (Noever, 2020). It was also noted that there were duplicate emails as well as multiple blank emails within the corpus.

Wang et al. (2010) suggest that the Enron email dataset has three main features: first, the emails have been grouped; second, there is no even spread of emails from the different senders, and third, due to the format the emails were supplied in, they can be classified into

threads. The threads serve to identify clusters of emails dealing with a common theme or purpose. Regarding the uneven distribution of emails in the dataset, most users in the database were found to have received far more mails than they had sent. The maximum number of sent items was 2000 mails, and only eight users who sent this number of emails were identified (Shetty & Adibi, 2004). Zaki et al. (2017) found that the Enron email dataset contains the full email body content as well as metadata and attachments.

Aery and Chakravarthy (2005) define an email as a text message with certain attributes. An email is made up of a header and a body. The header includes the metadata of the email message such as who the sender and the recipient are, the subject of the email, as well as the date (Tang, Pei, & Luk, 2014). Different Internet standards such as the Multipurpose Internet Mail Extensions (MIME) dictate different email formats and the data that is shown in the header. The body of the email contains the actual message content. Email message content differs from traditional text in the sense that messages are short and brief, and they may also contain non-textual data such as URL links and images (Tang et al., 2014). The format and structure of the Enron email dataset is discussed in detail below.

The Enron email dataset, which is available from Carnegie Mellon University, is neatly arranged into a hierarchical folder structure, with a labelled folder for each executive and employee, consisting of individual text files, each containing one email message (Brown et al., 2013). Shetty and Adibi (2004) from the University of Southern California stored the Enron email dataset into a MySQL database. This allows for queries to be easily executed so as to have quick access to emails regarding various topics and actors. A CSV file was also created to store the dataset and was made available on Kaggle.com (Cukierski, 2015), consisting of two columns. The first column of the CSV dataset, titled 'file', contains the file name or identification key of the email. It shows the sender name, email location (such as 'Sent Items'), as well as the number of the email. The following is an example of a file name:

allen-p/_sent_mail/1000.

The second column of the dataset, titled 'message', contains the actual email, which has two parts – header metadata and the actual email content. Each email is shown in the following format (Cukierski, 2015):

Message-ID:
Date:
From:
To:
Subject:
Mime-Version:
Content-Type:
Content-Transfer-Encoding:
X-From:
X-To:
X-cc:
X-bcc:
X-Folder:
X-Origin:
X-FileName:

Email body content

The format shown above that was used to dictate the structure of each Enron email, has various fields that require certain types of data. Each field in the structure is described below (Media Temple, 2020).

Message-ID: This field contains the identification key that is unique to each email and that is assigned to the email upon its creation.

Date: The *date field* refers to the actual date when the email was sent.

From: The *from* field indicates the person, otherwise known as the sender, who constructed the email.

To: This field shows the recipient of the mail, specifically, the person for whom the email is intended.

Subject: The *subject* contains a brief keyword summary to give the recipient an idea of what the actual email content body is about.

Mime-Version: This field refers to the version of the Multipurpose Internet Mail Extensions (MIME), which stipulates additions to the email format.

Content-Type: This field refers to the content that is maintained in the email body, for example plaintext or HTML. In addition, a character set, such as ASCII, can be included in this field.

Content-Transfer-Encoding: Some email data is in a format that cannot be transmitted over specific transport protocols. As such, a standardised encoding is required to transform the content into a format (with the required number of bits) that is acceptable to the given transport protocol. Therefore, an encoding mechanism such as BASE64 or BINARY is assigned to this attribute.

The following fields are referred to as x-headers, as they contain data that has been appended to the traditional header by the mailbox provider or email service provider for purposes such as monitoring and reporting. Below are the x-header fields used in the Enron email format (Cukierski, 2015).

X-From: This field contains the name of the sender of the email.

X-To: This field contains the name of the recipient of the email.

X-cc: The carbon copy (CC) field contains the name of additional recipients of the mail. They are not the primary recipients, but are also required or permitted to view the email content. As such, they have a relation to the content sent. The primary recipient can view the cc field and identify additional recipients of the mail.

X-bcc: The blind carbon copy (BCC) field contains the name of the additional recipients who are permitted to view the content of the email. Again, these recipients are not the primary recipients. In this case the primary recipient does not see who the additional recipients of the mail are.

X-Folder: This field contains the folder path to the email message file in the hierarchical structure where the emails were initially placed.

X-Origin: This field contains the IP address of the sender of the email.

X-FileName: This field contains the actual name of the email message file.

The content of the email which appears after the aforementioned header metadata would contain text formatted according to the Content-Type field, for example plain text. The above details clarify the structure of the Enron emails and provide the reader with a clearer picture of how the emails were stored when they were released in 2001.

In summary, this section presented the background of the Enron Email Dataset that was posted publicly on the internet by the Federal Energy Regulatory Commission. The discussion included the scope of the emails, as well as a brief comment on where the corpus is used nowadays. The structure of the emails as well as how the emails were organised and stored was also discussed in this section. Since the focus of this work was mainly on the content of the emails, knowledge of where this information was stored in the email format structure was essential. The next section focuses on the Enron email dataset, but specifically on the investigations into the email content that revealed insider threats such as the known fraudulent activities lurking within these communications.

2.5 Insider threats detectable in Enron email communications

As previously stated, the Enron email dataset is an example of an email corpus that contains a large number of examples of insider threat (Jiang et al., 2018). A number of these emails, which revealed the activities and opinions of the executives during the scandal, are discussed in the following sub-sections.

2.5.1 Disgruntlement leading to insider threats

Vincent Kaminski was a Managing Director at the Enron Corporation at the time of its bankruptcy. He was known for voicing his disapproval of the fraudulent activities to his fellow executives during meetings and via phone calls, but never took further measures to expose the illegal activities of Enron, perhaps because he was also enjoying financial benefits. The email communications in Kaminski's mailbox are included in the publicly available corpus and some of them are discussed below to determine whether, as the managing director, he might also have posed an insider threat. The discussion is based on the researcher's interpretation of the mails and aims to demonstrate the possibility of insider threats lurking in email datasets.

Inspection of Kaminski's emails revealed that this director apparently endured an extended wait to receive a promotion, even though he had worked hard for it (Leber, 2013). If an employee feels that they are not duly compensated or recognised for their effort, it could trigger them to become an insider threat to the organisation. An employee in a high position, such as Kaminski, would have had privileged access to confidential information as well as the credentials to access systems that would not be accessible to a lower-level employee. In the case of Kaminski, he could have used this feeling to take revenge on the corporation by selling his insider knowledge regarding the unscrupulous activities taking place at Enron to journalists and external parties. Even though this was not the case with Kaminski, employers should always consider such possibilities regarding potential threats to their organisation.

Additional emails were found containing evidence that Kaminski had expressed his dislike of a certain colleague. The latter's supposed feelings of disgruntlement perhaps spurred on the director to demonstrate his dislike of employees who might have exceeded him in terms of salary and title. Alternatively, Kaminski might have expressed his feelings of disgruntlement as anger and thus imposed this anger on fellow colleagues, indulging in disagreements and arguments. The director's mailbox furthermore contained evidence that he had sent a large number of emails dealing with human resource complaints and voicing his opinions on possible job candidates (Leber, 2013). This is another example to suggest that Kaminski frequently observed or experienced events at the organisation that contributed to his unhappiness.

Another instance of disgruntlement is evident from an email sent by Pamela Allison, a former employee, to Kenneth Lay, the Founder of Enron, on 15 August 2001. It demonstrates that the environment at Enron was challenging and hostile, and that employees were uncomfortable about the illegal activities taking place in their midst. The email read as follows:

"Mr. Lay, I am not writing this in malice but in hopes that it helps get Enron back the way it used to treat their employees and makes it the number one employer of choice again. I hope you can get back the feeling that I had when I first started there and get the stress level down in your organization for the sake of your employees."

In addition, this disgruntled former employee quoted the following reasons for her feelings of unhappiness:

“During the last 5 years I was there, I noticed a change in direction in the way employees were treated by upper management – and upper management was getting away with it. Not only were they getting away with it, these people were being rewarded for this behavior. I have heard stories of lower level employees being screamed at and in one instance, one of the VP s who was brought down from Canada was heard in his office screaming and pounding his telephone on his desk. Heaven only knows how he treats his subordinates.” (Cukierski, 2015).

These examples strengthen the belief that the environment at Enron caused employees to observe feelings of disgruntlement, which could have resulted in a potential insider threat. Allison was graceful in her feelings towards the situation at Enron, but one of the many other employees who had had similar experiences to Allison, might have been triggered into causing harm to the organisation.

2.5.2 Fraudulent activities as insider threats

Another well-known insider threat that was evident at Enron was the clandestine fraudulent activities that took place. Fraud is in fact one of the main reasons for the fall of Enron and as such the email dataset contains a large number of possible fraud-related emails. One example of insider fraud is found in email exchanges between two senior executives, John Lavoreto and Tim Belden (Tribolet, 2016). The emails show that both of the executives were informed of Enron’s altering the Canadian energy stock market in late 2000. In addition, the executives were hiding major debt and used special purpose entities to do so. One email which makes reference to 29 of these entities, was sent to Tim Belden on 1 November 2001 when external investigations into the illegal activities were being conducted. It read as follows:

“If you have any e-mails that relate in any way to the LJM Deal or Chewco Investments L.P., including any accounting issues related to these transactions, please forward the e-mails to LJM.Litigation@enron.com. 4. If you have any e-mails that relate in any way to Enron s public statements regarding EBS, Azurix, New Power Co., or any e-mail regarding financial transactions involving these matters, including accounting issues related to these matters, please forward the e-mails to ClassAction.Litigation@enron.com.”

Lavoreto and Belden's email communications also contained references to Project Stanley, which was the name of the group coordinating scandalous operations at Enron (Tribolet, 2016). These executives had one primary motivation, namely a substantial personal financial benefit. An example mail containing the subject line *"Project Stanley - History of the Design of the Alberta Power Pool"* was sent by Nella Cappelletto on 7 June 2000 to 11 recipients, of whom four were not Enron employees. The mail contained the following snippet regarding the non-compliance of Project Stanley:

"While I appreciate the determination of whether an offence occurred under the Competition Act would be independent of the compliance, or not, with the Power Pool Rules, I think it is noteworthy that the Power Pool incorporated these entirely new set of Participant Behavior Guidelines only recently, and after the acts that are the subject of Project Stanley had occurred."

In addition to the manipulation of stock markets, it was found that the Enron executives were trying to obtain major profitable deals for their own financial benefit. The following snippet of an email from William Giuliani sent to Andrew Fastow, CFO of Enron, on 7 June 2001 showed that a major deal was being made by the Enron executives to obtain a large amount of coal at below market prices. In addition, it could provide them with additional funding from marketing:

"In addition to redeeming part of our equity interest, the deal provides us 900,000 tons of coal priced below market, an option which could lead to a very profitable synfuel project, and the potential for more marketing fees from other Cline entities."

Another email sent to CFO Andrew Fastow from Rex Rogers on 12 October 2000 shows that an insider trading rule was being introduced at Enron. This could have differing effects on the private stock held by the CFO – some of these positive in terms of flexibility. As such, he was given the option to consider some alternatives regarding his own stock with regard to this rule. This is shown in the email below.

"I have been asked to make a brief presentation at next Mondays Executive Committee meeting addressing a new S.E.C. insider trading rule. Although the new rule may increase exposure to liability for insider trading, certain provisions of the new rule may actually provide

for greater flexibility in the timing of your personal trades in Enron Corp. common stock. Attached is a short memo addressing our current Company procedures and policies for trading, the new S.E.C. rule, and some suggestions for alternatives that you may want to consider concerning your personal trades in Enron Corp. common stock.”

Furthermore, emails were located that contained evidence regarding the California stock price manipulation. An email sent in this regard from an external legal advisor to 15 employees at Enron on 29 January 2001 stated the following (Noever, 2020):

“Steve thinks he might be asked about whether the market was manipulated. Please provide information on whether this was the case and who the participants likely were.”

A further case of fraud detected within the Enron email dataset was the planning of shutdowns for various power plants in California, involving Chris Germany and Victor Lamadrid. As a result, the demand and price of power would increase, in order to increase the profits of these executives (Cukierski, 2015). The email showing the list of outages that were planned was sent from Victor Lamadrid to Chris Germany on 4 April 2000 and is shown below.

*“OUTAGE REPORT FOR GATHERING Received 4/3/2000 2:35 p.m. Posted 4/3/2000 2:35 p.m.
STATE: PAFACILITY: Cherry Tree Station*

*PLANNED WORK: Overhaul Unit #2OUTAGE DATE: April 10-20FLOW REDUCTION: 2.4
MMCFDCONTACT: Kevin Miknis STATION PHONE: 724-468-3731
STATE: PAFACILITY: Stoney Run Station*

*PLANNED WORK: Overhaul Unit #3OUTAGE DATE: April 24 – May 5FLOW REDUCTION: 2.0
MMCFDCONTACT: Kevin Miknis STATION PHONE: 724-468-3731
STATE: WVFACILITY: Jones Station*

*PLANNED WORK: Overhaul Unit #1OUTAGE DATE: April 17-20FLOW REDUCTION: 1.0
MMCFDCONTACT: Larry Wade STATION PHONE: 304-477-3366
FACILITY: TL-263*

*PLANNED WORK: Replacement tie-ins OUTAGE DATE: April 15FLOW REDUCTION: 34.0
MMCFD CONTACT: Steve Searls STATION PHONE: 304-595-1270*

Notice to all Appalachian Pool Operators: Received 3/29/00 2:00 pm Reclassification Notice Posted 3/29/00 2:00 pm

On February 10, 1999, the FERC approved Docket No. CP97-549, granting CNG Transmission s (CNGT s) request for reclassification of various transmission lines to gathering lines.”

The emails of Enron founder Kenneth Lay, as well as Jeffrey Skilling, the CEO at the time of the collapse, were found to contain discussions on unscrupulous activity (Cukierski, 2015). Even though the executives were aware of the hidden losses, Kenneth Lay and Jeffrey Skilling encouraged investors and their own employees to purchase shares – while they themselves began clandestinely selling their own shares (Segal, 2019). Lay, with his calm and composed nature, reassured the public that the stock price would increase shortly after. This was the main reason why both Skilling and Lay were convicted of fraud, conspiracy and insider trading (Segal, 2019). An email sent from Larry Izzo to Kenneth Lay on 21 September 2000 regarding the problematic situation at Enron as well as the potential for downsizing, employee retrenchments and eliminating outsourced and third-party work, stated the following:

“I think it is important to agree on a clear plan and brief our employees, all of whom are stressed by the uncertainty of where they're going. This will have a negative impact on the company's performance, unless addressed.”

An email sent to all Enron employees from the general Enron Announcements/Corp/Enron@Enron stated the following:

“November 20 the Savings Plan system re-opens with great new features”.

This was probably one of the attempts agreed on by the executives to ensure that the employees would not panic about the fluctuating share price and their retirement funds. According to Noever (2020) an email was located in which the CEO Jeff Skilling motivated employees to *‘trade aggressively’* and denied that prices were manipulated in California. He further told employees that if they were unwilling to cooperate, they could *‘find another job’*.

The Enron email dataset contains a large number of emails relating to the topic ‘meetings’, and an increasing number of ‘meeting’ emails were sent and received close to the date of Enron’s collapse (Cukierski, 2015). The founder and the CEO were both linked to these emails.

It is possible that the 'meetings' were created to discuss the plan on how to dispose of their shares, and align the information in this regard that they would share with the public.

2.5.3 Insiders benefiting from confidential information whilst concealing the truth from investors

Another Enron executive, Paula Rieker, obtained Enron shares in her private capacity at a low price of \$15.51 per share and then sold them at \$49.77 each, to make a personal profit (NBC News, 2004). It should be noted that she sold these shares whilst being well aware of the secret million-dollar losses incurred by Enron (NBC News, 2004). Rieker thus abused confidential information of the organisation's debt (of which the public was unaware) and secretly sold shares for her own gain, and so she was also charged with insider trading when the scandal was discovered (NBC News, 2004). NBC News (2004) reported that she eventually admitted that she had assisted the executives to provide false, manipulated information to the public regarding Enron's earnings. It is thus possible that emails sent by Rieker might have included discussions with senior executives using keywords such as 'share price' or 'stock price' in addition to advice she may have provided. Rieker sent the following email to Mark Koenig on 23 October 2001:

"I have talked to Frevert, and we agreed to let the Mgmt. Comm. offsite on Wed. proceed, test the temperature of the management feedback there (see if it mirrors the MD session) and then decide how/when/if to summarize management feedback to the BoD."

This mail indicates that Rieker was responsible for presenting feedback to the Board of Directors. As such it is noted that she not only considered whether to send feedback, but also how the feedback should be summarised, which indicates her involvement with the information made public to investors. On the same day, the following mail was sent from Mark Koenig to Paula Rieker:

"P – an idea. You should consider summarizing the MD meeting we had yesterday for the Board. A lot more meaningful feedback than press reports and analyst summaries. The tone from investors today in San Diego was very hostile re: Andy and I don't think this is getting through to the Board. This is a little "venting" given the amount of time we have all put in to defend Andy, with very little help from him. Thanks again for your help this morning. MEK Mark Koenig." (Cukierski, 2015)

This email indicates that Rieker was beneficial to the organisation in how the corporation was projected to its investors. Furthermore, a Managing Director (MD) meeting was mentioned in the email – potentially a meeting during which the fraudulent activities and methods to disguise them and boost investor confidence were discussed. It appears that this mail was sent around the time that ‘Andy’ or Andrew Fastow, the Chief Financial Officer of Enron at the time, was being questioned. A snide comment indicates that Koenig felt little support from the CFO, which might suggest that Koenig was a potentially disgruntled employee.

Lastly, on 19 November 2001 Rieker sent an email to Greg Caudell and Raymond Bowen Jr stating the following:

“Ray – The Board has requested an “continual update” on a few subjects, one of which is liquidity. This information would be sent weekly to a secure, private fax machine, Ken and I are also discussing a weekly phone update, which would still require a schedule similar to this one. An alternative would be to send a weekly report that summarized only KEY CHANGES to the projected year-end cash balance and set forth a revised projected balance.”

This message confirms Rieker’s involvement during the critical time when the scandal was unfolding and how a report with the changes made to the year-end cash balance would need to be presented.

2.5.4 Insider threats accidentally constituted by non-malicious employees

Besides the unethical activities conducted by the executives of Enron, there are other means by which employees of the corporation might have posed an insider threat. For example, there are various ways in which employees could use their email to launch a phishing attack. Inspection of the Enron email dataset revealed that several suspicious emails contained URL links and attachments, as well as advertisements with URL links. If these emails had been sent by a phishing scam attacker, they could have been a means to lure a negligent employee into providing system credentials or other confidential information. Furthermore, the attachments might have launched malicious background processes that attempted to steal confidential information, or to access or disrupt the current user’s session. According to Noever (2020) an example of malware that is evident in the Enron corpus is known as ‘JokeStressRelief’ and 231 such executables were identified in the corpus.

Credit card details were obtained by malicious attackers and there are emails within the corpus that suggest as much. The following mail sent to a Hotmail account serves as evidence.

“Here are the list of roommates for each room with one credit card... Unless otherwise specified, all credit cards are in the name as listed below.” (Noever, 2020)

An email was also found providing credentials via email which, apart from being a major security risk, could potentially have been sent to a malicious party:

“it is a Mastercard and the number is XXXX XXXX XXXX XXX and the exp is 11/01”.

Zaki et al. (2017) argue that there are various ways in which phishing scam attackers could have executed attacks on Enron. The phisher could have used information about the type of business and sector in which Enron operated to create targeted emails relevant to those within the organisation, in order to lure employees to provide certain information. Zaki et al. (2017) who ran various tools on the email dataset, found main topics that emerged from the emails and identified employees who had been associated with these topics. One of these topics was ‘company image’. In the case of Enron, if a phisher was aware of the company’s reputation and customer relations, they could have used this information to damage the company’s image via the email platform. Malicious emails sent to negligent Enron employees could easily have led to harm within the organisation.

To conclude, it is evident that the Enron email dataset is rich with examples and cases of insider threats posed by employees as well as executives. As such, this dataset is a good fit for research that aims to study insider threat lurking within emails.

2.6 Conclusion

This chapter began by setting the scene of the Enron Corporation scandal, which led to the corporation’s large email corpus being made publicly available. This was followed by a discussion of the actual structure of the corpus, as well as the format of the actual emails of the Enron employees. Chapter 2 also provided several examples of typical incidents of insider threat that were lurking within Enron’s email dataset. Various types of insider threat were discussed; however, the different types are quite broad and there can be many different motivations for an attacker to execute the attacks. While the emails and activities discussed

in this chapter were not specifically grouped under the different categories of insider threat types, this would be an important aspect going forward.

The chapter provided the case study to be referred to in the rest of the research reported on in this dissertation. It should be noted that this chapter assumed the reader's knowledge and understanding of the term insider threat. The next chapter focuses on defining insider threats within the context of this work.

Chapter 3

Defining Insider Threats

3.1 Introduction

In Chapter 1 of this dissertation, potential threats lurking in employee email communications were introduced as insider threats. Furthermore, insider threats were stated to be a main theme of this research. A brief explanation was provided for this concept, as well as some examples that are relevant to the context of the research at hand. The purpose of this chapter is therefore to provide a suitable and detailed definition of the concept of insider threats.

It is necessary to provide a definition in this chapter to enable the reader to progress through the remainder of the dissertation with a clear understanding of what is meant with the concept, insider threat. Chapter 3 critically analyses various definitions from standards bodies and existing research to discern the elements in the definitions that are relevant to the work conducted in this dissertation. The final section of this chapter adopts an adapted definition of insider threats that will apply to the rest of this dissertation.

3.2 Definition of an insider threat

It must be acknowledged that the term insider threat consists of two separate words, insider and threat, each with their own meaning. As such, these two words need to be defined individually in order for the essence of each word to be captured and fully understood. The sub-sections that follow present a discussion of the various definitions for each word, insider(s) and threat, as well as for the two words combined, specifically, insider threat. Various definitions obtained from standards authorities and organisations such as ISO (ISO, 2019), NIST (NIST, 2015) and CERT (Cappelli et al., 2012) as well as those from relevant authors are explored.

3.2.1 An introduction to insiders

According to the Merriam-Webster dictionary, an insider is defined as “a person recognized or accepted as a member of a group, category, or organization: such as a person who is in a position of power or has access to confidential information” (Merriam-Webster, 2019). An

insider is defined by the Oxford Dictionary as “a person within a group or organization, especially someone privy to information unavailable to others” (Oxford, 2017).

There is a common denominator in the definitions provided above. The Merriam-Webster Dictionary (Merriam-Webster, 2019) refers to a person having certain powers or access to classified information, and the Oxford Dictionary (Oxford, 2017) mentions a person with access to certain information that is not available to others. It seems that such access is based on a trust relationship and perhaps required by the individual to fulfil a job. These specific aspects of the general meaning of the word insider are an important distinguishing characteristic of an insider.

The above definitions provide a good foundation for the explanation of an insider in terms of this study. An aspect worth noting is that an insider exists within the context of an organisation, which is relevant to the research at hand. In addition, these definitions clearly state the insider to be a person (Merriam-Webster, 2019; Oxford, 2017), and not a system or bot. Again, this is a necessary distinction for this research, as the focus is on human behaviour. It is important to consider that insider threats in the context of cybersecurity can be driven by bots that enter the network or other components of the information technology (IT) infrastructure to execute malicious commands. This research however excludes bots as the driver of insider threats, as a different approach would be required to detect bots.

The general definitions of an insider as discussed above have been diagrammatically captured in Figure 3.1 to highlight the important aspects that should be considered when defining an insider. This diagram, which was created using a basic UML notation structure, should be read from the top down, where each $n+1$ level expands and describes the properties of level n . The diagram is presented in a hierarchical structure and all diagrams in this chapter are created with this same approach.

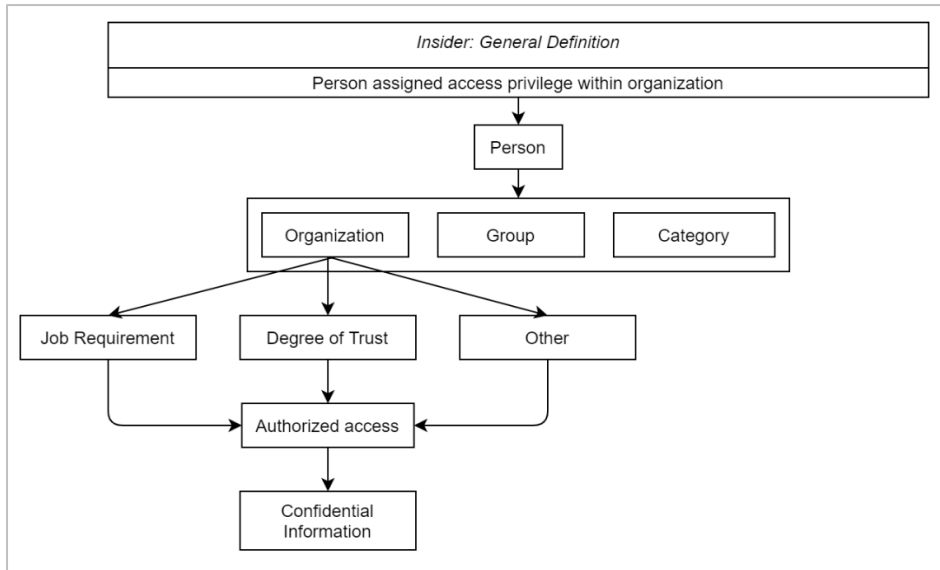


Figure 3.1: Diagrammatic structure of the insider definition

Figure 3.1 will serve as the foundation for this cybersecurity-specific definition. As such, the next section will consider definitions of insiders from a cybersecurity perspective.

3.2.2 Defining insiders from a cybersecurity perspective

According to the *Computer Security Handbook*, an insider is one who has obtained authorised access to the organisation's systems (Bosworth et al., 2014). The concept of authorised access stated in the Handbook is echoed by Cappelli, Moore and Trzeciak (2012), who state that an insider is a person within an organisation who is assigned certain privileges in respect of organisational systems such as the network, data and IT systems.

In addition to privileges, the insider is assigned special modes of access, such as the ability to read, write, modify and delete certain artefacts (Pfleeger & Pfleeger, 2012). Salem et al. (2007) agree and define an insider as a legitimate user with authorised access to company systems and data. As such, the insider possesses the necessary credentials to access confidential data (Alawneh & Abbadi, 2011).

Young et al. (2014) suggest that insiders have authorised access to the data and systems within the organisation due to its being a requirement of their daily job. Hence, they are allocated role-based access, in other words they have certain access privileges based on their job profile (Pfleeger & Pfleeger, 2012). This also means that the insider does not violate company security policies or the law to obtain access. Often, however, employees are

provided with more access rights than what is required to perform their daily jobs (Pfleeger & Pfleeger, 2012).

Insiders have often been found to exploit or misuse their privileged access to information within the organisation (Liu et al., 2019; Bell et al., 2019). Chinchani et al. (2005) define an insider as an individual in the organisation who combines these access privileges with knowledge of the company's data and systems to cause harm to the organisation. Their definition introduces the notion that an insider can act with malicious intent to inflict harm on the organisation. Intent is defined by Merriam-Webster (2019) as an end goal or purpose on which one's energy is focused.

Based on the discussion of definitions above and currently available cybersecurity research, the following aspects apply to the definition of an insider:

- An insider is working from inside the organisation, as opposed to an external party.
- Insiders have privileges and modes of access that are required for their daily work. These are known as role-based permissions (Pfleeger & Pfleeger, 2012).
- Privilege includes the access rights to various organisational resources, specifically sensitive content, and as such privilege should also be covered in the definition.
- The intent of the insider has been shown in this section to be either malicious or negligent.

A definition encompassing some of these factors and additional considerations is offered by Bishop et al. (2008). In their cybersecurity-specific definition, an insider is defined by their association with a given resource, which results in a level of 'insiderness':

- The person had authorised access to resources at some point in time.
- There is some degree of trust with this individual.
- The person is a system user who might exploit their access rights.

Bishop et al.'s (2008) definition introduces levels of 'insiderness' and varying degrees of trust, which could indicate that a person with greater access to a resource would be more of an insider than a person with minimal, restricted access. This assumption is valid, because restricted access would be better controlled and would be assigned to employees who would not normally be allowed to view certain confidential information. Alawneh and Abbadi (2011)

argue that when a potential employee applies for a job where they will be accessing confidential information, background checks are often performed and secrecy agreements are signed.

The definitions discussed so far have not yet made it clear which specific types of employees within an organisation are considered to be insiders. It is essential to identify the types of insiders so that incorrect assumptions or exclusions are not made when reading this dissertation. According to CERT, a cybersecurity definition for an insider is “a current or former employee, contractor, or business partner of the victim organization” (Cappelli et al., 2012). In the case study, it was evident that the insiders who threatened the Enron Corporation were senior executives Victor Lamadrid and Chris Germany, who manipulated the stock markets and planned powerplant shutdowns for their own financial benefit.

In the business world it is common practice for a company to outsource a specific function for a period of time. Since various access privileges are assigned to sub-contractors or contractors who work at the company during this period, they are considered to be insiders. Hunker and Probst (2011) include the software engineer who created the system in their definition of an insider, because even though the engineer is not a system user, or in some cases not even an employee of the organisation, he or she is fully aware of how to access the system. According to Alawneh and Abbadi (2011), temporary staff are a similar type of insider.

Alawneh and Abbadi (2011) mention another example of a type of insider, where a non-malicious internal employee may share login credentials or confidential information with someone who is not affiliated with the organisation. Hunker and Probst (2011) add to the list and mention a ‘masquerader’ – a user who does not have authorised credentials to access a given system, but stumbles upon this system that has already been logged into and so obtains access. According to Cappelli et al. (2012), a former employee who still retains their system credentials is also considered to be an insider. Consequently, when defining insiders, there should be a certain level of detail surrounding who is considered an insider.

The necessary components that should be considered from a cybersecurity perspective when creating a definition for an insider, as well as the specific types of insiders, are summarised in Figure 3.2.

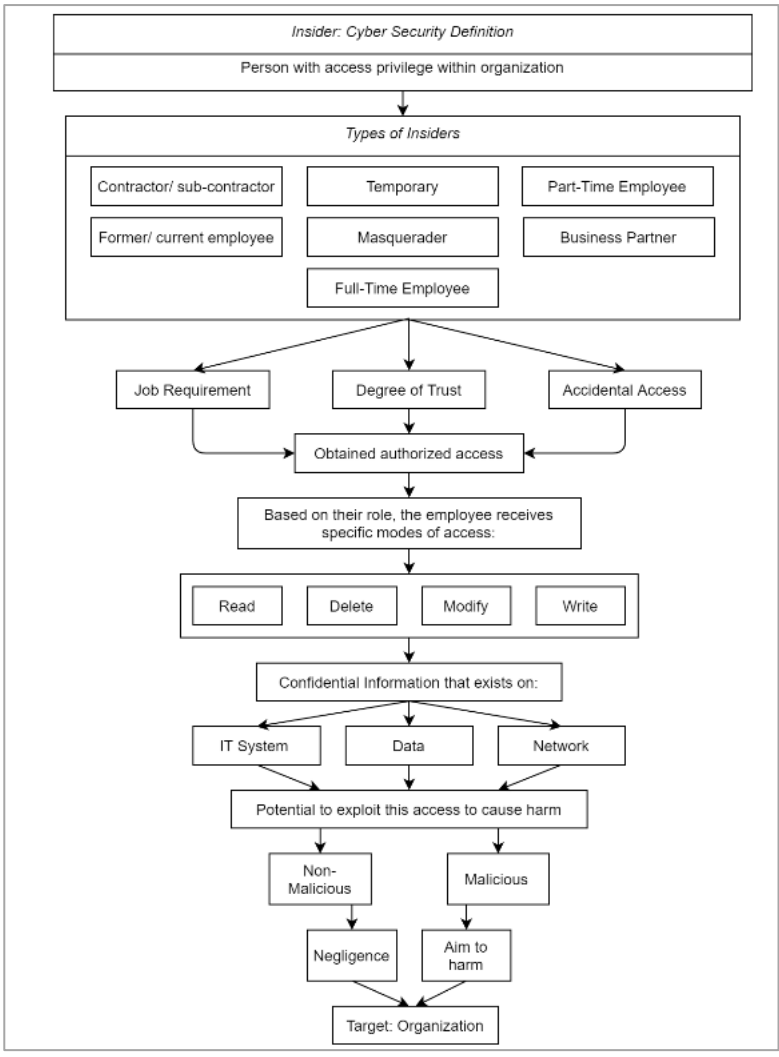


Figure 3.2: Diagrammatic structure for defining an insider from a cybersecurity perspective

Thus far, an overview was provided of the elements that the definition of an insider in the context of cybersecurity should include to ensure the reader’s understanding of the term. The next section discusses the definition of the term threats.

3.2.3 An introduction to threats

The Oxford Advanced American Dictionary (2020) defines a threat as “the possibility of trouble, danger, or disaster”. A threat is also defined as “an expression of intention to inflict evil, injury, or damage; one that threatens; an indication of something impending” (Merriam-Webster, 2019). There is a common denominator between these two definitions of threats, with the Oxford Dictionary referring to a ‘possibility’ and the Merriam-Webster Dictionary referring to ‘something impending’ – thus, the likelihood that something will happen, specifically that an undesirable event will occur.

The definitions however do not cover who or what is responsible for the existence of the threat and why the threat exists, which is an essential factor in this research where the focus is on who carries out the threat. The definitions assume that any process, person or event can potentially cause harm. In order for potential harm to be actualised, a trigger is required – specifically a person, process or event with the intention to cause harm, having set up the unfavourable circumstances.

It has, however, not explicitly been stated in the definitions that the threat is aimed at the organisation and as such this cannot be assumed. To align the research to the main topic, the target should be clearly defined as the organisational environment. In the case study, email evidence showed that harm was caused to the organisation by concealing debt and losses from investors and manipulating the financial statements shown to the public. This organisation was harmed to the extent that it was declared bankrupt during December 2001.

It is clear that some important factors should be considered when defining a threat, as was discussed so far. Figure 3.3 therefore covers the aspects required for a general, yet clear definition of a threat.

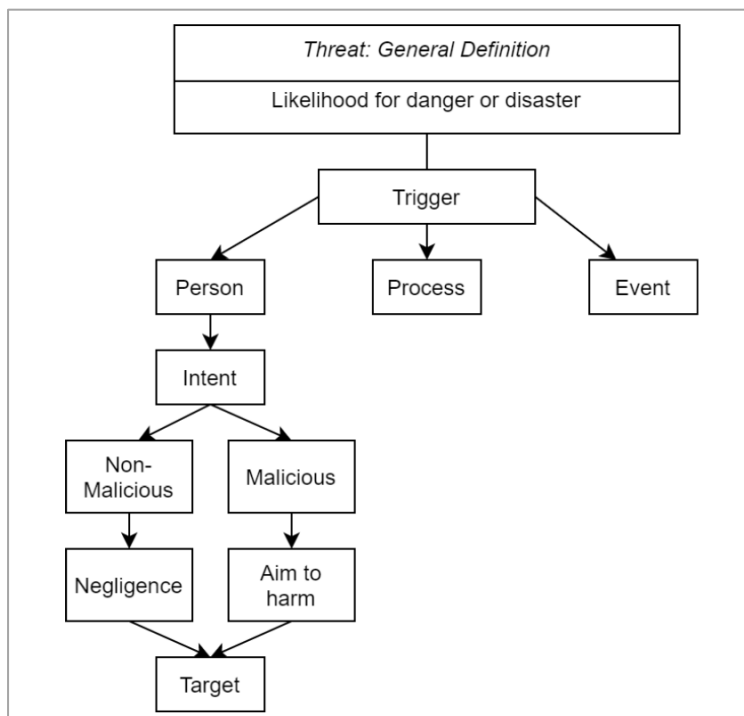


Figure 3.3: Diagrammatic structure of the threat definition

The next section discusses various definitions of threats, but the definitions and literature referred to will be specific to a cybersecurity perspective.

3.2.4 Defining threats from a cybersecurity perspective

Pfleeger and Pfleeger (2012) define a threat in a computing system as “a set of circumstances that has the potential to cause loss or harm”. This definition aligns with those in the Oxford Advanced American Dictionary (2020) and Merriam-Webster Dictionary (2019) as it also focuses on the result of the threat, specifically the possibility of a negative impact. This definition therefore serves as a good baseline for the definitions that follow in this chapter.

ISO 704:2009 (ISO, 2019) and NIST SP 800-128 (NIST, 2015) define a threat, specific to the cybersecurity domain, as an occurrence that has the likelihood of causing damage to organisational operations such as reputation and functions, organisational assets or people. This threat propagates via an information system where various malicious actions are inflicted, for instance unauthorised access, causing harm to the confidentiality, integrity and availability of information, and the exploitation of known vulnerabilities in the systems (NIST, 2015; ISO, 2019). It is clear that there is more detail in this definition than in the two dictionary definitions quoted above, as well as in the definition by Pfleeger and Pfleeger (2012). It includes the target of the threat, the reason that the threat is triggered, how it threatens, as well as what damage can be caused. It should be noted that the context of this definition is an organisation, as can be seen in the references to organisational operations and organisational assets.

The ISO (2019) and NIST (2015) definition firstly indicates that a threat is a circumstance or event that can have a negative impact on the target, but does not indicate who or what inflicts the threat. Pfleeger and Pfleeger (2012) suggest that a threat to an asset involves two aspects: firstly, who or what inflicts the damage, and secondly, what type of harm can be caused. Thus, both aspects are essential.

The ISO (2019) and NIST (2015) definition aligns with the dictionary definitions by indicating that a threat is the likelihood or potential of causing harm. It includes several types of harm caused by a threat, such as harm to organisational assets, operations and individuals, and is key to the definition to be used in the research at hand that will be created in this chapter. To

obtain a comprehensive list of types of organisational harm, literature relevant to the topic of this research was investigated and these are covered in the next paragraph.

Kowalski et al. (2008) identified financial and reputational losses, disruption of daily business operations, and damage to specific individuals as the main types of harm caused by insider threat within an organisation. Serious financial consequences and million-dollar losses were also identified by Greitzer et al. (2019). Hunker and Probst (2011) agreed and referred to financial loss, the disruption of organisational operations, reputational damage and damage to the organisational culture. Cappelli et al. (2012) compiled a list of the types of organisational harm that included financial loss, operational damage, harm caused to operations in other sectors, reputational damage, harm to individuals, and a tainted organisational image – leading to the retrenchment of employees and in some cases, the shutting down of business operations.

Harm caused by insiders in the public sector must also be considered, as it may hinder the delivery of essential services and have a negative impact on the daily life of members of the public (Bell et al., 2019).

The ISO (2019) and NIST (2015) definitions name malicious actions that facilitate the threat, especially unauthorised access and harm to the availability, integrity and confidentiality of information held in information systems. The effects of compromising each of these three aspects are explained next. Firstly, compromising the integrity of a system or data would result in the system being unable to guarantee that only an authorised user modified its information (Pfleeger & Pfleeger, 2012). The information would therefore become unreliable, inaccurate and unusable. Secondly, compromising its availability would mean that the full data or system is not readily accessible when an authorised user requires it (Pfleeger & Pfleeger, 2012). Time is crucial in a business day and if operations are not available when required, the delays could be costly. Lastly, when the confidentiality of data or a system has been compromised, there is no guarantee that only authorised individuals have had access to or were able to view and modify information (Pfleeger & Pfleeger, 2012). When organisational information such as customers' personal information is obtained by an unauthorised individual, it could cause serious reputational damage to and distrust in the organisation.

The diagram in Figure 3.4 presents a summary of the important aspects discussed in this section regarding the definition of a threat from a cybersecurity perspective.

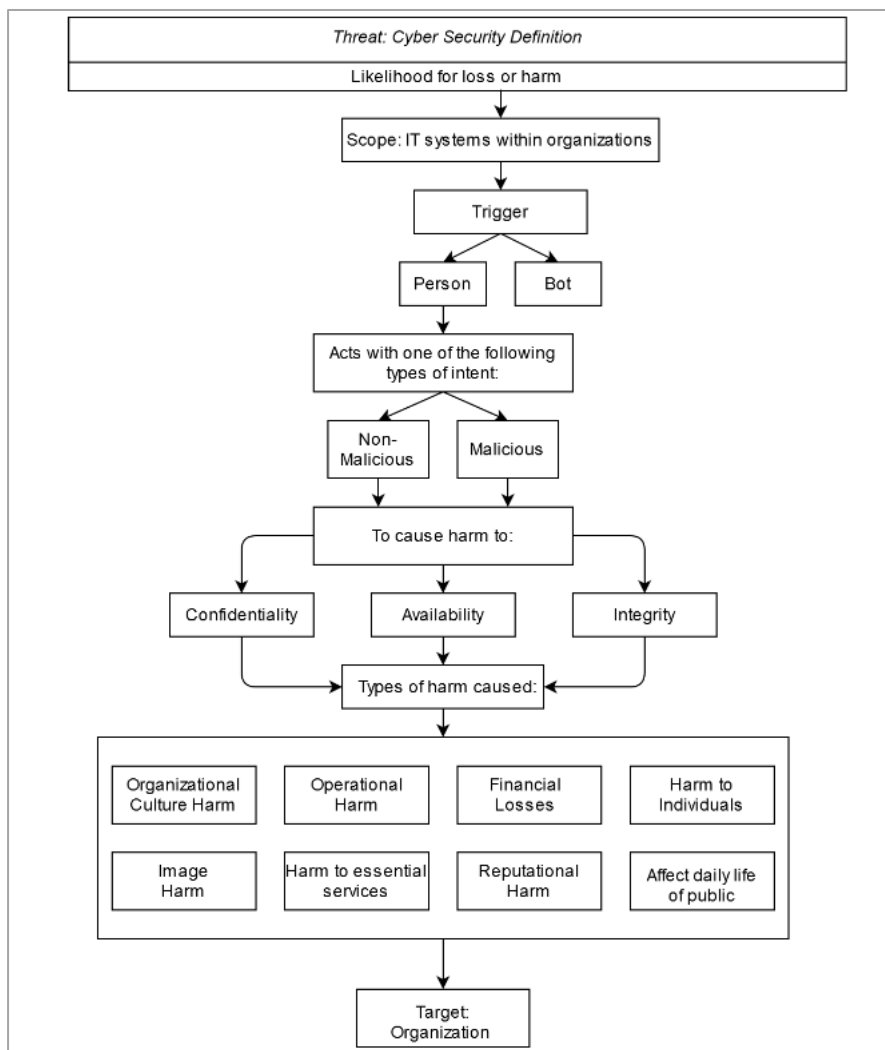


Figure 3.4: Diagrammatic structure of the definition of threat from a cybersecurity perspective

The concepts insider and threat were defined and dissected and therefore the section that follows will present definitions for the combined term insider threat within cybersecurity research.

3.2.5 Defining insider threat from a cybersecurity perspective

NIST (2019) defines insider threat as follows: “An entity with authorized access that has the potential to harm an information system through destruction, disclosure, modification of data, and/or denial of service”. The first detail that can be gleaned from this definition is that the subject that poses the threat is an “entity with authorized access”. This differs from the definitions studied in the previous section for the term threat, where a subject was not

provided. Section 3.2.1 covered the general definition of an insider and described the various types of insiders, which all constitute authorised entities. The second detail that can be noted is that the NIST (2019) definition, as well as the definitions in the previous section, refer to the likelihood that the insider acts with malicious intent, thus causing harm. The third takeaway from the NIST (2019) definition is that the types of harm shown, namely “destruction, disclosure, modification of data, and/or denial of service” refer to the compromising of the confidentiality, integrity and availability of data within the organisation. This clearly links to the cybersecurity definition of a threat by ISO (2019) and NIST (2015).

The NIST (2019) definition however omits important details that are essential to defining an insider threat. It is noted that an insider can be distinguished as an employee who has rightfully obtained access to various resources as part of the job. This was reflected in Figure 3.2, which covered the definition of insiders from a cybersecurity perspective. A main threat to an organisation is when the insider abuses their access privilege (Salem et al., 2008).

Costa (2017) suggests a structure that considers various factors to provide a more specific definition of insider threats. The structure shown in Figure 3.5 was designed to include four main components, each with a list of options. It must be noted, however, that specifying a list of options for each component might be too rigid. Therefore, it is essential to impose some flexibility and to indicate that the diagram is simply a guideline.

The definition provided by CERT aligns to some extent with the structure in Figure 3.5 and reads as follows: “A malicious insider threat is a current or former employee, contractor, or business partner who has or had authorized access to an organisation’s network, system, or data and intentionally exceeded or misused that access in a manner that negatively affected the confidentiality, integrity, or availability of the organisation’s information or information systems” (Cappelli et al., 2012). This definition addresses the numerous shortfalls of the previous definitions by supplying the necessary detail deemed as important for defining insider threat in this chapter. However, it does not include the types of harm that could be caused.

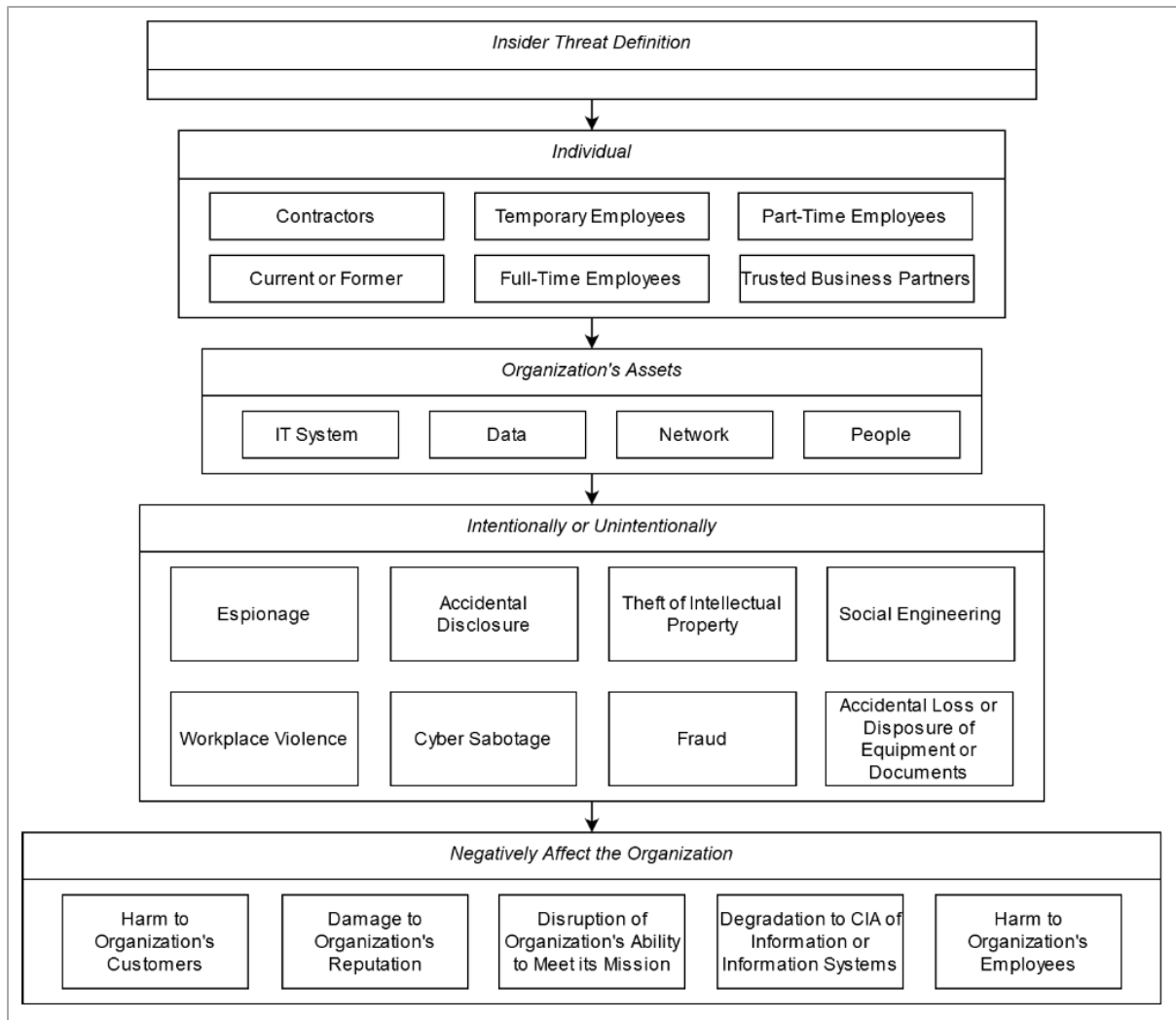


Figure 3.5: Structuring the definition of insider threats in the cybersecurity domain (Costa, 2017)

Kowalski et al. (2008) formulated a similar definition and described an insider threat as a current or former employee, or someone who was affiliated with the organisation, who has had legitimate access to the company's network, system and non-public data, and who exploited this access to compromise the confidentiality, integrity or availability (CIA) of the organisation's data or systems. These authors extended their definition to include negligent employees (Kowalski et al., 2008) and thus introduced the notion that there are categories of insider threats who are all triggered by different motivations.

Spooner et al. (2018), Claycomb et al. (2013), Munshi et al. (2012) and Cappelli et al. (2012) suggest three main types of insider threats. An additional type of insider threat, negligence,

is suggested by Young et al. (2014). Brief explanations of these types are provided below and will be expanded on in greater detail later in this dissertation.

3.2.5.1. Insider IT sabotage

Insider IT sabotage refers to a scenario where an employee has started to resent the company and becomes disgruntled for a specific reason, perhaps a poor performance review, and therefore desires to inflict harm on the company (Cappelli et al., 2012). An example from the case study shows Vincent Kaminski who waited a significantly long time for his promotion, which was intimated as a circumstance that could have caused disgruntlement.

3.2.5.2. Insider fraud

Insider fraud refers to a scenario where an employee is involved in illegal activity for various reasons such as to harm the organisation or for personal gain (Young et al., 2014). Fraud was referred to on numerous occasions in the case study, where top level executives within Enron, John Lavoireto and Tim Belden, were using deceptive accounting techniques to conceal the firm's debt from investors.

3.2.5.3. Insider intellectual property (IP) theft

Insider intellectual property theft refers to the theft of information that is created and owned by an organisation, such as client data, product data, business-related data or software and source code, to name a few (Kowalski et al., 2008). An employee who is guilty of insider theft feels entitled to the work that he or she has produced for the organisation and is adamant to obtain due credit for the work done.

3.2.5.4. Negligent insider

A negligent insider is an employee who does not act with malicious intent, but who is careless and does not read or follow proper security procedures (Young et al., 2014). Evidence was located within the Enron email dataset (as per the case study), where employees sent the company's credentials and bank details to unscrupulous recipients.

Although each of these four types of insider threat has different motivations, the same result is achieved, namely causing harm to the organisation.

3.3 Preparing the context-specific insider threat definition

In the context of this research, the way in which the insider threats are carried out or detected is a key focus. In this research, only insider threats that originate and lurk in corporate email communications are studied. This is essential to note for the following section.

3.4 Proposed definition of insider threat for use in this research

The diagram shown in Figure 3.5 created by Costa (2017) has been adapted based on the findings in this chapter, and the updated model is shown in Figure 3.6. The titles of the components were altered to better relate to the research at hand. Figure 3.6 includes only Information Technology assets, because the focus of this research is cybersecurity and the email platform must be used to detect threats. The diagram was adapted to include only the four high-level types of insider threat that serve as umbrella terms for most malicious and non-malicious actions in this dissertation. The negative effects to the organisation, shown within the diagram, were also expanded to include financial losses and operational damages.

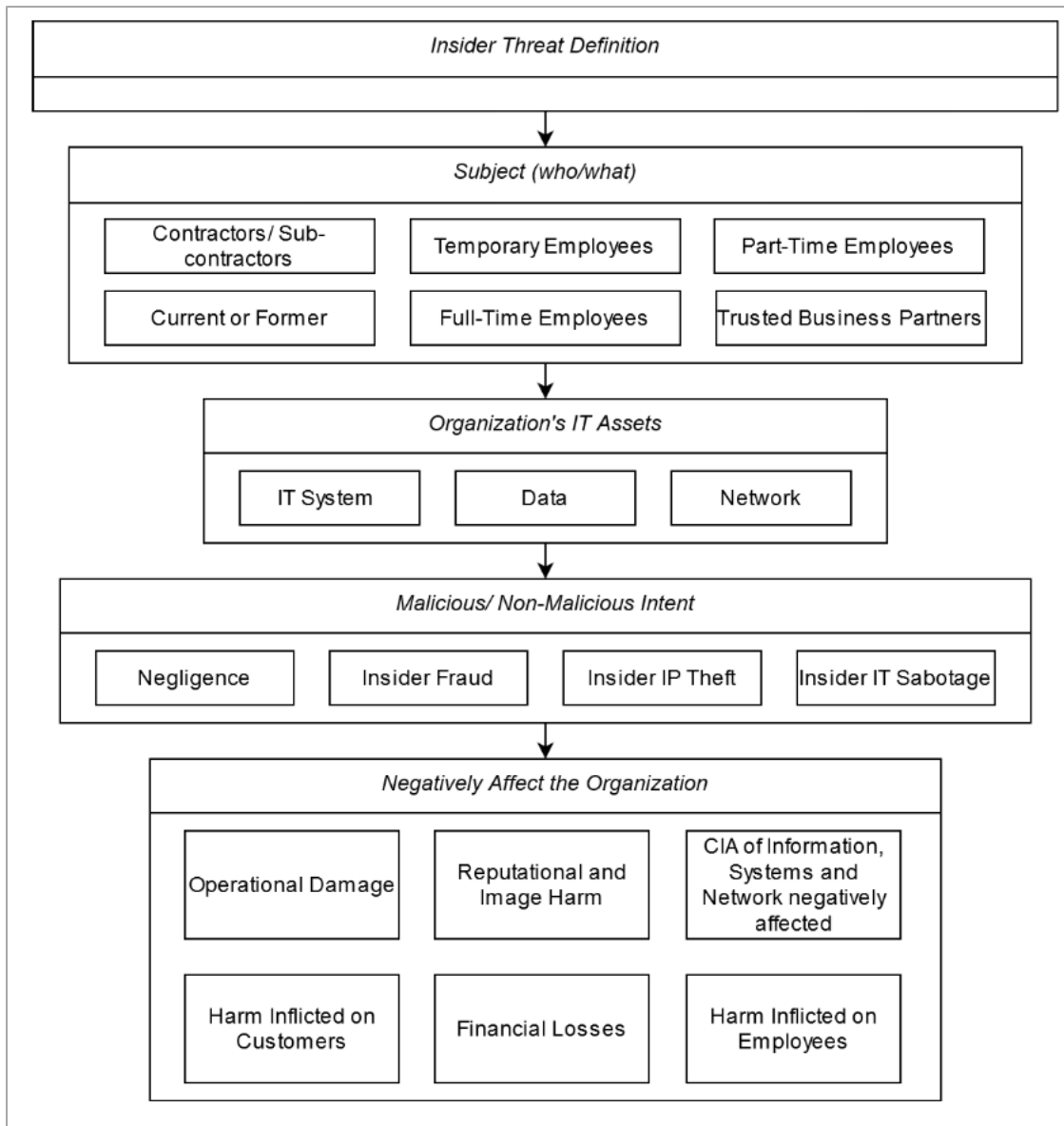


Figure 3.6: Adapted model aligned to the context of this research (Costa, 2017)

The definitions for insider threat provided in past research, as well as the guideline structure provided by Costa (2017) as adapted in Figure 3.6 (Costa, 2017) have been assimilated to define insider threat as intended in the context of this research.

For the purpose of this research, the following definition is proposed for the term insider threat:

An insider threat is a cybersecurity threat that is carried out by a current, former, full-time, part-time, temporary, contracting employee or trusted business partner who was legitimately assigned privileged access to perform his or her daily work but exploited such access to the

organisation's network, systems and data to compromise the confidentiality, availability or integrity of these assets. This individual acts with either malicious or non-malicious intent and their actions can be classified as either insider IT sabotage, insider fraud, insider theft of intellectual property, or negligence. The individual's actions might have a negative impact on some critical parts of the organisation such as the organisation's employees, customers, reputation, image, and information technology, or these can cause financial and operational damages. Lastly, this specific type of insider threat is propagated by the individual using their company email account.

Throughout this dissertation, the above definition for insider threat will be applicable to the research conducted and is depicted in Figure 3.7. The inclusion of the organisation's IT assets in Figure 3.7 is extracted from Figure 3.2 and Figure 3.6 (Costa, 2017).

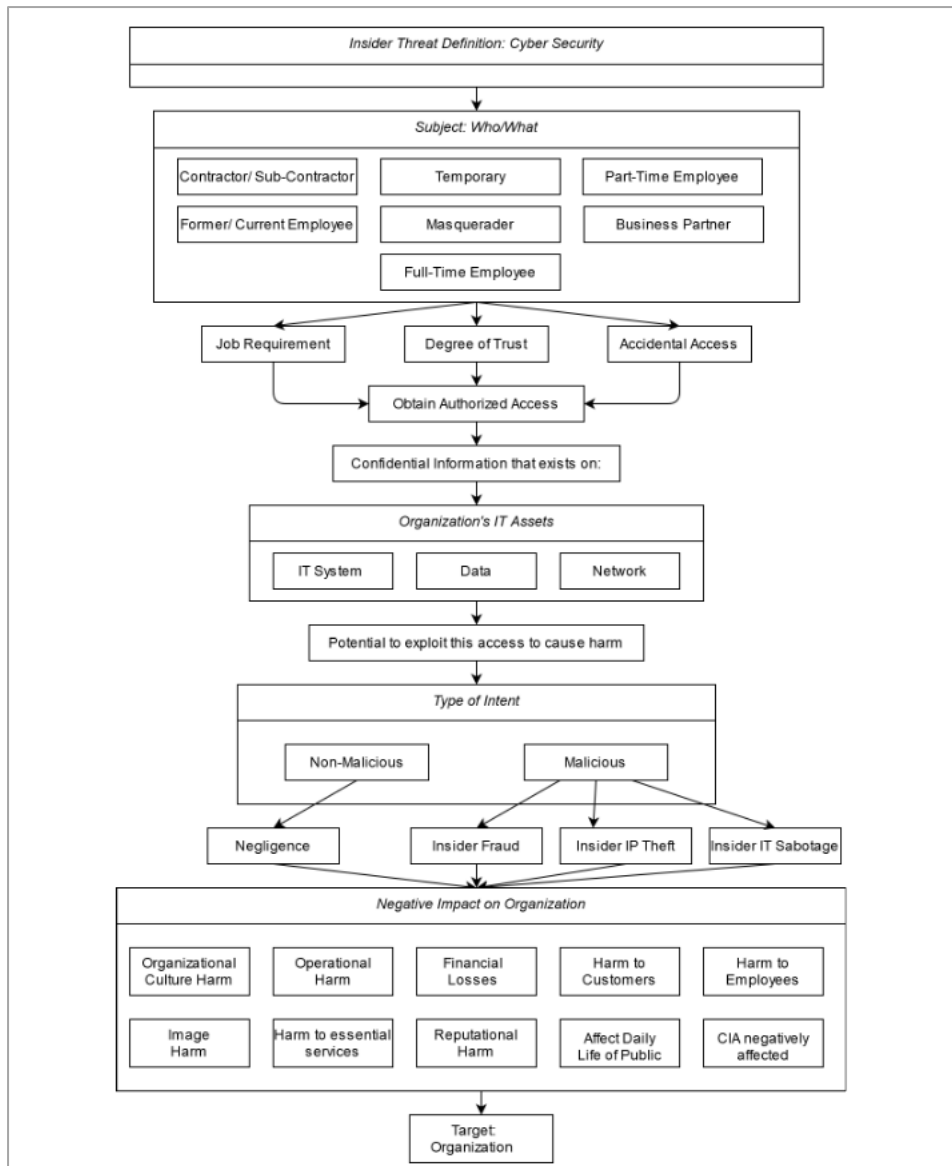


Figure 3.7: Diagrammatic structure of the definition for insider threat

The two types of intent that were introduced previously are expanded on in Figure 3.7 to include the threats stated in the studied literature as the main types of insider threat (Young et al., 2014; Cappelli et al., 2012; Kowalski et al., 2008). The component 'Negative Impact on Organisation' shows various types of harm that can be caused to an organisation and these were extracted from Figure 3.4. In addition, the 'Target: Organisation' component was extracted from the same figure.

Thus, the definition that was created for insider threat as used in this study includes the individual definitions for insider and threat (both the general definitions and those from a

cybersecurity perspective), as well as definitions provided for the combined term insider threat.

3.5 Conclusion

In conclusion, the purpose of Chapter 3 was to obtain various definitions of the individual terms, insider and threat, as well as of the term insider threat. These definitions were then analysed in conjunction with the scope of this dissertation to compile a relevant definition for the term insider threat to be used specifically in the research at hand. Various standards authorities, textbooks and journal papers were consulted and referenced in this chapter to garner key elements from their definitions that were relevant and necessary for the definition created in this chapter. A diagrammatic structure, consisting of the main components and sub-components of the definition, was used to guide the structure of the final definition.

Chapter 4

Insider Threat Behaviours

4.1 Introduction

In the previous chapter, it was noted that a critical threat encountered in organisations is the threat facilitated by people within the organisation – be it the contractors, employees, business partners (to name a few) – and this threat is known as an insider threat. The concept of insider threat was thoroughly defined in the context of the research at hand.

Chapter 4 presents an in-depth literature review of some of the main components of this dissertation. Specifically, the chapter includes a discussion based on how the email platform can be used to facilitate insider threats. The chapter also delves into the insider threat problem within organisations, the characteristics of insider threats, as well as various (human) behaviours displayed by these insiders. This is done to create a foundation for and expand on some of the components in the problem statement of this research.

4.2 Emails as a platform for attacks

Nowadays, most organisations are highly dependent on their IT infrastructures and email platform, and simply cannot function without these (Brown et al., 2013). This is because of their efficiency, ease of use, low cost, speed and lack of time-zone barrier (Zaki et al., 2017). Email infrastructures, however, have become a main cybersecurity vulnerability within an organisation that often allows malicious activities to be conducted (Butkovic et al., 2013). The focus of this dissertation is to determine means in which evidence of insider threat can be discovered within the email dataset of an organisation. According to Michael and Eloff (2019), the threat posed by insiders within organisations is that they can exploit email platform vulnerabilities, either intentionally or unintentionally, within their organisations.

The examples provided in the case study also make it clear that email infrastructures can host a number of cybersecurity threats in organisations and specifically insider threats. This chapter, going forward, will focus on employees constituting insider threats through their use of email communications. The subsequent section discusses the damage that insider threats can cause within organisations, as well as the identifiable characteristics of insider threats.

4.3 Insider threats in organisations

Attacks on organisations caused by insiders are increasing by the day and the consequences of these present a substantial threat to organisations (Michael & Eloff, 2019). In addition, organisations' increased use of and reliance on information technology also enhance the risk of insiders exploiting this platform to commit crimes within organisations (Bell et al., 2019).

There are various characteristics of an insider can contribute to insider threat activity. If properly understood, these characteristics can aid organizations in the detection of insider threats. The list of characteristics in the following section was compiled based on commonly identified attributes of insiders discussed in past research (Cappelli et al., 2012; Munshi et al., 2012; Young et al., 2014).

4.4 Characteristics of insider threats

4.4.1 System access and trust

The first characteristic that distinguishes an insider involves the correctly authorised access that the insider has obtained from the organisation to access the company system, network and data (Kowalski et al., 2008). Homoliak et al. (2018) agree that the insider has legitimately obtained access to confidential company resources and information. The employee does not have to force entry, as they have the right to access authorised systems to conduct the crime (Noever, 2019; Chi et al., 2016; Munshi et al., 2012). The insider must therefore be distinguished from an external party due to their having access to some level of the security layers of the organisation (Hunker & Probst, 2011).

System access is associated with a degree of trust that the employer puts in the employee (Greitzer et al., 2019). The employee has earned the trust by having fulfilled a certain function, or by successfully completing a background check (Alawneh & Abbadi, 2011). It is clear that most literature similarly associates the property of system access or of access privileges with insiders.

4.4.2 Technical skills and organisational knowledge

Insiders possess some degree of technical skill that allows them to conduct malicious activity that results in a more dangerous threat (Hunker & Probst, 2011). Insiders might spend a sufficient amount of time in their job working with a given system and as such, have obtained knowledge regarding the location of certain information and data. They might have limited

technical abilities, but the mere fact that the insider is able to retrieve certain information in specific locations, that are not available to the public, is a factor that benefits the insider. An employee with a strong technical aptitude might be able to run commands and perform sophisticated attacks that an ordinary employee would not be able to, such as executing a logic bomb (Munshi et al., 2012).

An insider is furthermore distinguished from an external person due to their knowledge of the organisation's policies, procedures and security measures, to name a few (Munshi et al., 2012). This knowledge can serve as an advantage to the insider who wants to conduct an attack. Furthermore, the insider may also be aware of means to conceal their malicious activities by using the combination of their technical skills and knowledge of the organisational processes (Furnell, 2004). This combination is an important characteristic of an insider threat, as it distinguishes the insider from an outsider.

4.4.3 Knowledge of system vulnerabilities

Past research aimed at detecting insider threat found that insiders are often aware of vulnerabilities within the organisation's systems and as such can easily exploit these to conduct an attack (Mills et al., 2017). In addition, if the employee is technically inclined, they might even build a vulnerability into a system that they are developing or currently utilising. Homoliak et al. (2018) argue that insider attacks are challenging to detect because the insider has knowledge of the 'weak points' within the deployed systems or business processes. A former employee can also set up backdoor accounts and use hacking software to obtain admin access to the systems. If the employee is not technically inclined, but has malicious intent and awareness of a vulnerability, they might well enlist an external party to assist with exploiting this vulnerability.

It is evident that this characteristic of insider threats (that is, exploitation of vulnerabilities) requires technical aptitude, and consequently not all insiders would be capable of performing an attack in this way. There is an association between knowledge of system vulnerabilities and the previous characteristic regarding the employee's technical skills and knowledge of organisational processes. As such, when classifying insider threats within a technical domain, knowledge of system vulnerabilities serves as a characteristic by which to identify a potential malicious insider.

4.4.4 Motivation

An insider would only be inclined to conduct an attack if they have strong motivation to do so. An employee would rationalise that the best course of action is an attack for reasons such as the benefits that can be obtained and the chance to go undetected. Several factors might serve as motivation. Bell et al. (2019) state that an employee is motivated either by a personal reason or a work-related reason. For example, an employee who is in financial need or has a strong sense of entitlement could easily be persuaded to exchange confidential company information with an external party (Homoliak et al., 2018). Furthermore, the employee might be awarded with an offer of better employment from the outsider after having conducted the attack (Cappelli et al., 2012). A work-related reason could motivate an employee who feels a sense of resentment against the organisation to use technology to seek revenge (Clark, 2016).

However, it cannot be assumed that only motivated employees can cause malicious attacks. As stated previously, an employee might simply be negligent or distracted, and unknowingly assist an external attacker to cause harm to the organisation. However, motivation is an important characteristic of an insider threat and works in conjunction with the next characteristic of changed human behaviour.

4.4.5 Human behavioural change

This specific characteristic of an insider is examined in greater detail in the rest of this chapter, as it is also a main theme of this dissertation. A brief introduction is given to human behaviour changes, as it was found that an insider is most likely to display a change in behaviour when planning or conducting an attack (Greitzer et al., 2012). According to Bell et al. (2019), two types of human behavioural change can be identified within organisations and both are discussed below.

4.4.5.1. *Change of physical attitude*

The first type of behavioural change relates to a change in physical attitude, such as an employee who suddenly becomes withdrawn and distant, even though this is not the typical nature of the employee (Azaria et al., 2014). Another change would be an employee who has started to exhibit anger management problems and as such is unable to use adequate self-control in the workplace.

If an employee displays a change in behaviour, he or she should be carefully monitored because it is likely that the employee might be plotting an attack.

4.4.5.2. Conducting activities unrelated to the employee's daily work

The second type of behavioural change would be an employee who suddenly conducts odd activities in the workplace, such as logging into a system or retrieving authorised information that is not relevant for this employee's daily work (Probst et al., 2010). Mills et al. (2017) found that there is a correlation between frequency of suspicious non-business-related employee activity and an employee being a malicious insider.

It is therefore evident that motivation and behavioural changes are closely related. In fact, both attributes are essential to classify an insider threat.

4.4.6 Outsourcing

It is common for third-party employees such as contractors and sub-contractors to enjoy a certain level of trust in and access to the organisation, for use in their job at the organisation. As such, they are trusted with knowledge on where certain information can be retrieved.

Munshi et al. (2012) argue that contractors and third-party employees will typically not have the necessary motivation to conduct attacks on their employer, and therefore only 20% of incidents are found to be carried out by these individuals. It is clear that the risk posed by this category of employees is lower than that of an employee who works full time at the organisation. However, the fact that there is still a risk, allows for outsourced employees to be classified as a potential characteristic of a type of insider threat.

4.4.7 Organisational culture

Munshi et al. (2012) suggest that changes in organisational culture, if not correctly communicated to employees, can spark concern, confusion, as well as ambiguity for employees. This would then have a negative impact on their attitude towards the organisation. It was found that a hostile or distrusting work environment, as well as an environment experiencing constant change and disruption in terms of reorganisation, might also trigger an employee to perform a malicious action (Cappelli et al., 2012). Insiders can act out of frustration that stems from new organisational policies, decisions made by management, and an overall culture that does not align with the employee.

However, from past research it was clear that very few insiders who had previously committed crimes, were motivated to act due to their unhappiness with the organisational culture (Munshi et al., 2012). However, even if the culture contributes only a small part to an employee's motivation to attack, it could add to their disgruntlement with the job and organisation, especially if the employee does not feel included or appreciated. These are feelings that may lead to an employee's disgruntlement, and since they have led to malicious attacks, they should not be disregarded (Greitzer et al., 2012).

4.4.8 Summary

The seven characteristics (discussed in Sections 4.4.1 to 4.4.7) that contribute to insider threat were found to be recurring themes in past research, and as such they were considered relevant to this study. They also established the essence of insider threats in this work.

Mills et al. (2017) suggest that system owners focus mainly on external threats such as hackers, and therefore they do not focus enough effort on the damage that could potentially be caused by insiders. A possible reason for an organisation's lack of attention to insider threats could be because it requires substantial research into possible detection methods (Greitzer et al., 2019). A countermeasure recommended by the United States Cybersecurity Magazine (Ali, 2018) is that organisations should start to emphasise the monitoring of employee activity and behaviour. A similar approach is considered in the research at hand, and as such human behaviours constitute a main component pertaining to the topic of this work.

The next section covers the changes in human behaviour that may be displayed by insiders and that could serve as evidence for organisations that their employees are indulging in malicious activity.

4.5 Insider threats driven by human behaviour

According to Jiang et al. (2018), recent research has included the analysis of human psychology and emotions of employees in the detection of insider threats. Various considerations regarding human behaviours are discussed in this section.

4.5.1 The window of opportunity

Regarding identification of the threats, there is often a window-of-opportunity period between when an employee starts to display visible behavioural changes and when he or she actually conducts the attack (Greitzer et al., 2012). If an organisation sets up a proper detection method, the indicators of the pending attack can potentially be discovered in this window.

There are sufficient research references to show that attacks are more often executed before the indicators can be identified by the organisation. Shaw and Fischer (2005) concluded that a large percentage of the attacks reported on in their study could have been minimised, had a proper plan been set up to manage the psychological and behavioural change factors revealed by insiders. However, they also noted that the poorly constructed intervention methods that were utilised by organisations in an attempt to curb the threats, in fact intensified the problem (Shaw & Fischer, 2005).

To successfully address the insider threat problem at hand, a study should be conducted not only of the employee's technical activity, but also of their verbal and physical behaviour and personality traits. This finding is echoed by Munshi et al. (2012) who indicated that the insider threat is both a technical and a human behavioural problem. Identifying an employee's changed behaviour might reveal the motivation that triggered the employee to conduct an attack, as well as the type of attack that might be conducted.

4.5.2 Behavioural indicators in the workplace

Bell et al. (2019) state that, in the traditional workplace setup, behavioural changes exhibited by an employee would strictly be a Human Resources (HR) manager's issue. However, it is becoming increasingly necessary for team members who identify such a problem to also get involved in reporting it and to assist with mitigating the threat (Bell et al., 2019). In the context of this work however, the focus is on equipping organisations with detection mechanisms to identify these threats via a technical platform and not through word of mouth from co-workers.

4.5.3 False positives

Human behaviours are known to be unpredictable and result in false positives and incorrect results (Probst et al., 2010; Munshi et al., 2012). That is why it is important when observing changes in behaviour to monitor factors such as whether certain actions undertaken by an

employee – suspicious as they may seem – are expected in his or her type of job (such as logging into a certain system many times a day). A suddenly stressed employee might be considered malicious if his or her behaviours are scrutinized, even though the employee is a well-intentioned, quality and loyal performer. Only when an adequate level of suspicion has been caused by a combination of factors and characteristics, such as poor work performance and dishonesty, can an employee confidently be classified as an insider.

Several types of human behaviour changes displayed by employees could be linked to different types of malicious or non-malicious harm that is caused to an organisation. These are discussed in the following section.

It should be noted that when detecting insider threats, the identified behavioural indicators should not only be combined with other factors, they should also be carefully considered (Azaria et al., 2014).

4.5.4 Types of behaviours

Various types of behaviours that are considered psychosocial predictors of insider threat activity were identified in past research and these will be discussed below (Greitzer et al., 2012). The behaviours shown in Table 4.1 were chosen as they were cited by various authors who conducted research that was quite similar to that of this dissertation (Ackerman & Mehrpouyan, 2016; Azaria et al., 2014; Greitzer et al., 2012; Hunker & Probst, 2011). Under the heading ‘Description’, a possible reason or motivation is given for the employee’s behaviour that led to an attack.

Table 4.1: Observable behaviours of insider threats

Observable behaviour	Description
Disgruntlement	There are several reasons that can cause employee dissatisfaction within an organisation. Firstly, an employee who might not have received a certain increase, promotion or recognition could begin to feel that his or her efforts are not appreciated (Greitzer et al., 2012). Secondly, the employee might feel that their current job and position are not aligned with their skills, knowledge, interests, or level of

	<p>expertise. This may lead to the employee's low morale and feelings of resentment towards the employer and the organisation.</p> <p>However, disgruntlement can also be caused by the employee's complaints being disregarded, such as a complaint about sexual harassment (Azaria et al., 2014). Therefore, the employee may feel justified to take revenge on the organisation. His or her emails might contain rants against the employer and show negativity towards the organisation.</p>
Difficulty accepting criticism	<p>This employee who has a sense of pride but also feels insecure, would become defensive and be personally offended when developmental feedback is given by superiors. Feedback can relate to the employee's performance, attitude, work quality or behaviour in the workplace. Greitzer et al. (2012) state that this employee would be too proud to acknowledge or admit to errors, and may use dishonesty to disguise these mistakes. This criterion is also associated with disgruntlement (Azaria et al., 2014). The employee's errors and dishonesty might lead him or her to unintentionally cause harm to the organisation.</p> <p>However, the hostile feelings regarding harsh feedback that was received might also trigger the employee to intentionally facilitate an attack. This might be evident in the emails that the employee has received regarding mistakes made, or in emails that were sent, rejecting feedback or containing hate speech against the organisation.</p>
Anger management problem	<p>An employee who displays this behaviour will have difficulty controlling their emotions when filled with rage (Greitzer et al., 2012). A trigger can be unhappiness with a work policy or new instruction, as well as a general feeling of unhappiness or disgruntlement in the job and work environment. Furthermore, the employee might have bottled these feelings for a long period and is no longer able to restrain them.</p> <p>An anger management problem can be observed in offensive correspondence such as emails being sent that contain insulting statements as well as vulgar language. Such an employee could be considered a risk to carry out a crime.</p>
Detachment	<p>This refers to an employee who appears withdrawn and distant in the workplace setting. The employee will keep interactions with other colleagues to a minimum (Azaria et al., 2014; Greitze et al., 2012). If</p>

	<p>someone who was previously an engaged employee, becomes distant, this is also a cause for concern (Bell et al., 2019). A detached employee might be concealing negative thoughts and feelings towards the organisation and as such may be plotting an attack. This employee might even have received an email of concern from HR regarding his or her sudden introversion.</p>
Disregard for the organisation's rules and authority	<p>A major threat can be posed by an employee who does not follow company policy and rules and who is often reported for a lack of regard for authority. In addition, the employee is unaware of or does not adhere to IT-related rules (Azaria et al., 2014). Greitzer et al. (2012) indicate that a possible reason for this is the employee having a sense of being above the law.</p> <p>It presents a risk to the organisation if such an employee works with sensitive information and systems that require adherence to the organisation's cybersecurity policies. The employee may commit a malicious action unintentionally, having received emails regarding outstanding security policy training, or warnings about security or other rule violations in the organisation.</p>
Poor performance	<p>An employee who is dissatisfied with the organisation will demonstrate a decline in work performance and may submit poor quality work. The employee may consequently have received a warning or some form of negative feedback (Greitzer et al., 2012) that can be located within the employee's email communications. This warning or feedback can be a motivating factor that causes the employee to commit an insider threat crime.</p>
Severe stress levels	<p>Due to the nature of the job, and the position and character of the employee, the latter will observe a high level of stress within the organisation. In this position, the employee might endure much stress and might not alert authorities, due to fear of job termination, poor appraisals, as well as a decrease in pay. Thus, the employee is unable to handle the stress properly (Greitzer et al., 2012).</p> <p>Furthermore, if an employee is stressed due to external factors such as financial or personal stress, they would be a good target for exploitation by external parties to partake in malicious activities. Emails might be sent to the employee regarding financial rewards that could be paid in exchange for revealing company information.</p>

	Furthermore, emails sent to a co-worker reflecting concerns regarding the employee's stress might be further evidence.
Argumentative or confrontational behaviour	This applies to an aggressive employee who often indulges in conflicts as well as in the bullying of fellow workers (Greitzer et al., 2012). These strong character traits possessed by the employee have an underlying motivation such as disgruntlement, as they are unhappy with certain elements of their job or of the organisation in general. Azaria et al. (2014) compare these traits to the anger management problem (Azaria et al., 2014), as they could drive an employee to carry out harmful actions. Emails reflecting conflicts with co-workers or employers might exist as evidence for this.
Inability to separate private and professional life	As stated in the severe stress levels category, stress can be caused in the employee due to an inability to separate their professional and private life (Azaria et al., 2014). This has a negative impact on the employee's work performance (Greitzer et al., 2012). As such, the employee might be vulnerable, careless and distracted when performing certain actions on the organisation's information system. An example of an email relating to this would be an outstanding bill or payment that adds to the employee's stress during work hours.
Egotistical	This employee prioritises his or her own needs and goals within the organisation and as such completely disregards the requests of employers and co-workers. An egotistical employee is not loyal to the organisation and therefore poses a high risk. Therefore, the employee would easily exchange company information for incentives offered by a malicious outsider (Azaria et al., 2014). Incentives could include a financial reward or a new offer of employment. Evidence of offers of employment or financial reward in exchange for benefits might be located in this type of employee's emails.
Unreliability	This employee cannot be depended upon in the organisation for timely delivery of work, quality of work, or for remembering certain details. Such unreliability may be fuelled by stress or being distracted and implies that the employee cannot be trusted (Azaria et al., 2014). This would be a problem as there would be a conflict between what the employee says and actually does. He or she may secretly partake in malicious activity. Emails from the employer sent to the employee regarding the latter's lack of focus and unreliability could signify this behaviour.

Absenteeism	An employee may show a lack of interest in the organisation through arriving late for work, not arriving at all or presenting a poor work ethic (Azaria et al., 2014). The employee offers poor excuses for the frequent absenteeism and cannot be depended upon. This individual can also be seen as a risk, as he or she is disloyal and can easily fall prey to a malicious external party. The employee may have received emails warning against tardiness and may have sent emails offering excuses for their absence from work.
-------------	---

These types of employee behaviour shown in Table 4.1 are grouped according to categories which are the main types of insider threats that have been developed by authors such as Cappelli et al. (2012) and Young et al. (2014). The next chapter explores these categories in detail and relates the different behaviours to the types of insider threat.

4.6 Conclusion

Chapter 4 focused on a theoretical discussion of some of the main themes of this research. Firstly, it was noted that the email platform, which is commonly used in organisations, is becoming a platform for facilitating malicious activities in an organisation. Characteristics and behaviours associated with insider threats were identified in this chapter and it was noted that they could be detected within an email platform. This finding showed that the human aspect is critical to the detection of insider threats.

The next chapter hones in on a discussion of the four main types of insider threat that can be discovered in email communications as well as their association with the identifiable human behaviours. In addition, Chapter 5 will look briefly at possible phrases that could signify an insider threat lurking in an employee's email communication. It will also consider detection techniques that utilise some of the theoretical aspects.

Chapter 5

Types of insider threat

5.1 Introduction

Chapter 4 highlighted characteristics associated with typical insiders lurking within organisations, as well as the different behaviours that can be exhibited by various types of insiders. Chapter 2 delved briefly into the types of insider threats and provided various examples of insider threats in the case study. This was critical to indicate that insiders can be classified into different groups based on certain behaviours, intent and motivations that are relevant to the research at hand. Therefore, continuing with the theoretical discussion, Chapter 5 will delve into the types of insider threats by discussing human behaviours, examples, as well as possible phrases that may be found lurking in employees' email communications.

5.2 Categories of insider threats

In this work, four main types of insider threats are studied, based on the categories defined by Cappelli et al. (2012) and Young et al. (2014).

The sub-sections that follow shed light on the four main insider threat types, so as to allow the reader to obtain a clear understanding of what they involve. This is facilitated through the use of real-world examples and the mapping of human behaviours. These paragraphs provide a description of each type of insider threat as well as the factors that may trigger or motivate an employee to pose such a threat. Real examples, specifically occurrences of this threat that were detected in the Enron email dataset, are also shown. The examples are linked to the behaviours that were found to be commonly exhibited by insiders. The discussions will also focus on the means by which the particular insider threat type can exist in email communications, as this is essential to validate the problem statement in this study. Each of the sub-sections conclude with a summary table that can be used for reference purposes.

5.2.1. Insider Information Technology (IT) sabotage

This first type of insider threat refers to an employee who intends to cause malicious harm to an organisation by making use of their own technical skills and privileged access (Chi et al.,

2016). An employee who is driven to cause insider IT sabotage requires a strong motivation to attack. Vengeance would be a common motivation when the employee feels diminished, undervalued, personally offended or has unmet expectations within the organisation (Clark, 2016; Young et al., 2014). In addition, conflicts with co-workers and an overall sense of dissatisfaction with regard to the type of work, culture and working environment can also fuel up feelings of hatred towards the organisation (Cappelli et al., 2012).

Clark (2016) indicates that this type of insider usually works outside normal business hours to plot and conduct the sabotage. This finding was confirmed by Maasberg et al. (2015), who stated that this type of employee not only performs the malicious actions outside of business hours, but usually is no longer employed in the organisation. This person's contract might have been terminated by the employer as a result of continual negative performance reviews.

Various reasons may trigger an employee to conduct such an attack, some of which will be described in the examples below.

5.2.1.1. Real-world examples of insider Information Technology (IT) sabotage

In the case study, Vincent Kaminski (Managing Director at Enron) was identified as an employee who had waited a significantly long period for his promotion. As such, he might have harboured feelings of negativity towards the company. It is known that he stayed mostly silent regarding the corruption at Enron, possibly to witness the company's downfall, or due to a lack of motivation to expose the illegality and save the company. Furthermore, this employee might actually have received negative feedback regarding his performance within the organisation, which caused feelings of disgruntlement (Cappelli et al., 2012). These feelings tie in with several of the human behaviours exhibited by insiders, with disgruntlement and difficulty accepting criticism among them.

Another employee from the Enron dataset, Pamela Allison, was also found to express feelings of disgruntlement during her time at the company. The Enron case study shows an email sent by her, indicating that the environment was hostile, and that employees were not treated well by top management and were constantly being scolded. In addition, Allison noted that the managerial staff were under immense stress. This employee merely resigned from the organisation, whereas any other employee, having experienced the same treatment or environment, might have felt compelled to inflict harm on the organisation.

A real-world example of harm caused by insider IT sabotage is an employee who was employed at the electric vehicle company Tesla. He received several poor performance reviews, and was even transferred to a lower position within the organisation (Schwartz, 2018). This employee was a strong technical candidate and therefore used his technical skills to wreak vengeance on the organisation by mainly tampering with the operating system code. Part of his attack included exporting a large amount of confidential data to outsiders (Schwartz, 2018). He also concealed his identity during the attack so that his username could not be traced, and as such, framed his colleagues. This attack had a major negative financial impact on Tesla. Afterwards, when investigations were carried out, it was noted that certain behavioural indicators had been present before the attack. For example, the employee was found to have been abrupt with his colleagues. This type of insider threat links to the anger management problem and argumentative or confrontational behaviour, which stem from unhappiness experienced within the organisation (Schwartz, 2018).

The next paragraph focuses on the use of email communications to detect behaviours relating to a potential insider IT sabotage threat.

5.2.1.2. Insider IT sabotage detected in employee email communications

Keeney et al. (2005) provide an example of an employee who was responsible for an insider IT sabotage attack on his organisation. Inspection of the employee's emails indicated that he had previously emailed his employer regarding his unhappiness in his current job. He threatened to produce lower quality work if the employer would not adhere to certain demands. Three of the main factors regarding insider IT sabotage can therefore be gleaned from this example. First of all, email content can reveal wording and behaviour that could suggest a potential insider threat. Secondly, the employee is strongly motivated to carry out the attack. Thirdly, there is a clear time window between becoming motivated to attack and conducting the actual attack, and as such, planning is involved in the attack.

5.2.1.3. Summary Table

Table 5.1 summarises known components of insider IT sabotage, as well as potential phrases that could be identified in email communications as evidence for this type of threat.

Table 5.1: Summary of insider IT sabotage

Component	Explanation
Description	An employee who is technically capable has received negative feedback or has an unfortunate experience within the organisation and therefore conducts an attack on the organisation (Cappelli et al., 2012). The employee is motivated to take revenge on the organisation due to feeling undervalued and unappreciated, lacking job satisfaction, as well as being unhappy with the organisation’s culture and management decisions (to name a few). The employee exploits and abuses their privileged system access and knowledge (Chi et al., 2016).
Observable (human) behaviour	<p>An employee who carries out an insider IT sabotage attack can potentially exhibit one or more of the following behaviours: disgruntlement; difficulty accepting criticism; struggling with anger management; detachment; poor work performance; being argumentative or confrontational; unreliability; absenteeism.</p> <p>These behaviours can be detected in email communications by inspecting an employee’s sent and received emails to find traces of co-worker conflict, rants regarding dissatisfaction with the employer or organisation, and work quality related feedback, for example. (Spooner et al., 2018). Furthermore, emails indicating tardiness with regard to late arrival at work and aggression can also be indicative of a potential attack (Cappelli et al., 2012).</p>
Phrases	Phrases located within email communications that relate to the behaviours shown in this summary, may include “wasted efforts”, “not happy at work” and “uncertain about the future” – to name a few (Whitman, 2016).

5.2.2. Insider intellectual property (IP) theft

Insider IP theft occurs when an employee feels inclined or a sense of entitlement to steal the organisation’s confidential information for their own benefit (Cappelli et al., 2012). Sensitive company information can include trade secrets, source code, as well as customer-, employee- and business-related data (Nurse et al., 2014). In some cases, when an employee has worked on a particular system or created new information for the organisation, they may insist on ownership regardless of the fact that they signed specific IP agreements (Moore et al., 2009). Research shows that IP theft attacks are most commonly conducted by technical employees who use technical means such as the company’s email platform (Nurse et al., 2014).

The employee who commits IP theft may display specific behaviours that will draw attention to the potential threat posed to the organisation. Firstly, this employee might display egotistical behaviour by placing their own needs above those of anyone else in the organisation (Greitzer et al., 2012). As such, the employee does not prioritise the requests of others in the organisation, unless it holds an opportunity for his or her benefit. If a situation

offers financial benefit or a benefit in terms of praise and promotion, the employee would be more inclined to get involved. This is also the type of employee who, when approached by an outsider for confidential information, would gladly exchange it for a reward (Azaria et al., 2014). This employee would therefore have no regard for the organisation. Nurse et al. (2014) suggest that an employee who displays traits such as narcissism and has a manipulative and an immoral disposition would most likely be the type to carry out an insider IP theft attack.

5.2.2.1. Real-world examples of insider intellectual property (IP) theft

In the Enron case study, an executive, Paula Rieker, used confidential company information regarding the decline in the company's stock price due to the manipulation of powerplants and the concealing of debt, for her personal gain. She obtained shares at a very low price but concealed this information from the investors. She subsequently sold her shares at a much higher price, even though she was acutely aware of the massive financial losses suffered by the company. Rieker was charged with insider trading once the scandal was investigated.

In a scenario provided by Cappelli et al. (2012), a sales employee was offered a decent employment opportunity by another company. However, the appointment was dependent on the condition that the employee would submit specific confidential information (such as, customer information, computer programmes and marketing information) to the potential new employer. This employee therefore spent the last two months at the first organisation forwarding all the requested information to their private email account. Once the information was handed over, the employee received a formal offer of employment from the second company. It is clear from the scenario that the employee was disloyal towards the original employer and placed their own needs above those of the organisation, regardless of any IP contracts signed. Moreover, the employee used the employer's email platform to accumulate the stolen information.

Such an employee would potentially be motivated by financial gain, as they might be experiencing personal financial issues that could tempt the employee to consider the offer of an attacker (Greitzer et al., 2012). Moore et al. (2009) also list possible motivations for stealing information, such as to help a new employer or the government, to be used in the employee's own direct competitor's business, or to exchange the information for a new job offer, as shown in the example from Cappelli et al. (2012). If the role of the employee in developing

the information was specifically important, they might feel an even greater sense of entitlement and wish to claim greater benefit from it (Moore et al., 2009). The employee may see the work as their intellectual property that they should rightfully own.

5.2.2.2. Insider intellectual property (IP) theft suggested in employee email communications

When inspecting an employee’s email communications, it is important to search for emails received from interested external parties in which they enquire about information within the organisation and propose monetary rewards. Furthermore, evidence of an employee sending large amounts of data via email to their private email account, or to that of an external party, should also be a cause for concern.

5.2.2.3. Summary Table

Table 5.2 summarises possible phrases that could be found in an email, indicating certain types of human behaviour displayed by an employee who is committing IP theft.

Table 5.2: Summary of insider IP theft

Component	Explanation
Description	A type of employee who has contributed to the development of a specific system or confidential information within the organisation, and who begins to feel a sense of entitlement to the particular work (Cappelli et al., 2012). As such, the employee feels a need to obtain further benefit from the organisation’s intellectual property, regardless of the fact that the employee has signed non-disclosure agreements with the employer.
Observable (human) behaviour	An employee who commits insider IP theft may exhibit one or more of the following malicious behaviours: disregard for the rules and authority, egotistical behaviour, detachment. These behaviours can be detected in email communications by inspecting an employee’s sent and received emails. One may find for example emails containing large amounts of data sent to external parties or to the employee’s private email, as well as offers received from external parties with regard to exchanging confidential information for monetary reward (Cappelli et al., 2012). In addition, emails containing offers of employment could also be examined for words suggesting financial benefits and conditions regarding projects the employee has worked on or developed.
Phrases	This employee might have email communications containing words such as “split the difference”, or “where did my money go” (Whitman, 2016).

5.2.3. Insider fraud

Insider fraud is committed by an employee who steals the financial information of an organisation by committing identity theft, or to obtain a financial benefit from an external party (Cappelli et al., 2012). This insider has access to confidential financial information such as credit card or personally identifiable information of customers and fellow employees. According to Nurse et al. (2014) this type of attack is most common and involves either direct theft or the selling of organisational data or services to attackers. Insider fraud involves insiders adding, modifying, deleting or sharing specific information located in the organisation with a view to committing fraudulent actions. Such actions are usually severely costly to organisations and can affect customers and fellow employees as well. To ensure that this insider threat is performed by misusing information technology (IT), only fraud cases facilitated via IT are considered in the work at hand.

A specific type of behaviour that can be exhibited by an employee who is likely to be fraudulent is (once again) disgruntlement. This employee probably feels wronged by the organisation or is unhappy about criticism, feedback, or the general work environment. The individual may pick conflicts with co-workers, or rant to fellow colleagues about the employer and organisation. This behaviour could also be triggered by the employee's private financial difficulties, which may be exacerbated by increasing debt, outstanding bills and possible addiction problems (Cappelli et al., 2012). The employee might thus be unable to separate private and professional life, and experience severe levels of stress. Albeit a risk, this person may feel justified to commit fraud as a result of these feelings of disgruntlement and stress (Shaw & Fischer, 2005).

Factors that assist the employee to conduct the attack with ease increase their motivation. These factors include weak security perimeters and controls, negligent managers, and the fact that the employee possesses authorised access (Cappelli et al., 2012). Moreover, the employee sometimes rationalises their actions by aiming to pay back the money to the organisation, once they are financially stable again (Cappelli et al., 2012). The employee therefore assumes their fraudulent act will go undetected.

5.2.3.1. *Real-world example of insider fraud*

An example from the Enron case study shows that senior executives were manipulating the stock market in Canada (Tribolet, 2016). Furthermore, there is evidence of the alteration of balance sheet data by senior executives, as well as of executives secretly selling their shares despite ensuring the public that the stock price would rise again (Segal, 2019). Accounts of bribery and many other instances of fraud were also evident in the email corpus.

5.2.3.2. *Insider fraud suggested in employee email communications*

The Enron example above indicates that the email communications of employees and even senior executives can reveal fraudulent activities. As such, specific phrases could be present in the emails that would allow the fraud to be detected.

5.2.3.3. *Summary table*

Table 5.3 includes examples of phrases occurring on the email platform that can be mapped to an employee’s behaviour to detect insider fraud.

Table 5.3: Summary of insider fraud

Component	Explanation
Description	An employee who commits insider fraud feels justified to steal confidential data of employees or customers of an organisation and use it to commit fraudulent actions (Cappelli et al., 2012; Young et al., 2014). The employee is usually driven by a deep financial need and in some cases also a feeling of disgruntlement towards the organisation.
Observable (human) behaviour	An employee who carries out an insider fraud attack exhibits one or more of the following malicious behaviours: disgruntlement; anger management problems; detachment; disregard for rules and authority; argumentative or confrontational behaviour; inability to separate private and professional life; egotistical behaviour. These behaviours can be detected in email communications by inspecting an employee’s sent and received emails. One may find for example emails containing financial terminology, emails containing sentiments regarding the employee’s own financial troubles (for example, an outstanding bill) or emails where large amounts of data are sent to an external email address. Evidence of collusion between co-workers can also be detected (Spooner et al., 2018).

Phrases	Various phrases and words can be used in a typical fraudulent mail, such as “money”, “share”, “percent”, as well as reference to “advocates” and “relations” (Nizamani et al., 2014).
---------	---

5.2.4. Negligence

This type of insider threat involves an insider who has authorised access to the information systems of an organisation, but whose careless act or omission – not aimed at causing malice – increases the possibility of causing harm to the confidentiality, integrity and availability of information and systems within the organisation (Zaytsev et al., 2017). This employee is unaware of possible risks involved due to negligence in the workplace. According to Young et al. (2014) the severity of damage that can be caused by a negligent insider is potentially on par with that caused by an insider IT sabotage attack. Furthermore, these inadvertent attacks are far more common than those caused by malicious insiders (Nurse et al., 2014). This is because it is not uncommon for a non-malicious employee to send a confidential mail to the wrong email address by mistake. Furthermore, this employee may also send a reply containing private or company-related information in response to a seemingly legitimate email.

Zaystev et al. (2017) suggest how an unintentional attack can transpire if the employee accidentally sends sensitive information to a malicious outsider by replying to a well-disguised phishing email or visiting a rogue website. In addition, the employee might download a harmful email attachment that contains destructive software. Such an employee might not take care in reading the organisation’s security policies and will therefore be unaware of security measures that should be implemented when working with sensitive data and credentials. This employee might assume that the cybersecurity policies are irrelevant to them because they do not occupy a specific technical position.

Wall (2013) also suggests that this type of employee might neglect implementing security measures so as to be more efficient, maximise their work time and simply to have less admin work. Furthermore, perhaps the organisation delivers the cybersecurity policies and training in a presentation-like or unhelpful format that does not relate to the employee for application in their everyday work life (Wall, 2013). For example, the training could focus on the problems

at hand, without suggesting practical steps that an employee could follow and incorporate in their daily work.

In the latter scenario, the organisation should expect that employees will be likely to cause unintentional harm, as they have not received adequate instructions. As such, any harm caused is also as a result of the organisation being negligent.

In today's working world, where it is common practice to work remotely, authorised employees might not have access to all the systems and data outside the workplace. Therefore, if they do not receive adequate security training and simply disregard the existing policies and rules, they might send confidential company information to their private email accounts, so as to have access to the data while working on it from a different location (Wall, 2013). Further reasons for storing confidential data could be for off-site meetings, or for safekeeping of the data for other uses in the organisation. Email communications should therefore be examined to obtain evidence of a negligent employee within an organisation. Nurse et al. (2014) indicate that misuse and negligent actions performed on the email platform are a main source of this type of insider threat.

Various types of behaviours could lead to an employee becoming a potential unintentional threat to the organisation. Firstly, he or she might have difficulty separating their private and professional life and as such can be overwhelmed and distracted at work, resulting in an inability to make rational, sound decisions (Azaria et al., 2014). Another behaviour that such a negligent employee could exhibit is unreliability. This person would deliver sub-standard work, often exceed deadlines and be forgetful and inaccurate with details. All of these are unlike the normal behaviour of the employee. Unreliability combined with stress could lead the employee to performing careless actions on the company's information systems and not being cognisant of specific details shown on targeted phishing emails. Several other types of behaviour that this type of employee could exhibit are shown in the summary in Table 5.4.

A negligent insider would not have any specific motivation to conduct an attack, and these attacks are virtually always unintentional.

5.2.4.1. Real-world examples of negligence

The Enron case study shows that a harmful email was sent to employees containing malware as an attachment under the guise of 'JokeStressRelief'. An employee who is unaware of security policies and concerns might well open such an attachment without a shred of doubt. Employees within the Enron corpus were also shown to have sent confidential credit card details to unknown email receivers, perhaps in response to a phishing scam – again due to negligence.

5.2.4.2. Negligence within an employee's email communications

Zaki et al. (2017) indicate that phishing emails can shrewdly be created by malicious attackers who have observed the organisation and conducted adequate research on the sector in which the organisation exists, as well as the company's image and reputation. From this information, an attacker can construct a believable and relatable email that would not easily be disregarded as spam by a negligent employee.

The insider threats studied in this work are those that can potentially be located within an employee's email communications.

5.2.4.3. Summary table

Table 5.4 presents various phrases that could be found within this type of insider's email communications, as well as potential behaviours that are associated with a negligent insider.

Table 5.4: A summary of negligence

Component	Explanation
Description	A negligent employee does not act with malicious intent, but disregards company security policies and rules. As such, the negligent actions of this employee can place the organisation at risk of an attacker accessing sensitive information (Young et al., 2014; Nurse et al., 2014).
Observable (human) behaviour	An employee who carries out a negligence attack potentially exhibits one or more of the following behaviours: disregard for the company's rules and authority; severe stress levels; inability to separate private and professional life; unreliability; absenteeism. These behaviours can be detected in email communications by inspecting an employee's sent and received emails. One may find for example phishing emails

	containing URL links and carefully worded business- and organisation-related emails from external senders (Zaki et al., 2017). Emails in which the employee sends credentials, data and other company-related information could also serve as evidence of this threat. Furthermore, a stressed employee who seems forgetful and inaccurate and sees no relevance in reviewing organisational policies, specifically those related to information security, displays behaviour that is associated with negligence (Young et al., 2014).
Phrases	Example phrases that would be located in an email that targets a negligent employee would include emergency words and phrases such as “urgent” or “as soon as possible” (Nizamani et al., 2014). A negligent employee might be found to send an email containing phrases and words such as “credentials”, “please find attached”, “difficulty to concentrate”, “apologies”.

5.3 Existing research on approaches to detect insider threats within emails

In order to develop the prototype for this dissertation, past research was considered, to discover various techniques that would be relevant and useful. Section 5.3.1 begins with a discussion of a commonly used technique called ‘clustering’, which is useful when pre-classified datasets are not available. Next, a combined approach of clustering and classification is introduced in 5.3.2. The focus in 5.3.3 is specifically on research regarding human behaviour detection. The discussion is continued in 5.3.4 with the key approaches identified and described in past research, and highlighting those that will be utilised in the design of the prototype.

5.3.1. Clustering approach

Using automated processes to detect insider threats in large corporate email communications is a complicated task, because the results can include a significant amount of false positives. However, it is completely unfeasible to label a large email dataset manually. One approach that aims to eliminate the manual labelling of emails and increase accuracy, is known as clustering (Alsmadi & Alhami, 2015). Okolica et al. (2007) employed a method of clustering to identify possible malicious emails. They began the process by tokenising every email into collections of words. Each root word was removed and when an occurrence of an inflection of this root word was identified, this was combined with the initial root word (Okolica et al.,

2007). This approach is referred to as stemming. For example, the word 'malice' is an inflection of the root word 'mal'.

Stemming is used in clustering for two reasons: firstly, the dataset size can be reduced, and secondly, the accuracy of clustering is enhanced (Okolica et al., 2007). A counter was created to contain the number of times a given email contained a certain word (Okolica et al., 2007). A tool called Author Topic (Rosen-Zvi et al., 2004) used this counter, such that 48 categories could be generated to serve as 'centroids' in a clustering algorithm. These centroids would be used to group similar emails together.

In the work done by Okolica et al. (2007), centroids were generated based on the frequency of words in the emails that were common to the rest of the dataset. For the purposes of this dissertation, however, where clustering should be used, it would be more relevant to construct specific centroids that map to each of the types of insider threat, such that specific keywords can be included in them. The problem with using an approach such as that of Okolica et al. (2007) to create centroids, is that the words in the dataset will be mostly business-related terminology. The malicious activity will not be as common and the wording that would raise suspicion might differ between emails.

Furthermore, since such a wide variety of possible phrases and emails could be sent, Okolica et al. (2007) demonstrated that clustering served well as a method for labelling a large dataset. Clustering is therefore an approach that is relevant to the research at hand.

5.3.2. Clustering and classification – a combined approach

An approach that combines both clustering and classification makes use of both supervised and unsupervised machine learning algorithms. A supervised machine learning algorithm refers to an algorithm that makes use of an input dataset and an output dataset, and it learns how to map the input dataset to the results of the output dataset (Furnell, 2004). If a new input dataset were to be provided to the algorithm, it would create an output dataset based on the mapping it had learnt (Wang et al., 2018). An unsupervised machine learning algorithm, in contrast, is provided with an input dataset and allowed to derive patterns without any intervention (Furnell, 2004). This means that an unlabelled dataset would be provided to the algorithm and the algorithm would be expected to learn behaviours and derive abnormal patterns from the large dataset (Wang et al., 2018).

Alsmadi and Alhami (2015) created a potential spam-detecting model that uses clustering and classification of a dataset made up of an individual's private emails. To achieve this, both supervised and unsupervised machine learning algorithms were included. The process started by utilising the 'bag of words' or vector space model to arrange emails according to the most frequently occurring words. The authors subsequently included processes such as data cleansing and stemming to ensure efficient, accurate clustering (Alsmadi & Alhami, 2015). Labelling was performed on the dataset through the use of the K-means algorithm (Alsmadi & Alhami, 2015). This algorithm randomly selected emails within the dataset to be used as centroids and to allow for the clustering of like emails to take place (Hussain & Qamar, 2014). Similar to the approach taken by Okolica et al. (2007), Alsmadi and Alhami (2015) also randomly selected a centroid based on the frequency of words within the email dataset.

After a labelled dataset was produced by the clustering algorithm, classification algorithms were executed in order to achieve better accuracy. These classification algorithms were run a specific number of times to improve accuracy. Alsmadi and Alhami (2015) also included metrics such as precision and accuracy indicators to measure the performance of each iteration.

It is important to note that the supervised learning approach can only produce results based on the dataset that was used for training. If this dataset contained incorrectly labelled data, then the dataset labelled by the supervised learning algorithm will simply label based on what it has learnt.

An approach shown by Mayhew et al. (2015) involves detecting malicious activity within various technical resources such as network activity, HTTP requests and email data. As with the approach taken by Alsmadi and Alhami (2015), Mayhew et al. (2015) also made use of a combination of supervised and unsupervised learning, specifically, unsupervised K-means clustering with supervised Support Vector Machine (SVM) algorithms. The authors believed that this was the best combination to ensure better quality, efficiency, as well as performance. After having inspected the available research, Mujtaba et al. (2017) actually noted that Support Vector Machine (SVM) and Naïve Bayes algorithms are among the most commonly used supervised machine learning algorithms in the big data research domain.

These algorithms were specifically used in research regarding email communications (Mujtaba et al., 2017).

Duc and Zincir-Heywood (2019) made use of another popular supervised learning algorithm, Logistic Regression, for experimentation to detect unknown insider threat cases. This approach was also mentioned by Zaytsev et al. (2017) as having been examined by Japanese scientists to detect indicators of insider threats. HaCohen-Kerner et al. (2020) chose Logistic Regression for their work as they found it to be one of the five most popular machine learning algorithms, and also included SVM and Naïve Bayes classifiers. For the purposes of this dissertation, a search was conducted to determine the most suitable algorithms used in similar research papers. The top 10 papers are discussed and shown in Table B.1 of Appendix B.

Furthermore, it was found that supervised machine learning techniques were more commonly used in this research domain, as opposed to unsupervised learning. The supervised machine learning algorithms SVM, Naïve Bayes and Logistic Regression, as well as the unsupervised machine learning K-means algorithm are also noted for use in this research. These machine learning algorithms are described in greater detail in the design chapter of this work. An evaluation of the algorithms is also shown in Table B.2 in Appendix B.

5.3.3. Human-behaviour-based approaches to insider threat discovery

An approach that considered human behaviours when classifying an email dataset was adopted by Brown et al. (2013), using the Enron email corpus for the experimentation. A number of dictionaries representing psychological traits such as “anger” and “anxiety”, containing words related to these traits were created. The words in each email body were then compared with the words in each dictionary (Brown et al., 2013). This approach of using a dictionary with words that could be present within an email, as opposed to the approach by Okolica et al. (2007) to use centroids, is more relevant in the scope of insider threat detection. This is because the dictionaries used in the approach by Brown et al. (2013) can be altered to include specific words and phrases linked to insider threat behaviours that could be present within email communications.

Furthermore, Brown et al. (2013) used a scoring approach, where if a match between the email and a word in one of the dictionaries was found, the score for the specific dictionary

was incremented. This approach of scoring is useful to enhance accuracy and avoid a large number of false positives. However, in the approach by Brown et al. (2013), if a given email obtained a high score for a number of dictionaries, it was not clear what label the email would be allocated, or if the email would be assigned multiple labels.

5.3.4. Key components of prototypes developed in past research

Past research introduced various techniques and approaches that will be incorporated in the design of the prototype for this dissertation. Firstly, it demonstrated the importance of using a combination of supervised and unsupervised machine learning methods to obtain better accuracy and performance. Furthermore, clustering has been shown to effectively divide a dataset into similar groups, based on centroids. Centroids were however created based on frequently occurring words in the email dataset used in past work. Centroids can nevertheless be useful to cluster emails as they can be fabricated to contain relevant words and phrases that would be associated with insider threats. The fabrication of centroids would involve constructing specific email text with sentences containing phrases that could pertain to an insider threat. For example, the phrase “I am unappreciated and my work is not valued”, could signify a potential insider IT sabotage contender. Such a phrase would be used in conjunction with other similar phrases to form the centroid. These centroids can then be used to group similar emails within the dataset to them.

The use of dictionaries, as has been suggested in past research, is still relevant to the research at hand. Besides using a dictionary for classification, scoring is used to control false positives and ensure accuracy. Furthermore, the email classification approach in past research was enhanced through running several iterations of the experiment and measuring this by means of performance metrics for accuracy and precision. These elements of the past approaches will be used to inform the model to be created in this work.

5.4 Conclusion

Chapter 5 focused on discussing the types of insider threats and mapping these to the identifiable human behaviours that were shown to be characteristic of insiders. The insider threat types, mapped to their human behaviours, would be useful for classifying an employee’s emails according to the proposed categories and for providing a more accurate identification – rather than just being classified as an insider threat. This is important because

not all insider threats are conducted with malicious intent and as such it was found necessary to distinguish between the types. The chapter also discussed certain approaches in past research towards insider threat detection conducted via the email platform, and focused on human behaviours that have relevance to this research.

Following on the theoretical discussion contained in this chapter, Chapter 6 will present the requirements for the model that will be developed in this dissertation.

Chapter 6

Requirements for Prototype Development

6.1 Introduction

Chapter 5 delved into the theoretical background of each of the components of the topic for this research. The chapter presented the different types of insider threat by discussing the potential human behaviours that could be exhibited by the insider. In addition, the chapter discussed the email platform as a means for potential attacks to originate from. The notion that specific phrases are associated with the human behaviours linked to specific types of insider threat was discussed. Possible phrases that could be present in an insider's email communications were also shown. Therefore, Chapter 5 introduced the context and components that are present when designing the prototype solution for the experimentation that is conducted in this research.

The purpose of Chapter 6 is to define the functional and technical requirements for the prototype developed in this work, based on the relevant theory.

6.2 Overview

According to the ISO/IEC/IEEE 29148:2018, a requirement is defined as a statement that describes a necessity, as well as its conditions and limitations (ISO, 2018). The requirements presented in this chapter are statements or conditions that the prototype must adhere to so as to address the objectives of and research problem in this study. In order to ensure that requirements are met, verification takes place. This implies that the requirements are confirmed through examining evidence and results, to detect whether the objectives were accomplished (ISO, 2018). The final prototype will be compared with the requirements in this chapter, as well as with the design specification in Chapter 7, to ensure that all these requirements have been fulfilled.

The requirements in Chapter 6 are summarised within diagrams that depict the requirements and the relationships that exist between them (Amyot et al., 2016). The importance of these diagrams lies mainly in the specific requirements that have been chosen.

The chapter is divided into two main sections, namely the functional and technical requirements for the prototype. The functional requirements are statements that describe expected behaviour and functionality (Zubcoff et al., 2019). They specifically show what the prototype should and should not do and are usually written with regard to input data. Technical requirements represent the quality attributes, as well as the limitations involved when developing the prototype (Zubcoff et al., 2019), and they can be used as the basis for verification. Technical requirements should also be considered when developing the functional requirements, to ensure the design is efficient and effective (Zubcoff et al., 2019).

6.3 Functional requirements

The diagram in Figure 6.1 depicts the main functional requirements as well as their relationship to the main requirement or main objective of this work (labelled FR001 in Figure 6.1). The five main functional requirements are shown to have the «contain» relation with this main requirement (FR001). This is simply a means of decomposing this abstract requirement into separate concrete requirements. It is also clear from the diagram that there are sub-requirements that have <<derives>> relationships with the concrete functional requirements. These sub-requirements are derived from the results or properties of the main functional requirements and are important for the functionality of the prototype.

Each requirement in the diagram is assigned an identity number that is used for referencing these requirements. Furthermore, the diagram shows that the verification method used is analysis. This will be done by evaluating and comparing the results of the prototype to ensure they adhere to the functional requirements.

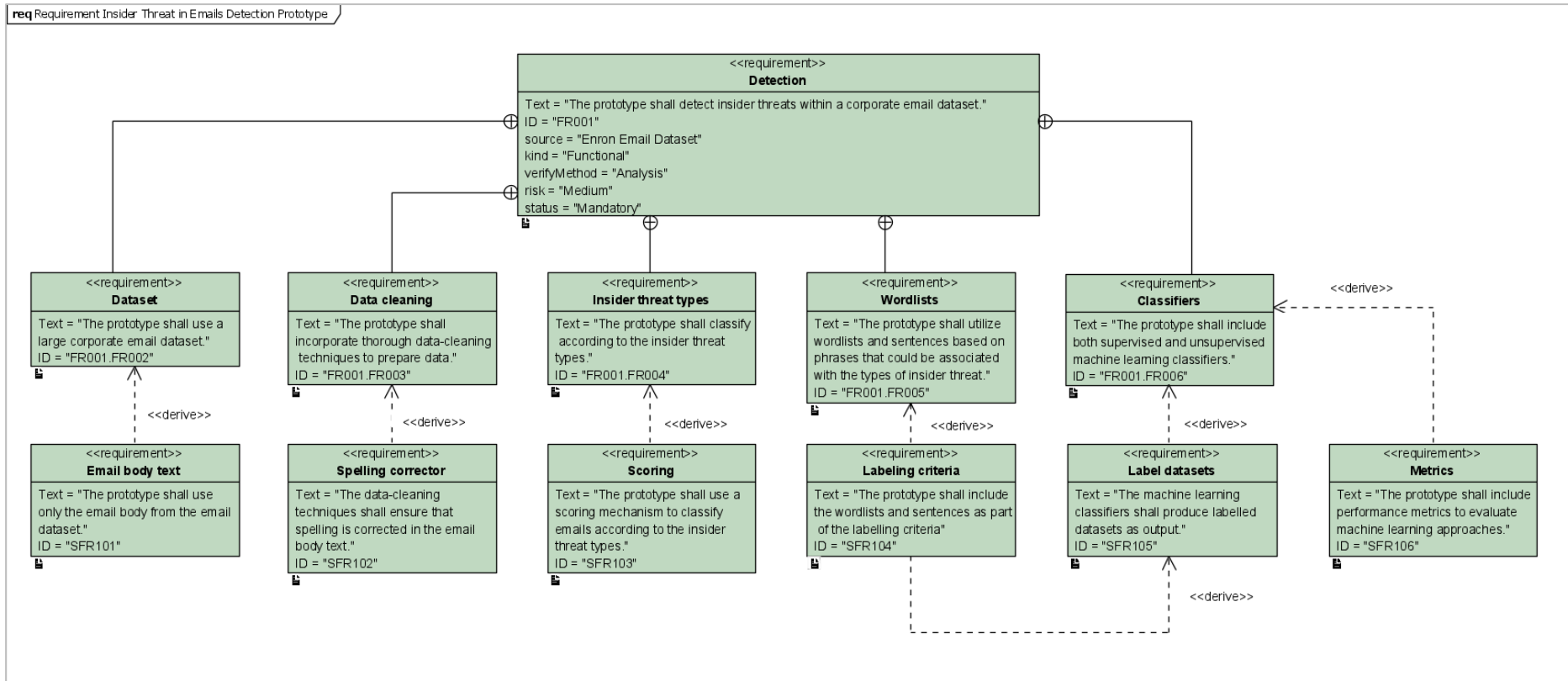


Figure 6.8: Functional requirements diagram

The requirements shown in the diagram in Figure 6.1 are expanded on below.

FR001: Detection

The prototype shall detect insider threats within a corporate email dataset. This is the main objective of this research and as such, all the functional requirements for the prototype stem from and decompose this main objective or requirement.

FR001.FR002: Dataset

The prototype shall use a large corporate email dataset. The dataset obtained must contain real-world data, specifically with known cases of insider threat (Leber, 2018) (Alsmadi & Alhami, 2015). The real-world dataset should therefore contain both malicious and non-malicious email data.

FR001.FR003: Data-cleaning

The prototype shall incorporate thorough data-cleaning techniques to prepare data. To enhance accuracy of the tool, it should involve a cleaning and normalisation process that removes duplicate mails, stems root words and removes redundancies such as contractions and stop words (Nizamani et al., 2014).

FR001.FR004: Insider threat types

The prototype shall classify according to the insider threat types. The prototype should ensure that the following types of insider threat associated with human behaviours are detected in the large corporate email corpus (Spooner et al., 2018; Young et al., 2014):

- Insider information technology sabotage
- Insider fraud
- Insider intellectual property theft
- Negligence

FR001.FR005: Wordlists

The prototype shall utilise wordlists and sentences based on phrases that could be associated with the types of insider threat. The prototype should make use of files containing specific phrases that might be found in an employee's email communications. These phrases should be associated with certain human behaviours that could signify the type of insider threat.

They should be used as data with which to compare the emails in the dataset and to classify the emails.

FR001.FR006: Classifiers

The prototype shall include both supervised and unsupervised machine learning classifiers. This requirement is in line with past research to enhance accuracy.

In addition to the main requirements shown above, the following are sub-requirements that derive from these main functional requirements:

SFR101: Email body text

The prototype shall only use the email body from the email dataset. Only the textual email body message should be used for the experiment, because the focus is on analysis of the text (Mujtaba et al., 2017) (Alsmadi & Alhami, 2015).

SFR102: Spelling corrector

The data-cleaning techniques shall ensure that spelling is corrected in the email body text. The normalisation process should correct spelling by using an appropriate online dictionary. It has been shown that spammers deliberately inject incorrectly spelled words into their email body in an attempt to go undetected (Nizamani et al., 2014).

SFR103: Scoring

The prototype shall use a scoring mechanism to classify emails according to the insider threat types. The mechanism created should score a given email on its prevalence of the different insider threat types (Young et al., 2014; Brown et al., 2013; Cappelli et al., 2012):

- Insider information technology sabotage
- Insider fraud
- Insider intellectual property theft
- Negligence

SFR104: Labelling Criteria

The prototype shall include the wordlists and sentences as part of the labelling criteria. The criteria for the labelling of the dataset should be based on the types of insider threat and should involve a comparison process between the dataset and specific phrases that are mapped to each type of insider threat. Labels should therefore be provided based on phrases

and words in given wordlists or dictionaries, or in fabricated emails that are mapped to each of the insider threat types.

SFR105: Label Datasets

The machine learning classifiers shall produce labelled datasets as output. The output of the classification and clustering processes should be email datasets in which each email has an assigned label.

SFR106: Metrics

The prototype shall include performance metrics to evaluate machine learning approaches. The prototype should include a means to measure accuracy of the supervised and unsupervised machine learning approaches for a certain number of repetitions (Mujtaba et al., 2017).

The functional requirements have been defined and discussed above. The following section will focus on the technical requirements for this work.

6.4 Technical requirements

This section presents the requirements that are not included in the functional requirements section and, as such, do not dictate the functionality of the prototype. The attributes in this section are a list of the most relevant and important qualities that the prototype should adhere to.

The diagram in Figure 6.2 shows the identity number and description of each of the main technical requirements.

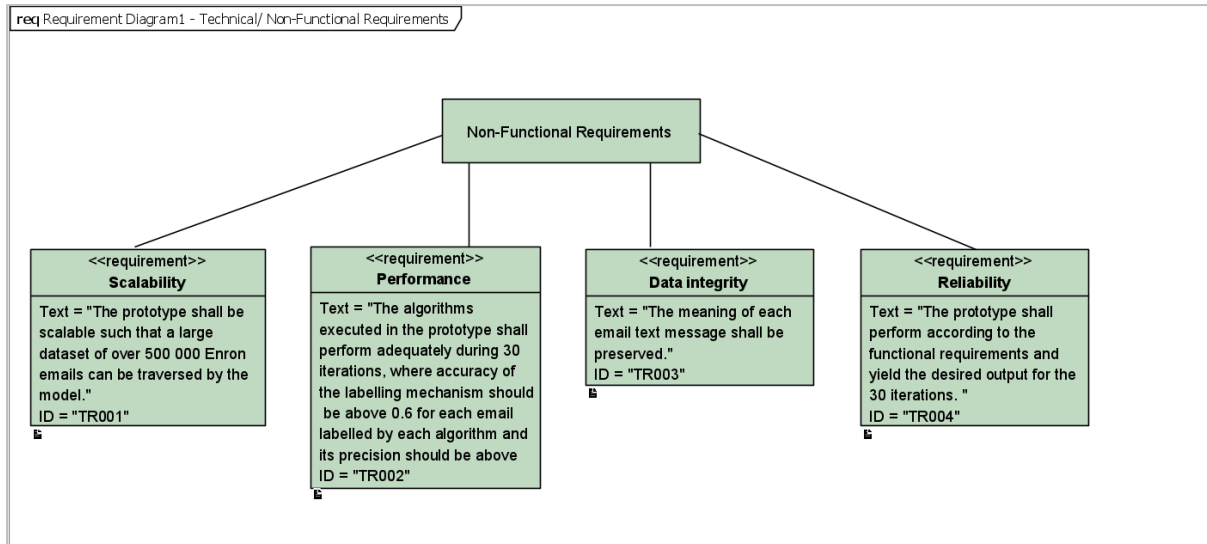


Figure 6.2: Detailed description of technical requirements

The technical requirements are described in detail below, based on the diagrams in Figures 6.1 and 6.2.

TR001: Scalability

The prototype shall be scalable such that a large dataset of over 500 000 Enron emails can be traversed by the model. To check that this quality attribute has been accomplished, the prototype should be able to run smoothly without halting or breaking during execution of the algorithms over the large dataset.

TR002: Performance

The algorithms executed in the prototype shall perform adequately during 30 iterations, where accuracy of the labelling mechanism should be above 0.6 for each email labelled by each algorithm and its precision should be above 0.6 (60%). The accuracy of the results obtained from each of the different machine learning algorithms executed in the prototype will be compared by using precision and accuracy metrics (Mujtaba et al., 2017). Note that the algorithms shall be executed 15 times for each of the labelled datasets.

TR003: Data integrity

The meaning of each email text message shall be preserved. The prototype will include rigorous pre-processing to clean the emails and place them in a format that can be standardised. Despite this pre-processing, the meaning of and phrases within the email must not be changed to different words or be removed entirely. It is essential to preserve the

meaning of the email so that accurate results can be obtained. To ensure that this quality attribute is successfully implemented, the accuracy of the labels assigned to the given emails will be manually inspected.

TR004: Reliability

The prototype shall perform according to the functional requirements and yield the desired output for the 30 iterations. This means that the prototype should include exception handling, so that the execution process will not be halted when an exception is caught. To ensure reliability, additional measures to eliminate bugs and errors should be put in place. This quality attribute will be achieved if all errors and exceptions are handled and do not affect the smooth running of the prototype.

In addition to fulfilling the critical qualities above, adherence to legal restrictions and privacy regulations is essential in the selection of the email dataset. The selected dataset shall as far as possible be subjected to minimal legal restrictions and requirements, so that the data can easily be used without implications.

6.5 Conclusion

Chapter 6 presented a detailed breakdown of the overall functional and technical requirements of the prototype, as well as the requirements for its individual components. The requirements discussed in this chapter were concrete components that were used to design a prototype that addresses the problem stated in this research.

Chapter 7 considers the requirements shown in this chapter to discuss how the prototype is designed and to outline which components of the design are included to fulfil a particular requirement.

Chapter 7

Prototype to Detect Insider Threats in Corporate Emails

7.1 Introduction

Chapter 6 defined the functional and technical requirements to guide the construction of the prototype. These were defined in a manner to ensure that the research problem in this work be addressed. The chapter also included diagrams that present an overview of the different requirements and a corresponding identity number for each requirement to be used as reference.

Chapter 7 defines the process that was followed to construct the prototype used for experimentation and to address the problem statement. Furthermore, the detailed steps of the process – based on the requirements and past research – are explored in this chapter to show their relevance and purpose.

7.2 Process overview

The process followed in this chapter to construct the prototype included three main components, which each served a unique purpose. The three components – data preparation, data discovery and detection – are briefly introduced in Figure 7.1. (Each of them is expanded on in greater detail later in this chapter.)

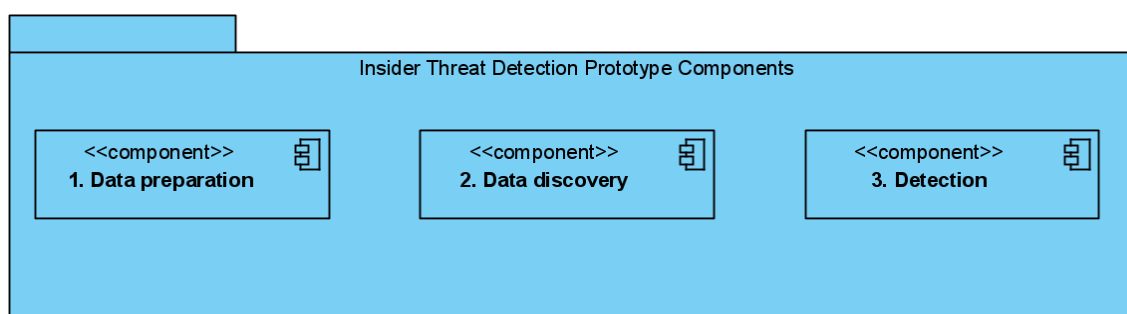


Figure 7.9: Components of the insider threat detection prototype

The first component of the prototype was the data preparation component. This involved the steps to clean and pre-process the Enron email data, so as to standardise the data that was used for the experimentation. This clean dataset was used as input for the second

component. A high-level process for the prototype is shown in Figure 2 below. The first block in this diagram presents an overview of the data preparation.

The data discovery component, the second part of the prototype, adopted two separate approaches to classify the pre-processed email data. Various labelling techniques and algorithms were used to produce two labelled email datasets as output. Performance metrics were also provided in this component. The steps as arranged in the data discovery component are shown in Figure 7.2.

The final component, detection, involved the execution of the prototype for experimentation. The Enron email dataset was used as input to the machine learning model that was created and as such two real labelled datasets and performance metrics were expected as output. The experiment was executed a certain number of times and as such, the performance metrics provided results for comparison and evaluation purposes. The detection component is also shown in Figure 7.2.

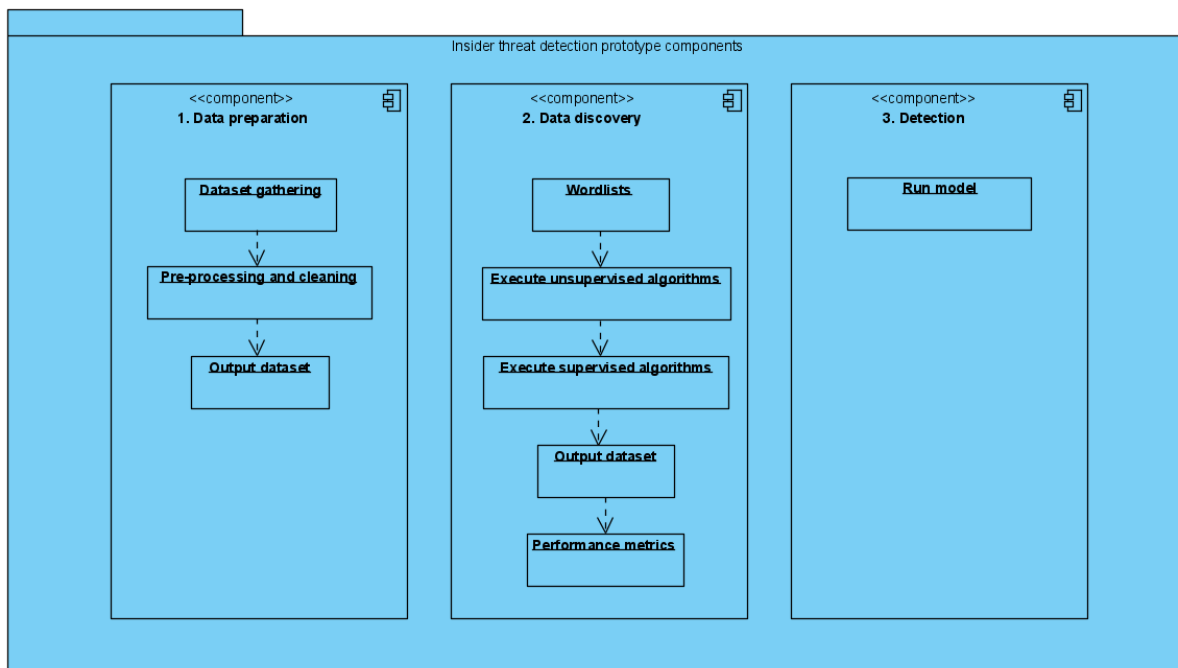


Figure 7.10: High-level diagram to show the prototype development process, based on work of Michael and Eloff (2019)

7.3 Data preparation

The data preparation component comprised two main parts. The first portion of the component involved obtaining a dataset that would suitably adhere to the requirements of the prototype. Therefore, the Enron email corpus (in CSV format) was used in this work. The second part of data preparation was the data cleansing and pre-processing phase, which focused on the email body content only. The data was pre-processed by using the various techniques discussed below. The output of this component was a dataset containing the pre-processed email body texts. The process that took place in this component is shown in Figure 7.3 and afterwards discussed in detail. The requirements that guided the inclusion of various steps in the process are shown in notes in the diagram.

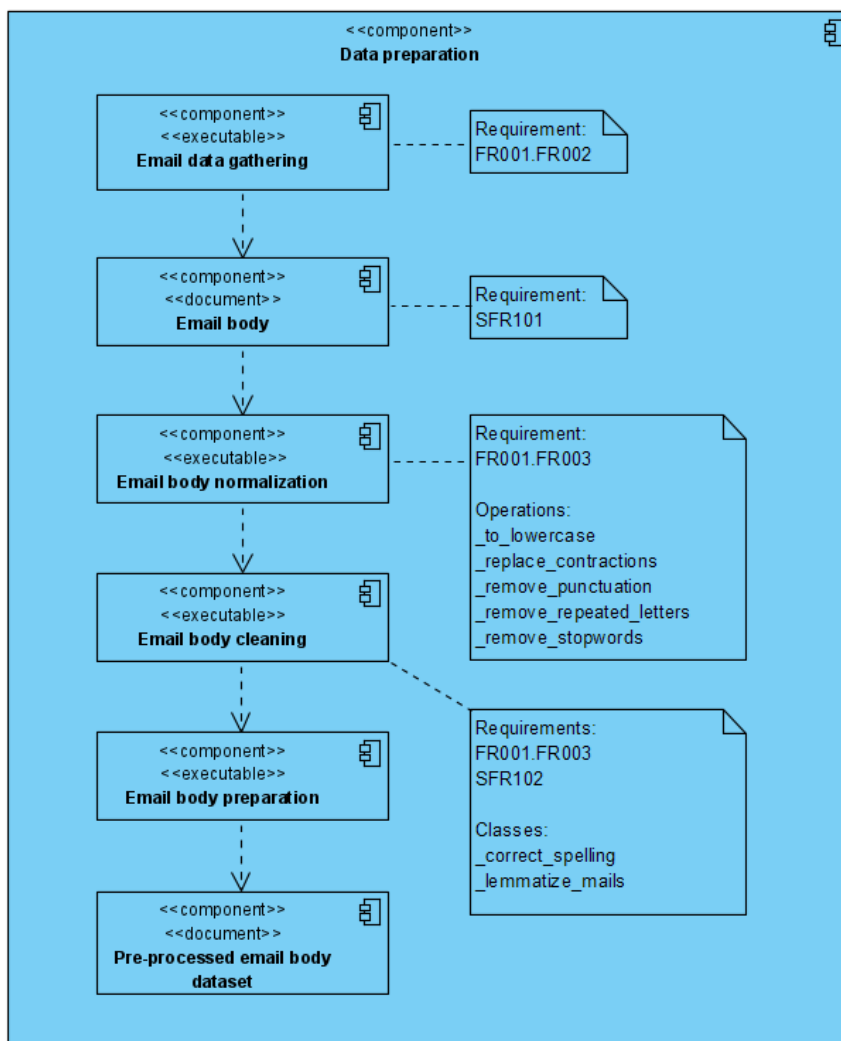


Figure 7.3: Details of the data preparation component

7.3.1. Executable: Email data gathering

In order to select a suitable dataset for this research, certain requirements had to be adhered to. Requirement FR001.FR002 stipulates that, firstly, the dataset shall contain existing real-world emails and should have a mix of private and business-related content. Furthermore, a large dataset is required to test the efficiency of the algorithms. The size of the real-world email dataset that exists within organisations is typically very large and can reveal various patterns, based on human behaviours and shown via computational means (Zaki et al., 2017).

A large publicly available dataset, the Enron email corpus obtained in CSV format from Kaggle.com (Cukierski, 2015), fulfilled all the aforementioned requirements and was selected for this research.

7.3.2. Document: Email body

Once the email corpus had been selected, the email body was extracted from the dataset for use in the experimentation, according to requirement SFR101. The Enron dataset was specifically important because (as shown in the case study) the email body content is known to contain real cases of insider threat.

7.3.3. Executable: Email body normalisation

In terms of the data-cleaning requirement, FR001.FR003, a standardised email dataset had to be presented to the algorithms to ensure that the most optimal and accurate results would be obtained. The normalisation executable was included to eliminate duplicate emails and other inconsistencies. After the first pre-processing step, a new dataset was created with changes made to the dataset. This dataset was then updated after each pre-processing step to reflect the latest changes made to the dataset.

7.3.3.1. Tokenisation

It is important to note that machine learning classifiers are unable to process raw text data, and as such, the text was represented as a vector of numbers. This technique, known as feature extraction, is shown in the work by Alkhereyf and Rambow (2017), where their model ensures that each email is represented as a vector that consists of word frequencies based on the rest of the dataset.

According to Mujtaba et al. (2017) and Okolica et al. (2007), tokenisation is used to break the email into a string of words, by making use of the 'bag of words' model. For the purpose of

natural language processing, each email in the current work is broken up into a ‘bag of words’ or tokenised words, so that the individual email can be analysed. Unlike the typical ‘bag of words’ approach which would tokenise the entire dataset, in this dissertation, each email was tokenised and analysed individually, and not compared with the rest of the dataset. The Natural Language Toolkit, NLTK (2020), available in Python, has a function that allows word tokenisation. The NLTK (2020) is a leading platform for constructing python programs that interact with human language data.

7.3.3.2. Conversion to lowercase: `_to_lowercase`

The conversion of text to lowercase is a very simple task; however, it is used for standardisation and for faster processing. In addition, it aims to remove unique words. This operation is performed on the tokenised email data, which is presented as input to this function. HaCohen-Kerner et al. (2020) conducted an experiment and found that the combination of conversion of text to lowercase and spelling correction yielded the best results for the dataset processed by the authors.

In Python, a simple built-in function was used to convert the case of the text to lowercase.

7.3.3.3. Contractions removal: `_remove_contractions`

The contractions removal process in this prototype is as follows. Once each tokenised email has been presented as input to the relevant function, each word within the email is compared to regular expression patterns, which in this case are contractions. If a match is located, the contraction in the email is replaced with the appropriate expanded word.

The reason for including this step was to ensure as far as possible that a given word would not be excluded because it was in contracted format. Words that were in this format were converted to their original full individual words, again with a view to eliminating unique words.

7.3.3.4. Punctuation removal: `_remove_punctuation`

Noever (2020) refers to punctuation removal as another technique that is used when normalising the data for use with machine learning algorithms. This was done mainly to reduce the size of the training data, as punctuation does not add any value to the experiment at hand. The punctuation removal step was conducted only after the removal of contractions

step had taken place. This was to ensure that the apostrophe used in the contraction was not removed before all contractions could be detected and expanded.

In the source code of the prototype, shown in Appendix C, regular expression, *re*, in Python were used to allow pattern-matching checks to be performed on the tokenised emails. These checks were done to determine whether a character in the tokenised email matched a punctuation mark that was defined in the Python String class. However, considering the context of this work, full stops were not removed from the tokenised emails. Full stops were essential in the classification process to mark the end of each individual sentence and to allow for sentences to be counted within an email. Furthermore, there could be full stops in URLs as well as email addresses, and these would need to be differentiated from other text in the email body.

7.3.3.5. Repeated letter replacement: `remove_repeated_letters`

This step involved checking whether there were repeated letters within a word, so as to improve the data sent to the stopwords removal process in the following step. This was done by using an NLTK function in Python within the WordNet dictionary called `synsets`. `Synsets` was used to test whether a given word actually exists within the English dictionary by checking for synonyms of the word. In the prototype design, if a given word was found to be valid, the next word in the tokenised email was presented to the `remove_repeated_letters` function. If a synonym was not found for the word, the word was assumed invalid due to the repeated characters. The replacement for the misspelt word was then found by means of suggestions from the NLTK WordNet dictionary, by using `symspellpy`.

7.3.3.6. Stopwords removal: `_remove_stopwords`

Noever (2020) refers to stopwords as common English terms; however, these words are mostly articles used within sentences that do not provide extra meaning to the sentence. Examples are 'a', 'the', 'in', and 'an'. These words were removed to free up some space within the dataset. HaCohen-Kerner et al. (2020) performed experimentation and found that the removal of stopwords was the only individual step within preprocessing that largely improved the data in the datasets used.

The NLTK in Python contains a list of stopwords for the English language, and this was used by the prototype to perform matching checks with the tokenised emails. If a word in a

tokenised mail was found to be a stopword, it was simply removed from the dataset. The changes made to this dataset were then saved.

It should be noted that the word 'not' as a stopword was the only exception in the work at hand and as such it was not removed from the tokenised emails. This is because 'not' in conjunction with other words conveyed important meaning in an email. The normalisation did not intend to be rigorous to the point where actual meaning would be removed from the email (as per requirement TR004). For example, the phrase in an email stating "I am not happy", has significant relevance to this work and as such the meaning of the sentence would be distorted and incorrect if the article 'not' was removed.

7.3.4. Executable: Email body cleaning

According to requirement SFR102, the data-cleaning techniques should include a spelling corrector. This falls under the requirement for data-cleaning, FR001.FR003, and as such, lemmatisation was included to conclude the data-cleaning process.

7.3.4.1. *Spelling correction: spelling_corrector*

During this step, each word within a tokenised email was assigned as input for the NLTK WordNet Synsets function. As described earlier, this function checks whether a word is valid or not. If the word was found to be invalid due to it not having a synonym, the Python symspellpy provided suggestions or possibilities for the correct spelling of the word. These suggestions originated from the WordNet dictionary. If the Python symspellpy was unable to retrieve a corrected spelling suggestion, the original word was returned.

In addition to the above individual word correction, another process segmentation function was executed to detach words that had been conjoined, even if the individual words were correctly spelled. For example, the conjoined words 'documentsubmitted' would be converted to 'document submitted'. The function works by taking in each word within the tokenised email as input. The symspellpy word_segmentation method was run on the individual word. If the individual word consisted of two conjoined words, the result would contain two separated words.

The length of the output was subsequently checked against the length of the input to determine whether the output length was a larger value. The output was then returned. If

the length of the input and output was the same, then no segmentation took place or was required. Specifically, if two words were not conjoined, the original word was returned. The token that was returned was then appended to the sentence to replace the original word.

Once the spelling had been corrected, the tokenised emails could be lemmatised.

7.3.4.2. Lemmatise words: lemmatiser

This was quite a crucial step within the normalisation process, because it aimed to reduce the number of words in the dataset (Mujtaba et al., 2017). Lemmatisation refers to reducing the number of inflectional forms by reducing words to their root or base form. Inflected words are words that are derived from other words within the English dictionary and these other words are the root words. This change of word form is usually necessary to grammatically change tenses, number or person. For example, the word 'going' and the word 'went' are inflections of the root word 'go'. Within Python, WordNet Lemmatizer from the NLTK was used to perform lemmatisation.

7.3.5. Executable: Email data preparation

Once the email dataset had been rigorously cleaned and normalised by the various individual processes stipulated above, the data was saved.

7.3.6. Document: Pre-processed email body dataset

This step signified that the normalised email data was the output of the data preparation component. The new dataset thus contained the pre-processed email bodies that would be provided as input for the second component of the prototype, namely the classification and clustering processes.

7.4 Data discovery

Once the dataset had been pre-processed, the next step of the process was data discovery (as shown in the second component of Figure 7.2). During this process, the researcher attempted to address the problem statement of this research. As such, the main purpose of the data discovery step was to produce a labelled dataset based on the different types of insider threat (Cappelli et al., 2012; Spooner et al., 2018; Young et al., 2014). Two different labelling tasks were included so that two separately labelled CSV datasets were produced. These are described in the sub-sections that follow.

The first process labelled by means of a Regular Expression Pattern Matching algorithm, whereas the second process labelled with an unsupervised K-means clustering algorithm. Both of the labelled datasets were afterwards required as training data for the supervised machine learning algorithms. The results for each of the datasets were compared through the use of performance metrics. These metrics served as the output of the data discovery component, in addition to the labelled datasets. Figure 7.4 includes the detailed description of the processes that made up this component and also refers to the relevant requirements.

It should be noted at this point that there was no existing email dataset, labelled with the four types of insider threat, that could be used as training data. Furthermore, it was not feasible to manually label or annotate 500 000 emails, although manual labelling would be more accurate. Alternative automated approaches using machine learning were therefore devised so that patterns could be retrieved from the dataset.

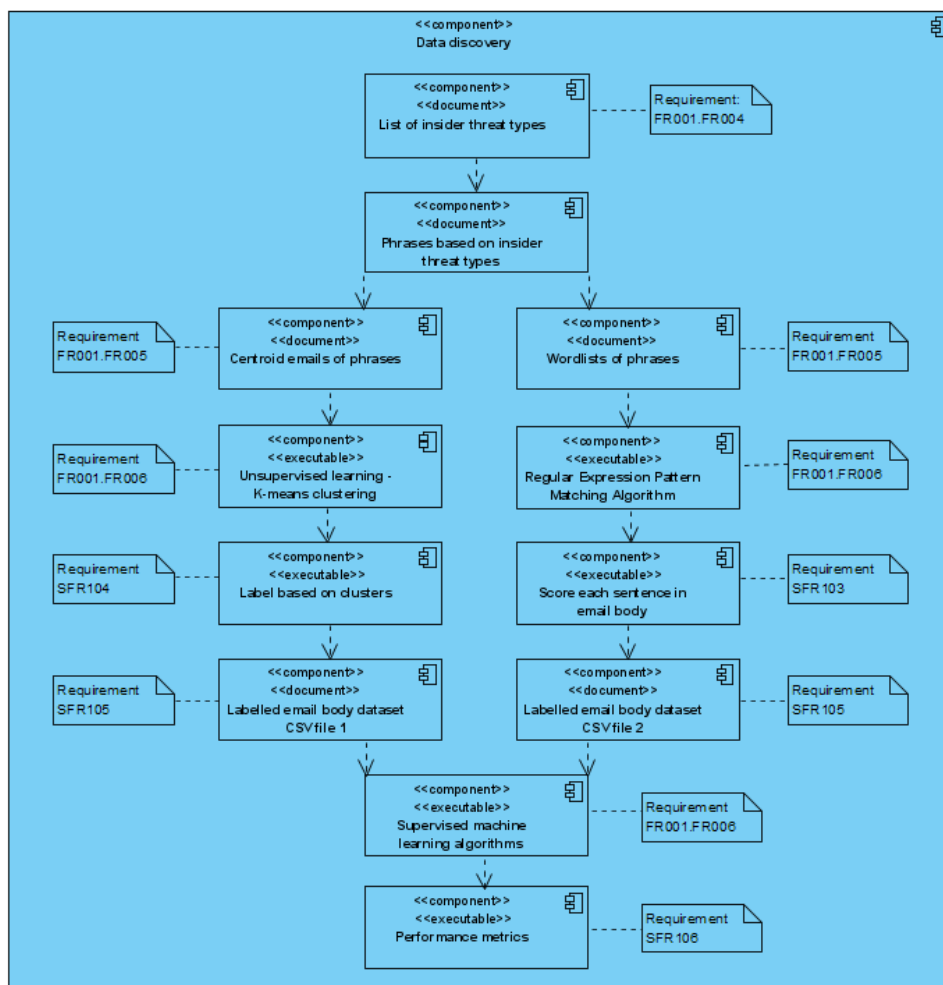


Figure 7.4: Details of the data discovery component

7.4.1. Document: List of insider threat types

The first step in this component involved manually performing a literature review of research within a similar domain. The various types of insider threat were identified based on the work done by Cappelli et al. (2012), Spooner et al. (2018) and Young et al. (2014). Furthermore, these were associated with actual examples from the Enron case study. The types were identified as insider IT sabotage, insider intellectual property theft and insider fraud (Spooner et al., 2018; Claycomb et al., 2013; Cappelli et al., 2012; Munshi et al., 2012). The fourth type of insider threat was identified as negligence and is discussed by Young et al. (2014). The listing of insider threat types is a step that was dictated by requirement FR001.FR004.

7.4.2. Document: Phrases based on insider threat types

The significance of the insider threat types in this research is their association with various human behaviours that are demonstrated in the workplace. The organisational behaviours were identified and associated with the types of insider threat. Lists were then compiled of phrases that could potentially be located within an employee's corporate email communications, and that were associated with a type of insider threat. These lists were created based on past research, examples provided from other companies who utilise similar surveillance systems, as well as through examples found via manual inspection of the Enron dataset. News articles and the email snippets regarding insider threat, found in the Enron email dataset, were also considered.

From the above it is clear that this step in the data discovery process was also a manual step.

7.4.3. Document: Centroid emails of phrases

This step was again a manual step and required the use of phrases that were obtained during the previous step, according to requirement FR001.FR005. These phrases were used to construct centroids. In order to describe a centroid, an overview is given of the clustering algorithm. The clustering algorithm, the K-means clustering algorithm, works by creating clusters of similar data within a dataset. This is done by identifying a certain number of points within the dataset that serves as the centroids. The rest of the data in the dataset is then allocated to these clusters, by determining the minimal sum of squares.

For the current prototype, however, the centroids that were used for clustering, were fabricated based on the phrases identified in the previous step. In addition, synonyms and

various different sentences with similar meanings were included in the centroids. Since each centroid represented a type of insider threat, there were four main centroids. One additional centroid that was fabricated included generic business email jargon and was labelled as 'non-malicious'. This was used to group similar harmless emails together for which no malicious phrases were detected.

The words within the centroids were presented in lowercase letters. All contractions had been removed and replaced with the two words that originally made up the contraction. To ensure standardisation, the centroids were also tokenised. Furthermore, before the centroids were used in the experimentation, they were also run through the lemmatisation process to ensure that only the root words were contained in the centroids. This was done to ensure that if a lemmatised word such as 'frustrate' was present in the pre-processed email, the inflected word 'frustrated' would not be contained in the insider IT sabotage centroid, but rather also the root word 'frustrate'.

7.4.3.1 Executable: Unsupervised learning – K-means clustering

The K-means clustering algorithm was used to create clusters of the email dataset, as per requirement FR001.FR006. This is an unsupervised learning approach and, as previously described, it works to find patterns within the dataset without leveraging off a training dataset. The algorithm worked by transforming each normalised email body into a vectorised format and using these as input to the K-means model.

The number of clusters was also specified within the Python K-means function. The dataset containing the fabricated centroids was sent as input to the K-means model so that it could be used as a type of training dataset. The clusters were provided to the K-means model. In each iteration, the predict function worked by predicting which centroid was most similar to the given email, and then assigning the email to the most suitable cluster. Alsmadi and Alhami (2015) argue that the main task handled by the K-means algorithm was to calculate the cosine similarity between the emails contained in the dataset and the centroids.

7.4.3.1.1 Executable: Label based on clusters

This step was based on requirement SFR104. A lookup structure was used so that if, for example, the key '3' was associated with the insider fraud cluster, when labelling occurs, the

label that corresponded with that key would be provided – in this case insider fraud. This lookup structure is referred to as the cluster map.

Once the algorithm was finished with its execution, the emails within the cluster were assigned the label that was assigned to the centroid that they had been clustered with. As such, these labels were saved with their corresponding emails in a newly created CSV file.

It should be noted that, once the labelling process was completed, a similarity percentage was computed between each labelled email and the centroid that had been used to provide the label. For example, if an email was labelled as 'negligence', a similarity percentage was calculated between the email and the negligence centroid. This was done, firstly, by creating an attribute vector from the centroids and the tokenised emails. Secondly, checks were performed to identify the emails that had the same attributes as the centroids. A similarity indicator was then provided to the email.

[7.4.3.1.2. Document: Labelled email body dataset CSV File 1](#)

As discussed in the previous step, the output of the K-means clustering algorithm was the labelled CSV dataset, which contains each email and its assigned label, in accordance with requirement SFR105. This file was examined and tested with the supervised machine learning steps discussed later in this chapter.

[7.4.4. Document: Wordlists of phrases](#)

Words and phrases were obtained for the types of insider threat, based on consultation of various sources. These phrases were placed within appropriately named wordlists, according to requirement FR001.FR005. Each wordlist was assigned the name of the type of insider threat that the phrases in the wordlist were referring to. Once the wordlists had been created, these were saved in an appropriately named folder that could be accessed by the Regular Expression Pattern Matching algorithm. Once again, this step included only a manual process.

The idea of constructing such wordlists was obtained from various past research, such as the work done by Alkhereyf and Rambow (2017). These authors provided various criteria, in a number of categories, that would be used to annotate or label a given email. As such, the purpose of the wordlists created in this work at hand was also to be used with machine learning classifiers to label the selected corporate dataset.

It should be noted that in addition to phrases and words, the wordlists also contained URL link formats, references to attachments and HTML tags, especially for emails that could be classified as phishing. Cohen et al. (2018) indicated that the malicious emails within the dataset contained HTML tags in the email body, as well as suspicious URLs and malicious attachments.

In terms of formatting, as with the words contained in the centroids, the words within the wordlists were also presented in lowercase letters. Furthermore, all contractions were replaced. Furthermore, these words were sent through the lemmatisation process, so that only the root words were contained in the wordlists.

7.4.4.1 Executable: Regular Expression Pattern Matching algorithm

The Regular Expression Pattern Matching algorithm, included in accordance with requirement FR001.FR006, was constructed as follows. Firstly, the initialisation function was created to navigate to the folder containing the wordlists. A regular expression pattern object of the values in the wordlists was created. These regular expression pattern objects were then placed in a dictionary, as key-value pairs, where the key of each value in the dictionary was the name of the given wordlist and the insider threat type. The value within the dictionary would be the list of words and phrases in the associated wordlist.

The second step of the process involved assigning the classification score and deciding whether to assign a label to a given email. Each sentence within the tokenised email was run through the pattern matching algorithm to determine whether a sentence in the tokenised email contained words that matched those in the wordlists. Every sentence was therefore checked against the regular expression pattern object that was created with a dictionary in the first step of the process. A score was subsequently assigned to the email based on a confidence score that determined whether the email should receive a given label. More detail regarding the scoring mechanism is described as part of the next step.

7.4.4.1.1 Executable: Score each sentence in email body

The scoring mechanism informed by requirement SFR103, within the Regular Expression Pattern Matching algorithm, worked as follows. For each email within the tokenised dataset, a JSON scoring dictionary object was initialised. The scoring dictionary contained four keys that represented each of the insider threat types. Their values were blank upon initialisation.

After a given email was assessed by the Regular Expression Pattern Matching algorithm and a match was found between the wordlist dictionary and the email words, the relevant field in the scoring dictionary for the given email was incremented. For example, if the phrase “I am unhappy” was found within an email and it matched the same phrase within the insider IT sabotage wordlist, the ‘sabotage’ score within the scoring dictionary for that email was incremented. As each sentence within the email was traversed, the scores for the insider threat types within the scoring dictionary were incremented when matches were found.

Once this process was completed for the given email, each of the scores within the scoring dictionary, for each type of insider threat, were divided by the total number of sentences within the email. This yielded a value between 0 and 1. This was done to ensure that an email with only one or two matches to a given insider threat type would not be assigned the incorrect label. The insider threat type with the highest score was then identified. However, just as with the work done by Aski and Sourati (2016), even though this label had the highest score for a given email, the score was compared with a chosen threshold before a label was printed within the new labelled dataset. This indicates that a level of certainty was required to assign a label to an email. The chosen threshold for scoring in the work at hand was 0.5.

This implies that, in the example above, if the email obtained the highest score for the sabotage insider threat type within its scoring dictionary and its overall score was above the 0.5 threshold value, then the email would be labelled as ‘sabotage’. If, however, the score was below the 0.5 threshold value, the email would be assigned a label of ‘non-malicious’. The email and its label were then written to a new Excel file that contained the dataset labelled by the Regular Expression Pattern Matching algorithm.

The results from the labelling process are shown in Chapter 8 and snippets of the scoring dictionaries for the given emails are also included.

[7.4.4.1.2. Document: Labelled email body dataset CSV File 2](#)

Once the Regular Expression Pattern Matching algorithm and the scoring mechanism had traversed through all the emails within the dataset and assigned labels to each of the emails, these emails and their labels were written to a CSV file. This was done in accordance with requirement SFR105. The CSV file was saved and named accordingly for use in the supervised

machine learning process. In the steps that follow, the dataset was compared with the dataset labelled by the K-means clustering algorithm.

7.4.5. Executable: Supervised machine learning algorithms

During the classification step of the prototype, supervised machine learning algorithms were used to predict the classes that each data value in the testing dataset belongs to, based on training data. This step adhered to requirement FR001.FR006. Due to their suitability with regard to the requirements at hand, the supervised algorithms that were selected for use in this work were a Naïve Bayes classifier, the Support Vector Machine (SVM) and the Logistic Regression model. The two CSV-labelled datasets that were labelled by the K-means clustering algorithm and the Regular Expression Pattern Matching algorithm respectively, were used as training data for the three supervised machine learning algorithms.

In Python, the scikit-learn toolkit was required to obtain each of the different classifiers or models for use in the prototype. The training data that is used by the classifiers, was provided to a vectoriser function, 'TfidfVectorizer', which is responsible for transforming the textual data into feature vectors that are suitable to be provided as input to the classifiers.

The pre-labelled dataset that was used for the experimentation was divided into two separate datasets, based on a given size ratio. The one portion of the dataset was the training dataset, which contains the rules and features that were applied when classifying the second dataset, namely the testing dataset. In the context of this work, the training dataset contained the pre-labelled emails, and the testing set was labelled based on these.

All the classifiers, therefore, required the same input in the form of two input arrays: the training data (array X) and the class labels that have been assigned to this data (array Y). In this work, these are labelled as 'x_train' and 'y_train' respectively, based on the first portion of the dataset. The second portion, as previously mentioned, was the dataset on which the model was tested. The first output array was array X , which represents the testing data or email bodies that were selected. These were used to determine the number of false classifications. The second output, array Y , contains the labels selected to be testing data. These are named 'x_test' and 'y_test' respectively in the work at hand.

7.4.5.1. *Naïve Bayes classifier*

The Naïve Bayes algorithm is used for classification tasks with very large training datasets (scikit-learn developers, 2019). In this work, multinomial Naïve Bayes was implemented, as the training data in this study contained more than two outcomes. This method was used to determine the probability of a given feature vector that is linked to a specific label. Furthermore, the classifier assumed that features within a given dataset have conditional independence (scikit-learn developers, 2019). The Naïve Bayes classifier is a generative model, specifically a model of the joint probability distribution of the feature variable X and the target variable Y . As such, it infers assumptions regarding the data in its predictions. In text classification problems the data is represented as word vector counts (scikit-learn developers, 2019).

The algorithm works by calculating the probability that a data value belongs in each of the given classes. The largest prediction value was used to determine the class. In the work at hand, the fit function was used to create the model that would be used to train the testing dataset. The predict function was then used to determine the class of the data value.

7.4.5.2. *Logistic Regression Model*

The Logistic Regression model is a statistical model that uses a set of independent input variables, denoted as $\mathbf{x} = (x_1, \dots, x_r)$, where r represents the number of input variables (Stojiljković, 2020). These input variables represent the training dataset. The goal of this model is to discover the logistic regression function $p(\mathbf{x})$ to ensure that the predictions, represented as $p(\mathbf{x}_i)$, correlate closely with the actual output y_i (Stojiljković, 2020). In essence, the model learns the probability of a given dataset corresponding to a particular class. The model attempts to locate the most optimal boundary that best divides the different classes. Unlike the Naïve Bayes model, the Logistic Regression model is a discriminative model that predicts strictly on the basis of the content of the training dataset and as such it does not make assumptions regarding its predictions.

In Python, the NumPy package was used, as it allows for scientific and mathematical computation. In addition, scikit-learn, which is used for data science and machine learning activities in Python, was included.

7.4.5.3. *Support Vector Machine (SVM) classifier*

The SVM classifier works by creating a line of separation or a hyperplane in a multidimensional area in order to distinguish between different clusters (Azaria et al., 2014). The points of data that are nearest to the hyperplane are called support vectors. These are used to better define the hyperplane. In Python, a polynomial kernel is used to work with a nonlinear space. The scikit-learn toolkit was used to include the SVM classifier in the Python code. The training dataset had to be split using the appropriate function. The SVM module also had to be imported, in order for the relevant SVM functions to be accessible.

The fit function uses the labelled dataset or the training dataset as it is known in supervised machine learning. This was used to create the model that was used to train the testing dataset. For each of the values in the testing dataset, the predict function identified where they would be classified, based on this model.

7.4.6. *Executable: Performance metrics*

Once the supervised machine learning classifiers had been executed, performance metrics were used to determine the accuracy and precision of the results produced by the classifiers. This was done in accordance with requirement SFR106 by comparing the percentage of emails containing the same label in the results dataset and the training dataset. This process was automated, and the results were printed as output when the process halted.

The Python metrics module in scikit-learn was used to access the metrics that were included for testing the different machine learning models.

7.5 **Detection**

Since the Enron dataset is quite a large dataset, the process of labelling it, as shown in the data discovery component, required automation. The source code files are contained in Appendix C.

The diagram in Figure 7.5 shows the main processes that were part of the detection component.

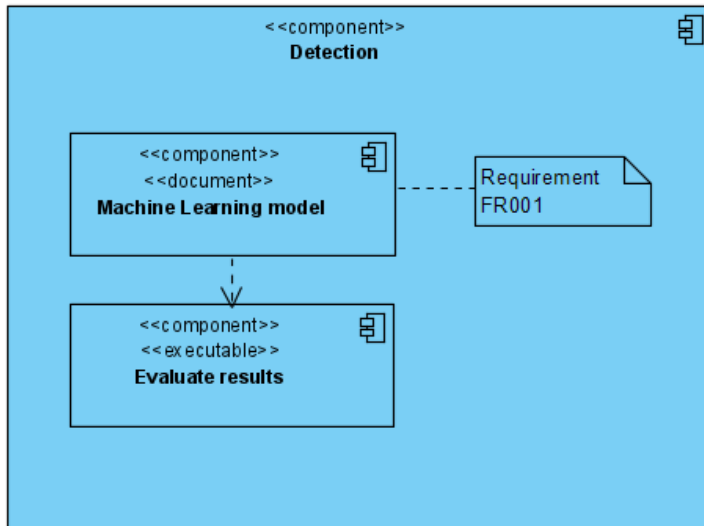


Figure 7.5: Details of the detection component

7.5.1 Document: Machine learning model

As described above, this step was required so that experimentation could take place. The selected dataset (the Enron email corpus) traversed through the steps shown in the second component (data discovery) in Figure 7.2 a given number of times. For each execution, the performance metrics were recorded for each of the supervised machine learning algorithms. Aski and Sourati (2016) utilised a similar process to execute the experiment a given number of times, and to compare the classification algorithms in terms of execution time and false positives. These metrics were then used to aid decisions and reach conclusions for the given research questions.

In addition, objects that were created in the main file, through the relevant constructors for the normalisation and classification processes, wrote the output in Excel and log files. The output was stored in the relevant folders so that data labelled by the different processes could easily be retrieved for analysis. Furthermore, log files were examined to determine if the steps of the various processes executed correctly. Therefore, in this step the entire prototype that addressed requirement FR001, was executed.

7.5.2 Executable: Evaluate results

During this step, the performance metrics – specifically measuring the accuracy, precision and false positives of each of the supervised machine learning classifiers during each of the experiment executions – were examined. The averages were compared to determine which

of the classifiers were best suited for the experimentation at hand. In addition to the comparison of the performance metrics, the actual datasets labelled by the two different approaches (the Regular Expression Pattern Matching algorithm and the K-means clustering algorithm) were compared. The labels within these datasets were compared to determine the percentage of similarly labelled emails. Graphs and other results from the execution of the prototype were also studied.

A discussion of future improvements and further research was also part of this evaluation step. The results of the experiment were analysed and substantiated based on existing strengths and weaknesses of the developed prototype.

7.6 Conclusion

This chapter presented the process that was followed to construct the prototype that aims to address the problem statement. The detailed process was shown, mapped to the relevant requirements, in order to explain the purpose of each step. Similar processes from past research were also included. In addition, outputs of the machine learning models were discussed to determine their benefit in the experimentation process.

The chapter that follows shows the actual results obtained once the prototype was developed and the experimentation process was executed. The purpose of Chapter 8 is to address the problem statement of this research. Specifically, this is done to demonstrate whether a prototype to detect insider threats in corporate emails, by focusing on email body content, is a viable solution.

Chapter 8

Results Obtained from the Prototype Experimentation

8.1 Introduction

The previous chapter discussed each component of the prototype developed as part of this research to show how it aimed to address the requirements set in the research question. Three main components were identified, and their relevance and importance within the prototype were described.

Chapter 8 presents the setup of the actual experimentation environment, as well as the results obtained after development and execution of the prototype. This was done to show whether or not the actual process proved effective in addressing the research question. This would thus serve as a useful tool for detecting insider threat behaviours within an organisation.

8.2 Research question

At this point, it is important to revisit the research question so as to enable the reader to contextualise the results that follow. The steps in and results of the experimentation process serve to address the research question of this work, based on the problem statement, which was as follows:

How can insider threats, caused by human behaviours, accurately be detected within a large corporate email dataset with the development of a comprehensive model?

8.3 Experimentation

This section contains the details of the experimentation that was run, using the Enron dataset of 517401 emails. It must be clarified that, at the start of the experimentation phase, no emails were removed or added to the dataset. During the data-cleaning phase however, duplicate emails were detected and removed from the dataset. As such, the actual number of emails that were processed by the classification and clustering steps – the data discovery component of the prototype – was smaller than the total number. This removal of duplicates was handled at the very start of the process when the mail array was transformed into a set.

This was just before the activities within the data preparation component of the prototype were executed.

8.3.1 Experimentation environment

The experiment was conducted in an environment consisting of an Intel® Core™ i7-8650U CPU @ 1.90GHz Processors, 16GB of RAM, and an x64-based processor.

PyCharm (PyCharm, 2020) is the platform that was chosen to read in the CSV data file, and the Pandas library (Pandas, 2019), was used to apply data normalisation and cleansing techniques so as to develop and run the classifier and clustering models. Not only was the experimentation executed on this platform, but the outputs, log files, resulting datasets and performance metrics were also obtained by using the PyCharm platform. Graphical results were created by using the relevant Python libraries. (The source code is stored in Appendix C for reference.)

8.3.2 Experimentation data

Before the experiment took place, the CSV file containing the Enron email corpus (Cukierski, 2015) was briefly analysed to establish a clear experimentation scope. The exact number of emails in the dataset, as well as the unique total number of sent and received mails was retrieved from the Enron email corpus. The results shown in Figure 8.1 represent a snippet from a log file that records the data upon initial analysis of the file. The number of duplicate emails that were removed from the file are included in the figure. The count for remaining mails shown in the log, 253787, is the total number of lines remaining in the CSV file and includes the header line. As such, 253786 emails remained to be used in the experiment.

```
Mail count before removing duplicated: 517401
Loaded in 4s
Total duplicates removed: 263614
Remaining mails: 253787
```

Figure 8.11: Detail of emails contained in Enron email corpus

Since it was essential to establish the time period spanned by the emails, a graph was created (see Figure 8.2) based on the emails in the CSV corpus, to depict the email activity per time period (Cukierski, 2015). The graph shows that the bulk of the emails were sent and received between 1999 and 2001. The peak of email communications – just over 150000 – was observed in 2001. As previously stated, the company was declared bankrupt in early

December 2001, which explains the massive decline in email communications between 2001 and 2004.

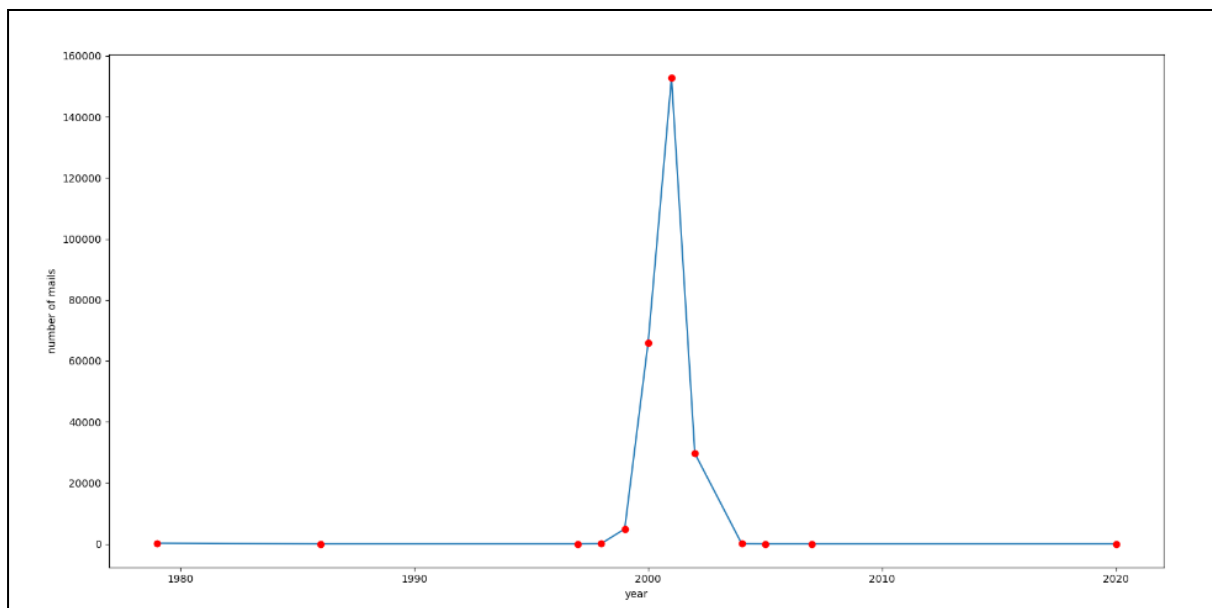


Figure 8.12: Time period spanned by the Enron emails

The number of emails, duplicates removed, as well as the time periods have now been clearly identified. The rest of this chapter focuses on the results obtained during each step of the prototype.

8.3.3 Experimentation iterations

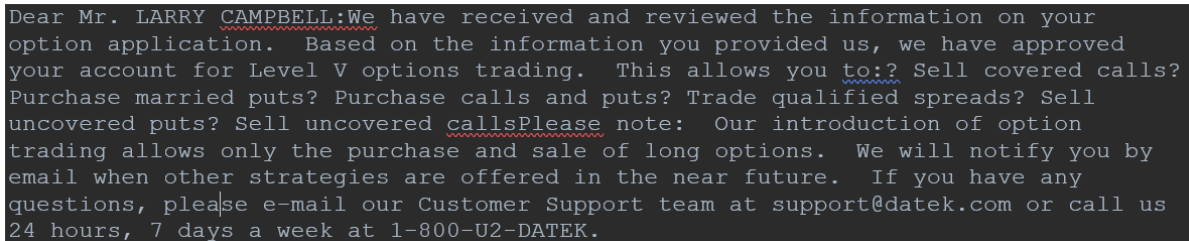
Once the data preparation and data discovery phases of the prototype phase were developed, the actual experimentation process, the detection phase, followed. A main program was created to automatically execute the process. The Enron email dataset was provided to this program and the first component, data preparation, was executed once.

The second component, the data discovery component, was executed next. The supervised machine learning algorithms within this component were executed a total of 15 times for each of the two labelled datasets produced as output from the Regular Expression Pattern Matching algorithm and the K-means algorithm. In total there were 30 executions for each supervised machine learning algorithm. For each run of the supervised machine learning algorithms, performance metrics such as the accuracy and precision of the results were measured.

8.3.4 Data preparation results

8.3.4.1 Attribute selection

The email body, which was selected as the main feature from the Enron email corpus, was extracted from the corpus, while other fields such as the subject and header data were disregarded for the research at hand. An example of one email body extracted from the corpus is shown in Figure 8.3.



```
Dear Mr. LARRY CAMPBELL: We have received and reviewed the information on your
option application. Based on the information you provided us, we have approved
your account for Level V options trading. This allows you to: Sell covered calls?
Purchase married puts? Purchase calls and puts? Trade qualified spreads? Sell
uncovered puts? Sell uncovered calls. Please note: Our introduction of option
trading allows only the purchase and sale of long options. We will notify you by
email when other strategies are offered in the near future. If you have any
questions, please e-mail our Customer Support team at support@datek.com or call us
24 hours, 7 days a week at 1-800-U2-DATEK.
```

Figure 8.13: Extract of email body obtained from the Enron email corpus

8.3.4.2 Feature engineering

Feature engineering methods were not applied to the selected attribute to maintain a simplistic experimentation domain.

8.3.4.3 Data cleaning and normalisation

The data was rigorously cleaned and transformed into a standardised format that would yield more accurate results. Table 8.1 shows the difference between the email bodies before and after the data preparation phase. Two emails have been included for demonstration purposes. It should be noted that the original emails were converted to lowercase before the snapshot was extracted. The examples in Table 8.1 further demonstrate that all the pre-processing steps were successfully applied to the email bodies. Tables D.1 – D.4 in Appendix D contain the snapshots of 10 original emails after the normalisation phase, spelling correction and lemmatisation, which is the final phase of the pre-processing step.

Table 8.1: Differences between the original emails and the pre-processed emails

Email body before normalisation and cleaning	Email body after normalisation and cleaning
<p>any conflicts? ----- forwarded by tana jones/hou/ect on 05/02/2001 09:52 am ----- mark s palmer/enron@enronxgate 05/02/2001 09:09 am to: tana jones/hou/ect@ect cc: subject: lehman nda can i come down and get this this am? thanks, mark p. input info for the nda ----- legal name of the company: lehman brothers inc. business contact name at the company (who will receive the nda): jarett wait business contact email & phone (& fax, if applicable): jwait@lehman.com company's address: 3 world financial center 10th floor new york, ny 10285 nature of discussion: project offline nda delivery method: hard-copy (return to mark s. palmer) other delivery needed? (hard copy sent by mail, fedex, fax): no legal contact at the company (optional but helpful):</p>	<p>conflict forward tana jones of ect 050201 52 mark palmer end on enronxgate 050201 90 tana jones of ect ect cc subject leman da come get thank mark p input info da legal name company leman brother inch business contact name company receive da caret wait business contact email phone fax applicable wait lehman.com company address 3 world financial center 10th floor new york ny 105 nature discussion project online da delivery method hardtop return mark st palmer delivery need hard copy send mail fever fax legal contact company optional helpful</p>
<p>hi guys, we don't expect to deal with the facility agreement until around 130, so there isn't a need for lee to listen to the rest unless he wants to. lee, you are more than welcome to come in person, in which case i can give you some exected change orders. kay ----- forwarded by kay mann/corp/enron on 11/10/2000 09:14 am ----- from: suzanne adams@ect on 11/09/2000 04:08 pm to: sheila tweed/hou/ect@ect, ben jacoby/hou/ect@ect, lisa bills/corp/enron@enron, roseann engeldorf/corp/enron@enron, scott dieball/enron_development@enron_developme nt, stephen.swift@ps.ge.com, michael.barnas@ps.ge.com, kent.shoemaker@ae.ge.com, kay.mann@enron.com, lee.johnson@ss.ps.ge.com cc: subject: ge conference call for november 10, 2000 the conference call will be held at 1:00 p.m. ct tomorrow, november 10, 2000. for the people in houston, i have reserved eb38c1. the dial-in</p>	<p>hi guy not expect deal facility agreement around 130 not need lee listen rest under want . lee welcome come person case give execute change order . kay forward kay mancorpenron 100 14 susan adam ect 100 108 pm sheika toe house ct ect ben jacob house ct ect list oil scorper on enrol rosea engels of cowpen ron enrol scott diet allen nondevelopment nondevelopment stephen.swift p sage com michael.barnas p sage com kent.shoemaker aec get com cayman enron.com leg johnson sep sage com cc subject ge conference call november 10 200 conference call hold 100 pom a ct tomorrow november 10 20 people houston reserve eb38c1 . violin information follow dial 84763757 participant 54 30 30 host 609623 sheika question please give call 7138537340. thank susan</p>

<p>information follows: dial in: 888-476-3757 participant: 543030 host: 609623 (sheila) if you have any questions, please give me a call at 713- 853-7340. thanks, suzanne</p>	
---	--

8.3.5 Data discovery and detection results

The results, inputs and outputs, as well as the various metrics obtained during execution of the experiment to test the prototype are discussed next.

8.3.5.1 Centroids based on insider threat types

The centroids constructed have been included below, in Table 8.2. These aim to show the results obtained from the compilation of the centroids and to illustrate the extensiveness of the synonyms and variety of words and phrases included.

Table 8.2: Centroids fabricated for use in the K-means classifier

Insider threat type	Centroid ID	Centroid
Insider Fraud	C01	so i have been hearing people talk about something that sounds an awful lot like our little deal we have going on. i think you should just send all further communications to my house email, because it could result in a possible lawsuit. let us just get this stuff out of the way as soon as possible before legal steps in and starts asking questions. i have a strong lawyer on hand, but i rather not involve him in this matter. you know, i am not exactly playing by the rules here, but then neither are you. so if you want to keep manipulating these energy prices, i suggest you keep a lid on it from now on. do not let your loose tongue be the cause of a legal conflict. if we play our cards right, it may not be illegal in the eyes of the stakeholders, we just play ignorant. there are a few things i just want to get out of the way right here and now. the first thing is that we cannot be leaving anything anywhere for people to start digging after. so no emails, no messages on company networks, nothing. secondly, there are a few deals and documents i had people sign into without the chance to negotiate. it is important that they remain in the dark as to what is actually going on. if you have any questions, contact me via personal means, like a

		<p>phone call or my own email. now i am sure this is still a light matter and that it will not end up as a federal court case, but that depends if we can keep it under the company's radars. i will let you know when my plans are in place, then you can proceed to advocate the increases in prices to the board members. when they see the numbers, they will buy it for sure.</p>
Negligence	C02	<p>it has come to our attention that you have did not adhere to the policy and guidelines regarding your password for the central system please change your password using our password replacing system. use of the system guarantees that your password complies with the criteria set out in your organization's password policy regarding your written warning, you have to make sure that you sign it by the end of the day this matter requires your immediate attention. welcome to true-tunes.com to make use of this service, please follow the link below to verify your registration details use of the service is subject to these terms and conditions there are several users complaining about receiving an email which says their computer is infected with a virus. we have determined the cause of the email. it is being sent automatically when you log in the central system to get you to enter your details on another site. please disregard this email and delete it immediately great offers for you at your favorite sports and outdoor retailer step into a land of adventure with these great deals on some of our best products www.greatoutdoors.com come and get these and other fantastic offers this is not a joke. you stand a chance to win large cash prizes simply click the link below to register yourself in the lucky draw. http://luc.ky.com/registration this mail is to inform you that your system has been infected with the wannacry virus. this virus locks down your computer and encrypts your files. click the link below to pay \$5000 to unlock your computer. we also have access to all your files. if the money is paid in less than two days we will release your computer and delete all the copies of your files that we have downloaded</p>
Insider IT Sabotage	C03	<p>hey man. i just wanted to talk to someone about some stuff. recently i noticed that i am not happy here. you know, i did not even complete that security training they have been complaining about so much. there are also a ton of stuff i am sitting with that are just thorns in my side. for instance, i have incomplete work, i am just so fed up with all this stuff. if i get one more email about people and their problems, i am losing my cool completely. i am angry, anxious and just plain sick of all documents i have to process all day. i applied for a different position and i could not believe when someone else got the position. It is like nobody here is seeing my hard work is not recognized here and i have given my all for this place for 2 horrible years. i am unappreciated and my work is not valued, although i would care to simply convert a pdf into a spreadsheet. i really want to resign here and i really want to leave. anyway, i have been checking out internet job boards and sent my resume around. i am just so uncertain about the future here. i cannot work under these orders any longer, because that is what they are, orders barked in my face. this is not a skill i</p>

		can market myself with and my frustration has hit a record high. i see no pay raise and no promotions on the horizon. so that is it man. i am really out of here by the end of this month or next month
Insider Intellectual Property (IP) Theft	C04	i have been working on this project for the last 6 year and most of the work done was of my own ingenuity. that is why i think i have a right to the product and formulas used in the software packages. i am not entirely willing to split the difference with anyone. i submitted an enquiry to finance to find out of someone embezzled the account to where the payments for my product went. see attached a copy of the email hi louis i gave been making enquiries this past week and so far only one has returned with an answer. i want to know where did my money go, as i am entitled to at 75% of the profits generated by the entrade-boom formula and the improvements i made to the spread sheet application please advise, as i am entitled to at 75% of the profits generated by the entrade-boom formula and the improvements i made to the spread sheet application please advise
Non-malicious	C05	that party last week was really good. i relaxed a ton. and it was also nice to see some of the managers and executives without their suits for a change. so are we still up for the meeting at four? because i see on our calendars that we are actually scheduled for another meeting at five. i do not know if an hour will be enough to discuss all the points mentioned in the agenda, anyway here is a list of my concerns. the client has not responded to my mails yet and i have sent several to him and his secretary. the latest numbers coming in all match up to the spread sheets generated, but i am missing a file here. could you forward the mail with all the attachments in. the new interns are coming in next week, we need to get their space ready. see you this afternoon

8.3.5.2 Results obtained from the execution of the K-means algorithm

This sub-section presents the results and outputs obtained after execution of the K-means clustering algorithm. During this process, each tokenised email within the pre-processed email dataset was compared with the centroids and classified with the centroid that it was most similar to.

8.3.5.2.1 Dataset labelled by K-means clustering algorithm

Figure 8.4 shows a snippet of the dataset labelled by this process. The link to the labelled dataset can be found in Appendix E.

```

label,body
negligence,filename philip platter 62 60 2.p s time run short . company prepare tag 17 minimum require step must co
negligence,filename golden salisbury 62 60 2.p st i media tel y delete not open email clam raf subject hi attach fi
negligence,filename immediately delete not open email clam raf subject hi attach file gone .s cr serious virus work
nonmalicious,filename first week gross revenue expense net p l 80 88 28 27 point bustle counterpart price deal hour
fraud,enrol ... 20 innovative company america five consecutive years' number one energy co mod it y house 20 top co
fraud,proud announce enrol one title sponsor 100 year energy special air locally abc channel 13 january 13th 20th 6

```

Figure 8.14: Snippet of dataset labelled by the K-means clustering algorithm (Michael & Eloff, 2019)

8.3.5.2.2 Percentage of emails assigned certain labels

Table 8.3 shows the distribution of labels assigned to the emails within the pre-processed Enron email dataset by applying the K-means clustering algorithm. The actual number of emails labelled is displayed, as well as the percentage that indicates the portion of the dataset that was classified according to the given insider threat type. The values were obtained from the log file result shown in Figure 8.5.

Table 8.3: Distribution of labels assigned to emails by using the K-means clustering algorithm

Label	Number of emails labelled	Percentage
Insider IT Sabotage	27753	10.94
Insider Fraud	79680	31.40
Insider Intellectual Property Theft	11182	4.41
Negligence	94930	37.41
Non-malicious	40241	15.86

```

kmeans counts
{'negligence': 94930, 'nonmalicious': 40241, 'fraud': 79680, 'sabotage': 27753, 'iphief': 11182}

```

Figure 8.15: Snippet of K-means clustering algorithm results from log file

8.3.5.2.3 Results of the execution of the K-means clustering algorithm

Figure 8.6 graphically displays the distribution of emails labelled with the K-means clustering algorithm, where the x axis indicates the insider threat types and the y axis shows the number of emails.

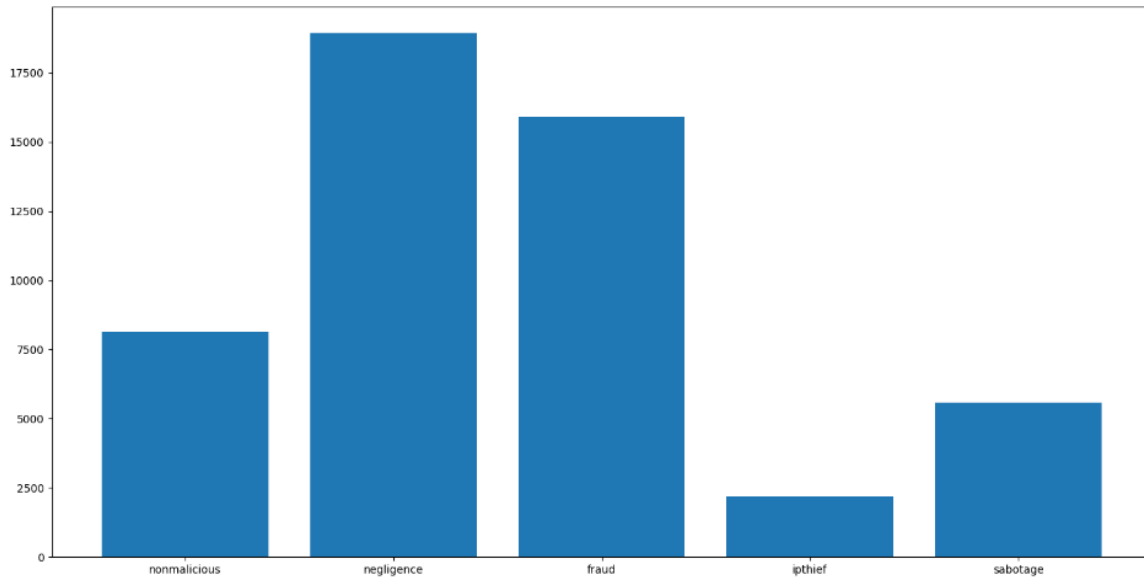


Figure 8.16: Distribution of labels after labelling with the K-means algorithm

8.3.5.2.4 Enron emails most similar to the centroids

Table 8.4 shows, for each centroid, a snippet of the most similar email in the dataset, as well as the similarity percentage for this. A larger snippet of the emails from the log file is shown in Table F.7 in Appendix F.

Table 8.4: The ten emails most similar to the centroids and their cosine similarity values

Insider threat type	Centroid ID	Closest email (snippet from actual email)	Similarity (%)
Insider Fraud	C01	<p>"inform john 20 friendship prize big trouble . situation like 20 rat 34 tick totally pathetic circumstance . hey john stick 20 not worry thing . go get this.20 let take inventory week crisis . electric 20 shortage crisis . center stage 3 alert zone thursday afternoon 20 early morning . natural gas ptyas you burn crisis . of natural gas seller determine p e might not secure 20 buyer . cash crisis . two major utility e ha vein liquid fund pay energy bill . mush three crisis together 20 get new word 1 gash like state motto eureka an 20 find exclaim gash . screw . 20 every one 's start feel pain . esp remain 20 class market participant inform p e payment 20 px credit direct acer customer bill esp due under 20 consolidate bill henceforth suspend . rate freeze end last 20 august logic go p e not need acknowledge px credit . 20 put serious breach contract kink esp utility relationship . 20 course</p>	70.15

		go write hundred million dollar 20 short payment write billion dollar short payment . mean care financial well”	
Negligence	C02	“. password protect secure store image . window 9598me20 suggest retail price e 69.95 image morton utility e 20 improve pc performance . speed disk optimize hard drive put mostneded file sea r front disk faster access . significantly speed load time application document . subsequent optimization n even faster easy keep hard drive work efficient y fix window problem . morton inductor diagnose solve wide run ge window problem software error hardware configuration cone list . protect work fix problem might otherwise lead data loss help window run better clean problem cu r everyday use computer . keep hard drive healthy . r hard drive develop problem make pc run poorly even dama ge valuable data .”	61.40
Insider IT Sabotage	C03	“also hear greg shock look leave . think put ask job opportunity e . oh well ... talk later susan original message christ germany enron.com pereirae houston rico m send thursday april 11 22 82 subject hey know estate team completely separate u . 6 4 not work anemone u however go lunch work wife judy 3 time week . move 6 old build next week .” “hope thing go well not know work stay home may non event . cm energy try develop presence north east . headhunter set interview cm last friday . speak curt liza vice president wholesale gas trade . not know agree see 30 min interview say not look . want someone contact cash trade experience . say hop look someone like year . never know . number 7132307205 . let know thing go . focus live sin pregnant girlfriend . later chris email property enrol corp candor relevant affiliate may contain confidential privilege material sole use intend recipient . review use distribution disclosure other strictly prohibit . not intend recipient authorize receive recipient please contact sender reply enrol corp nero name age administration enron.com delete copy message . email attachment hereto not intend offer acceptance not create evidence bind enforceable contract enrol corp affiliate intend recipient party may not rely anemone basis contract estoppel otherwise . thank .”	70.05
Insider Intellectual	C04	“rustworthiness report process account profession call question . click read . http wpm ulex investor com article	52.04

Property (IP) Theft		asp dock de 58 nd0131 2 investment idea marvell set sight broadloom brim broadloom still dominate communication chip sector marvell mail grow torrid clip . wave sherman equity research columnist firm able grow quickly broadloom brim . communication chip maker saw sale rise 42 million 17 1 billion 20 investor get early make huge profit share broadloom rise twentyfold spring 19 ipo 200 peak . broadloom look maintain base revenue communication chip sector experience cyclical lull previously obscure competitor manage grow downturn set sight industry number one spot”	
Non-malicious	C05	“today edition daily update muller investor director investment research marc epstein explain math theory behind stock price explain investor least know basic buy . also equity research columnist ben marlin discuss pair wireless telecom stock may potential bargain . also today feature couple broker report network stock well report morgan stanley network associate beta reader access free charge register firm free research trial . see . setscrew center help novice well experience investor develop investment idea . click see http wpm ulex investor comb page wasp target stock advisor home dock 50 9 nd0123 receive mail register muller investor”	59.87

8.3.5.3 Wordlists based on insider threat types

The wordlists that were created for the Regular Expression Pattern Matching algorithm are included in Table G.8 in Appendix G. The results of this process are shown next.

8.3.5.4 Results obtained from the execution of the Regular Expression Pattern Matching algorithm

For this classification technique, all wordlists were compared with each pre-processed email to obtain matches and to ensure that scoring could take place to determine which labels should be assigned.

The scoring dictionaries of the 10 highest scored emails shown in Table 8.5 were extracted from the log file printed during the execution of the Regular Expression Pattern Matching classifier. It is evident that the Insider IP Theft counts were significantly low.

Table 8.5: The scoring dictionaries for the 10 highest scored emails

Rank	Dictionary scores
------	-------------------

1	{'total score': 568} {'negligence': 248, 'fraud': 194, 'sabotage': 125, 'ipthief': 1}
2	{'total score': 513} {'negligence': 163, 'fraud': 188, 'sabotage': 162, 'ipthief': 0}
3	{'total score': 485} {'negligence': 215, 'fraud': 160, 'sabotage': 110, 'ipthief': 0}
4	{'total score': 449} {'sabotage': 161, 'negligence': 162, 'fraud': 126, 'ipthief': 0}
5	{'total score': 430} {'fraud': 177, 'sabotage': 105, 'negligence': 148, 'ipthief': 0}
6	{'total score': 385} {'negligence': 181, 'fraud': 101, 'sabotage': 103, 'ipthief': 0}
7	{'total score': 383} {'negligence': 151, 'sabotage': 132, 'fraud': 98, 'ipthief': 2}
8	{'total score': 370} {'negligence': 97, 'fraud': 159, 'sabotage': 114, 'ipthief': 0}
9	{'total score': 357} {'fraud': 154, 'negligence': 128, 'sabotage': 75, 'ipthief': 0}
10	{'total score': 350} {'negligence': 136, 'fraud': 106, 'sabotage': 108, 'ipthief': 0}

8.3.5.4.1 Dataset labelled by Regular Expression Pattern Matching algorithm

Figure 8.7 contains a snippet of the dataset labelled by this process. A link to the labelled dataset can be found in Appendix E.

```
label,body
nonmalicious,full list article send monday initial coverage yesterday today ... money enrol energy trader spinmeiste
negligence,next con ten type text plain reward s news let er december 20 20 1 i sue number dear brad earn 10000 poin
nonmalicious,sample article original message schmidt m sent thursday october 25 20 18 subject 29 enrol mention s enr
negligence,image inform es aging web preview membership reward october travel update subset name 3d ful name inform
negligence,c l c k z tuesday december 12 20th internet lead resource http w. click . com for business online email c
sabotage,forward vine j kam in ski ho u ect 12 19 20 30 pm alliance energy supplier alliance viborg ls. viborg 12 18
```

Figure 8.17: Snippet of dataset labelled with by the Regular Expression Pattern Matching algorithm

8.3.5.4.2 Percentage of emails assigned certain labels

Table 8.6 shows the distribution of labels assigned to the emails within the pre-processed Enron email dataset by using the Regular Expression Pattern Matching algorithm. The actual number of emails labelled is displayed, in addition to the percentage, which indicates the portion of the dataset classified according to the given insider threat type. These values were obtained from the log file result shown in Figure 8.8.

It should be noted that emails were only assigned labels if the overall score, divided by the number of sentences within the email, exceeded a certain threshold.

Table 8.6: Distribution of labels assigned to emails by using the Regular Expression Pattern Matching algorithm

Label	Number of emails labelled	Percentage
Insider IT Sabotage	27411	10.80
Insider Fraud	16796	6.62

Insider Intellectual Property Theft	0	0.00
Negligence	28636	11.28
Non-malicious	180943	71.30

```
regex counts
{'nonmalicious': 180943, 'negligence': 28636, 'fraud': 16796, 'sabotage': 27411}
```

Figure 8.18: Snippet of Regular Expression Pattern Matching algorithm results from log file

8.3.5.4.3 Results of the execution of the Regular Expression Pattern Matching algorithm

Figure 8.9 graphically displays the distribution of emails labelled by the Regular Expression Pattern Matching algorithm, where the x axis indicates the insider threat types and the y axis shows the number of emails.

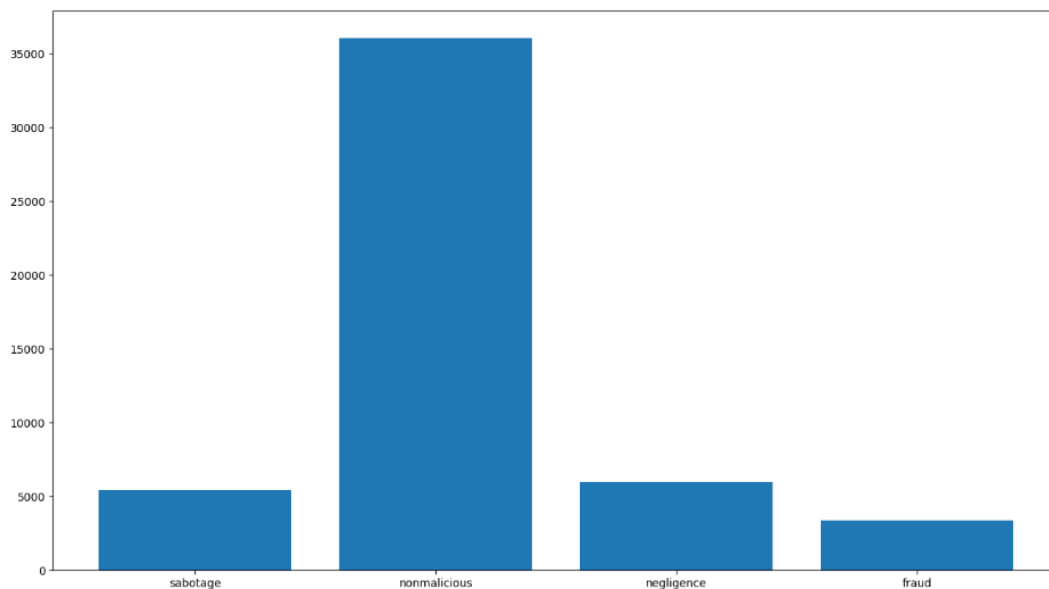


Figure 8.19: Distribution of labels after labelling with the Regular Expression Pattern Matching algorithm

8.3.5.4.4 Percentages of emails that received the same label in both labelling processes

The overall number of similarly classified emails were 64413 out of 253786 emails (25%). These were assigned the same label by both the K-means and Regular Expression Pattern Matching algorithms. This finding is discussed in further detail in the observation section. Both of the algorithmic approaches would need refining to yield a higher percentage for this comparison.

8.3.5.5 Supervised machine learning algorithm results

This sub-section presents the results obtained from the execution of the four supervised machine learning algorithms that used the two labelled datasets from the K-means clustering algorithm and the Regular Expression Pattern Matching algorithm.

It should be noted that each supervised machine algorithm was executed 15 times for each of the datasets.

8.3.5.5.1 Results obtained from the execution of the Support Vector Machine algorithm

The SVM algorithm was first executed using the dataset labelled by the K-means clustering algorithm.

Table 8.7: Performance metrics for the SVM algorithm using the dataset labelled by the K-means clustering algorithm

Iteration number	Precision	Accuracy	False classified (%)
1	0.928	0.928	1.802
2	0.925	0.925	1.870
3	0.925	0.925	1.870
4	0.927	0.927	1.833
5	0.927	0.927	1.822
6	0.926	0.926	1.852
7	0.926	0.926	1.851
8	0.927	0.927	1.826
9	0.926	0.926	1.804
10	0.915	0.915	1.920
11	0.921	0.921	1.870
12	0.928	0.928	1.831
13	0.926	0.926	1.827
14	0.925	0.925	0.925
15	0.926	0.926	1.849

The SVM algorithm was also executed using the dataset labelled with the Regular Expression Pattern Matching algorithm.

Table 8.8: Performance metrics for the SVM algorithm using the dataset labelled by the Regular Expression Pattern Matching algorithm

Iteration number	Precision	Accuracy	False classified (%)
1	0.912	0.914	2.157

2	0.911	0.913	2.175
3	0.912	0.914	2.158
4	0.910	0.912	2.197
5	0.913	0.914	2.140
6	0.910	0.912	2.200
7	0.909	0.911	2.226
8	0.911	0.913	2.171
9	0.922	0.924	2.117
10	0.894	0.875	2.191
11	0.922	0.924	2.158
12	0.910	0.911	2.187
13	0.914	0.915	2.123
14	0.910	0.912	2.200
15	0.910	0.911	2.222

8.3.5.5.2 Results obtained from the execution of the Naïve Bayes algorithm

The Naïve Bayes algorithm was executed using the dataset labelled with the K-means clustering algorithm.

Table 8.9: Performance metrics for the Naïve Bayes algorithm using the dataset labelled by the K-means clustering algorithm

Iteration number	Precision	Accuracy	False classified (%)
1	0.690	0.578	10.542
2	0.650	0.576	10.596
3	0.694	0.578	10.540
4	0.696	0.576	10.594
5	0.702	0.576	10.599
6	0.695	0.576	10.609
7	0.687	0.575	10.634
8	0.697	0.577	10.584
9	0.655	0.576	10.594
10	0.696	0.579	10.535
11	0.678	0.579	10.534
12	0.700	0.577	10.580
13	0.701	0.580	10.492
14	0.694	0.575	10.619
15	0.659	0.583	10.425

The Naïve Bayes algorithm was also executed using the dataset labelled with the Regular Expression Pattern Matching algorithm.

Table 8.10: Performance metrics for the Naïve Bayes algorithm using the dataset labelled by the Regular Expression Pattern Matching algorithm

Iteration number	Precision	Accuracy	False classified (%)
1	0.701	0.737	6.581
2	0.676	0.736	6.591
3	0.708	0.737	6.570
4	0.648	0.736	6.596
5	0.691	0.737	6.578
6	0.697	0.737	6.569
7	0.672	0.740	6.489
8	0.706	0.738	6.555
9	0.699	0.738	6.560
10	0.704	0.735	6.615
11	0.681	0.735	6.617
12	0.725	0.738	6.561
13	0.729	0.738	6.543
14	0.678	0.738	6.562
15	0.704	0.736	6.597

8.3.5.5.3 Results obtained from the execution of the Logistic Regression algorithm

The Logistic Regression algorithm was executed using the dataset labelled with the K-means clustering algorithm.

Table 8.11: Performance metrics for the Logistic Regression algorithm using the dataset labelled by the K-means clustering algorithm

Iteration number	Precision	Accuracy	False classified (%)
1	0.927	0.926	1.848
2	0.926	0.925	1.865
3	0.927	0.926	1.856
4	0.927	0.926	1.852
5	0.926	0.925	1.875
6	0.928	0.927	1.820
7	0.927	0.927	1.831
8	0.928	0.927	1.827
9	0.927	0.926	1.854
10	0.927	0.926	1.841
11	0.928	0.927	1.832
12	0.928	0.927	1.814
13	0.927	0.927	1.835

14	0.926	0.925	1.877
15	0.928	0.927	1.814

The Logistic Regression algorithm was also executed using the dataset labelled with the Regular Expression Pattern Matching algorithm.

Table 8.12: Performance metrics for the Logistic Regression algorithm using the dataset labelled by the Regular Expression Pattern Matching algorithm

Iteration number	Precision	Accuracy	False classified (%)
1	0.898	0.901	2.473
2	0.896	0.899	2.532
3	0.898	0.900	2.488
4	0.896	0.899	2.516
5	0.897	0.900	2.510
6	0.898	0.901	2.474
7	0.897	0.900	2.501
8	0.900	0.902	2.441
9	0.896	0.900	2.512
10	0.898	0.901	2.479
11	0.896	0.899	2.527
12	0.897	0.900	2.509
13	0.898	0.901	2.478
14	0.898	0.901	2.471
15	0.897	0.900	2.500

8.3.5.5.4 Performance metrics from the 15 iterations of the supervised machine learning algorithms

Table 8.13 shows the averages computed, based on the values in Tables 8.7 to 8.12.

Table 8.13: Performance metrics from the execution of the supervised ML algorithms

Supervised ML algorithm	CSV file labelled by Regular Expression Pattern Matching algorithm			CSV file labelled by K-means clustering algorithm		
	Precision	Accuracy	False classified	Precision	Accuracy	False classified
Support Vector Machine	0.911	0.912	2.175	0.925	0.925	1.783
Multinomial Naïve Bayes	0.695	0.737	6.572	0.686	0.577	10.565

Logistic Regression	0.897	0.900	2.494	0.927	0.926	1.843
---------------------	-------	-------	-------	-------	-------	-------

8.4 Observations

One of the interesting findings that emerged from the experimentation conducted with regard to the K-means clustering algorithm, was that the email that was found to be most similar to the negligence centroid, was one that could potentially originate from an IT support consultant or anti-virus advertisement (see Table 8.4). This signifies that several of the words in the centroids are contained in this mail, but when these are arranged in a different manner, they have a different meaning that is not malicious or negligent. As such, it could be considered a false positive. This finding indicates that further refinement of the centroids and process would be required.

Regarding the Regular Expression Pattern Matching algorithm, it is clear that the scoring mechanism ensured that emails were not just assigned a label without the level of surety. Table 8.5 shows that the insider IP theft label did not get assigned to any email. A possible explanation could be that the words in the wordlist are not commonly used words, for example, the word ‘embezzled’ is very specific to this type of attacker. Figure 8.6 also indicates that the K-means clustering algorithm identified a minimal number of Insider IP Theft cases. The centroid, however, contained more detail than the wordlist used by the Regular Expression Pattern Matching algorithm.

It is evident from Figure 8.9 that the Regular Expression Pattern Matching algorithm, with the scoring mechanism, classified most emails that contained typical business jargon as non-malicious, and sorted few emails within the other categories. This differs from the K-means clustering algorithm, which sorted most of the emails into the negligence category. Only 25% of the emails were assigned the same label by both processes, which indicates that the processes require much refinement before the results can be deemed trustworthy. However, the purpose was to demonstrate the proof of concept.

With regard to the supervised machine learning algorithms, the Support Vector Machine yielded the highest scores for accuracy and precision during its iterations. It also yielded the lowest false positive values when using the dataset labelled by the Regular Expression Pattern Matching algorithm. The Logistic Regression model yielded higher results when using the

dataset labelled by the K-means clustering algorithm, while the SVM yielded fewer false positives with this dataset.

A key observation during the early stages was that it was essential to apply identical pre-processing techniques to the centroids and word lists, in order to ensure that contracted, misspelled or uppercase words would not get disregarded. This problem was encountered in the work of Michael and Eloff (2019), and was therefore to be avoided in this work. However, with regard to pre-processing, it was found that the spelling corrector was unable to identify names. It often changed the spelling of names to that of similar words, which in some cases could also be similar to words in the wordlists or centroids.

A detailed discussion of the research questions and results is included in Chapter 9.

8.5 Verification

The results of the experimentation were verified through means such as manual inspection of the labels assigned, the use of multiple supervised machine learning algorithms that tested for false positives, and several iterations of these algorithms. Furthermore, two unique unsupervised labelling processes were used with similar words and phrases, but different processes, and the results of these were compared by means of a similarity score.

Graphs and tables were used to graphically display the results and allow for easier comparison of the effectiveness of the methods. Various log files were also used to capture results throughout the process. For example, the most similar centroids in Table 8.4 and the top dictionary scores in Table 8.5 were obtained.

As part of the pre-processing of data, large snippets of the dataset were compared at different stages of the normalisation and cleaning process, to ensure that they were altered correctly and according to the requirements.

The experiment steps and result snippets were evaluated to ensure that they carefully corresponded with and adhered to the requirements of this work.

8.6 Conclusion

Chapter 8 presented the detailed results that had emerged during each step of the execution of the prototype to demonstrate whether the carefully selected processes would actually be

effective in addressing the problem statement. A detailed discussion of the results is required to determine their contribution to the solution of the problem at hand. Furthermore, attention should be given to the drawbacks of the process and its necessary enhancements.

Chapter 9 therefore provides an in-depth discussion of the various benefits and shortfalls of the process, allowing the researcher to provide adequate answers to the research question stated in this work.

Chapter 9

Discussion and Conclusion

9.1 Introduction

This dissertation focused on detecting the insider threat lurking in corporate email datasets. In order to address the problem statement, research was conducted to establish a definition of insider threats to be used in the work at hand. Furthermore, an overview was presented of insider threats in organisations, the characteristics of insider threats, emails as a platform for attack, human-behaviour-driven insider threats, categories of insider threat and various phrases that can be associated with these in corporate emails. Various requirements were devised for the prototype that was eventually constructed in this work. The development of the prototype ensured that machine learning would be incorporated in labelling the large, pre-processed Enron email dataset and in identifying potential insider threats. Experiments were conducted to determine the effectiveness and accuracy of this prototype.

This concluding chapter therefore aims to determine whether the research and experimentation adequately addressed the criteria set in the research questions and contributed to solving the problem statement.

9.2 Problem statement and main objective of the research

The main objective of this research was to determine whether human-behaviour-driven insider threats could be accurately detected within a large corporate email dataset. A multi-faceted prototype that incorporates machine learning would be developed to determine the validity of this objective. It is thus important to highlight that the objective was not to consider measures to prevent insider threats, but to focus on their detection. The problem statement was formulated as follows:

Human-behaviour-driven insider threats can be accurately detected in a large corporate email communication dataset with the development of a comprehensive model.

Various research questions were subsequently constructed based on this problem statement in order to determine the main objective of this dissertation. The main research question that was devised, read as follows:

How can insider threats caused by human behaviours be accurately detected in a large corporate email dataset by developing a comprehensive model?

To answer the above research question and address the problem statement, the various components of these statements would need to be explored:

- The types of insider threat that exist in a corporate environment
- Human-driven behaviours that trigger these insider threats
- Machine learning as a possible solution to detecting insider threats in emails
- The possibility to detect insider threats in a corporate email dataset by using machine learning algorithms

The above components facilitated the creation of sub-questions that relate to the main research question. Therefore, the next section presents a discussion of each of the research sub-questions to determine whether they were effectively addressed in this dissertation.

9.2.1 Addressing the sub-questions associated with the main research question

The researcher's attempts to address the different sub-questions, based on research and experimentation were presented in the various chapters. The attempts are discussed for each of the three sub-questions as indicated below.

Sub-question (i) What are the main types of insider threat found in a corporate environment and what are the human behaviours that drive these insider threats?

The first part of this question was addressed in chapters 2, 3 and 5. Firstly, by using real examples from the Enron dataset, Chapter 2 introduced the notion of different types of insider threat. Next, four main types of insider threat were briefly introduced in Chapter 3. More detailed discussions of these types, as well as various Enron examples that corresponded with each type, were provided in Chapter 5.

This study focused on identifying specific groupings of insider threat. There might have been merit in specifically identifying scenario-based groupings of insiders, such as the 'ambitious

leader', 'fraudster', 'saboteur' and 'rager' (to name a few) (Young et al., 2014), as this would allow for more categories by means of which to streamline results. In the work at hand, however, using only four groupings meant that characteristics of different insiders had to be grouped together as a single insider threat, where these could actually refer to separate types of insiders. For example, the insider threat type 'sabotage' houses traits from the 'rager' and 'saboteur' insider types.

Inclusion of the non-malicious grouping (see the requirements in Chapter 6) was essential so as to avoid incorrectly forcing emails that had few or no references to insider threat into a category of insider threat.

The second part of sub-question (i) was addressed in Chapters 4 and 5. Human behaviours had to be examined, as past research established that they are the source of various different types of insider threat. It was found that a study of behaviours would make it easier to describe the traits of the different insider threat types and thus allow for a more targeted detection process.

An important point made in Chapter 4 involved the identification of a window of opportunity, during which an employee would often display visible changes in physical behaviour that could allow for an imminent insider threat to be detected, before the attack is executed. Technical behaviour changes, associated with actions that were not required for the employee's daily tasks on the systems or network, were also considered. Identifying this window of opportunity was an important finding and highly relevant to the development of the prototype for this work.

In Chapter 6, the requirements for the prototype were devised. The main functional requirement would be to attempt to detect – from the large corporate dataset – the four main types of insider threat associated with human behaviours. The prototype that was designed in Chapter 7 and executed in Chapter 8, reflected these detection efforts and confirmed the researcher's successful handling of sub-question (i).

Sub-question (ii) How can machine learning be used to detect insider threats with specific reference to corporate email systems?

Chapter 5 studied past research to determine whether it contained approaches to detecting insider threats and which approaches were considered. It became evident that classification and clustering techniques were employed for similar tasks involving insider threats as well as the labelling of large datasets. Furthermore, the chapter looked into identifying the most suitable algorithms for this work.

The prototype that was developed and presented in Chapters 7 and 8, was shown to effectively eliminate manual classification processes. It was also found that machine learning was effective in classifying emails according to specific groupings (see further discussion in the response to sub-question (iii)). However, using the machine learning approach of this work in isolation to classify and cluster a dataset, exhibited shortcomings, as the results contained false positives and incorrect labels.

To enhance the results obtained from the machine learning algorithms, a confidence score should be implemented, which considers various metadata features such as the email address of the sender and recipient, as well as subject line detail. Furthermore, the date or time when the mail was sent might also play a role because it was found that attackers usually conduct attacks after hours. Similar to the dictionary score used in this work, the confidence score would assign a label only if a certain threshold was exceeded.

The current study also found that there was definite room for improvement in respect of the data provided to the machine learning algorithms, the wordlists and centroids. The words used were limited to the extent of research conducted regarding the behaviours of insiders within the four categories (further discussed in response to sub-question (iii)).

Finally, it became clear that an existing labelled big dataset would provide the ideal training data for using machine learning to accurately detect insider threat in a corporate email environment.

Sub-question (iii) How can insider threats be identified in a given corporate email dataset, based on a set of phrases that link to different insider threat types and that are related to specific behaviours associated with insiders?

This question was addressed by the research conducted in the literature review chapter (Chapter 4) as well as in the design and execution of the prototype (Chapters 7 and 8). The

results reported in Chapter 8 showed that pre-processing was very necessary if a detection tool were to prove even slightly effective at classifying a dataset. Standardisation of the Enron emails as well as the phrases within the wordlists and centroids was essential. It was also essential to increase the number of iterations of the experimentation in order to determine overall accuracy. A comparison was made to determine the similarity of the two labelled datasets and the result showed quite a significant difference between the two sets.

Manual inspection of the emails in the dataset revealed that labels that did not correspond with the content of the email had been assigned and some were false positives. The use of scoring also allowed for more accurate results, because far fewer emails were classified as threats. It was found that incorrect labelling had occurred due to the large number of unique terms in the emails that had not been included in the wordlists. In addition, the words in the wordlists or centroids might have been detected in an email – which caused the mail to be classified as a threat – despite the fact that the words had been used in a different and harmless context. Furthermore, where the wordlist or centroid contained uncommon words or phrases, these would often not be sufficiently evident in the dataset, and few labels would be assigned.

It should be noted that while detecting physical changes in human behaviour is doable, detecting changes in technical behaviour via the email platform is not a simple task because it often involves the use of systems that might not leave email evidence. However, a possible way of checking for these changes would be to check if a given user has contacted persons from other departments who do not usually engage in business matters or work together. This would require using features such as the sender and recipient (see the work by Zaki et al., 2017) to create social graphs and map the topics of discussion between the various parties to detect unusual interactions. The latter would be a possible enhancement to the work at hand.

The above responses to the research sub-questions have shown that it is feasible to identify insider threats in a given corporate email dataset, based on a set of phrases that link different insider threat types to specific insider behaviours. However, enhancements would definitely be required for detection to be more accurate in confidently labelling emails reflecting potential insider threat.

9.2.2 Summary and conclusion

This work presented a novel approach that focuses on detecting the insider threat risk in corporate email systems. Detection involved approaches such as the normalisation, cleaning, classification and clustering of data through the use of supervised and unsupervised machine learning algorithms.

It was found that the results of the supervised machine learning algorithms can only be as accurate as the training data they use, which is generated by the unsupervised machine learning algorithms. Thus, the processes selected and executed for insider threat detection – as theoretically valid as they seemed – could not be guaranteed to accurately discover insiders.

The approach adopted in the study at hand seemed to be a good proof of concept, but various enhancements would be necessary in future research (see suggestions in Section 9.5). The key contributions brought forward in this work are discussed next.

9.3 Main contributions

The key contributions that were made by this dissertation are summarised below:

- An insider threat detection prototype that can identify potential insiders – based on human behaviours and types of insider threat – was developed to be used with corporate CSV email datasets. Methods such as normalisation, supervised and unsupervised machine learning, scoring and performance metrics were included in this approach.
- Since a labelled dataset, classified according to types of insider threat, did not exist upon commencement of this study, the researcher facilitated the development of such a labelled big data corporate email dataset to be used for training data in further applications.
- The study validated the need for insider threat detection software in organisations, based on the evidence in corporate email data found in this work.
- The researcher demonstrated that machine learning is a valid approach to detect insider threats within organisations, based on the classification of specific human behaviours and insider threats.

- The work contained in this dissertation can also be tailored to apply to various insider threat related problems within organisations. One example is that the insider threat types discussed in this work, as well as the examples detected within the Enron email corpus, could be used to formulate insider threat scenarios. These can be reconciled against existing insider threat scenarios within an organisation's qualitative risk assessment, which are hypothetical instances used to estimate probability and impact. These reconciled scenarios can then serve as instruction material for incident response teams, as well as to assess the appropriateness of company policies and SETA (security education, training and awareness) programmes.

9.4 Work flowing from this research

An earlier version of this prototype was described and executed for the research presented in a conference paper by Michael and Eloff (2019). The current work was also used to implement a focused Insider IT Sabotage Detection Model as described in the paper by Michael and Eloff (2020).

9.5 Future research

The research reported on in this dissertation served as the proof of concept for a model to detect human-behaviour-based insider threats in corporate emails. However, various drawbacks were identified and as such there are certain areas that would require enhancement in future research:

- In order for the prototype developed to yield better results, a labelled dataset needs to be created according to the types of insider threats. Without such a dataset, training data has to be constructed through the use of wordlists, centroids and machine learning algorithms. These need to be as detailed and complete as they can, in order to ensure the best possible results.
- Different approaches to the wordlists could be devised to incorporate different dictionaries that consist of words and phrases relating to specific human behaviours and emotions (for example 'anger') in the classification process.
- Header data and email metadata could be included in the dataset to enhance the accuracy of the labels assigned. For example, results could be enhanced by considering the sender's email address(es) to determine if the latter represented known

suspicious accounts and to check for email addresses that did not belong to Enron employees. Furthermore, examining the dates and times when emails were sent, as well as the frequency of mails, could provide additional indications of insider threat. These could be incorporated with a scoring mechanism that only assigns labels if they exceed a given threshold.

- Detecting social networks between the different employees in the dataset could be considered to enable the researcher to determine normal business relationships, as well as strange connections between unlikely candidates, due to differing job roles, departments and positions.
- During the pre-processing tasks, only the Python `symspelly` English WordNet dictionary was used for spelling correction, which obviously excludes the detection of emails that consist of words written in other languages. Additional languages could be added to the spelling corrector and the wordlists to account for cases where the company conducts business or engages with people from other countries. The Synsets would require updates because synonyms would not be provided for words written in other languages.
 - Another potential change to the spelling correction task would be the inclusion of English slang and jargon words.
 - Currently the spelling correction mechanism in the prototype uses the first best match as the new spelling to replace the misspelled word. This can be changed to utilise the overall best match as the replacement.
- Due to the size of the labelled dataset and the lack of existing labelled datasets, checking whether correct labels have been assigned is currently a manual process. Future research to consider developing an automated label-checking process could be conducted.

Appendix A

Ethics approval clearance



Faculty of Engineering, Built Environment and Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le TheknoloSi ya Tshedimošo

Reference number: EBIT/132/2019

Miss A Michael
Department: Computer Science
University of Pretoria
Pretoria
0083

Dear Miss A Michael

FACULTY COMMITTEE FOR RESEARCH ETHICS AND INTEGRITY

Your recent application to the EBIT Research Ethics Committee refers.

Conditional approval is granted.

This means that the research project entitled "A MACHINE LEARNING APPROACH TO DETECT INSIDER THREATS IN EMAILS CAUSED BY HUMAN BEHAVIOURS" is approved under the strict conditions indicated below. If these conditions are not met, approval is withdrawn automatically.

Conditions for approval

The applicant does not imply that public data is free to use when the corresponding papers/dissertation is written. The applicant indicated that "dataset that was leaked available from Kaggle.com (Cukierski, 2015). " leaked implies it is still illegally provided on Kaggle. Although this is not the case with regards to the Enron data set, but please rephrase that it was "made" public.

This approval does not imply that the researcher, student or lecturer is relieved of any accountability in terms of the Code of Ethics for Scholarly Activities of the University of Pretoria, or the Policy and Procedures for Responsible Research of the University of Pretoria. These documents are available on the website of the EBIT Ethics Committee.

If action is taken beyond the approved application, approval is withdrawn automatically.

According to the regulations, any relevant problem arising from the study or research methodology as well as any amendments or changes, must be brought to the attention of the EBIT Research Ethics Office.

The Committee must be notified on completion of the project.

The Committee wishes you every success with the research project.

Prof K.-Y. Chan

Chair: Faculty Committee for Research Ethics and Integrity

FACULTY OF ENGINEERING, BUILT ENVIRONMENT AND INFORMATION TECHNOLOGY

Appendix B

Research detail on the selection of the algorithms to detect insider threats within corporate email datasets

The algorithms and approaches shown in Table B.1 were selected by conducting a search in various academic search engines such as DBLP, IEEE, Science Direct and Google Scholar. These search engines were selected because they are commonly used in the Computer Science domain and are known for hosting a large number of journal articles. Furthermore, these were recommended by the research supervisor.

The searches were performed by using various combinations of keywords such as ‘insider threats’, ‘email’, ‘big data’, ‘machine learning’, ‘text analysis’ and ‘sentiment’, and by restricting the dates to the period between 2015 to 2020. Table B.1 shows the 10 most relevant papers found in the searches, with a summary of the processes and approaches used by the authors involved.

Table B.1: Algorithms and processes followed in top 10 similar research papers

Name of Paper	Machine Learning Algorithm	Description of Process or Prototype Developed by the Authors	Available in Python?
Prediction and Detection of Malicious Insiders’ Motivation based on Sentiment Profile on Webpages and Emails (Jiang et al., 2018)	Convolutional Neural Network (CNN) (Supervised learning in paper)	Prediction of Insider Threats based on user’s email content and network URL browsing activity. Based on this content, daily and weekly sentiment profiles are developed for the users. A dictionary with words related to feelings of revenge and disgust is used for sentiment analysis. A score is then determined for each user to indicate the likelihood of a threat posed by the user. The changes in the wording used in their email and browsing history are compared on a daily and weekly basis and those users whose changes exceed a specific threshold are labelled as potential insider threats. The detection model is implemented using the Convolutional Neural Network approach.	Yes

<p>Use of Machine Learning in Big Data Analytics for Insider Threat detection (Mayhew et al., 2015)</p>	<p>Support Vector Machine (SVM) (Supervised learning)</p> <p>K-Means Clustering (Unsupervised learning)</p>	<p>The approach examines network information, HTTP requests, server records and email and message content. A tool developed by the authors called Behaviour-Based Access Control (BBAC) takes in large datasets and creates a standardised representation of the multiple sources of information via a feature enrichment process. K-Means clustering is used to group specific actors that are in turn mapped to other clusters. In each cluster, the Support Vector Machine algorithm is executed. The output will show whether the actor poses a potential insider threat and, if so, suggest a suitable course of action, for example, to notify the administrator.</p>	<p>Yes</p>
<p>Determining Predisposition to Insider Threat Activities by using Text Analysis (Chi et al., 2016)</p>	<p>K-Means Clustering (Unsupervised learning)</p>	<p>The authors consider the email and social media content of employees and use an approach of linguistic analysis as well as a personality model to determine whether the employee displays similar characteristics as past known insider threat attackers. The Enron dataset is used in conjunction with other synthetic data. The first step is to detect and remove non-malicious emails from the dataset. The emails are then scored against categories of personality traits, based on known insider threat cases. The K-Means algorithm is used to cluster the dataset and detect outliers. If the score assigned to a given actor exceeds a specific threshold, the actor is considered an insider threat.</p>	<p>Yes</p>
<p>The Insider Threat: Behavioral Indicators and Factors influencing Likelihood of Intervention (Bell et al., 2019)</p>	<p>N/A</p>	<p>A large Critical National Infrastructure organisation within the energy domain is used as the case study. Machine learning is not used in this paper. The authors used a survey to conduct the experimentation and sent a questionnaire to various levels of employees within different departments of the organisation. The questions were structured to obtain information regarding behavioural changes, extreme attitude changes, unusual working hours and distasteful content regarding the organisation posted on social media. In addition, questions regarding the willingness of employees to report unethical behaviour were added to detect whether different statuses and motivations of the employee play a role in this.</p>	<p>N/A</p>

		It must be noted that even though the paper does not use an algorithmic approach, important factors regarding employee behaviour are relevant and essential to the prototype that will be developed in the research at hand.	
Security Threats for Big Data: A Study on Enron e-Mail Dataset (Zaki et al., 2017)	<p>Social Network Analysis tool that consists of the following:</p> <p>Natural Language Processing (can include both supervised and unsupervised methods)</p> <p>Data Mining Algorithm</p>	The software categorises the Enron emails into different topics that emerged from the data, the network of the organisation and the network of the individual users. The communication network of the organisation is shown graphically based on topic categories and individual users. The email topics that emerged are discussed to show how these could be used by malicious actors. The communication network was created to show whether any strange connection existed between two parties that would not engage for business-related discussions, and to show their topic of discussion. A graph network model using edges and nodes was used for this.	Yes
Clustering and Classification of Email Contents (Alsmadi & Alhami, 2015)	<p>Vector Space Model, which includes K-Means algorithm (Unsupervised learning)</p> <p>Support Vector Machine (SVM) (Supervised learning)</p> <p>N-gram algorithm</p>	Five predetermined categories are created based on typical email folders such as 'job' and 'personal'. The K-Means algorithm is used to provide labels to the dataset based on clusters. A random selection is made of documents in the dataset and these are used as centroids. A similarity check is performed in order to cluster like emails. For accuracy, classification algorithms such as SVM are run. Several iterations of the classification algorithm experiment are run to obtain performance metrics.	Yes

	(Supervised learning)		
Email Classification Research Trends: Review and Open Issues (Mujtaba et al., 2017)	Support Vector Machine (SVM) (Supervised learning) Decision Tree (Supervised learning) Naïve Bayes (Supervised learning) K Nearest Neighbour (Supervised learning) Random Forest (Supervised learning) K-Means clustering (Unsupervised learning)	The authors examine the most commonly used email classification techniques by examining a large array of research in the domain. Email classification is commonly used to detect spam and phishing and specifically to label emails as either 'ham' or 'spam', according to the authors. An email dataset is used, and the Enron email dataset is one of the most commonly used datasets. The emails are cleaned with means such as stemming and tokenisation, which break the email into a string of words, using the 'bag of words' model. Features are then extracted based on user behaviour or activity, in order to improve the machine learning accuracy. Features include the email header and email body. A classification algorithm is subsequently run on the data. The most commonly used one is the SVM technique, followed by the Decision Tree and the Naïve Bayes algorithm. Clustering can also be used to enhance accuracy, and the authors indicate that K-Means is the most commonly used unsupervised algorithm. To ensure that the classification technique yields accurate results, metrics such as true positive, false positive, true negative and false negative cases are considered.	Yes
Novel Set of General Descriptive Features for Enhanced Detection of Malicious Emails Using Machine	Supervised learning approaches: J48	The authors obtained a large collection of real-world malicious and non-malicious emails. The email data was run through a feature extractor that considers the metadata and content. The feature vectors were collected from this extraction to form a CSV file. Nine machine algorithms were applied to the dataset via the Weka data mining tool. The algorithmic results showed that the use of various HTML tags in the email	Yes

Learning Methods (Cohen et al., 2018)	Random Forest Naïve Bayes Bayesian Network Logistic Regression LogitBoost Sequential Minimal Optimisation Bagging AdaBoost	body, as well as links with differently named URLs and suspicious attachments were common to the malicious emails within the dataset. Features were also extracted for the non-malicious emails. The precision metric was obtained for the experimentation. One of the experiments conducted by the authors was to draw a comparison between features shown in past research to be associated with malicious emails, and the features extracted by the algorithms.	
Proposed Efficient Algorithm to Filter Spam using Machine Learning Techniques (Aski & Sourati, 2016)	Decision Tree (C4.5) (Supervised learning) Multilayer Perceptron (Supervised learning)	The authors detected electronic spam by using a scoring method based on rules that considered the email header, the email body, as well as keyword matching. Various classification machine learning algorithms were used to facilitate this. Each email was assessed by 23 rules or criteria and the email was assigned a score for each rule. Rules included, for example, an email address being on a black or whitelist, specific characters in the email body, a vague subject field and empty spaces in the email body. The total sum of scores was then computed. The email score was compared to a threshold value to determine whether the email would be assigned the email score and be labelled as 'junk'. The	Yes

	Naïve Bayes (Supervised learning)	classification algorithms were compared in terms of execution time and false positives.	
Insider Threat Detection Using Characterising User Behavior (Wang et al., 2018)	Support Vector Machine (SVM) (Supervised learning)	The authors consider user behaviour to detect possible insiders. Behaviours and user actions within the organisation network, such as users' shell command sequences, keystrokes and other interactions with the GUI, are considered. This information is given to the developed system so that a profile can be built for the user. The system computes feature values based on the behaviours and a model is subsequently trained for the features. An example of a feature is the time a command was entered in the shell to conduct a search. When a user behaviour profile is constructed, the authors can calculate a value to show the extent to which an action performed by the user corresponds to expected behaviour. If the value exceeds a threshold, the user is classified as a potential threat. During experimentation, the F-measure is used to test the effectiveness of the detection algorithm.	Yes

Having scoped 98 big data research articles that delve into email content classification, Mujtaba et al. (2017) stated that the Support Vector Machines and Naïve Bayes classifier are among the most commonly used supervised machine learning algorithms. Furthermore, it was found that the supervised machine learning algorithm, Logistic Regression, is also a popular choice for the detection of insider threat classifiers. The most commonly used unsupervised machine learning algorithm is the K-Means classifier (Mujtaba et al., 2017). These algorithms (shown in Table B.1) have been used in other insider threat or email-related big data research. The algorithms are aligned with the requirements stipulated in this work and therefore they are suitable for use.

Table B.2, however, includes a discussion of the benefits and shortfalls of the selected algorithms so as to show their suitability to the requirements of this study.

Table B.2: Benefits and shortfalls of the selected algorithms

Algorithm	Benefits	Shortfalls
Naïve Bayes	1. Popular choice for email spam detection (Hussain & Qamar, 2014).	1. Results of probability outputs are not very sophisticated or specific (Gupta, 2020).

	<ol style="list-style-type: none"> Works well for classification problems and to make predictions (Nizamani et al., 2014). Found to yield the most accurate detection results when compared to other supervised machine learning classifiers such as SVM, Multilayer Perception and Random Forest after experimentation was conducted (Cohen et al., 2018; Hussain & Qamar, 2014). 	<ol style="list-style-type: none"> Processing time is longer than for the other supervised machine learning classifiers (Hussain & Qamar, 2014).
Support Vector Machine	<ol style="list-style-type: none"> A cut above the other classification algorithms when it comes to text classification (Nizamani et al., 2014). Performs better than the other machine learning algorithms on a multi-dimensional dataset (Nizamani et al., 2014). It can also transform data that is not linearly separable into data that can be divided by a linear line into two classes (Nizamani et al., 2014). Used for a wide variety of machine learning solutions such as image classification and speech recognition (Mayhew et al., 2015). Latencies are lower and performance of classification is much faster due to the training of the algorithm with available data (Mayhew et al., 2015). When combined with K-Means algorithm, SVM provides an optimal balance of quality and efficiency (Mayhew et al., 2015). 	<ol style="list-style-type: none"> When classes overlap, the performance of the classifier is diminished (Gupta, 2020). Difficult for a human to understand the meaning of the classifier results obtained, therefore it is less user friendly than decision trees (Mayhew et al., 2015). Heavily reliant on CPU for classification speed and can take a specific amount of time based on complexity and size of dataset. (Mayhew et al., 2015).
K-Means Clustering	<ol style="list-style-type: none"> Most commonly used unsupervised learning method, and easiest to use (Tsipenyuk & Crowcroft, 2017). Can easily create clusters and groupings based on specific similarities or behaviours in the dataset and not just an arbitrary, meaningless grouping (Mayhew et al., 2015). 	<ol style="list-style-type: none"> Results are not as reliable and effective as those from a supervised algorithm that has been trained from a pre-classified dataset (Alsmadi & Alhami, 2015). The larger the dataset, the less accurate the results will be, due to more unique terms being present in

	<ol style="list-style-type: none"> 3. These clusters can help users understand emerging themes within the data and make it easier for users to work with the grouped dataset (Mayhew et al., 2015). 4. In addition, the resulting clusters can improve the speed of classification and training as well as enhance the accuracy of results (Mayhew et al., 2015). 5. Does not require any training data, which is useful when working in a domain that does not have pre-existing labelled datasets. 	<p>the dataset, (Alsmadi & Alhami, 2015).</p>
<p>Logistic Regression</p>	<ol style="list-style-type: none"> 1. Used to minimise noisy data by filtering with a threshold for false negative data during predictions. Thus, it is shown to be an effective means of distinguishing between bogus data and useful data within the dataset (Wijaya & Bisri, 2016). More true positive results can be obtained through the use of this model. 2. Found to be an excellent classifier for email spam detection (Wijaya & Bisri, 2016). 3. It is known to be simple to implement. 	<ol style="list-style-type: none"> 1. When the Logistic Regression model was compared with Naïve Bayes, SVM, K Nearest Neighbour, and C4.5 Decision Tree algorithms (based on their performance in handling three datasets, using two models), the best results were yielded by Naïve Bayes and SVM and not by Logistic Regression (HaCohen-Kerner et al., 2020). 2. Performance is diminished when data in various classes overlap (Gupta, 2020). 3. A significant amount of time is required for the classifier to process a larger dataset (Gupta, 2020).

Appendix C

Python scripts in accordance with requirements to yield results in Chapter 8

1. Initial Setup

```
1 |
2 import os
3
4 import nltk
5
6 # from wordlist_normalizer import WordListNormalizer
7
8 nltk.download('wordnet')
9 nltk.download('punkt')
10 nltk.download('stopwords')
11 nltk.download('omw')
```

2. Normaliser

```
2 import re
3 import string
4
5 from nltk import word_tokenize
6 from nltk.corpus import stopwords
7
8 from replacers import RegexpReplacer, RepeatReplacer
9
10
11 class Normalizer(object):
12     _regex_replacer = RegexpReplacer()
13     _repeat_replacer = RepeatReplacer()
14     _punct_pattern = re.compile('[%s]' % re.escape(string.punctuation.replace('.', '')))
15     _stopwords = stopwords.words('english')
16     _stopwords.remove('not')
17
18     def normalize_mails(self, mails):
19         i = 0
20         total = len(mails)
21         for mail in mails:
22             i += 1
23             print('\rNormalizing: %d%%' % (i / total * 100), end='')
24             try:
25                 new_body = self.normalize_text(mail.get_body())
26                 mail.set_body(new_body)
27             except Exception as e:
28                 print('error normalizing %s' % e)
29                 break
30             print('')
31
32     def normalize_text(self, text):
33         new_text = self._to_lowercase(text)
34         new_text = self._replace_contractions(new_text)
35         new_text = self._remove_punctuation(new_text)
36         new_text = self._remove_repeated_letters(new_text)
37         new_text = self._remove_stopwords(new_text)
38         return new_text
39
40     def _to_lowercase(self, text):
41         return text.lower()
```

```

42
43 ▼ def _replace_contractions(self, text):
44     new_text = Normalizer._regex_replacer.replace(text)
45     return new_text
46
47 ▼ def _remove_punctuation(self, text):
48     tokenized = word_tokenize(text)
49     new_text = []
50 ▼     for token in tokenized:
51         new_token = Normalizer._punct_pattern.sub('', token)
52         if new_token != '':
53             new_text.append(new_token)
54     new_text = ' '.join(new_text)
55     return new_text
56
57 ▼ def _remove_stopwords(self, text):
58     new_text = ' '.join([word for word in word_tokenize(text) if word not in Normalizer._stopwords])
59     return new_text
60
61 ▼ def _remove_repeated_letters(self, text):
62     new_text = ' '.join([Normalizer._repeat_replacer.replace(word) for word in word_tokenize(text)])
63     return new_text
64

```

3. Lemmatiser

```

1
2 from nltk import WordNetLemmatizer, word_tokenize
3
4
5 ▼ class Lemmatizer(object):
6 ▼     def __init__(self):
7         self._lemmatizer = WordNetLemmatizer()
8
9 ▼     def lemmatize_mails(self, mails):
10 ▼         for mail in mails:
11             mail.set_body(self.lemmatize_text(mail.get_body()))
12
13 ▼     def lemmatize_text(self, text):
14         new_text = []
15         tokens = word_tokenize(text)
16 ▼         for token in tokens:
17             new_token = self._lemmatizer.lemmatize(token, pos='v')
18             new_token = self._lemmatizer.lemmatize(new_token, pos='n')
19             new_text.append(new_token)
20         return ' '.join(new_text)
21

```

4. Mail

```

1
2 # the code in this file is made to work with the data found in the
3 # /data/enron_large.csv file. The _parse method is tightly
4 # coupled with the format of the enron_large.csv file
5
6 # it is assumed that there are 15 headers in each mail, refer to _meta_headers
7 # for the headers used in this program, also to add or remove headers
8
9 import re
10 import datetime
11
12 from datetime import datetime
13
14
15 ▼ class Mail(object):
16     _header_start_pattern = re.compile('[a-z-]*:')
17     _meta_headers = ['message-id', 'date', 'from', 'to', 'subject']
18     _default_label = 'nonmalicious'
19     _date_pattern = re.compile('([a-zA-Z]*), ([0-9]*) ([a-zA-Z]*) ([0-9]*)')
20 ▼     _month_mapping = {
21         'jan': 1,
22         'feb': 2,
23         'mar': 3,
24         'apr': 4,
25         'may': 5,
26         'jun': 6,
27         'jul': 7,
28         'aug': 8,
29         'sep': 9,
30         'oct': 10,
31         'nov': 11,
32         'dec': 12
33     }

```

```

34
35 ▼ def __init__(self, header, content):
36     self._meta_info = {}
37     self._body = ''
38     self._header = header
39     self._original_body = ''
40     self._parse(content)
41     self._scores = {}
42     self._num_sentences = 0
43     self._regex_label = None
44     self._kmeans_label = None
45
46 ▼ def __setitem__(self, index, value):
47     if index not in self._scores:
48         self._scores[index] = 0
49 ▼     else:
50         self._scores[index] = value
51
52 ▼ def __getitem__(self, index):
53     if index in self._scores:
54         return self._scores[index]
55     return 0
56
57 ▼ def __lt__(self, other):
58     return self.get_total_scores() < other.get_total_scores()
59
60 ▼ def __gt__(self, other):
61     return self.get_total_scores() > other.get_total_scores()
62
63 ▼ def __eq__(self, other):
64     if type(self) != type(other):
65         return False
66     return self.get_body() == other.get_body()
67
68 ▼ def __hash__(self):
69     return hash(self._body)
70
71 ▼ def __str__(self):
72     return '%s\nTotal score: %d' % (self._meta_info['subject'], self.get_total_scores())
--

```

```

73
74 ▼ def get_total_scores(self):
75     total = 0
76     for key in self._scores:
77         total += self._scores[key]
78     return total
79
80 ▼ def get_regex_label(self):
81     if self._regex_label is not None:
82         return self._regex_label
83     label = Mail._default_label
84     highest_score = 0
85 ▼     for key in self._scores:
86 ▼         if self._scores[key] > highest_score:
87             label = key
88             highest_score = self._scores[key]
89     self._regex_label = label
90     return label
91
92 ▼ def get_kmeans_label(self):
93     return self._kmeans_label
94
95 ▼ def set_kmeans_label(self, label):
96     self._kmeans_label = label
97
98 ▼ def get_num_sentences(self):
99     return self._num_sentences
100
101 ▼ def set_num_sentences(self, value):
102     self._num_sentences = value
103
104 ▼ def get_original(self):
105     return self._original_body
106
107 ▼ def get_body(self):
108     return self._body
109
110 ▼ def set_body(self, new_body):
111     self._body = new_body
112
113 ▼ def get_meta_data(self, header):
114     try:
115         return self._meta_info[header]
116 ▼     except KeyError:
117         return ''
118
119 ▼ def set_meta_data(self, header, value):
120     self._meta_info[header] = value
121
122 ▼ def get_scores(self):
123     return self._scores
124

```

```

125 ▼ def _parse(self, content):
126     num_headers = 15
127     header_index = 0
128     line_index = 0
129     lines = [line.lower().strip() for line in content.split('\n')]
130     current_header = ''
131 ▼     while header_index < num_headers:
132         line = lines[line_index]
133         line_index += 1
134 ▼         if Mail._header_start_pattern.match(line):
135             header_index += 1
136             colon_index = line.index(':')
137             current_header = line[:colon_index]
138 ▼             if current_header in Mail._meta_headers:
139                 current_header_content = line[colon_index + 1:].strip()
140 ▼                 if current_header == 'date':
141                     match = Mail._date_pattern.match(current_header_content)
142 ▼                     try:
143                         self._meta_info[current_header] = datetime.strptime(
144                             '%s-%s-%s' % (match[2], Mail._month_mapping[match[3].lower()], match[4]), '%d-%m-%Y')
145                     )
146                     except:
147                         self._meta_info['date'] = None
148                 else:
149                     self._meta_info[current_header] = current_header_content
150             else:
151                 continue
152 ▼         else:
153 ▼             if header_index < num_headers:
154                 if current_header in Mail._meta_headers:
155                     self._meta_info[current_header] += ' %s' % line
156             if line_index >= len(lines):
157                 break
158             self._body = ' '.join(lines[line_index:])
159             self._original_body = ' '.join(lines[line_index:])
160

```

5. Replacers

```

2   import re
3
4   from nltk.corpus import wordnet
5
6 ▼ replacement_patterns = [
7       ('won\t', 'will not'),
8       ('can\t', 'cannot'),
9       ('i\m', 'i am'),
10      ('ain\t', 'is not'),
11      ('\w+)\ll', '\g<1> will'),
12      ('\w+)\n\t', '\g<1> not'),
13      ('\w+)\ve', '\g<1> have'),
14      ('\w+)\s', '\g<1> is'),
15      ('\w+)\re', '\g<1> are'),
16      ('\w+)\d', '\g<1> would')
17  ]
18
19
20 ▼ class RegexpReplacer(object):
21 ▼     def __init__(self, patterns=replacement_patterns):
22         self.patterns = [(re.compile(regex), repl) for (regex, repl) in patterns]
23
24 ▼     def replace(self, text):
25         s = text
26         for (pattern, repl) in self.patterns:
27             (s, count) = re.subn(pattern, repl, s)
28         return s
29
30

```



```

20 class RegexpReplacer(object):
21     def __init__(self, patterns=replacement_patterns):
22         self.patterns = [(re.compile(regex), repl) for (regex, repl) in patterns]
23
24     def replace(self, text):
25         s = text
26         for (pattern, repl) in self.patterns:
27             (s, count) = re.subn(pattern, repl, s)
28         return s
29
30
31 class RepeatReplacer(object):
32     def __init__(self):
33         self.repeat_regex = re.compile(r'(\w*)(\w)\2(\w*)')
34         self.repl = r'\1\2\3'
35
36     def replace(self, word):
37         try:
38             if wordnet.synsets(word):
39                 return word
40             repl_word = self.repeat_regex.sub(self.repl, word)
41             if repl_word != word:
42                 return self.replace(repl_word)
43             else:
44                 return repl_word
45         except Exception as ex:
46             pass
47         return ''
48
49
50 class WordReplacer(object):
51     def __init__(self, word_map):
52         self.word_map = word_map
53
54     def replace(self, word):
55         return self.word_map.get(word, word)

```

6. Spelling correction main

```

1 import os
2 import sys
3 import pickle
4
5 from spelling_correcter import SpellingCorrecter
6
7 if not os.path.exists('temp'):
8     os.mkdir('temp')
9
10 first_index = int(sys.argv[1])
11 second_index = int(sys.argv[2])
12 data_name = sys.argv[3]
13
14 f = open('pickled/normalized_%s.pkl' % data_name, 'rb')
15 mails = pickle.load(f)
16 f.close()
17
18 second_index_clipped = second_index if second_index < len(mails) else len(mails) - 1
19 mails = mails[first_index:second_index_clipped]
20
21 c = SpellingCorrecter()
22 c.correct_mail_spelling(mails)
23 print('done')
24 f = open('temp/%s_%d_%d.tmp' % (data_name, first_index, second_index), 'wb')
25 pickle.dump(mails, f)
26 f.close()

```

7. Spelling corrector

```

1 import os
2
3 from nltk import word_tokenize
4 from nltk.corpus import wordnet
5
6 from symspellpy.symspellpy import SymSpell, Verbosity
7
8
9 class SpellingCorrecter(object):
10
11     _file_name = 'data/wordnet_dictionary.txt'
12     if not os.path.isfile(_file_name):
13         _words = [word for word in wordnet.words('eng')]
14         f = open(_file_name, 'wt')
15         for word in _words:
16             f.write('%s\n' % word)
17         f.close()
18     _symspell = SymSpell(2, 7)
19     _in_error = False
20     if not _symspell.create_dictionary(_file_name):
21         _in_error = True
22     wordnet.ensure_loaded()
23
24     def __init__(self):
25         self._synset = wordnet.synsets
26
27     def correct_mail_spelling(self, mails):
28         i = 0
29         total = len(mails)
30         for mail in mails:
31             i += 1
32             print('\rProcessing spelling: %d%% %d of %d' % (i / total * 100, i, total), end='')
33             new_body = self.correct_text_spelling(mail.get_body())
34             mail.set_body(new_body)
35         print('')
36
37     def correct_text_spelling(self, text):
38         SpellingCorrecter._symspell._max_dictionary_edit_distance = 2
39         new_text = self._process_spelling(text)
40         SpellingCorrecter._symspell._max_dictionary_edit_distance = 1
41         new_text = self._process_segmentation(new_text)
42         return new_text
43
44     def _process_spelling(self, text):
45         tokenized = word_tokenize(text)
46         new_text = []
47         for token in tokenized:
48             if not self._synset(token) and token != '.,':
49                 try:
50                     suggestions = SpellingCorrecter._symspell.lookup(token, Verbosity.ALL)
51                     new_text.append(self._find_best_suggestion(token, text, [suggestion.term for suggestion in suggestions]))
52                 except:
53                     print('error while processing %s' % token)
54             else:
55                 new_text.append(token)
56         return ' '.join(new_text)
57
58     def _process_segmentation(self, text):
59         new_text = []
60         for word in word_tokenize(text):
61             try:
62                 suggestion = SpellingCorrecter._symspell.word_segmentation(word)
63                 if len(suggestion.corrected_string) > len(word):
64                     for item in word_tokenize(suggestion.corrected_string):
65                         new_text.append(item)
66             else:
67                 new_text.append(word)
68             except:
69                 print('error processing segmentation %s' % word)
70         return ' '.join(new_text)
71
72     def _find_best_suggestion(self, word, sentence, suggestions):
73         best_match = suggestions[0] if len(suggestions) > 0 else word
74
75         return best_match

```

8. Classifier

```

2 import os
3 import re
4 import pandas
5
6 from nltk import sent_tokenize
7
8 from sklearn.cluster import KMeans
9 from sklearn.feature_extraction.text import TfidfVectorizer
10
11 from normalizer import Normalizer
12
13
14 class Classifier(object):
15     def __init__(self, required_score=0.5, lbl_prefix='lb_'):
16         self._required_score = required_score
17         self._lbl_prefix = lbl_prefix
18         self._in_error = False
19         self._error_msg = 'classifier: no error'
20         self._labels = None
21         self._tfidf_vectorizer = None
22         self._kmeans = None
23         self._cluster_map = None
24         self._centroid_data = None
25         self._normalizer = Normalizer()
26         self._has_regex = self._init_regex()
27         self._has_kmeans = self._init_kmeans()
28     def in_error(self):
29         return self._in_error
30
31     def get_error_msg(self):
32         return self._error_msg
33
34     def classify_mails(self, mails):
35         self._regex_classify(mails)
36         self._kmeans_classify(mails)
37
38
39
40

```

9. Main file

```

1
2 import os
3 import time
4 import math
5 import pandas
6 import pickle
7 import threading
8 import subprocess
9 import multiprocessing
10
11 from mail import Mail
12 from math import floor
13 from random import random
14 from results import Results
15 from functools import reduce
16 from classifier import Classifier
17 from normalizer import Normalizer
18 from lemmatizer import Lemmatizer
19 from spelling_correcter import SpellingCorrecter
20
21
22 class Runner(threading.Thread):
23     def __init__(self, task, mail_data):
24         self._task = task
25         self._mails = mail_data
26         threading.Thread.__init__(self)
27
28     def run(self):
29         if len(self._mails) == 0:
30             return
31         self._task(self._mails)
32
33
34 class ProcessRunner(threading.Thread):
35     def __init__(self, task, index):
36         self._task = task
37         self._index = index
38         threading.Thread.__init__(self)

```

```

39
40 ▼ def run(self):
41     self._task(self._index)
42
43
44 ▼ def get_mail_by_id(mails, id):
45     mail = [mail for mail in mails if mail.get_meta_data('message-id') == id]
46     return mail[0]
47
48
49 ▼ def write_normalize_step(step_name, mails, ids):
50     if not os.path.exists('./output/normalize_steps'):
51         os.mkdir('./output/normalize_steps')
52     f = open('./output/normalize_steps/%s.txt' % step_name, 'wt')
53     for id in ids:
54         f.write('%s\n\n' % get_mail_by_id(mails, id).get_body())
55     f.close()
56
57
58 ▼ def normalize_mails(mails):
59     n = Normalizer()
60     n.normalize_mails(mails)
61
62
63 ▼ def correct_spelling(mails):
64     c = SpellingCorrecter()
65     c.correct_mail_spelling(mails)
66
67
68 ▼ def lemmatize_mails(mails):
69     l = Lemmatizer()
70     l.lemmatize_mails(mails)
71
72

```

```

73 ▼ def run_process(index):
74     first_index = index * mail_set_size
75     second_index = first_index + mail_set_size
76     s = subprocess.Popen(
77         ['python3', './spelling_correction_main.py', str(first_index), str(second_index), data_name]
78     )
79     file_names.append('temp/%s_%d_%d.tmp' % (data_name, first_index, second_index))
80     (out, err) = s.communicate()
81
82
83 ▼ def log(text):
84     print(text)
85     f = None
86     if not os.path.exists('small_outputs/zz_run_log.txt'):
87         f = open('small_outputs/zz_run_log.txt', 'wt')
88     else:
89         f = open('small_outputs/zz_run_log.txt', 'at')
90     str_val = '%s' % text
91     f.write('%s\n' % str_val.replace('\n', ''))
92     f.close()
93
94
95 num_threads = multiprocessing.cpu_count()
96 data_name = 'enron_large'
97 file_names = []
98
99 ▼ if not os.path.exists('pickled'):
100     os.mkdir('pickled')
101
102 ts_start = time.time()
103

```

```

104 mails = []
105 pickle_file = 'pickled/%s.pkl' % data_name
106 ▼ if os.path.isfile(pickle_file):
107     f = open(pickle_file, 'rb')
108     mails = pickle.load(f)
109     f.close()
110 ▼ else:
111     log('reading mails...')
112     file_name = 'data/%s.csv' % data_name
113     raw_data = pandas.read_csv(file_name)
114     data = zip(raw_data['file'], raw_data['message'])
115     mails = [Mail(file, content) for (file, content) in data]
116     f = open(pickle_file, 'wb')
117     pickle.dump(mails, f)
118     f.close()
119
120 ts_end = time.time()
121
122 total_duplicates = len(mails)
123
124 log('Mail count before removing duplicated: %d' % len(mails))
125 log('Loaded in %ds' % (ts_end - ts_start))
126
127 mails_set = list(set(mails))
128 num_mails = len(mails_set)
129 total_duplicates -= num_mails
130 mail_set_size = math.ceil(num_mails / num_threads)
131
132 random_indices = list(set([floor(random() * num_mails) for i in range(50)]))
133 random_ids = [mails_set[i].get_meta_data('message-id') for i in random_indices]
134
135 log('Total duplicates removed: %d' % total_duplicates)
136 log('Remaining mails: %d' % num_mails)
137 log('Number of mails per thread: %d' % mail_set_size)
138
139 ts_start = time.time()
140
141 write_normalize_step('original', mails_set, random_ids)
142
143 normalized_mails = []

```

```

144 normalized_pickle_file = 'pickled/normalized_%s.pkl' % data_name
145 ▼ if os.path.isfile(normalized_pickle_file):
146     f = open(normalized_pickle_file, 'rb')
147     normalized_mails = pickle.load(f)
148     f.close()
149 ▼ else:
150     log('normalizing mails')
151     mail_sets = []
152     for i in range(num_threads):
153         mail_sets.append(mails_set[i * mail_set_size:(i + 1) * mail_set_size])
154     threads = []
155 ▼ for i in range(num_threads):
156     threads.append(Runner(normalize_mails, mail_sets[i]))
157     threads[i].start()
158     for i in range(num_threads):
159         threads[i].join()
160     normalized_mails = reduce(lambda x, y: x + y, mail_sets)
161     f = open(normalized_pickle_file, 'wb')
162     pickle.dump(normalized_mails, f)
163     f.close()
164
165 ts_end = time.time()
166
167 write_normalize_step('normalizing', normalized_mails, random_ids)
168
169 log('Normalized %d mails in %ds' % (len(normalized_mails), ts_end - ts_start))
170
171 ts_start = time.time()
172

```

```

173 corrected_mails = []
174 corrected_pickle_file = 'pickled/corrected_%.pkl' % data_name
175 ▼ if os.path.isfile(corrected_pickle_file):
176     f = open(corrected_pickle_file, 'rb')
177     corrected_mails = pickle.load(f)
178     f.close()
179 ▼ else:
180     threads = []
181 ▼ for i in range(0, num_threads):
182     r = ProcessRunner(run_process, i)
183     r.start()
184     threads.append(r)
185 for t in threads:
186     t.join()
187 ▼ for f_name in file_names:
188     f = open(f_name, 'rb')
189     m_temp = pickle.load(f)
190     f.close()
191     corrected_mails = corrected_mails + m_temp
192     f = open(corrected_pickle_file, 'wb')
193     pickle.dump(corrected_mails, f)
194     f.close()
195
196 ts_end = time.time()
197
198 write_normalize_step('spelling_correction', corrected_mails, random_ids)
199
200 log('Corrected spelling in %ds' % (ts_end - ts_start))
201
202 ts_start = time.time()
203
204 lemmatized_mails = []
205 lemmatized_pickle_file = 'pickled/lemmatized_%.pkl' % data_name

```

```

206 ▼ if os.path.isfile(lemmatized_pickle_file):
207     f = open(lemmatized_pickle_file, 'rb')
208     lemmatized_mails = pickle.load(f)
209     f.close()
210 ▼ else:
211     mail_sets = []
212     for i in range(num_threads):
213         mail_sets.append(corrected_mails[i * mail_set_size:(i + 1) * mail_set_size])
214     threads = []
215 ▼ for i in range(num_threads):
216     threads.append(Runner(lemmatize_mails, mail_sets[i]))
217     threads[i].start()
218 for i in range(num_threads):
219     threads[i].join()
220     lemmatized_mails = reduce(lambda x, y: x + y, mail_sets)
221     f = open(lemmatized_pickle_file, 'wb')
222     pickle.dump(lemmatized_mails, f)
223     f.close()
224
225 ts_end = time.time()
226
227 write_normalize_step('lemmatizing', lemmatized_mails, random_ids)
228
229 log('lemmatized mails in %ds' % (ts_end - ts_start))
230
231 ts_start = time.time()
232
233 classified_mails = []
234 classified_pickle_file = 'pickled/classified_%.pkl' % data_name

```

```

235 ▼ if os.path.exists(classified_pickle_file):
236     f = open(classified_pickle_file, 'rb')
237     classified_mails = pickle.load(f)
238     f.close()
239 ▼ else:
240     log('classifying mails')
241     classifier = Classifier(0.8)
242     classifier.classify_mails(lemmatized_mails)
243     classified_mails = lemmatized_mails
244     f = open(classified_pickle_file, 'wb')
245     pickle.dump(classified_mails, f)
246     f.close()
247
248 ts_end = time.time()
249
250 log('Classifications done in %ds' % (ts_end - ts_start))
251
252 regex_counts = {}
253 kmeans_count = {}
254 ▼ for mail in classified_mails:
255     if mail.get_regex_label() in regex_counts.keys():
256         regex_counts[mail.get_regex_label()] += 1
257     else:
258         regex_counts[mail.get_regex_label()] = 1
259     if mail.get_kmeans_label() in kmeans_count.keys():
260         kmeans_count[mail.get_kmeans_label()] += 1
261 ▼     else:
262         kmeans_count[mail.get_kmeans_label()] = 1
263
264 log('regex counts')
265 log(regex_counts)
266 log('kmeans counts')
267 log(kmeans_count)
268 f = open('output/%s_classification_counts.txt' % data_name, 'wt')
269 f.write('regex classification counts')
270 for k, v in regex_counts.items():
271     f.write('%s: %d\n' % (k, v))
272 f.write('\nkmeans classification counts')
273 for k, v in kmeans_count.items():
274     f.write('%s: %d\n' % (k, v))
275 f.close()

```

```

276
277 results = Results(data_name)
278 results.write_output(classified_mails)
279
280 mails_original_processed_count = 10
281 ▼ for i in range(mails_original_processed_count if len(classified_mails) > mails_original_processed_count else
len(classified_mails)):
282     f = open('output/%s_processed_mail_%d.txt' % (data_name, i), 'wt')
283     f.write(classified_mails[i].get_original())
284     f.write('\n')
285     f.write(classified_mails[i].get_body())
286     f.close()
287

```

10. Test

```

1
2
3 import pickle
4
5 normalized_file = 'pickled/normalized_enron_large.pkl'
6 corrected_file = 'pickled/classified_enron_large.pkl'
7
8 f = open(normalized_file, 'rb')
9 normalized_mails = pickle.load(f)
10 f.close()
11
12 f = open(corrected_file, 'rb')
13 corrected_mails = pickle.load(f)
14 f.close()
15
16 for i in range(len(corrected_mails)):
17     if normalized_mails[i]['message-id'] == corrected_mails[i]['message-id']:
18         corrected_mails[i].set_meta_data('date', normalized_mails[i].get_meta_data('date'))
19         print(normalized_mails[i].get_meta_data('date'))
20     else:
21         print('DOUBLE WTF!!!')
22
23 f = open(corrected_file, 'wb')
24 pickle.dump(corrected_mails, f)
25 f.close()
26

```

11. Results

```

1
2 from normalizer import Normalizer
3
4 import pandas
5
6 from sklearn.feature_extraction.text import TfidfVectorizer
7
8
9 class Results(object):
10     def __init__(self, data_name, top_n=10):
11         self._data_name = data_name
12         self._top_n = top_n
13         self._normalizer = Normalizer()
14         self._vectorizer = TfidfVectorizer()
15
16     def write_output(self, mails):
17         self._write_labeled(mails)
18         self._write_mutual(mails)
19         self._write_scores(mails)
20         self._write_centroid_similar(mails)
21         self._write_per_centroid_similar(mails)
22         self._write_top_n_scores(mails)
23
24     def _write_top_n_scores(self, mails):
25         mails.sort(key=lambda m: (m.get_total_scores()), reverse=True)
26         f = open('output/%s_%d_scores.txt' % (self._data_name, self._top_n), 'wt')
27         for i in range(self._top_n):
28             total_score = {
29                 'total': mails[i].get_total_scores()
30             }
31             f.write('%s %s\n' % (total_score, mails[i].get_scores()))
32         f.close()
33

```



```

34 ▼ def _write_labeled(self, mails):
35     f = open('output/%s_regex.csv' % self._data_name, 'wt')
36     f.write('label, body\n')
37     for mail in mails:
38         f.write('%s, %s\n' % (mail.get_regex_label(), mail.get_body()))
39     f.close()
40     f = open('output/%s_kmeans.csv' % self._data_name, 'wt')
41     f.write('label, body\n')
42     for mail in mails:
43         f.write('%s, %s\n' % (mail.get_kmeans_label(), mail.get_body()))
44     f.close()
45
46 ▼ def _write_mutual(self, mails):
47     f = open('output/%s_similarities.csv' % self._data_name, 'wt')
48     f.write('mail_id, regex_label, kmeans_label\n')
49     for mail in mails:
50         f.write("%s, %s, %s\n" % (mail.get_meta_data('message-id'), mail.get_regex_label(), mail.get_kmeans_label()))
51     f.close()
52     f = open('output/%s_similarities_counts.txt' % self._data_name, 'wt')
53     total_mails = len(mails)
54     total_common = len([mail for mail in mails if mail.get_regex_label() == mail.get_kmeans_label()])
55     f.write('total mails labeled: %d\n' % total_mails)
56     f.write('total mutual labels: %d\n' % total_common)
57     f.close()
58

```

```

59 ▼ def _write_scores(self, mails):
60     mails.sort(key=lambda m: (m.get_total_scores()), reverse=True)
61     f = open('output/%s_scores.csv' % self._data_name, 'wt')
62     f.write('mail_id, score\n')
63     for mail in mails:
64         f.write("%s, %s\n" % (mail.get_meta_data('message-id'), mail.get_total_scores()))
65     f.close()
66     f = open('output/%s_label_scores' % self._data_name, 'wt')
67     f.write('mail_id, scores\n')
68     f2 = open('output/%s_label_scores' % self._data_name, 'wt')
69     index = 0
70     for mail in mails:
71         f.write("%s, %s\n" % (mail.get_meta_data('message-id'), mail.get_scores()))
72     if index < 20:
73         f2.write('%s\n\n' % mail.get_body())
74         index += 1
75     f.close()
76     f2.close()
77

```

```

78 ▼ def _write_centroid_similar(self, mails):
79     data = pandas.read_csv('data/kmeans/kmeans_training.csv')
80     centroids = [self._normalizer.normalize_text(centroid) for centroid in data['body']]
81     y_centroids = self._vectorizer.fit_transform(centroids).todense()
82     y_data = self._vectorizer.transform([mail.get_body() for mail in mails])
83     similarity_mat = (y_centroids * y_data.T).A
84     j_row = 0
85     j_col = 0
86     highest_score = 0.0
87     highest_coords = []
88     for i in range(self._top_n):
89         for j in range(len(similarity_mat)):
90             for k in range(len(similarity_mat[j])):
91                 if similarity_mat[j][k] > highest_score:
92                     j_row = j
93                     j_col = k
94                     highest_score = similarity_mat[j][k]
95                 highest_coords.append((j_row, j_col, highest_score * 100.0))
96                 similarity_mat[j_row][j_col] = 0.0
97                 highest_score = 0.0
98     f = open('output/%s_centroid_similar.csv' % self._data_name, 'wt')
99     f.write('score, centroid, mail\n')
100     for coord in highest_coords:
101         f.write("%.4f, %s, %s\n" % (coord[2], centroids[coord[0]], mails[coord[1]].get_body()))
102     f.close()
103

```

```

104 ▼ def _write_per_centroid_similar(self, mails):
105     data = pandas.read_csv('data/kmeans/kmeans_training.csv')
106     centroids = [self._normalizer.normalize_text(centroid) for centroid in data['body']]
107     most_similar = []
108 ▼     for centroid in centroids:
109         y_centroid = self._vectorizer.fit_transform([centroid])
110         y_data = self._vectorizer.transform([mail.get_body() for mail in mails])
111         similarity_map = (y_centroid * y_data.T).A
112         highest_score = 0.0
113         highest_index = 0
114 ▼         for i in range(len(similarity_map[0])):
115             if similarity_map[0][i] > highest_score:
116                 highest_score = similarity_map[0][i]
117                 highest_index = i
118         most_similar.append((centroid, highest_index, highest_score * 100.0))
119     f = open('output/%s_per_centroid_similar.csv' % self._data_name, 'wt')
120     f.write('score, centroid, mail\n')
121     for coord in most_similar:
122         f.write('%.4f, %s, %s\n' % (coord[2], coord[0], mails[coord[1]].get_body()))
123     f.close()
124

```

Supervised machine learning models are shown below.

12. Support Vector Machine

```

2   from sklearn.svm import SVC
3
4   from ml_model import MLModel
5
6
7 ▼ class SupportVectorMachine(MLModel):
8 ▼     def __init__(self, mails):
9         super().__init__(mails)
10        self._model = SVC(C=1.0, kernel='linear', degree=3, gamma='auto')
11        self._name = 'support vector machine'
12
13 ▼     def _calculate_data(self, labels, data):
14        super()._calculate_data(labels, data)
15        self._y_train_dtm = self._y_train
16
17        def _fill_metrics(self, data_class):
18            super()._fill_metrics(data_class)

```

13. Logistic Regression

```

2   from sklearn.linear_model import LogisticRegression
3
4   from ml_model import MLModel
5
6 ▼ class LogRegression(MLModel):
7 ▼     def __init__(self, mails):
8         super().__init__(mails)
9         self._model = LogisticRegression(solver='lbfgs', multi_class='auto', max_iter=4000)
10        self._name = 'logistic regression'
11
12 ▼     def _calculate_data(self, labels, data):
13        super()._calculate_data(labels, data)
14        self._y_train_dtm = self._y_train
15
16        def _fill_metrics(self, data_class):
17            super()._fill_metrics(data_class)
18

```

14. Naïve Bayes

```

3 from sklearn.naive_bayes import MultinomialNB
4
5 from ml_model import MLModel
6
7
8 class NaiveBayes(MLModel):
9     def __init__(self, mails):
10        super().__init__(mails)
11        self._model = MultinomialNB()
12        self._name = 'naive bayes'
13
14    def _calculate_data(self, labels, data):
15        super()._calculate_data(labels, data)
16        self._y_train_dtm = self._y_train
17
18    def _fill_metrics(self, data_class):
19        super()._fill_metrics(data_class)
20
21
22

```

15. ML Model

```

2 from sklearn import metrics
3 from sklearn.model_selection import train_test_split
4 from sklearn.feature_extraction.text import TfidfVectorizer
5
6
7 class MLModel(object):
8     def __init__(self, mails):
9         self._mails = mails
10        self._vectorizer = TfidfVectorizer()
11        self._num_labels = 0
12        self._name = 'generic'
13        self._model = None
14        self._fitted_result = None
15        self._y_train = None
16        self._y_test = None
17        self._x_train = None
18        self._x_test = None
19        self._y_train_dtm = None
20        self._y_test_dtm = None
21        self._x_train_dtm = None
22        self._x_test_dtm = None
23        self._prediction = None
24        self._probabilities = None
25        self._metrics = {}
26        self._false_classifications = []
27        self._target = set([mail.get_regex_label() for mail in mails])
28
29    def __iter__(self):
30        for metric in self._metrics:
31            yield metric
32
33    def __getitem__(self, metric):
34        if metric in self._metrics:
35            return self._metrics[metric]
36        return -1
37
38    def process(self):
39        self.process_regex()
40        self.process_kmeans()
41

```

```

42 ▼ def process_regex(self):
43     print('running regex for %s' % self._name)
44     if self._model is None:
45         return
46     labels = [mail.get_regex_label() for mail in self._mails]
47     bodies = [mail.get_body() for mail in self._mails]
48     self._num_labels = len(set(labels))
49     self._calculate_data(labels, bodies)
50     self._run_model()
51     self._fill_metrics('regex')
52
53 ▼ def process_kmeans(self):
54     print('running kmeans for %s' % self._name)
55     if self._model is None:
56         return
57     labels = [mail.get_kmeans_label() for mail in self._mails]
58     bodies = [mail.get_body() for mail in self._mails]
59     self._num_labels = len(set(labels))
60     self._calculate_data(labels, bodies)
61     self._run_model()
62     self._fill_metrics('kmeans')
63
64 ▼ def get_predictions(self):
65     return self._prediction
66
67 ▼ def calculate_probabilities(self, x_dims):
68     try:
69         return self._model.predict_proba(x_dims)
70     except:
71         return None
72
73 ▼ def get_target(self):
74     return self._target
75
76 ▼ def metrics(self):
77     return self._metrics
78

```

```

79 ▼ def _calculate_data(self, labels, data):
80     print('calculating training and testing data for %s' % self._name)
81     self._prediction = None
82     self._x_train, self._x_test, self._y_train, self._y_test = train_test_split(data, labels)
83     self._x_train_dtm = self._vectorizer.fit_transform(self._x_train)
84     self._x_test_dtm = self._vectorizer.transform(self._x_test)
85
86 ▼ def _run_model(self):
87     print('fitting %s' % self._name)
88     self._fitted_result = self._model.fit(self._x_train_dtm, self._y_train_dtm)
89     print('predicting with %s' % self._name)
90     self._prediction = self._model.predict(self._x_test_dtm)
91
92 ▼ def _fill_metrics(self, data_class):
93     try:
94         self._metrics[data_class] = {}
95         accuracy = float('%0.3f' % metrics.accuracy_score(self._y_test, self._prediction))
96         precision = float('%0.3f' % metrics.precision_score(self._y_test, self._prediction, average='weighted'))
97         self._metrics[data_class]['accuracy'] = accuracy
98         self._metrics[data_class]['precision'] = precision
99         y_test_list = list(self._y_test)
100        x_test_list = list(self._x_test)
101        self._false_classifications = []
102        for i in range(len(self._prediction)):
103            if y_test_list[i] != self._prediction[i]:
104                self._false_classifications.append((y_test_list[i], self._prediction[i], x_test_list[i]))
105        percentage_false_classifications = float('%0.3f' % (100.0 * len(self._false_classifications) / len(self._mails)))
106        self._metrics[data_class]['false_classified'] = percentage_false_classifications
107    except Exception as err:
108        print(err)

```

16. Time data

```

2 import datetime
3
4 import matplotlib.pyplot as plt
5
6
7 class TimeData(object):
8     _min_year = 1970
9     _max_year = 2005
10
11 def __init__(self, mails):
12     self._mails = mails
13
14 def graph(self):
15     frequency = {}
16     for mail in self._mails:
17         date = mail.get_meta_data('date')
18         if date is None:
19             continue
20         try:
21             if date.year < TimeData._min_year or date.year > TimeData._max_year:
22                 continue
23             if date.year not in frequency.keys():
24                 frequency[date.year] = 1
25             else:
26                 frequency[date.year] += 1
27         except:
28             pass
29     years = sorted(frequency.items())
30     plt.plot([year[0] for year in years], [year[1] for year in years])
31     plt.plot([year[0] for year in years], [year[1] for year in years], 'ro')
32     plt.ylabel('number of mails')
33     plt.xlabel('year')
34     plt.show()
35

```

17. K-means histogram graph

```

2 import numpy as np
3 import matplotlib.pyplot as plt
4
5
6 class KmeansHistogramGraph(object):
7     def __init__(self, mails):
8         self._mails = mails
9
10    def graph(self):
11        labels = list(set([mail.get_kmeans_label() for mail in self._mails]))
12        groups = {labels[i]: [] for i in range(len(labels))}
13        for mail in self._mails:
14            groups[mail.get_kmeans_label()].append(1)
15        totals = {key: len(groups[key]) for key in groups.keys()}
16        x = np.arange(len(labels))
17        counts = [totals[key] for key in totals.keys()]
18        tags = [key for key in totals.keys()]
19        fig, _ = plt.subplots()
20        plt.bar(x, counts)
21        plt.xticks(x, tags)
22        plt.show()
23

```

18. Regex histogram graph

```

2 import numpy as np
3 import matplotlib.pyplot as plt
4
5
6 class RegexHistogramGraph(object):
7     def __init__(self, mails):
8         self._mails = mails
9
10    def graph(self):
11        labels = list(set([mail.get_regex_label() for mail in self._mails]))
12        groups = {labels[i]: [] for i in range(len(labels))}
13        for mail in self._mails:
14            groups[mail.get_regex_label()].append(1)
15        totals = {key: len(groups[key]) for key in groups.keys()}
16        x = np.arange(len(labels))
17        counts = [totals[key] for key in totals.keys()]
18        tags = [key for key in totals.keys()]
19        fig, _ = plt.subplots()
20        plt.bar(x, counts)
21        plt.xticks(x, tags)
22        plt.show()
23

```

19. Main

```

2 import pickle
3 import threading
4
5 from datetime import datetime
6 from naive import NaiveBayes
7 from svm import SupportVectorMachine
8 from dt import DT
9 from logistic_regression import LogRegression
10
11 from time_data import TimeData
12
13 from regex_histogram_graph import RegexHistogramGraph
14 from kmeans_histogram_graph import KmeansHistogramGraph
15
16 from svm_graph import SvmGraph
17
18
19 class Runner(threading.Thread):
20     def __init__(self, task, model):
21         self._task = task
22         self._model = model
23         threading.Thread.__init__(self)
24
25     def run(self):
26         print('running thread\n')
27         self._task(self._model)
28         print('*** done ***')
29
30
31 def write_matrix(ml_model, file_name):
32     f_name = 'output/%s_%s.txt' % (file_name, datetime.now())
33     rf = open(f_name, 'wt')
34     rf.write('NaiveBays : %s\n' % ml_model.metrics())
35     rf.close()
36
37
38 def run_naive_bayes(ml_nb):
39     ml_nb.process()
40     write_matrix(ml_nb, 'naive_bayes')

```

```

41
42
43 ▼ def run_svm(ml_svm):
44     ml_svm.process()
45     write_metrix(ml_svm, 'svm')
46
47
48 ▼ def run_logistic_regression(ml_lg):
49     ml_lg.process()
50     write_metrix(ml_lg, 'logistic_regression')
51
52

```

```

58 ▼ def work(mails):
59     percentage = .05
60     subset_index = int(len(mails) * percentage // 1)
61
62     mails = mails[:subset_index]
63
64     print('running with %s mails\n' % len(mails))
65
66     log = LogRegression(mails)
67     lg = Runner(run_logistic_regression, log)
68
69     n_bayes = NaiveBayes(mails)
70     nb = Runner(run_naive_bayes, n_bayes)
71
72     svm = SupportVectorMachine(mails)
73     sv = Runner(run_svm, svm)
74

```

```

86     print(log.metrics())
87     print(n_bayes.metrics())
88     print(svm.metrics())

```

```

71
92 set_name = 'enron_large'
93 pickle_file = 'pickled/classified_%s.pkl' % set_name
94 f = open(pickle_file, 'rb')
95 mails = pickle.load(f)
96 f.close()
97
98 work_array = []
99
100 for i in range(1):
101     work_array.append(Runner(work, mails))
102 for t in work_array:
103     t.start()
104 ▼ for t in work_array:
105     t.join()
106

```

20. Test

```

2 import pickle
3 import numpy as np
4 import matplotlib.pyplot as plt
5
6 from logistic_regression import LogRegression
7
8 f = open('pickled/classified_enron_large.pkl', 'rb')
9 mails = pickle.load(f)
10 f.close()
11
12 svm = LogRegression(mails)
13 svm.process_kmeans()
14
15 print(svm.metrics())
16 print(svm.get_predictions())
17
18 xx = np.linspace(3, 9, 100)
19 yy = np.linspace(1, 5, 100).T
20
21 xx, yy = np.meshgrid(xx, yy)
22 xfull = np.c_[xx.ravel(), yy.ravel()]
23 y_pred = svm.get_predictions()
24 num_classes = np.unique(y_pred).size
25 probabilities = svm.calculate_probabilities(xfull)
26
27 print(probabilities)

```


Appendix D

Results from experimentation with the prototype to detect insider threats in a corporate email dataset

The section below includes a snapshot of a subset of 10 original emails that have traversed through the normalising, spelling correction and lemmatising phases. The emails were converted to lowercase before the snapshots were captured.

Table D.1: Original subset of 10 emails

Nr	Original Email
1	under separate cover, i will fax to you a copy of the february 6, 1998 legal opinion which, from my research, created the 'issue' between sg&e and enron north america corp., formerly enron capital & trade resources corp. it might be easiest if sd&g reissues the opinion without the qualification contained in paragraph 5. please let me know what you think. regards. sara
2	jeff: i think we may be past due on when we need to play by, but thats fine. i think saturday works for me as well. i know things are quite busy right now, so good luck with everything. ben -----original message----- from: donahue, jeff sent: monday, october 29, 2001 9:25 pm to: rogers, benjamin subject: re: tennis? things are a little busy now - when do we have to play by? i normally play in the morning but maybe next saturday would work. see ya. -----original message----- from: rogers, benjamin sent: monday, october 22, 2001 1:11 pm to: donahue, jeff subject: tennis? jeff: it looks like we are playing eachother in the met's consolation round. the best time i can play is either after work or some time on the weekend. let me know what works for you. ben
3	x-origin: lavorado-j x-filename: jlavora.pst executive committee weekly meeting date: monday, june 11th time: 11:00 a.m. (cdt) location: 50th floor boardroom video: connections will be established with remote locations upon request. conf call: at&t lines have been reserved. please contact sherri sera (713-853-5984) or bill dorsey (713-646-6505) for the weekly dail-in number and passcode. please indicate below whether or not you plan to attend this meeting and through what medium. yes, i will attend in person ----- by video conference from ----- by conference call ----- no, i will not attend ----- please return this e-mail to me with your response by 12:00 p.m., friday, june 8th. thank you, bill dorsey.

4	<p>x-origin: kaminski-v x-filename: vkamins.nsf martin, let me read it friday. we run our papers by our pr department to review for any potential conflict with the company line. i shall fwd it to them. i think you should submit it as an enron employee with a note that it was developed when you were at ut. vince from: martin lin on 04/02/2001 11:59 am to: vince j kaminski/hou/ect@ect cc: subject: publication submission question my supervising professors from ut and i are finishing a paper (finally!) that is based on work done for my phd. all of the research was done while i was a grad student. i have a couple of questions regarding submission of this paper (to iee transactions on power systems). 1. should i submit it with my affiliation as the university of texas at austin or as enron (ena, corp, etc)? 2. what legal or other reviews/clearances would i need? a draft of the paper for your review is attached. thanks, martin</p>
5	<p>from the birmingham sunday mercury (7th jan 2001) worker dead at desk for 5 days bosses of a publishing firm are trying to work out why no one noticed that one of their employees had been sitting dead at his desk for five days before anyone asked if he was feeling okay. george turklebaum, 51, who had been employed as a proof-reader at a new york firm for 30 years, had a heart attack in the open-plan office he shared with 23 other workers. he quietly passed away on monday, but nobody noticed until saturday morning when an office cleaner asked why he was still working during the weekend. his boss elliot wachiaski said "george was always the first guy in each morning and the last to leave at night, so no one found it unusual that he was in the same position all that time and didn't say anything. he was always absorbed in his work and kept much to himself. "a post mortem examination revealed that he had been dead for five days after suffering a coronary. ironically, george was proofreading manuscripts of medical textbooks when he died. you may want to give your co-workers a nudge or kick occasionally.</p>
6	<p>matt, attached is the entergy-koch/ena isda master agreement you requested. let me know if you have trouble accessing the document. stephanie http://edms.livelink.enron.com/ena/livelink.exe/96057022_base_contract_02_01_01_a.pdf?func=doc.fetch&nodeid=10191237&doctitle=96057022+base+contract+02%2f01%2f01+a</p>
7	<p>the approval status has changed on the following report: status last changed by: automated administrator expense report name: jeff shankman report total: \$14,591.49 amount due employee: \$14,591.49 amount approved: \$14,591.49 amount paid: \$0.00 approval status: pending payment status: pending to review this expense report, click on the following link for concur expense. http://expense.ms.enron.com</p>
8	<p>any conflicts? ----- forwarded by tana jones/hou/ect on 05/02/2001 09:52 am ----- mark s palmer/enron@enronxgate 05/02/2001 09:09 am to: tana jones/hou/ect@ect cc: subject: lehman nda can i come down and get this this am? thanks, mark p. input info for the nda ----- legal name of the company: lehman brothers inc. business contact name at the company (who will receive the nda): jarett wait business contact email & phone (& fax, if applicable): jwait@lehman.com company's address: 3 world financial center 10th floor new york, ny 10285 nature of discussion: project offline nda delivery method: hard-</p>

	<p>copy (return to mark s. palmer) other delivery needed? (hard copy sent by mail, fedex, fax): no legal contact at the company (optional but helpful):</p>
9	<p>you're probably right, sharen, since the counterparty continues to send confirmation letters with fpl energy power marketing instead of florida power & light. unfortunately, the marketer who contacts our trader to transact identifies himself as being with florida power & light, which is where i believe the confusion is coming from. i'll let the traders know of the correct counterparty name, though. thanks for the heads up, kate</p> <p>from: sharen cason 04/20/2001 02:55 pm to: kate symes/pdx/ect@ect cc: kimberly hundl/corp/enron@enron, amy smith/enron@enronxgate subject: #587574 can you check with the trader on this deal. we have had to change several deals input with this cp name. i believe it should be fpl energy power mktg. not florida power & light company. i think one is a utility which would be local and one is a marketer, who can trade anywhere. thanks!</p>
10	<p>paula, i am forwarding you the resume of marilyn mielke. she is a very highly educated and accomplished person. she has a background in science but no skills required in my group. i was impressed with her excellent communication and social skills - maybe she can help in your area. she was knocking on the enron's door for a long time. vince</p> <p>-----original message----- from: marilyn mielke <mmielke@bcm.tmc.edu>@enron sent: monday, october 29, 2001 11:08 am to: vkamins@enron.com subject: thank you. dear dr. kaminski, thank you so very much for your time on wednesday. it has been challenging to find out, from the outside, exactly what your group does on a daily basis and what type of people make up your team. i do not feel that there is any substitute for the information, and feel, that one can get directly from the source itself, thank you for making your team available to meet with me as well. you mentioned that the associate program at enron could be a great starting point for me to acquire the business skills and experience that i seek. i would be a most enthusiastic candidate as i have looked into the program, and i too, feel that it would be an exciting and logical starting point for me. it is my understanding that in order for me to enter the program i would require sponsorship. would you be able to sponsor me for this program? if so, i have attached my cv (as a word '97 document) and understand that the person to be contacted with a nomination would be mr jeff davis in human resources. please let me know if there is anything further that i can do to realize my goal of working within enron while gaining business skills and experience. it was a pleasure speaking with you. thank you again for your help and encouragement. sincerely, dr. marilyn mielke - mmielke_cv_sept01.doc</p>

Table D.2: Emails after normalisation

Nr	Normalisation
1	separate cover fax copy february 6 198 legal opinion research created issue sg e enron north america corp. formerly enron capital trade resources corp. might easiest sd g reissues opinion without qualification contained paragraph 5. please let know think . regards . sara
2	jef think may past due need play thats fine . think saturday works well . know things quite busy right good luck everything . ben original message donahue jef sent monday october 29 201 925 pm rogers benjamin subject tennis things little busy play normally play morning maybe next saturday would work . see ya . original message rogers benjamin sent monday october 22 201 11 pm donahue jef subject tennis jef looks like playing eachother met consolation round . best time play either work time weekend . let know works . ben
3	xorigin lavoradoj xfilename jlavora.pst executive committee weekly meeting date monday june 11th time 110 a.m. cdt location 50th floor boardroom video connections established remote locations upon request . conf call lines reserved . please contact sheri sera 7138535984 bill dorsey 713646505 weekly dailin number pascode . please indicate whether not plan attend meeting medium . yes attend person video conference conference call not attend please return email response 120 p.m. friday june 8th . thank bill dorsey .
4	xorigin kaminskiv xfilename vkamins.nsf martin let read friday . run papers pr department review potential conflict company line . shal fwd . think submit enron employee note developed ut . vince martin lin 040201 159 vince j kaminskihouect ect cc subject publication submission question supervising professors ut finishing paper finally based work done phd . research done grad student . couple questions regarding submission paper ie transactions power systems . 1. submit affiliation university texas austin enron ena corp etc 2. legal reviewsclearances would need draft paper review attached . thanks martin
5	birmingham sunday mercury 7th jan 201 worker dead desk 5 days bosses publishing firm trying work one noticed one employees sitting dead desk five days anyone asked feeling okay . george turklebaum 51 employed proofreader new york firm 30 years heart attack openplan office shared 23 workers . quietly passed away monday nobody noticed saturday morning office cleaner asked still working weekend . boss eliot wachiaski said george always first guy morning last leave night one found unusual position time not say anything . always absorbed work kept much . post mortem examination revealed dead five days suffering coronary . ironically george proofreading manuscripts medical textbooks died . may want give coworkers nudge kick occasionally .
6	matt attached entergykochena isda master agreement requested . let know trouble accessing document . stephanie http edms.livelink.enron.comenallivelink.exe9605702basecontract020101a.pdf funcdoc.fetch nodeid10191237 doctitle9605702basecontract02 2f01 2f01a
7	approval status changed following report status last changed automated administrator expense report name jef shankman report total 14591.49 amount due employee 14591.49

	amount approved 14591.49 amount paid 0.0 approval status pending payment status pending review expense report click following link concur expense . http expensxms.enron.com
8	conflicts forwarded tana joneshouect 050201 0952 mark palmerenron enronxgate 050201 0909 tana joneshouect ect cc subject lehman nda come get thanks mark p. input info nda legal name company lehman brothers inc. business contact name company receive nda jaret wait business contact email phone fax applicable jwait lehman.com company address 3 world financial center 10th floor new york ny 10285 nature discussion project offline nda delivery method hardcopy return mark s. palmer delivery needed hard copy sent mail fedex fax legal contact company optional helpful
9	probably right sharen since counterparty continues send confirmation letters fpl energy power marketing instead florida power light . unfortunately marketer contacts trader transact identifies florida power light believe confusion coming . let traders know correct counterparty name though . thanks heads kate sharen cason 0420201 025 pm kate symespdxect ect cc kimberly hundlcorpenron enron amy smithenron enronxgate subject 587574 check trader deal . change several deals input cp name . believe fpl energy power mktg . not florida power light company . think one utility would local one marketer trade anywhere . thanks
10	paula forwarding resume marilyn mielke . highly educated accomplished person . background science skills required group . impressed excellent communication social skills maybe help area . knocking enron door long time . vince original message marilyn mielke mielke bcm.tmc.edu enron sent monday october 29 201 108 vkamins enron.com subject thank . dear dr. kaminski thank much time wednesday . challenging find outside exactly group daily basis type people make team . not feel substitute information feel one get directly source thank making team available meet well . mentioned associate program enron could great starting point acquire business skills experience seek . would enthusiastic candidate looked program feel would exciting logical starting point . understanding order enter program would require sponsorship . would able sponsor program attached cv word 97 document understand person contacted nomination would mr jef davis human resources . please let know anything realize goal working within enron gaining business skills experience . pleasure speaking . thank help encouragement . sincerely dr. marilyn mielke mielkecvsept01.doc

Table D.3: Emails after spelling correction

Nr	Spelling Correction
1	separate cover fax copy february 6 19 legal opinion research created issue sg e enrol north america corps formerly enrol capital trade resources corps might easiest sd g reissues opinion without qualification contained paragraph 5 please let know think . regards . sara

2	jet think may past due need play that fine . think saturday works well . know things quite busy right good luck everting . ben original message donatus jet sent monday october 29 20 25 pm rogers benjamin subject tennis things little busy play normally play morning maybe next saturday would work . see a . original message rogers benjamin sent monday october 22 20 11 pm donatus jet subject tennis jet looks like playing each other met consolation round . best time play either work time weekend . let know works . ben
3	origin lavoradoj filename flavor arp st executive committee weekly meeting date monday june 11th time 110 a.m. cat location 50th floor boardroom video connections established remote locations upon request . cone call lines reserved . please contact shari sera 7138535984 bill horsey 713646505 weekly sailing number cascade . please indicate wether not plan attend meeting medium . yes attend person video conference conference call not attend please return email response 120 pom a friday june 8th . thank bill horsey .
4	origin gamin skin filename kam inst nsf martin let read friday . run papers pr department review potential conflict company line . seal fed . think submit enrol employee note developed ut . vine martin lin 040201 15 vine j gamin skin of ect ect cc subject publication submission question supervising professors ut finishing paper finally based work done phd . research done grad student . couple questions regarding submission paper ie transactions power systems . 1 submit affiliation university texas austin enrol vena corp etc 2 legal reviews clearances would need draft paper review attached . thanks martin
5	birmingham sunday mercury 7th jan 20 worker dead desk 5 days bosses publishing firm trying work one noticed one employees sitting dead desk five days anemone masked feeling okay . george turtle baum 51 employed proofreader new york firm 30 years heart attack open plan office shared 23 workers . quietly passed away monday nobody noticed saturday morning office cleaner masked still working weekend . boss eliot wachiaski said george always first guy morning last leave night one found unusual position time not say anything . always absorbed work kept much . post mortem examination revealed dead five days suffering coronary . ironically george proofreading manuscripts medical textbooks died . may want give workers nudge kick occasionally .
6	matt attached energy koch vena sida master agreement requested . let know trouble accessing document . stephanie http eds olive link enrol come native link exec 60 70 base contract 200 a fpd f fund dock fetch nodeid10191237 doc title 60 70 base contract 2 20 2f01a
7	approval status changed following report status last changed automated administrator expense report name jet shankar report total 14591.49 amount due employee 14591.49 amount approved 14591.49 amount paid 20 approval status pending payment status pending review expense report click following link concur expense . http expensxms.enron.com
8	conflicts forwarded tana jones of ect 050201 52 mark palmer end on enronxgate 050201 90 tana jones of ect ect cc subject leman da come get thanks mark p input info da legal name company leman brothers inch business contact name company receive da caret wait business contact email phone fax applicable wait lehman.com company address 3 world

	financial center 10th floor new york ny 105 nature discussion project online da delivery method hardtop return mark st palmer delivery needed hard copy sent mail fever fax legal contact company optional helpful
9	probably right shares since counterpart continues send confirmation letters fal energy power marketing instead florida power light . unfortunately marketer contacts trader transact identifies florida power light believe confusion coming . let traders know correct counterpart name though . thanks heads kate shares mason 0420201 25 pm kate some sad sect ect cc kimberley hundlcorpenron enrol amy smitten ron enronxgate subject 587574 check trader deal . change several deals input up name . believe fal energy power moth . not florida power light company . think one utility would local one marketer trade anywhere . thanks
10	paul forwarding resume marilyn milk . highly educated accomplished person . background science skills required group . impressed excellent communication social skills maybe help area . knocking enrol door long time . vine original message marilyn milk milk bum atm cue du enrol sent monday october 29 20 108 vkamins enron.com subject thank . dear dr. babinski thank much time wednesday . challenging find outside exactly group daily basis type people make team . not feel substitute information feel one get directly source thank making team available meet well . mentioned associate program enrol cold great starting point acquire business skills experience seek . would enthusiastic candidate looked program feel would exciting logical starting point . understanding order enter program would require sponsorship . would able sponsor program attached cv word 97 document understand person contracted nomination would mr jet davis human resources . please let know anything realize goal working within enrol gaining business skills experience . pleasure speaking . thank help encouragement . sincerely dr. marilyn milk milk eck sept 1 doc

Table D.4 Emails after lemmatisation

Nr	Lemmatisation
1	separate cover fax copy february 6 19 legal opinion research create issue sg e enrol north america corp formerly enrol capital trade resource corp might easiest sd g reissue opinion without qualification contain paragraph 5 please let know think . regard . sara
2	jet think may past due need play that fine . think saturday work well . know thing quite busy right good luck evert . ben original message donatus jet send monday october 29 20 25 pm rogers benjamin subject tennis thing little busy play normally play morning maybe next saturday would work . see a . original message rogers benjamin send monday october 22 20 11 pm donatus jet subject tennis jet look like play each other meet consolation round . best time play either work time weekend . let know work . ben
3	origin lavoradoj filename flavor arp st executive committee weekly meet date monday june 11th time 110 a.m. cat location 50th floor boardroom video connection establish remote location upon request . cone call line reserve . please contact shari serum 7138535984 bill

	<p>horsey 713646505 weekly sail number cascade . please indicate wether not plan attend meet medium . yes attend person video conference conference call not attend please return email response 120 pom a friday june 8th . thank bill horsey .</p>
4	<p>origin gamin skin filename kam inst nsf martin let read friday . run paper pr department review potential conflict company line . seal feed . think submit enrol employee note develop ut . vine martin lin 040201 15 vine j gamin skin of ect ect cc subject publication submission question supervise professor ut finish paper finally base work do phd . research do grad student . couple question regard submission paper ie transaction power system . 1 submit affiliation university texas austin enrol vena corp etc 2 legal reviews clearances would need draft paper review attach . thank martin</p>
5	<p>birmingham sunday mercury 7th jan 20 worker dead desk 5 day bos publish firm try work one notice one employee sit dead desk five day anemone mask feel okay . george turtle baum 51 employ proofreader new york firm 30 year heart attack open plan office share 23 worker . quietly pas away monday nobody notice saturday morning office cleaner mask still work weekend . bos eliot wachiaski say george always first guy morning last leave night one find unusual position time not say anything . always absorb work keep much . post mortem examination reveal dead five day suffer coronary . ironically george proofread manuscript medical textbook die . may want give worker nudge kick occasionally .</p>
6	<p>matt attach energy koch vena sida master agreement request . let know trouble access document . stephanie http ed olive link enrol come native link exec 60 70 base contract 200 a fpd f fund dock fetch nodeid10191237 doc title 60 70 base contract 2 20 2f01a</p>
7	<p>approval status change follow report status last change automate administrator expense report name jet shankar report total 14591.49 amount due employee 14591.49 amount approve 14591.49 amount pay 20 approval status pending payment status pending review expense report click follow link concur expense . http expensxms.enron.com</p>
8	<p>conflict forward tana jones of ect 050201 52 mark palmer end on enronxgate 050201 90 tana jones of ect ect cc subject leman da come get thank mark p input info da legal name company leman brother inch business contact name company receive da caret wait business contact email phone fax applicable wait lehman.com company address 3 world financial center 10th floor new york ny 105 nature discussion project online da delivery method hardtop return mark st palmer delivery need hard copy send mail fever fax legal contact company optional helpful</p>
9	<p>probably right share since counterpart continue send confirmation letter fal energy power market instead florida power light . unfortunately marketer contact trader transact identify florida power light believe confusion come . let trader know correct counterpart name though . thank head kate share mason 0420201 25 pm kate some sad sect ect cc kimberley hundlcorpenron enrol amy smite ron enronxgate subject 587574 check trader deal . change several deal input up name . believe fal energy power moth . not florida power light company . think one utility would local one marketer trade anywhere . thank</p>

10	<p>paul forward resume marilyn milk . highly educate accomplish person . background science skill require group . impress excellent communication social skill maybe help area . knock enrol door long time . vine original message marilyn milk milk bum atm cue du enrol send monday october 29 20 108 vkamins enron.com subject thank . dear dr. babinski thank much time wednesday . challenge find outside exactly group daily basis type people make team . not feel substitute information feel one get directly source thank make team available meet well . mention associate program enrol cold great start point acquire business skill experience seek . would enthusiastic candidate look program feel would excite logical start point . understand order enter program would require sponsorship . would able sponsor program attach cv word 97 document understand person contract nomination would mr jet davis human resource . please let know anything realize goal work within enrol gain business skill experience . pleasure speak . thank help encouragement . sincerely dr. marilyn milk milk eck sept 1 doc</p>
----	---

Appendix E

Larger subsets of datasets labelled by the unsupervised algorithms

1. Link to dataset labelled by K-means clustering algorithm:

<https://drive.google.com/file/d/1tD8HoKnYxZPg38ttzWGThxGuWtNe7lqk/view?usp=sharing>

2. Link to dataset labelled by Regular Expression Pattern Matching algorithm:

<https://drive.google.com/file/d/1tD8HoKnYxZPg38ttzWGThxGuWtNe7lqk/view?usp=sharing>

Appendix F

The email in the dataset most similar to each centroid, as well as the similarity percentage

Table F.1: Larger Subset of Emails Most Similar to Centroids

Insider Threat Type	Centroid ID	Closest Email (Snippet from actual Email)	Similarity (%)
Insider Fraud	C01	friday burrito 20 think nobody care try miss couple payment . 20 bigger miss payment care . californianus 20 care potentially big test miss payment century hey ca 20 say tell truth . . john burka nin coo px 20 tell end phone conversation yesterday know gary 20 might friend leave anymore . inform john 20 friendship prize big trouble . situation like 20 rat 34 tick totally pathetic circumstance . hey john stick 20 not worry thing . go get this.20 let take inventory week crisis . electric 20 shortage crisis . center stage 3 alert zone thursday afternoon 20 early morning . natural gas ptyas you burn crisis . of natural gas seller determine p e might not secure 20 buyer . cash crisis . two major utility e ha vein liquid fund pay energy bill . mush three crisis together 20 get new word 1 gash like state motto eureka an 20 find exclaim gash . screw . 20 every one 's start feel pain . esp remain 20 class market participant inform p e payment 20 px credit direct acer customer bill esp due under 20 consolidate bill henceforth suspend . rate freeze end last 20 august logic go p e not need acknowledge px credit . 20 put serious breach contract kink esp utility relationship . 20 course go write hundred million dollar 20 short payment write billion dollar short payment . mean care financial well remain 20 competitor duc . let fold tent . california customer 20 not want choice energy provider not know . vein 20 easy cynical.20 commissioner carl wood also hand couple late christmas gift . h 20 rule of standard offer contract cold pay 67.45mwh regardless formula energy payment . of to buy wintertime natural gas market price production cost ar 20 easily range 200th . great idea . force 20 w 20 remain of capacity voluntarily shut time that capacity desperately need 20 know	70.15

	<p>pick mr. wood lot . week serve 20 fm radio talk show qed forum . hate admit carl 20 really nice person . not like disagree people basinal lynx decent . would rather gently tease people like mike florio 20 rather out handout disagree . figure deep inside the 20 misguide soul cross wire give wrong answer 20 every major policy question . carl start talk good place to get corn beef sandwich leave studio waste positively effusive delicatessen frequent orange county . mask carl jewish say yes grow 20 jewish section baltimore . carl confess not look jewish 20 talk like communist . story laugh th at chat away walk car radio station park lot 20 walk mine . mask carl think mark 20 twain quote use dance deregulation dead statement 20 say rumor deregulation death greatly exaggerate h 20 real ied moment utter word dais puck hear room fear someone go tag twain quote . 20 20 retie rogue panelist fm show . speak desire 20 corenoncore split default customer hereby small customer would core large customer would honore . duc 20 would require procure core not honore . the larger customer ability resource cold negotiate longer contract supplier duc . take herb platform tweak little get something look much like 20 moro bay principle release greater 10 connect load 20 customer duc default service.20 wrinkle turn plan however . want keep retain edge generate asset egg . p e hydra plant core procurement portfolio . not think customer honore group go 20 like . contribute revenue requirement retain plant 20 . want credit contribution . hand 20 p e ice transition corenoncore split bean assure utility company charge core honore class fo 20 carry strand cost under collection revenue today isle freeze rate debacle . saw cut ways.20 speak saw governor simply one bang job state head state message last monday night . semiprofessional talk 20 head give u quote line four news reporter prior the address standby opinion governor speech . waste diplomatic sort press . let tell really think . 20 stink joint . cold not believe nervous the delivery . understand slur word mispronunciation oral two . gray grim . rock and role gut his comment sound like mr . rogers . say eminent 20 domain say outset not go assign blame point 20 finger . little hopeful would conciliatory tone . 20 not . rant rave forty minute half bash 20 folk well folk . sell power state . 20 wrong people portland base vista lunda hamilton rot 20 note aside mask 1 quite entertain governor . 20 somehow yet put together intend solve problem . well not get either lunda . mystery.20 okay gash head . let get on.20 thing people republic california px writ it dc</p>	
--	---	--

Negligence	C02	<p>morton 20 software suite offer 5 product 3d 300 value get 5 ... image free ship 29 image image image image image 29 powerful solution protect computer terrific suite proud act morton combine world 1 utility suite advance tool f pc expert . take look get image morton antiviral 200 1 morton antiviral 20 world trust antiviral solution . n of repair common virus infection automatically without interrupt work . scan clean incoming outgo email defend script base virus love even virus define ilion update . protect pc today award din morton antiviral window 9598me20 suggest retail price 49.95 image morton ghost 20 powerful solution system up grade backup recovery . north ghost 20 provide high performance utility e fast safe system up grade backup recovery . write disk image directly many po polar ccrc dry drive make easy back valuable data.files add previously create image elimination g need recline entire disk back new content . fast ectopic clone use high speed parallel u home network ip connection provide versatile sir ect connectivity . support linux ex microsoft pc file system in cloud fat 16 fat 32 ntis let clone elder newer system . multiple clone method allow choose st approach situation . automatic size destination partition streamline clone process . g disk provide command line partition functionality disk . g disk gallows completely erase hard drive safe thorough wipe . create notable diskette ca n include driver network card irritable cd drive u port . fly bible norris interface give precise control clone image process . build error check image comparison ensure the store image exactly duplicate original disk . password protect secure store image . window 9598me20 suggest retail price e 69.95 image morton utility e 20 improve pc performance . speed disk optimize hard drive put mostneded file sea r front disk faster access . significantly speed load time application document . subsequent optimization n even faster easy keep hard drive work efficient y fix window problem . morton inductor diagnose solve wide run ge window problem software error hardware configuration cone list . protect work fix problem might otherwise lead data loss help window run better clean problem cu r everyday use computer . keep hard drive healthy . r hard drive develop problem make pc run poorly even dama ge valuable data . morton disk doctor detect repair variety hard drive problem . not check disk physical sur face also</p>	61.40

		<p>director re . stop trouble start . run morton system doctor continuously background keep pc work shoot fly every minute . morton system doctor find potential disk sus tem problem many case automatically take proactive measure prevent minor problem become serious . morton protect recy le bin help recover file accidentally deplete . complete erase unwanted file . want make sure file erase really g onespecialy one contain confidential information wipe info perca gently delete content select file folder hard dr ice . use peace mind give old pc hard drive family member charity . window 9598me20 suggest retail price 49.95 image morton cleanser 20 tire hard drive c cluster web page curl graphic cooky active control plain down load program leave internet session internet sweep feature remove unwanted file safely easily without cause problem program . clean interne buildup award din morton cleanser semantic . improve pc performance re move unwanted program file wa st disk space while protect accidentally de let important fi lens . remove unwanted file computer click mouse . safe easy way clean computer gain back disk space improve performance . trust morton cleanser 20 safe easy complex te hard drive cleanup . window 9598me20 suggest retail price 99 5 image morton infix pro 20 basic edition infix lead fax management software small business bestseling fax of ware product june 19 june 20 high quality fax help project professional image client customer . image car city vital generate photo quality fax . even send fo ward fax via email people fax hardware software . r equine free viewer download able semantic web site . infix pro let entire work group send receive high quality fax without buy additional modem phone line . infix pro easy set learn use . integrate smoothly key business application simplify management client customer interaction . window 9598me20 0 suggest retail price 49.95 29 limit time get great suite 5 product 300 value ... image free ship apply standard ship order limit stock hand expire 103 101 29 image</p>	
Insider IT Sabotage	C03	<p>thank update chris . hear joe park take job ridgeline trade name . also hear greg shock look leave . think put ask job opportunity e . oh well ... talk later susan original message christ germany enron.com pereirae houston rico m send thursday april 11 22 82 subject hey know estate team completely separate u . 6 4 not work anemone u however go lunch work wife judy 3 time week . move 6 old build next week</p>	70.05

	<p>. go move next week 5 week think really go time . bath kelly estate team might work u update curve . estate team folk leave ed mcmiachael bath kelly louis carlo side bridge ruth concavo maria gaza robin bare troy genet bosie paul gregory phil polska chris figure of yuan tin victoria verse susan scott ... think that . eric boat leave last week midland work guy new business buy sell royalty secure ng lease think . joe park leave last week . bob hall announce quit last week well . go stay home awhile . oldest daughter start college national merit scholar . say really help tuition . judy fine . every one 's u play name . come back full time not like much . little work . chat later chris original message pereirae pereirae houston rico m enrol send tuesday april 29 22 43 pm germany chris subject hey bad news robin . hope start feel better . get older stuff bird hope evert go well birth . please boy girl sure . andrew christopher nice name . way sister judy not judith dad responsible one . call scott h cm job might good match . last email say go take time do energyusa . work estate team not think cold go back . know horrible thing might say scott seal lie incredibly spineless tell not part team back december not think cold stomach interact level different floor right probably still see u people though . judy not hear long time . email funny joke back never hear anything else . well get ta go . take care say hello ingrid . susan original message christ germany enron.com pereirae houston rico m send tuesday april 29 22 145 pm subject hey tell robin give curt name number . not know follow not . estate team . back surgery 3 4 week ago . rid car one day leg go numb . go couple specialist say need surgery right away . herniated disc disk back . aorta like leg never go numb hurt like crazy numbness . start work half day week . ingrid pretty good . problem far due date may 7th . man hang move together late february seem go well . not want know sex not sure male female . think little girl adorable smart enough realize female pain rear end . child mother exception . want last name baby middle name . not sound good check . think go andrew timer christopher timer . always want christopher plain of chris . ed michael keep say not enough people need do . hire chris cigar of work fred laurasia craig back contract . not know feel might something consider . number 71 38 37 57 . take care keep touch . keep post hear anything . chance hear though . original message pereirae pereirae houston rico m enrol send monday april 8 22 106 pm germany chris subject hey hi chris good hear . thank think . saw cm ad sunday paper position . send resume response ad back december hr folk call early january interview . situation</p>	
--	---	--

		<p>must change hr lady act like not really position fill . discover hr people really lame nicest thing say . call curt liza however think qualify trade ne physical . probably better suit gulf coast . talk reliant may gulf coast position available soon . not thrill people know duke may gulf coast position available someone recently quit however rebato work sol . good interview late january not spot . since rebato start origination group . time evert guess rebato work eric gonzales formerly enrol fame fortune . day old buddy . robin might interest cm job . still work estate team enronubs not speak since december . little contact anemone . last email go unanswered decide give . good luck baby due date know excite . real change lifestyle not negative one believe . take care susan original message germany chris christ germany enron.com pereirae houston rico m send monday april 8 22 105 subject hey susan get email address scott godel . hope thing go well not know work stay home may non event . cm energy try develop presence north east . headhunter set interview cm last friday . speak curt liza vice president wholesale gas trade . not know agree see 30 min interview say not look . want someone contact cash trade experience . say hop look someone like year . never know . number 7132307205 . let know thing go . focus live sin pregnant girlfriend . later chris email property enrol corp candor relevant affiliate may contain confidential privilege material sole use intend recipient . review use distribution disclosure other strictly prohibit . not intend recipient authorize receive recipient please contact sender reply enrol corp nero name age administration enron.com delete copy message . email attachment hereto not intend offer acceptance not create evidence bind enforceable contract enrol corp affiliate intend recipient party may not rely anemone basis contract estoppel otherwise . thank .</p>	
<p>Insider Intellectual Property (IP) Theft</p>	<p>C04</p>	<p>Sabxvada01 content type explain charles iso8591 content transfer encode bite today edition daily update frequent contributor rick cayman cha explain balance sheet shenanigan first get start head culture investor like react . also wave sherman look performance one marvelous chip maker communication chip sector general . muller investor member like read ovenware system opv report morgan stanley download free charge provide sign firm free research trial . link directly research page feature synopsis broker free report also register morgan stanley trial click morgan stanley logo</p>	<p>52.04</p>

	<p> http wpm ulex investor com article asp dock 65 89 nd0131 receive mail register muller investor . subscribe see bottom message . sponsor get w tax online h r block . let program select form math . answer simple question . fast easy accurate . http wpm ulex investor comp wasp idol mind 110 21 investment idea broker third party research online advice chat free sponsor report investment idea 1 investment idea elongate financial engineer bubble finally burst rick cayman cha think safe invest elongate roil financial water investor barely recover hangover cause myriad cause include doctor bubble sept 11th recession face much bigger challenge loss faith core underpin king invest decision financial report . elongate apt name like watergate white watergate scandal problem initially appease d isolate spread like cancer presidency encompass much larger universe player . every day another company announce restate operate result . trustworthiness report process account profession call question . click read . http wpm ulex investor com article asp dock de 58 nd0131 2 investment idea marvell set sight broadloom brim broadloom still dominate communication chip sector marvell mail grow torrid clip . wave sherman equity research columnist firm able grow quickly broadloom brim . communication chip maker saw sale rise 42 million 17 1 billion 20 investor get early make huge profit share broadloom rise twentyfold spring 19 ipo 200 peak . broadloom look maintain base revenue communication chip sector experience cyclical lull previously obscure competitor manage grow downturn set sight industry number one spot . marvell technology group mail grow torrid clip already surpass industry next largest player pac sierra pecs . click read . http wpm ulex investor com article asp dock 65 92 nd0131 sponsor pay much auto insurance take test drive insurance.com america best know carrier compete help save . get obligation quote . compare policy . font pay . seek best policy money insurance.com http wpm ulex investor comp wasp idol mind 110 22 broker third party report 1 investor choice day favorite bear stearns publish top ten list internet security prediction 22 page general sector commentary firm discus list 3 prediction call sophisticate virus attack 6 prediction biometric hype fizzle 10 predict security vendor stock outperform . page report purchase 10 http wpm ulex investor comedown load wasp dock 25 96 46 nd0131 2 today special report leman brother comment 22 growth potential telecom equip . vendor china . leman discus </p>	
--	--	--

Non-Malicious	C05	<p>“today edition daily update muller investor director investment research marc epstein explain math theory behind stock price explain investor least know basic buy . also equity research columnist ben marlin discuss pair wireless telecom stock may potential bargain . also today feature couple broker report network stock well report morgan stanley network associate beta reader access free charge register firm free research trial . see . setscrew center help novice well experience investor develop investment idea . click see http wpm ulex investor comb page wasp target stock advisor home dock 50 9 nd0123 receive mail register muller investor”</p>	59.87
---------------	-----	--	-------

Appendix G

Wordlists for each insider threat type for Regular Expression Pattern Matching algorithm

Table G.1: Wordlists for Each Insider Threat Type

Insider Threat Type	Wordlist
Insider Fraud	.*\\ssend stuff house email\\s.* .*\\sadvocate\\s.* .*\\srelation\\s.* .*\\slegal\\s.* .*\\sdispute\\s.* .*\\sfederal court\\s.* .*\\slawyer\\s.* .*\\slawsuit\\s.* .*\\ssoon possible\\s.* .*\\sproject stanley\\s.* .*\\senergy change\\s.* .*\\smanipulate\\s.* .*\\senronemissions\\s.* .*\\sdig\\s.* .*\\sillegally\\s.* .*\\soutage\\s.* .*\\ssign without chance negotiate\\s.* .*\\smanipulate energy price\\s.* .*\\sillegal\\s.*

	<p>.*\sfail\s.*</p> <p>.*\sdeal\s.*</p> <p>.*\snasty\s.*</p> <p>.*\sconspiracy incompetence\s.*</p> <p>.*\senergy price manipulate\s.*</p> <p>.*\smay not illegal\s.*</p> <p>.*\swhole matter whether action violate law\s.*</p> <p>.*\sdecline say whether consult advance\s.*</p> <p>.*\slegal conflict\s.*</p>
Negligence	<p>.*\snot adhere policy\s.*</p> <p>.*\swrite warn\s.*</p> <p>.*\snegligence\s.*</p> <p>.*\snot follow policy\s.*</p> <p>.*\scompany information phone\s.*</p> <p>.*\sfree\s.*</p> <p>.*\sprize\s.*</p> <p>.*\swin\s.*</p> <p>.*\sreward\s.*</p> <p>.*\spayment\s.*</p> <p>.*\smiss\s.*</p> <p>.*\se-mail not spam\s.*</p> <p>.*\sverify registration detail\s.*</p> <p>.*\strip lifetime\s.*</p> <p>.*\sgreat additional prize\s.*</p> <p>.*\sfind cool site\s.*</p> <p>.*\snb\s.*</p> <p>.*\surgent\s.*</p>

	<p>.*\sattention\s.*</p> <p>.*\simmediate\s.*</p> <p>.*\shelp\s.*</p> <p>.*\scongratulation\s.*</p> <p>.*\saccount\s.*</p> <p>.*\ssuspend\s.*</p> <p>.*\slog\s.*</p> <p>.*\sreceive virus\s.*</p> <p>.*\sdelete not open email\s.*</p> <p>.*\scommission pay daily !!!\s.*</p> <p>.*\sgrow faster microsoft\s.*</p> <p>.*\saggie virus\s.*</p> <p>.*\schange password\s.*</p> <p>.*\sclick\s.*</p> <p>.*\schange [a-zA-Z]*password\s.*</p> <p>.*\scomplete follow\s.*</p> <p>.*\snot joke\s.*</p>
<p>Insider IT Sabotage</p>	<p>.*\sanxious\s.*</p> <p>.*\snot happy\s.*</p> <p>.*\snot complete security train\s.*</p> <p>.*\sincomplete work\s.*</p> <p>.*\sincomplete train\s.*</p> <p>.*\spay raise\s.*</p> <p>.*\ssalary increase\s.*</p> <p>.*\snot get promote\s.*</p> <p>.*\swant resign\s.*</p> <p>.*\swant leave\s.*</p>

.*\someone else get position\s.*
.*\shard work not recognize\s.*
.*\sangry\s.*
.*\sfrustrate\s.*
.*\sfrustration\s.*
.*\supset\s.*
.*\sirritate\s.*
.*\sfeed\s.*
.*\slope cool\s.*
.*\sembarrassment\s.*
.*\sprofessional manner\s.*
.*\sattach resume\s.*
.*\sinternet job board\s.*
.*\sgive\s.*
.*\sgive\s.*
.*\snot value\s.*
.*\sunappreciated\s.*
.*\swaste effort\s.*
.*\suncertain future\s.*
.*\sfail\s.*
.*\sperformance relation issue\s.*
.*\sfeel rather ill\s.*
.*\sill today\s.*
.*\sput resume internet job board ?\s.*
.*\swant resume\s.*
.*\sknow anxious frustrate time\s.*
.*\sembarrassment\s.*

	<p>.*\snot work\s.*</p> <p>.*\sorder\s.*</p> <p>.*\spatronize\s.*</p> <p>.*\shate job\s.*</p>
<p>Insider Intellectual Property (IP) Theft</p>	<p>.*\sworry take care\s.*</p> <p>.*\ssplit difference\s.*</p> <p>.*\smoney go\s.*</p> <p>.*\sembezzle account\s.*</p>

BIBLIOGRAPHY

- Ackerman, D. & Mehrpouyan, H., 2016. *Modeling Human Behavior to Anticipate Insider Attacks via System Dynamics*. Pasadena, CA, 2016 Symposium on Theory of Modeling and Simulation (TMS-DEVS).
- Aery, M. & Chakravarthy, S., 2005. *eMailSift: Email Classification Based on Structure and Content*. s.l., IEEE Computer Society, pp. 1-8.
- Agarwal, A., Omuya, A., Zhang, J. & Ranbow, O., 2014. *Enron Corporation: You're the Boss if People get Mentioned to You*. s.l., Social Com '14.
- Alawneh, M. & Abbadi, I. M., 2011. *Defining and Analyzing Insiders and their Threats in Organizations*. s.l., IEEE Computer Society, pp. 785-794.
- Ali, Z., 2018. *Insider Threats – 2018 Statistics*. [Online] Available at: <https://www.uscybersecurity.net/insider-threats-2018-statistics/> [Accessed 14 March 2019].
- Alkhereyf, S. & Rambow, O., 2017. *Proceedings of TextGraphs-11: The Workshop on Graph-based Methods for Natural Language Processing*. Vancouver, Canada, Association for Computational Linguistics, pp. 57-65.
- Alsmadi, I. & Alhami, I., 2015. Clustering and classification of email contents. *Journal of King Saud University – Computer and Information Sciences*, 27, pp. 46-57.
- Altman, W., 2003. What went wrong at Enron?. *Engineering Management*, 12(6), pp. 251-254.
- amaBhungane & Scorpio, 2017. *#GuptaLeaks*. [Online] Available at: <http://amabhungane.co.za/article/2017-11-10-guptaleaks-released-to-journalists-worldwide>
- Anon., 2015. *Information Security Breaches Survey: Technical Report*. [Online] Available at:

https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/432412/bis-15-302-information_security_breaches_survey_2015-full-report.pdf

Anon., 2019. *Help Net Security*. [Online] Available at: <https://www.helpnetsecurity.com/2019/04/05/detect-insider-threats/>

Aski, A. S. & Sourati, N. K., 2016. Proposed efficient algorithm to filter spam using machine learning techniques. *Pacific Science Review A: Natural Science and Engineering*, 18, pp. 145-149.

Azaria, A., Richardson, A., Kraus, S. & Subrahmanian, V. S., 2014. Behavioral analysis of insider threat: A survey and bootstrapped prediction in imbalanced data. *IEEE Transactions on Computational Social Systems*, 1(2), pp. 135-155.

Bell, A. J., Rogers, M. B. & Pearce, J. M., 2019. The insider threat: Behavioral indicators and factors influencing likelihood of intervention. *International Journal of Critical Infrastructure Protection*, 24, pp. 166-176.

Bishop, M., Gollman, D., Hunker, J. & Probst, C. W., 2008. *Countering Insider Threats*. Santa Monica, California, RAND Corp., pp. 1-18.

Bosworth, S., Kabay, M. E. & Whyne, E., 2009. *Computer Security Handbook*. New Jersey: Wiley (5th edition).

Brown, C. R., Watkins, A. & Greitzer, F. L., 2013. Predicting insider threat risks through linguistic analysis of electronic communication. *46th Hawaii Int. Conf. Syst. Sci*, pp. 1849-1858.

Butkovic, A., Mrdovic, S. & Mujacic, S., 2013. IP geolocation suspicious email messages. *21st Telecommunications Forum, TELFOR 2013*, pp. 881-884.

Cappelli, D., Moore, A. & Trzeciak, R., 2012. *The CERT Guide to Insider Threats: How to Prevent, Detect, and Respond to Information Technology Crimes (Theft, Sabotage, Fraud)*. s.l.:Addison-Wesley Professional.

Chi, H., Prodanoff, Z. G., Scarllet, C. & Hubbard, D., 2016. Determining predisposition to insider threat activities by using text analysis. *Future Technologies Conference*, pp. 985-990.

- Chinchani, R., Ngo, I. H. & Upadhyaya, S., 2005. *Towards a Theory of Insider Threat Assessment*. Washington, DC, IEEE Computer Society.
- Clark, J. W., 2016. *Threat from within: Case Studies of Insiders who Committed Information Technology Sabotage*. Austria.
- Claycomb, W. R., Huth, C. L., Phillips, B., Flynn, L., & McIntire, D., 2013. Identifying indicators of insider threats: Insider IT Sabotage. *IEEE*.
- Cohen, A., Nissim, N. & Elovici, Y., 2018. Novel set of general descriptive features for enhanced detection of malicious emails using machine learning methods. *Expert Systems with Applications*, 110, pp. 143-169.
- Costa, D., 2017. *CERT Definition of 'Insider Threat' - Updated*. [Online] Available at: <https://insights.sei.cmu.edu/insider-threat/2017/03/cert-definition-of-insider-threat---updated.html>
- Cukierski, W., 2015. *The Enron Email Dataset*. [Online] Available at: <https://www.kaggle.com/wcukierski/enron-email-dataset> [Accessed 18 January 2018].
- Da Veiga, A., 2016. *A Cybersecurity Culture Research Philosophy and Approach to Develop a Valid and Reliable Measuring Instrument*. s.l., SAI Computing Conference London UK.
- Dasgupta, T. & Dey, L., 2016. *Enterprise Risk Analytics: Automatic Analysis of Risk Factors from Textual Feedbacks*.
- Duc, L. C. & Zincir-Heywood, A. N., 2019. *Machine Learning-Based Insider Threat Modelling and Detection*. Arlington, VA, USA, USA. IEEE.
- Furnell, S., 2004. Enemies within: The problem of insider attacks. *Computer Fraud & Security*, Volume 7, pp. 6-11.
- Greitzer, F. L., Kangas, L. J. & Noonan, C. F., 2012. *Identifying At-risk Employees: Modeling Psychosocial Precursors of Potential Insider Threats*. Hawaii, 45th Hawaii International Conference on System Sciences.

Greitzer, F. L., Moore, A. P., Cappelli, D. M., Andrews, D. H., Carroll, L. A., & Hull, T. D., 2008. Combating the insider cyber threat. *IEEE Security & Privacy*, 6(1), pp. 61-64.

Greitzer, F. L., Purl, J., Leong, M. Y. & Sticha, P. J., 2019. Positioning your organization to respond to insider threats. *IEEE Engineering Management Review*, June, 47(2), pp. 75-82.

Gural, N., 2013. *Banks like Goldman Sachs are going too far cyber-spying on their employees.* [Online]

Available at: <https://news.efinancialcareers.com/us-en/150631/banks-are-going-too-far-cyber-spying-on-their-employees>

HaCohen-Kerner, Y., Miller, D. & Yigal, Y., 2020. The influence of preprocessing on text classification using a bag-of-words representation. *PLoS One*, 15(5).

Homoliak, I., Toffalini, F., Guarnizo, J., Elovici, Y., & Ochoa, M., 2018. Insight into insiders and IT: A survey of insider threat taxonomies, analysis, modeling, and countermeasures. *ACM Computing Surveys*, 99(99), pp. 2-54.

Hunker, J. & Probst, C. W., 2011. Insiders and insider threats: An overview of definitions and mitigation techniques. *Journal of Wireless Mobile Networks, Ubiquitous Computing and Dependable Applications*, pp. 4-27.

Hussain, R. & Qamar, U., 2014. An approach to detect spam emails by using majority voting. *Proceedings of the International Conference on Data Mining, Internet Computing, and Big Data*, Kuala Lumpur, Malaysia, pp. 76-83.

IBM, 2019. *IBM.* [Online]
Available at: <https://www.ibm.com/za-en/>
[Accessed 2019].

ISO, 2012. *ISO/IEC 27032:2012(en) Information technology — Security techniques — Guidelines for cybersecurity.* [Online]
Available at: <https://www.iso.org/obp/ui/#iso:std:iso-iec:27032:ed-1:v1:en>
[Accessed 07 July 2019].

ISO, 2019. *ISO 704:2009 Terminology work — Principles and Methods*. [Online] Available at: <https://www.iso.org/standard/38109.html>

Jiang, J., Chen, J., Choo, K.-K. R., Liu, K., Liu, C., Yu, M., & Mohapatra, P., 2018. Prediction and detection of malicious insiders' motivation based on sentiment profile on webpages and emails. *Milcom 2018 Track 3 - Cyber Security and Trusted Computing*, pp. 225-230.

Keeney, M., Kowalski, E., Cappelli, D., Moore, A., Shimeall, T., & Rogers, S., 2005. *Insider Threat Study: Computer System Sabotage in Critical Infrastructure Sectors*, Pittsburgh: US Secret Service and CERT Coordination Center/SE.

Kowalski, E., Cappelli, D. & Moore, A., 2008. US Secret Service and CERT/SEI Insider Threat Study: Illicit cyber activity in the Information Technology and Telecommunications Sector. *US Secret Service and CERT Program Software Engineering Institute*.

Leber, J., 2013. *The Immortal Life of the Enron Emails*. [Online] Available at: <https://www.technologyreview.com/s/515801/the-immortal-life-of-the-enron-e-mails/> [Accessed 7 February 2018].

Leber, J., 2018. *The Immortal Life of the Enron Emails*. [Online] Available at: <https://www.technologyreview.com/s/515801/the-immortal-life-of-the-enron-e-mails/> [Accessed 7 February 2018].

Lepinsky, R., 2013. *Analyzing Keywords in Enron's Email*. [Online] Available at: Rodger's Notes: <https://rodgersnotes.wordpress.com/2013/11/24/analyzing-keywords-in-enrons-email/> [Accessed 18 January 2019].

Liu, L. et al., 2019. Detecting and preventing cyber insider threats: A survey. *IEEE Communications Surveys & Tutorials*, 20(2), Second Quarter 2018, pp. 1397-1417.

Maasberg, M., Warren, J. & Beebe, N. L., 2015. *The Dark Side of the Insider: Detecting the Insider Threat Through Examination of Dark Triad Personality Traits*. Hawaii, 48th Hawaii International Conference on System Sciences.

- Mayhew, M., Atighetchi, Adler, A. & Greenstadt, R., 2015. Use of machine learning in big data analytics for insider threat detection. *Milcom 2015 Track 3 - Cyber Security and Trusted Computing*, pp. 915-921.
- Media Temple, 2020. *Understanding an Email Header*. [Online] Available at: <https://mediatemple.net/community/> [Accessed 03 May 2020].
- Merriam-Webster, 2019. *Merriam-Webster*. [Online] Available at: <https://www.merriam-webster.com/dictionary/insider> [Accessed March 2019].
- Michael, A. & Eloff, J. H., 2019. *A Machine Learning Approach to Detect Insider Threats in Emails Caused by Human Behaviours*. Nicosia, Cyprus, HAISA, pp. 34-49.
- Michael, A. & Eloff, J., 2020. Discovering “Insider IT Sabotage” based on human behaviour. *Information and Computer Security*.
- Mills, J. U., Stuban, S. M. & Dever, J., 2017. Predict insider threats using human behaviours. *IEEE Engineering Management Review*, 45(1), pp. 39-48.
- Moore, A. P., Cappelli, D. M., Caron, T. C., Shaw, E., & Trzeciak, R. F., 2009. *Insider Theft of Intellectual Property for Business Advantage: A Preliminary Model*. West Lafayette, 1st International Workshop on Managing Insider Security Threats (MIST 2009), Purdue University.
- Mujtaba, G., Shuid, L., Raj, G. R., Majeed, N., & Al-Garadi, M. A., 2017. Email classification research trends. *IEEE Access*, pp. 9044-9064.
- Munshi, A., Dell, P. & Armstrong, H., 2012. Insider threat behavior factors: A comparison of theory with reported incidents. *45th Hawaii International Conference on System Sciences*, pp. 2402-2411.
- NBC News, 2004. *Former Enron Executive Pleades Guilty*. [Online] Available at: <http://www.nbcnews.com/id/5020783/ns/business->

[corporate scandals/t/former-enron-executive-pleads-guilty/#.XouhCogzY2w](https://www.foxnews.com/corporate-scandals/t/former-enron-executive-pleads-guilty/#.XouhCogzY2w)

[Accessed 06 April 2020].

NIST, 2015. *Computer Security Resource Centre*. [Online]

Available at: <https://csrc.nist.gov/glossary/term/Cyber-Threat>

NIST, 2019. *Information Technology Laboratory, Computer Security Resource Center*. [Online]

Available at: <https://csrc.nist.gov/glossary/term/insider-threat>

Nizamani, S., Memon, N., Glasdam, M. & Nguyen, D. D., 2014. Detection of fraudulent emails by employing advanced feature abundance. *Egyptian Informatics Journal*, pp. 169-174.

Noever, D. A., 2019. Classifier suites for insider threat detection. *Computing Research Repository*, Vol. abs/1901.10948.

Noever, D., 2020. The Enron Corpus: Where the Email Bodies are Buried? *ArXiv*, Vol. abs/2001.10374.

Nurse, J. R., Legg, P. A., Buckley, O., Agrafiotis, I., Wright, G., Whitty, M., Upton, D., Goldsmith, M., Creese, S., 2014. *A Critical Reflection on the Threat from Human Insiders – its Nature, Industry Perceptions, and Detection Approaches*. s.l., Springer, Cham, pp. 270-281.

Okolica, J., Peterson, G. & Mills, R., 2006. Using PLSI-U to detect insider threats by datamining email. *International Journal of Security and Networks*, 1(1/2/3), pp. 66-74.

Okolica, J., Peterson, G. & Mills, R., 2007. Using Author Topic to detect insider threats from email traffic. *Digital Investigation*, 4, pp. 158-164.

Oxford Advanced American Dictionary, 2020. *Threat*. [Online]

Available at:

https://www.oxfordlearnersdictionaries.com/definition/american_english/threat

Oxford, 2017. *English Oxford Living Dictionaries*. Oxford: Oxford University Press.

Pandas, 2019. *Python Data Analysis Library*. [Online] Available at: <https://pandas.pydata.org/>

[Accessed 12 March 2019].

- Patil, S., Jangra, A., Bhale, M., Raina, A., & Kulkarni, P., 2017. *Ethical Hacking: The Need for Cyber Security*. Chennai, India, IEEE, pp. 1602-1606.
- Pfleeger, C. P. & Pfleeger, S. L., 2012. *Analysing Computer Security*. Michigan: Pearson Education International.
- Probst, C. W., Hunker, J., Gollmann, D. & Bishop, M., 2010. Aspects of insider threats. *Insider Threats Cybersecurity*, pp. 1-15.
- PyCharm, 2020. *PyCharm*. [Online] Available at: <https://www.jetbrains.com/pycharm/> [Accessed 16 July 2020].
- Quinlan, J. R., 1996. Improved use of continuous attributes in C4.5. *Journal of Artificial Intelligence Research*, 4, pp. 77-90.
- Reid, R. & Van Niekerk, J., 2014. From information security to cyber security cultures. *IEEE 2014*.
- Robertson, T., 2017. *FINRA Continues Scrutiny of Financial Industry for Improper Electronic Communications*. [Online] Available at: <https://blogs.thomsonreuters.com/financial-risk/risk-management-compliance/finra-continues-scrutiny-financial-industry-improper-electronic-communications/>
- Rosen-Zvi, M., Griffiths, T., Steyvers, M. & Smyth, P., 2004. The author-topic model for authors and documents. *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*. s.l., s.n.
- SAIC & McAfee, 2011. *Underground Economies. Intellectual Capital and Sensitive Corporate Data now the latest Cybercrime Currency*. [Online] Available at: <http://freepdfs.net/rp-underground-economies-national-defense-industrial-association/dd8111e5875c655df5def936e6110948/>
- Salem, M. B., Hershkop, S. & Stolfo, S. J., 2008. *A Survey of Insider Attack Detection Research*, New York: s.n.

Sashikanth, D. B., 2015. *Analysis of Communication Patterns with Scammers in Enron Corpus*. [Online]

Available at: <https://arxiv.org/abs/1509.00705>
[Accessed 18 January 2018].

Schwartz, M., 2018. *Tesla Accuses Insider of Stealing Gigabytes of Data*. [Online]
Available at: <https://www.bankinfosecurity.com/tesla-lawsuit-alleges-insider-stole-gigabytes-data-a-11118>

[Accessed 01 October 2019].

scikit-learn developers, 2019. *1.9. Naive Bayes*. [Online]

Available at: https://scikit-learn.org/stable/modules/naive_bayes.html
[Accessed 14 July 2020].

Segal, T., 2019. *Enron Scandal: The Fall of a Wall Street Darling*. [Online]

Available at: <https://www.investopedia.com/updates/enron-scandal-summary/>
[Accessed 06 April 2020].

Shaw, E. D. & Fischer, L. F., 2005. Ten tales of betrayal: The threat to corporate infrastructures by information technology insiders. *Report 1 - Overview and General Observations*, Monterey, CA, Defense Personnel Security Research Center.

Shetty, J. & Adibi, J., 2004. *The Enron Email Dataset Database Schema and Brief Statistical Report*. s.l., s.n.

Spooner, D., Silowash, G., Costa, D. & Albrethsen, M., 2018. Navigating the insider threat tool landscape: Low-cost technical solutions to jump start an insider threat program. *2018 IEEE Symposium on Security and Privacy Workshops*, pp. 247-257.

Stojiljković, M., 2020. *Logistic Regression in Python*. [Online]

Available at: <https://realpython.com/logistic-regression-python/>
[Accessed 15 July 2020].

Tang, G., Pei, J. & Luk, W.-S., 2014. Email mining: Tasks, common techniques and tools. *Knowledge and Information Systems*, pp. 1-31.

Tribolet, M., 2016. *Investigating Enron's Email Corpus: The Trail of Tim Belden*. [Online] Available at: <https://linkurio.us/blog/investigating-the-enron-email-dataset/> [Accessed 18 January 2018].

Tsipenyuk, G. & Crowcroft, J., 2017. *An Email Attachment is Worth a Thousand Words, or is it?* New York, Association for Computing Machinery, pp. 1-10.

Van der Walt, E. & Eloff, J., 2018. Are attributes on social media platforms usable for assisting in the automatic detection of identity deception? *HAlSA 2018*, pp. 56-66. University, Abertay (Dundee): s.n.

Varone, M., Mayer, D. & Melegari, A., 2019. <https://www.expertsystem.com/machine-learning-definition/>. [Online] Available at: <https://www.expertsystem.com/machine-learning-definition/>

Verizon, 2019. *2019 Data Breach Investigations Report*. [Online] Available at: <https://enterprise.verizon.com/resources/executivebriefs/2019-dbir-executive-brief.pdf> [Accessed 20 January 2020].

Wall, D. S., 2013. Enemies within: Redefining the insider threat. *Security Journal*, pp. 107-124.

Wang, M., He, Y. & Jiang, M., 2010. *Text Categorization of Enron Email Corpus Based on Information Bottleneck and Maximal Entropy*. s.l., 2010 IEEE 10th International Conference.

Wang, X., Tan, Q., Shi, J., Su, S., & Wan, M., 2018. Insider threat detection using characterizing user behavior. *2018 IEEE Third International Conference on Data Science in Cyberspace*, pp. 476-482.

Warren, M., 2015. Modern IP theft and the insider threat. *Computer Fraud & Security*, 6, pp. 5-10.

White, J. & Panda, B., 2009. Implementing PII honeytokens to mitigate against the threat of malicious insiders. *IEEE*, p. 233.

Whitman, E., 2016. *Goldman Sachs Employee Email Surveillance: Which Terms Trigger Review Amid Concerns over Losses and Insider Trading?*. [Online]

Available at: <https://www.ibtimes.com/goldman-sachs-employee-email-surveillance-which-terms-trigger-review-amid-concerns-2383065>

[Accessed 16 February 2018].

Widup, S., 2018. *New Report Puts Healthcare Cybersecurity Back under the Microscope.*

[Online]

Available at: <https://securityboulevard.com/2018/03/employees-are-biggest-threat-to-healthcare-data-security/>

Wilson, G. & Banzhaf, W., 2009. *Discovery of Email Communication Networks from the Enron Corpus with a Genetic Algorithm using Social Network Analysis.* Trondheim, Norway, CEC 2009.

Young, W. T., Memory, A., Goldberg, H. G. & Senator, T. E., 2014. Detecting unknown insider threat scenarios. *2014 IEEE Security and Privacy Workshops*, pp. 277-288.

Zaki, T., Uddin, M. S., Hasan, M. M. & Islam, M. N., 2017. *Security Threats for Big Data: A Study on Enron E-mail Dataset.* Langkawi, Malaysia , 2017 International Conference on Research and Innovation in Information Systems (ICRIIS).

Zaytsev, A., Malyuk, A. & Miloslavskaya, N., 2017. Critical analysis in the research area of insider threats. *IEEE 5th International Conference on Future Internet of Things and Cloud*, pp. 288-296.