

# **Faculty of Information and Communication Technology**

# INSIDER THREATS DETECTION MODEL FOR EMAIL CONTENT USING STATISTICAL ANALYSIS

اونيونرسيتي تيكنيكل مليسيا ملاك UNIVERSITI TEKNIKAL MALAYSIA MELAKA

Nur Ameera Natasha Binti Mohammad

Master of Science in Information and Communication Technology

# INSIDER THREATS DETECTION MODEL FOR EMAIL CONTENT USING STATISTICAL ANALYSIS

## NUR AMEERA NATASHA BINTI MOHAMMAD

A thesis submitted in fulfillment of the requirements for the Master of Science in Information and Communication Technology



**Faculty of Information and Communication Technology** 

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

## **DECLARATION**

I declare that this thesis entitled "Insider Threats Detection Model For Email Content Using Statistical Analysis" is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature : Nur Ameera Natasha Binti Mohammad
Date : 1/3/2022

## **APPROVAL**

I hereby declare that I have read this thesis and in my opinion this thesis is sufficient in terms of scope and quality for the award of Master of Science in Information and Communication Technology.

MALAYSIA MALAYSIA	Wani
Signature S	:
Supervisor Name	: Dr. S.M Warusia Mohamed Bin S.M.M.
MINO	Yassin
Date	: 1/3/2022
UNIVERSITI TEKN	IKAL MALAYSIA MELAKA

## **DEDICATION**

To my beloved mother and father, Mohammad Bin Ismail and Enisabiren Binti Basir and those who have always by my side during my upside down, without whom none of my success would be possible.



#### **ABSTRACT**

An insider threat has become one of the most challenging malicious activities in cybersecurity defence systems in a contrast to outsider threats recently. Usually, IP theft, fraud and sabotage against legal information are three well-known types of insider threat. Since an insider threat usually expands and spread internally, no one could predict what, when and how exactly malicious insider launched their attacks. This is with a view of fact that an email becomes one of the primary targets of an internal threat as this medium is widely used by everyone to communicate, share, and exchange confidential information. Therefore, it is extremely important to understand the nature of insider threat behavior beforehand and construct an accurate detection model. Furthermore, every single keyword used in an email can reflect the behavior of an individual and can be used to determine their intentions, such as having a motive to launch an insider threat or not. Henceforth, an innovative approach is proposed in modelling insider threat detection in this work. In addition, various approaches such as scoring, Friedman, linear regression (R<sup>2</sup>) and correlation coefficient applied to analyse an insider threat relationship between historical insider threats behavior and relevant extracted keywords from email content. Firstly, the email content filtered into three different factors that influence the characteristics of an insider such as motive, opportunity and capability, before calculating the scores for the entire insider's keywords. Next, the Friedman statistical used to determine the minimum differences between each extracted insider threats keywords that represent different insider threat factors (motive, opportunity, capability). Besides, linear regression applied to estimate the relationship of an insider threat from training keywords and testing keywords with allocating an anomaly score. Finally, the correlation coefficient approach used to determine how strong a relationship is between extracted insider threats keywords and insider threat behavior in this research. The proposed modelling approach has been evaluated using the benchmark dataset known as CERT that comprises a malicious email file. Throughout the experiment, the proposed insider threats detection approach has achieved a higher attack detection rate as well as minimized undetectable insider threats behavior as compared to the previous researcher works.

## MODEL PENETAPAN ANCAMAN ORANG DALAM UNTUK KANDUNGAN E-MEL MENGGUNAKAN ANALISIS STATISTIK

#### **ABSTRAK**

Ancaman dalaman telah menjadi salah satu aktiviti berbahaya yang paling mencabar dalam sistem pertahanan keselamatan siber berbanding dengan ancaman luaran baru-baru ini. Biasanya, pencurian, penipuan dan sabotaj IP terhadap maklumat sulit adalah tiga jenis ancaman dalaman yang terkenal. Oleh kerana ancaman dalaman biasanya berkembang dan tersebar secara dalaman, tidak ada yang dapat meramalkan apa, bila dan bagaimana pelaku yang jahat melancarkan serangan mereka. Kenyataan ini dinyatakan bersama fakta bahawa e-mel menjadi salah satu sasaran utama ancaman dalaman kerana media ini digunakan secara meluas oleh setiap individu untuk berkomunikasi, berkongsi, dan bertukar maklumat sulit. Oleh itu, adalah sangat penting untuk memahami sifat tingkah laku ancaman dalaman terlebih dahulu dan membina model pengesanan yang tepat. Selanjutnya, setiap kata kunci yang digunakan dalam e-mel dapat menggambarkan tingkah laku individu dan dapat digunakan untuk menentukan niat mereka, iaitu melalui motif untuk melancarkan ancaman dalaman atau tidak. Sejauh ini, pendekatan inovatif dicadangkan dalam memodelkan pengesanan ancaman dalaman dalam karya ini. Di samping itu, pelbagai pendekatan iaitu pemarkahan, Friedman, regresi linear  $(R^2)$  dan pekali korelasi digunakan untuk menganalisis hubungan antara tingkah laku ancaman dalaman dengan kata kunci relevan yang diekstrak dari kandungan e-mel. Pertama, kandungan e-mel disaring menjadi tiga faktor berbeza yang mempengaruhi ciri ancaman dalaman seperti motif, peluang dan kemampuan, sebelum mengira skor untuk kata kunci keseluruhan ancaman dalaman. Seterusnya, statistik Friedman digunakan untuk menentukan perbezaan minimum antara setiap kata kunci ancaman dalaman yang diambil yang mewakili faktor ancaman yang berbeza (motif, peluang, kemampuan). Selain itu, regresi linear digunakan untuk menganggar hubungan ancaman dalaman dari kata kunci latihan dan kata kunci pengujian dengan memperuntukkan skor anomali. Akhirnya, pendekatan pekali korelasi digunakan untuk menentukan seberapa kuat hubungan antara kata kunci ancaman dalaman yang diekstrak dan tingkah laku ancaman dalaman melalui penyelidikan ini. Pendekatan pemodelan yang dicadangkan telah dinilai menggunakan set data penanda aras yang dikenali sebagai CERT yang terdiri daripada fail e-mel yang berniat jahat. Sepanjang eksperimen, pendekatan pengesanan ancaman dalaman yang dicadangkan telah mencapai kadar pengesanan serangan yang lebih tinggi serta meminimumkan tingkah laku ancaman dalaman yang tidak dapat dikesan dibandingkan dengan karya penyelidik sebelumnya.

#### ACKNOWLEDGMENTS

First and foremost, I would like to take this opportunity to express my sincere acknowledgement to my supervisor, Dr. S.M Warusia Mohamed Bin S.M.M Yassin and my co-supervisor, Professor Ts. Dr. Rabiah Binti Ahmad for all the supervision, support and suggestion towards the completion of this thesis and contributed to its technical content. I am also very grateful to authors of various articles on the internet who make me aware of the research currently ongoing in this field. Special thanks to my beloved parents, sister, family members and all my friends who have always been there to give support and helped me to complete this research work. Last but not least, I would like to thank everyone who had been to the crucial parts of the realization of this research.

اونيوسيتي تيكنيكل مليسيا ملاك UNIVERSITI TEKNIKAL MALAYSIA MELAKA

# TABLE OF CONTENTS

APPI DEDI ABST ACK TABI LIST LIST LIST	ROVA ICAT: IRAC IRAC INOW LE OI IOF T IOF F	ATION AL ION	i ii iii iv vi viii x
СНА	PTER		
1.		RODUCTION	1
	1.1	Introduction	1
	1.2	Motivation	2
	1.3	Problem Statement	2 3 3
		Research Question	
		Research Objective	4
	1.6	Research Scope	5
		1.6.1 Specific Tools	5 5
	1.7	1.6.2 Specific Data Used Research Contribution	5
	1.7	Thesis Organization	<i>5</i>
	1.9	Summary	8
2.	LIT	ERATURE REVIEW KNIKAL MALAYSIA MELAKA	9
	2.1	Introduction	9
	2.2	Insider Threats and The Impact	9
	2.3	Characteristic of Insider Threats	21
	2.4	Keyword-based Sentiment Analysis	35
	2.5	Insider Threats Detection Approach	40
		<ul><li>2.5.1 Insider Threats Detection Approach Through Email Content</li><li>2.5.2 Previous Implementation of Scoring And Data Division Method</li></ul>	53 62
		<ul><li>2.5.2 Previous Implementation of Scoring And Data Division Method</li><li>2.5.3 Correlation Coefficient</li></ul>	66
		2.5.4 Similar Study	75
	2.6	Research Gap and Summary	79
3.	RES	SEARCH METHODOLOGY	81
	3.1	Introduction	81
	3.2	Requirement Analysis of the Proposed Insider Threats Detection Model	
		Approach	81
	3.3	Keyword Analysis and Extraction	82
	3.4	Designing the Proposed Insider Threats Detection Model Approach 3.4.1 Profiling Process	82 82

		3.4.2 Scoring Process	83
		3.4.3 Statistical Analysis Process	84
	3.5	Implementation of The Proposed Insider Threats Detection Model	
		Approach	85
	3.6	Evaluation of the Proposed Insider Threats Detection Model	
		Approach	88
	3.7	Summary	89
4.	TH	E PROPOSED INSIDER THREATS DETECTION MODEL	
	API	PROACH	91
	4.1	Introduction	91
	4.2	CERT Insider Threats Dataset	91
		4.2.1 Data Pre-Processing	92
	4.3	The Proposed Insider Threats Detection Model Approach	93
		4.3.1 Keyword Analysis and Extraction	93
		4.3.2 Training Stage	95
		4.3.3 Testing Stage	98
	4.4	Friedman Statistical	101
		4.4.1 Linear Regression	111
		4.4.2 Correlation Coefficient	115
	4.5	Summary	127
_	DEG		120
5.	5.1	SULT Introduction	<b>128</b> 128
	5.2	Performance Evaluation	128
	3.2	5.2.1 Evaluation of Factor Indicator	129
		5.2.2 Attack detection rate	133
	5.3	Result and Discussion	137
	3.3	6.3.1 Friedman Statistical Approach	139
		6.3.2 Linear Regression Approach	142
		6.3.3 Linear Regression (R <sup>2</sup> ) with Correlation Coefficient Approach	144
	5.4	Overall Finding	147
	5.5	Comparison of Performance Evaluation	150
	5.6	Summary	151
6.	CO	NCLUSION AND FUTURE WORK	156
	6.1	Conclusion	156
	6.2	Contribution of the Work	158
	6.3		159
REF	EREN	ICES	160
APP	ENDI	CES	176

# LIST OF TABLES

TABLE	TITLE	PAGE
1.1	Summary of Problem Statement	3
1.2	Summary of Research Objective	4
1.3	Summary of Problem Statement, Research Question, Research	l
	Objective and Research Contribution	5
2.1	Characteristic of Insider Threats	31
2.2	The keyword-based sentiment analysis studies	38
2.3	The insider threats detection approach	49
2.4	The insider threats detection approach through email content	58
2.5	Related work of previous applied scoring and data division method	65
2.6	The insider threats detection using correlation coefficient	72
2.7	Previous Similar Study	78
4.1	Example of generate anomaly score	97
4.2	Example of Friedman statistical Case 1	106
4.3	Example of Friedman statistical Case 2	107
4.4	Example of Friedman statistical Case 3	108
4.5	Example of Friedman statistical Case 4	116
4.6	R <sup>2</sup> value for Case 1	117
4.7	R <sup>2</sup> value for Case 2	118
4.8	R <sup>2</sup> value for Case 3	119

4.9	R2 value for Case 4	123
4.10	Example of correlation coefficient for Case 1	124
4.11	Example of correlation coefficient for Case 2	125
4.12	Example of correlation coefficient for Case 3	126
4.13	Example of correlation coefficient for Case 4	126
5.1	Five different critical value for defining null hypothesis in Friedman	
	as threshold	133
5.2	The R <sup>2</sup> value of linear regression	134
5.3	Correlation coefficient threshold value	134
5.4	Previous insider threats detection methods	154



## LIST OF FIGURES

FIGURE	TITLE	PAGE
2.1	CMO model for insider threat detection adapted from Maasberg et al.	
	(2015)	12
2.2	Multi-dimension taxonomy of insider threats adapted from	
	Alhanahnah et al. (2016)	14
2.3	Framework for characterizing insider threats adapted from Nurse et al.	
	(2014)	16
2.4	Impacts of insider threats	20
3.1	Research Process	81
3.2	Design of proposed insider threats detection model	83
3.3	Process of data collection and analysis methodology LAKA	86
3.4	CERT insider threats dataset Glassier and Lindalier (2013)	87
4.1	The framework of proposed Insider Threats Detection Model	
	Approach Using Statistical Analysis	94
4.2	Distint data in SQLyog	95
4.3	Anomaly score table for attack profiling in SQLyog	97
4.4	Example of matching field	99
4.5	Output of allocating attack score	112
4.6	Attack score in testing table	116
4.7	Flowchart of Friedman statistical approach	117

4.8	Loosely coupled	112
4.9	Tightly coupled	1192
4.10	Linear regression of Case 1	116
4.11	Linear regression of Case 2	117
4.12	Linear regression of Case 3	118
4.13	Linear regression of Case 4	119
5.1	Poorly detected (motive factor) using Friedman statistical approach	129
5.2	Poorly detected (capability factor) using Friedman statistical	
	approach	130
5.3	Poorly detected (opportunity factor) using Friedman statistical	
	approach	131
5.4	Attack detection rate (correctly and falsely detected) of proposed	
	insider threats detection model approach	135
5.5	Correctly and falsely detected result of Friedman statistical approach	140
5.6	Correctly and falsely detected result of linear regression (R <sup>2</sup> )	
	approachERSITI TEKNIKAL MALAYSIA MELAKA	146
5.7	Correctly and falsely detected result of linear regression (R <sup>2</sup> ) with	
	correlation coefficient approach	149
5.8	Overall finding of insider threats relationship between insider threats	
	behaviour and relevant extracted insider threats keyword	152
5.9	Detection result based on factor for each statistical appraoch	153

# LIST OF APPENDICES

APPENDIX	TITLE	PAGE
A	SQLyog Query for Friedman Statistical Approach	176
В	SQLyog Query for Linear Regression (R <sup>2</sup> ) Approach	179
С	SOLvog Ouery for Correlation Coefficient Approach	180



#### LIST OF ABBREVIATION

CITD - Corporate Insider Threat Detection

CMO - Capability Means Opportunity

DLP - Data Loss Prevention

GRU - Gate Recurrent Unit

HIE - Heterogeneous Information Ensemble

IP - Intellectual Property

IT - Information Technology

ITS - Insider Threats Score

K-NN - K-Nearest Neighbours

NMF - Non-negative Matrix Factorization

PCA - Principal Component Analysis

PS - Problem Statement

RO - Research Objectives

SOFIT - Sociotechnical and Organizational Factor for Insider Threats

SQL - Structured Query Language

SVM - Singular Value Decomposition

SVM - Singular Value Decomposition

TF-IDF - Term Frequency-Inverse Document Frequency

#### **CHAPTER 1**

#### INTRODUCTION

#### 1.1 Introduction

Intellectual property and private information have become more difficult to protect as malicious threats nowadays emanate more from internal sources Gavai, Sricharan, Gunning, Hanley and Rolleston (2015). Known as insider threats, they are launched by malicious insiders who have circumvented any cybersecurity defence systems of organizations. Such malicious insiders come from within organizations themselves and could be current employees, retired or resigned employees, business collaboration partners, and anyone having knowledge of computer systems including the vulnerabilities and weaknesses of the organizations. IP theft, fraud and sabotage are types of insider threats launched by malicious insiders that create trust and financial issues as well as loss of reputation of the targeted entity (Jang-Jaccard, 2014);(UK, 2015);(Heneke et al., 2016); (Ekran, 2016);(Wood et al., 2016);(Scott and Spaniel, 2017);(Huff et al., 2019);(Biringer et al., 2019);(Lykou, 2019);(Ambrogi, 2019).

Therefore, it is important to understand and analyse insider threats from internal sources such as email platforms to mitigate the damage. Common platforms, such as emails are familiar and used by millions including malicious insiders to easily perform daily tasks. As such, communicating and sharing or exchanging legal and private information through emails make it easy for malicious insiders to access important information. In addition, emails also contain various keywords used for communicating. Thus, their content also becomes a medium for people, including potential malicious parties, to express and convey

their feelings. This research proposes modelling an insider-threat detection approach for understanding insider-threat behaviour as reflected in relevant extracted keywords and in detecting the relationships between each of the elements involved using Friedman's statistical, linear regression, and correlation coefficients. In addition, insider threat behaviour is divided into three factor indicators, namely motive, capability, and opportunity to enable a better understanding of which behaviour will indicate the particular insider threat factor.

#### 1.2 Motivation

Addressing insider threats has become a major cybersecurity issue in mitigating malicious threats. As malicious insiders are individuals who benefit from access to legal and private information and assets as well as have knowledge about the vulnerabilities in an organization, they can easily launch attacks without being detected by cybersecurity defence. This situation is worsened if an organization relies only on traditional security cybersecurity systems that exclude any insider threat detection mechanisms. Consequently, more updated and efficient cybersecurity defence is required especially in detecting and mitigating insider threats. Additionally, organizations rely heavily on email as an important and ubiquitous platform to facilitate their daily tasks. Therefore, email content is potentially open to malicious insiders for gathering legal and private information for their own benefit. Also, keywords used in email content can directly portray individuals, including insider threat, behaviour. Hence, this research applies various approaches that focus on detecting insider threat relationships and behaviour and relevant keywords used through Friedman's statistical, linear regression, and correlation coefficients. This research proposes a model for identifying the high attack-detection rate to detect insider threats.

## 1.3 Problem Statement

The number of insider threat issues to cybersecurity is growing rapidly and of major concern to the data and privacy of organizations. Due to the lack of studies on insider threat detection, proposing a solution could be difficult as the influencing factor of insider threats is difficult to be determine. This is because understanding and acknowledging such malicious behaviour needs more studies to define such threats more accurately. As this research focuses on insider threat detection on email content, it is important to understand and determine such activities and the relevant keywords that are used. However, the difficulty to determine the factor that influences insider threats by analysing extracted insider keywords is one of the issues that need to be overcome.

Moreover, detecting insider threats behaviour more accurately as insider threats keywords effected by the outlier is difficult as this is a time consuming and complex process. This phase needs to be done with accurate references from various resources to determine what indicators of insider threat behaviour is specifically reflected in each malicious keyword.

Table 1.1: Summary of Problem Statement

PS	Problem Statement.
	Difficult to determine the factor that influences insider threat by analyzing
PS1	extracted insider keywords.
	Difficult to detect insider threat behaviour more accurately as insider
PS2	threat keywords effected by the outlier.

## 1.4 Research Question

- i. What are the current issues regarding insider threats?
- ii. What method should be used to detect the behaviour of insider threats from relevant extracted insider threat keywords?
- iii. How to accurately detect insider threat relationships?

## 1.5 Research Objective

Insider threats can be determined with the proposed relevant detection method. The process of understanding such malicious behaviour and the characteristics, types and impacts of insider threats can contribute towards proposing effective models for insider threat detection. Besides that, current issues of insider threats also should be investigated as such threats are daily found to have various ways and platforms to create harm. Besides, by investigate critical success factor of insider threats detection model, previously proposed detection approaches could be used as a reference to provide better and more effective insider threat detection. Therefore, the first research objective is to investigate critical success factors of insider threat detection model.

Through a deeper investigation of insider threats issues, it will be possible to have a better understanding of insider threat behaviour and lead to efficient detection approaches. In this research, several previously proposed insider threat detection approaches and mechanisms have been used to accurately detect insider threats. However, such attempts face many drawbacks as malicious insiders use various means to launch insider threats. Therefore, this research proposes modelling an insider threat detection approach by the application of a scoring and statistical-based mechanism that can analyse and accurately detect insider threat behaviour. This second research objective proposes modelling an insider threat detection approach that is updated and more effective in mitigating current insider threat issues. Table 1.2 provides a summary of the research objectives.

Table 1.2: Summary of Research Objective

RO	Research Objective.
RO1	To investigate critical success factors of insider threats detection model.
RO2	To propose insider threats detection model using scoring and statistical analysis to improve detection accuracy that influenced by outlier.

## 1.6 Research Scope

This research applies various approaches to detect insider threats.

## 1.6.1 Specific Tools

- i. Use of Microsoft Excel to gather related insider threat data.
- ii. Use of SQLyog to execute SQL query.
- iii. Weka 3.8.3.

## 1.6.2 Specific Data Used

i. CERT dataset.

## 1.7 Research Contribution

Table 1.3 provides a summary of the research problem statement, research questions, research objectives, and research contribution of the proposed model of the insider threat detection approach.

Table 1.3: Summary of Problem Statement, Research Questions, Research Objectives and Research Contribution

<b>Problem Statement</b>	<b>Research Questions</b>	Research Objectives	Research
			Contribution
Difficult to	What are the	To investigate	The proposed
determine the	current issues	critical success	modelling of
factor that	regarding insider	factors of insider	the insider
influences insider	threats?	threats detection	threat
threat by		model.	detection
analyzing	What method		approach is
extracted insider	should be used to	To propose insider	able to analyze
keywords.	detect the	threats detection	and detect
	behaviour of	model using	insider threat
Difficult to detect	insider threats	scoring and	behaviour as
insider threat	from relevant	statistical analysis	reflected from
behaviour more	extracted insider-	to improve	relevant
accurately as	threat keywords?	detection accuracy	extracted

insider threat	How to	that influenced by	insider threat
keywords	accurately detect	outlier.	keywords.
effected by the	insider threat		
effected by the outlier.	relationships?		The scoring method applied in the proposed insider threat detection model can produce anomaly scores for attack profiling which is important to mark the detected insider threat
H H	>		behaviour
E			more
2400			accurately.

In this research, the proposed modelling for the insider threats detection approach is based on earlier studies and covers three major factors, namely motive, capability and opportunity. Each of the relevant insider threats keywords is extracted from email content and indicates different factors representing the existence of insider behavior. When the relationship between insider threat behaviour and each extracted keyword is determined, the threat is defined as detected. In addition, the attack profiling and scoring method are also used in the proposed modelling approach to marking insider behaviour more accurately. The applied Friedman statistical, linear regression and correlation coefficient in this proposed model is able to better detect the relationship between insider threat behavior and the extracted insider threats keywords compared to previous studies.

## 1.8 Thesis Organization

This dissertation is divided into seven chapters.

## **Chapter 1: Introduction**

This chapter describes the research and the activities to be developed. It covers the research motivation, problem statement, research questions, research objectives, research scope, and research contribution.

## **Chapter 2: Literature Review**

This chapter discusses related work in a number of earlier studies and includes journals and books on the subject matter from 2010 to 2021. It explores current drawbacks, identifies gaps, and proposes solutions. The factors indicating insider threat behaviour (motive, capability and opportunity) and the impacts of insider threats are also covered.

## **Chapter 3: Research Methodology**

This chapter discusses the framework methodology and each stage of the processes involved in conducting the research. In addition, insider threat dataset collection is also explained briefly.

## **Chapter 4: The Proposed Insider Threats Detection Model Approach**

This chapter introduces the implementation phase of the proposed model on insider threat detection including the SQL query and data pre-processing process following data collection. The insider threats dataset is divided into the training and testing stages to produce better accuracy in attack detection. The framework for the proposed model of keyword analysis and extraction of relevant insider threats keywords, the Friedman statistical, linear regression, and correlation coefficient are also discussed.

## **Chapter 5: Results**

Chapter 5 covers the performance evaluation including evaluation of factor indicator (motive, capability and opportunity), attack detection rates of the proposed model on insider

threat detection, results and discussion of applied approaches, overall findings of the relationship between insider threats, and comparison of performance evaluation.

## Chapter 6: Conclusion and Future Work

Chapter 6 includes the conclusion and future work on the proposed threat detection model approach and the contributions of the work.

## 1.9 Summary

This chapter included an introduction of insider threats which to express the importance of understanding and analyzing insider threats from internal sources such as email. As one of the common communication platforms, emails are used by millions including malicious insiders to perform daily tasks smoothly. Moreover, communicating and exchanging legal and private information through emails is able to make malicious insiders access important information. As well as emails containing various keywords used for communicating, the content becomes a medium to express and convey their feelings. Therefore, each keyword used represent emotion of many people. Furthermore, most organizations relied on email as important platform to simplify their daily tasks. This situation is worsened if organization relies only on traditional security cybersecurity systems that exclude insider threat detection mechanisms because malicious insiders are individuals who have benefit to access any private information and assets. With knowledge about the vulnerabilities in an organization, they can easily launch attacks without being detected by cybersecurity defense. Prior objectives of this research detecting insider threats from email, malicious behaviour of insiders that reflect from relevant extracted insider threats keywords are difficult as this is a time consuming and complex process. Therefore, various studies need to be done with accurate references from a number of resources in order to determine what indicators of insider threat behaviour is specifically reflected in each malicious keyword.