MALWARE DETECTION USING ENSEMBLE METHOD

AMIZAH AIDA BINTI AHMAD

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

MALWARE DETECTION USING ENSEMBLE METHOD

BORANG PENGESAHAN STATUS TESIS

JUDUL: **MALWARE DETECTION USING ENSEMBLE METHOD**

SESI PENGAJIAN: **SEMESTER II 2016/2017**

Saya **AMIZAH AIDA BINTI AHMAD (B031510056)**mengaku membenarkan tesis (PSM/~~Sarjana/ DoktorFalsafah~~) ini disimpan di Perpustakaan Fakulti Teknologi Maklumat danKomunikasi dengansyarat-syarat kegunaan seperti berikut:

1. Tesis dan projek adalah hakmilik UniversitiTeknikal Malaysia Melaka.
2. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat Salinan untuk tujuan pengajian sahaja.
3. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat Salinan tesis ini sebagai bahan pertukaran antara institusi pengajian tinggi.
4. ** Sila tandakan (/)

_____	SULIT	(Mengandungi maklumat yang berdarjah keselamatanatau kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA RASMI 1972)
_____	TERHAD	(Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi/badan di mana penyelidikan dijalankan)
_____	TIDAK TERHAD	

(TANDATANGAN PENULIS) (TANDATANGAN PENYELIA)

Alamat tetap:
No 34C , Blok F Jalan 17/1A'
Seksyen 17,46400 Petaling jaya

Tarikh: Tarikh: 18/8/2017

MALWARE DETECTION USING ENSEMBLE METHOD

AMIZAH AIDA BINTI AHMAD

This report is submitted in partial fulfillment of the requirements for the

Bachelor of Computer Science (Computer Networking) With Honors

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2017

DECLARATION

I hereby declare that this project report entitled

MALWARE DETECTION USING ENSEMBLE METHOD

is written by me and is my own effort and that no part has been plagiarized without citations.

STUDENT: _____ Date: _____

(AMIZAH AIDA BINTI AHMAD)

I hereby declare that I have read this project report and found this project report is sufficient in term of the scope and quality for the award of

Bachelor of Computer Science (Interactive Media) with Honors

SUPERVISOR: _____ Date: 18/8/2017

(EN. AZMAN BIN MAT ARIFF)

DEDICATION

This final project is dedicated to all generous people who have been aided us and supported, encouraged and help me during this project.

Especially to my beloved parents Ahmad bin Jasa and Maimon binti Siron for their endless support and love , both of my siblings Amirah Arina binti Ahmad and Armin Azrai bin Ahmad for their motivation throughout the year in completing my studies and always pray the best for me.

Special dedication to my supervisor, En. Azman bin Mat Ariff who has guided and never lose hope by supporting , helping and advising me while making the progress for this project.

Not forget to my evaluator, P.M. DR. Faizal bin Abdollah for her positive advices, feedback and comment on this project,

.

Last but not least, to all my beloved friends who always help me from

the beginning of this project until the end of it.

ACKNOWLEDGEMENT

Bismillahirahmanirrahim,

First and foremost, I would like to express my gratitude to Allah S.W.T for giving me strength and for keeping me in the path of righteous while completing the whole process of this project. Without His blessings, I would not have been able to be as I am today and completing this project.

Also a million thanks to my supervisor, En. Azman bin Mat Ariff for his guidance, constant supervision and kindness in completing this project. Without his guide, I am surely would not know the right trail.

I would also like to special thanks to my family members , my dear father and mother for who always support and pray the best for me and constantly give me motivation throughout my project. Besides, thanks to my friends that really play a big role who have been very helpful by giving ideas and suggestion especially Nor Anis Syazana and Sarah Shamimi the during project development.

Thank you.

ABSTRACT

In today's technology driven world, the increasing of malware in the cybercriminals that exploiting the internet and always create and distribute harmful malware has become a serious threat. Malware significantly impact computer's performance and often go unnoticed in our systems and causes several problems to the user. Hence, It's imperative to take the precautions necessary to detect and prevent malware infections. One of the way to detect malware detection is by using machine learning techniques. Malware detection is detected by looking at its behavioural. Behavioural malware detection is a field where malware is detected by its behaviour and the machine learning will look at the pattern of the behavioural. Then it will be analyzed and a report will be generate from the data. Thus, in this project, the behavioral of malware is analyzed and ensemble method is applied in detecting malware. Firstly, the data is collected by a multiple categories of system log and parser chooses from application. Then from the dataset it will classify it to 5 type of n-gram. Secondly, the best features from each of the n-gram are extracted using three feature selection techniques, namely Information Gain, Symmetrical Uncertainty and Chi-Square. SVM classifier is used to train the feature vectors and create a model for each n-gram. Finally, every model from 1-gram to 5-gram is combined using ensemble method. The significant contribution of this project is the effectiveness and efficiently of malware prediction using the state-of-the art techniques named ensemble method.

ABSTRAK

Dalam dunia teknologi terkini, peningkatan malware dalam penjenayah siber yang mengeksploitasi internet dan sentiasa mencipta dan mengedarkan malware berbahaya telah menjadi ancaman yang serius kepada dunia. Malware secara ketara telah memberi kesan kepada prestasi komputer dan sering kali tidak disedari oleh sistem sedia ada yang menyebabkan beberapa masalah kepada pengguna. Oleh itu, sangat penting untuk mengambil langkah berjaga-jaga yang diperlukan untuk mengesan dan mencegah jangkitan malware. Salah satu cara untuk mengesan pengesanan malware adalah dengan menggunakan teknik pembelajaran mesin. Pengesanan perisian hasad dikesan dengan melihat perilakunya. Pengesanan malware tingkah laku adalah bidang di mana malware dikesan oleh tingkah laku dan pembelajaran mesin akan melihat corak tingkah laku. Kemudian ia akan dianalisis dan laporan akan menjana dari data. Oleh itu, dalam projek ini, perilaku malware dianalisis dan kaedah ensemble digunakan untuk mengesan malware. Pertama, data dikumpulkan oleh pelbagai kategori log sistem dan parser yang dipilih dari aplikasi. Kemudian dari dataset ia akan mengklasifikasikannya kepada 5 jenis n-gram. Kedua, ciri-ciri terbaik dari setiap n-gram diekstrak dengan menggunakan tiga teknik pemilihan ciri, iaitu *Information Gain*, Ketidaktentuan Simetris dan Chi-Square. Pengelas SVM digunakan untuk melatih vektor ciri dan membuat model untuk setiap n-gram. Akhir sekali, setiap model dari 1 gram hingga 5 gram digabungkan menggunakan kaedah ensemble. Sumbangan besar projek ini adalah keberkesanan dan kecekapan ramalan malware menggunakan teknik ensiklopedia state-of-the-art yang dinamakan ensemble method.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

INTRODUCTION

1.1 Introduction

Malware, short for malicious software, is is any PC program software used to mess up PC or portable operations, collect sensitive data, have access to privatepc framework , show undesirable promoting. Malware may be stealthy, expected to take data or keep an eye on PC clients for a period time of without their insight and designed to infiltrate and damage computers and often go unnoticed in our systems and causes several problem . Thus, malware is a software programmed to do bad thing for user or user computer, and user don't want it. One of the way to detect malware detection is by using machine learning techniques. Malware detection is detected by looking at its behavioural. Behavioural malware detection is a field where malware is detected by its behaviour and the machine learning will look at the pattern of the behavioural Then it will be analyzed and a report will be generate from the data. Thus, in this project, the behavioral of malware is analyzed and ensemble method is applied in detecting malware. Other than that, The increasing of latest malware and various of malware that exploiting the internet and always create and

distribute harmful malware has become a serious issue. Thus, in this project we will purposed a new malware classification technique by analyzing and look at the pattern of the behavioral of malware and how to detect the the malware detection using an ensemble method . Using this technique we can analyze the malware behavioral and from the results showed that this technique approach can effectively classify and differentiate the malware.

1.2 Problem Statement

The problem that has been identified is summarized in Table 1.1 below

Table 1.1 Problem Statements

PS	Problem Statement
PS1	Malware detection has increasing day by day thus it is needed to detect malware s by using machine learning techniques. Malware detection is detected by looking at its behavioural .

a) PS1: Malware should be detected by looking at its behavioral

Malware be identified and detected so that the proposed to take data or spy on users pc without their knowledge for a period of time will not happen. One of the method to detect the malware is by using Machine Learning techniques. This technique will be used in this project to detect the behavioral attack of malware

1.3 Project Question

In this project, there is three Project Question (PQ) needs to be answered in this project. The summary of project question is shown in Table 1.2.

Table 1.2 Summaries of Project Questions

PQ	Project Questions
PQ1	What are the purposes on this project?
PQ2	What are the approach techniques that will use to detect the malware ?
PQ3	What is the contribution when applying the process and techniques?

PQ1: What are the purposes on this project?

Identify the purposes and goals of this project.

PQ2: What are approach or techniques that will use to malware ?

Distinguish the approach and procedures that will use to recognize malware to have a greater accuracy.

PQ3: What is the contribution when applying the process and techniques?

The enhance on applying the process and techniques on detecting malware to be more accurate .

1.4 Project Objective

Based on this project, the Project Objectives (PO) are developed as follows.

Table 1.3 Summaries of Project Objectives

PQ	Project Objectives
PO1	To study the performance of ensemble method that combine 1-gram to 5-gram feature vector
PO2	To measure the effectiveness of three different combination methods namely majority voting, product rules and mean rules.
PO3	To study the performance of ensemble method that combine specific n-gram feature vectors using different classifier model

1.5 Project Scope

The scope of these projects is to detect the behavioural of malware using ensemble method so that can easily classify it to malware or non-malware and Identify the technique and algorithm that is used to detect the malware. This project is to test the accuracy of the detection of malware using the machine learning technique and to find result of which technique is the most accurate.

1.6 Project Contribution

The contribution of this technique are by doing this project , it will evaluate from five n-gram of feature selection. Each of the n-gram will be evaluate and test based on the support vector machine classifier. Then, from the result each of the model will be combine to predict the accurancy and find out which is the best andmore accurate. The contribuition of this finding is it will contribute an idea in finding the vest method to detect the malware and which n-gram has the high accurancy.

1.7 Thesis Organization

Chapter 1: Introduction

In this chapter 1 section it will review the history of project . Other than that, In this chapter 1 also contain the problem statement, project question, project scope statement , project contribution, thesis organization and conclusion.

Chapter 2: Literature Review

In this chapter 2 it will review about about past work that are knit together to the project and review of previous researcher by study at the literature review. Each of the literature review result will be consider to help in generate this project. In addition . in this chapter also include the conclusion that are clear up to run this project .

Chapter 3: Project Methodology

In this chapter 3 section it will review about the methodology that are used in these projects. This chapter explain step by step of the methodology in and the project milestone.

Chapter 4: Analysis and Design

In this chapter 4 it will review the discussion about the analysis of the project and the design of the project that had been used in the experiment that consists of the software and hardware requirement to run this project

.

1.8 Conclusion

This project (Malware detection using ensemble method) is a research project that detect the malware by looking at its behavioural . Malware will significantly impact on your computer performance and often go unnoticed in your system. Thus its imperative to take the precautions necessary to detect and prevent malware infections . In this reasearch it used the machine learning technique to detect the malware using an algorithm. From the algorithm the data will be identified whether it is malware or non-malware. The objective of the project is to help users to prevent from being attack by the user by analysis the behavioural of malware. Thus in next chapter will explains the previous research about malware detection and how the data is conducted.

CHAPTER II

LITERATURE REVIEW

2.1 Introduction

The goals of this Chapter 2 is to summarize, explained and increase knowledge and provide better understanding about malware detection and ensemble method. In this chapter 2 we will explain the definition , approach and elaborate from a technique that were used from previous research from a literature review. From the literature review we can compare the findings and evaluate based on the information. The explanation of literature review are categorize to several topics which consist of malware , detection technique and ensemble method. In malware will focus more on the definition of malware , effect of malware ,type of malware and the use of malware. Next for the

detection technique will be focus more on machine learning technique where will explain the definition and benefits of machine learning technique . Lastly in this topic also will explain the detection that are use in this research which is ensemble method. In the ensemble method will focus more on definition of ensemble method and why ensemble method is use in conducting this research. The hierarchy of this topics is based on the figure 2.1 and consists of three phase . Phase one is about malware , phase 2 is detection technique and lastly phase 3 is detection technique.

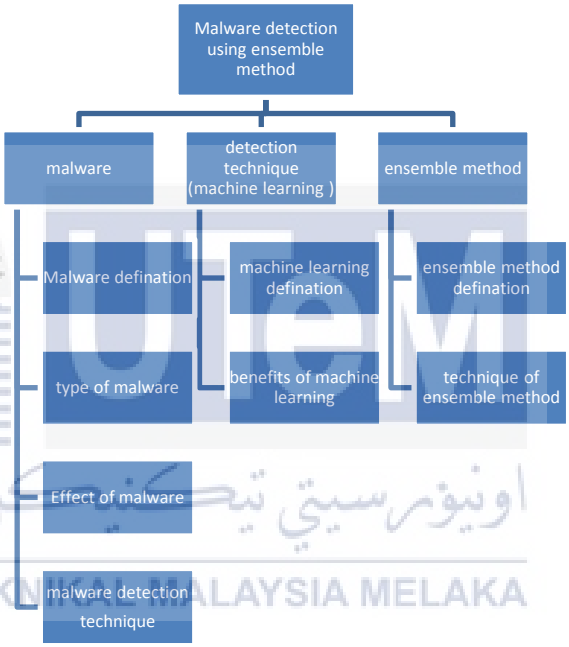

Figure 2.1 Hierarchy

2.2 Related Work / Previous Work

Based on figure 2.1 , this section 2.2.1 is an explanation about malware , section 2.2.2 explaination about detection technique name machine learning and 2.2.3 definition about ensemble method and responsibility of ensemble method.In addition in this chapter also discuss the related work of previous researcher . All of the section will explain based on understanding and previous research..