MALWARE DETECTION USING N-GRAM WITH TF-IDF WEIGHTING

NATASHA BINTI ZAINAL

BACHELOR OF COMPUTER SCIENCE

UNIVERSITI MALAYSIA PAHANG

SUPERVISOR'S DECLARATION

I hereby declare that I have checked this thesis and in my opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Bachelor of Computer Science in Computer System and Networking.

_____

(Supervisor's Signature)

Full Name : DR NOORHUZAIMI@KARIMAH BINTI MOHD NOOR
Position : SENIOR LECTURER
Date : 12/12/2018

_____

(Co-supervisor's Signature)

Full Name :

Position :

Date :

STUDENT'S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Universiti Malaysia Pahang or any other institutions.

_____

(Student's Signature)

Full Name : NATASHA BINTI ZAINAL

ID Number : CA15128

Date : 12/12/2018

MALWARE DETECTION USING N-GRAM

NATASHA BINTI ZAINAL

Thesis submitted in fulfillment of the requirements
for the award of the degree of
Bachelor of Computer Science (Computer System & Networking) with Honors

Faculty of Computer System & Software Engineering

UNIVERSITI MALAYSIA PAHANG

DECEMBER 2018

ACKNOWLEDGEMENTS

First and for most, I would like to express my special appreciation and thank you to the people whom have helped me to complete this thesis mentally, emotionally. They have contributed their support and thoughts towards me during the completion of this thesis.

Firstly, thank you to my supervisor, Dr Noorhuzaimi@Karimah binti Mohd Noor for giving me a chance and believing in me that I can complete this thesis successfully. Thank you for giving me continuous motivation, guidance, patience and limitless of knowledge during my research. Without her motivation and patience, my research could not be completed and presented.

I would also like to thank to my parents Zainal bin Abdul Rahman and Khadijah binti Lot and families who have always give me mentally and spiritually support throughout the process in completing this research.

Last but not least, I would like to thank to all my friends especially to Nornadzirah Hafizah for always giving me moral support, motivations and ideas during the completion of this thesis.

ABSTRAK

Di dalam era technology ini, komputer dan rangkaian terdedah kepada perisian perosak. Perisian perosak juga dikenali sebagai perisian yang merbahaya. Perisian perosak ini di cipta untuk menggangu, memusnahkan atau mendapat kebenaran akses di dalam system komputer. Terdapat pelbagai perisian dan kaedah yang telah di cipta untuk mengesan pelbagai jenis perisian perosak. Perisian perosak yang sangat kuat telah dicipta dan ia tidak dapat di kesan. Terdapat pelbagai anti-virus dan kaedah yang telah di cipta, walaubagaimanpun, cara ini tidak dapat mengesan malware di mana perisian perosak pada masa kini tidak dapat dikesan. Objektif kajian ini adalah untuk mengesan atribut perisian perosak, untuk mencipta model perisian perosak menggunakan *n-gram* dan *TF-IDF*, dan untuk menilai model untuk mengesan perisian perosak. Skop untuk kajian ini adalah set data, cara-cara dan ujian penilaian dan ukuran. Methodologi kajian ini adalah berdasarkan kajian melalui kajian yang sudah berlalu, mengenalpasti attribute perisian perosak, membina konsep model dan akhir sekali menilai konsep moel. Model ini akan dilaksanakan dengan menggunakan bahasa pengaturcaraan Python. Dengan menggunakan kaedah ini, jangkaan keputusan oleh sistem ini adalah berdasarkan *n-gram* dan *TF-IDF*, perosak perisian boleh dikesan.

ABSTRACT

In this era of technology, computers and networks are exposed to malwares. Malwares are also known as malicious software. Malwares are created to disrupt, destroy or to gain authorization in access in a computer system. There are different types of software and methods that have been implemented that are used to detect different types of malware. Powerful malware that was implemented may not get easily detected. Different kinds of anti-virus and methods were used, nevertheless the problem is that this may not fully detect the malware as malwares now a days are hard to detect. The objectives of this research is to identify the attributes of malware, to develop a conceptual model of malware detection using *n-gram* and *TF-IDF* and to evaluate the model of malware detection. The scope for this research are dataset, method and evaluation testing and measurements. The methodology are literature review based on previous research, identifying the attributes of malware, developing the conceptual model and lastly, evaluating the conceptual model. The model is implemented by using Python programming language. By using this method, the expected result of this system is based on the *n-gram* and *TF-IDF*, thus malware could be detected.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

RLE	Run Length Encoding
TF	Term Frequency
TF-IDF	Term Frequency- Inverse Document Frequency
HMM	Hidden Markov Model
OPS	Operational Codes
TP	True Positive
FP	False Positive
TN	True Negative
FN	False Negative
AS	Attribute Similarity
API	Application Programming Language
NLTK	Natural Language ToolKit

CHAPTER 1

INTRODUCTION

1.1 Introduction

In the increasing era of technology nowadays, even a high-tech technology can get easily infected by known or unknown malware. Malware is also known as malicious software where it is designed to damage, disrupt or gain unauthorized access into a computer. The types of malware could be Trojan Horses, Worms, Virus, Spyware and more. Even there are many different types of highly secured security or protocols and policies that are implemented to protect the cyber, with the development of malware from time to time, not every cyber system could be protected. Furthermore, malwares are created to either steal personal information without knowing or malware can destroy a particular system to achieve its objectives.

Malwares that could not be recognized has been increasing drastically, security software that exist could not identify the malwares effectively (Fuyong & Tiezhu, 2017). Powerful malwares that have their different types of intention may not get easily detected especially to intrude in their security system, even there are many different types of method to detect malware that have been designed and implemented into existing systems. Malware is a boundless problem and although familiar use of anti-virus software, the diversification is still ascending (Mira, Huang, & Brown, 2017).

Nevertheless, there are many different types of system methods that have been developed by the researcher from time to time. For example, the methods that have

been developed are by using the complex-flows, neural network, Run Length Encoding (RLE) and more. Different kinds of anti-virus were also implemented to defend the computers from being attacked by the malwares. Nevertheless, these anti-virus gave some limitations where they have limited detection techniques where they scans the computer for known virus pattern, but gives a false alarm where the pattern matches normal file's code. Even worse, it will not detect the code of the virus. But then, the database of the anti-virus program should be updated periodically. All of these methods have their own excellent result analysis and some gave several limitations. Thus, we want to make a deep research on malware detection should be done. Based on this introduction, in this chapter, we will discuss on problem statement, objectives, scope, expected result and thesis organization where this project will conduct on the investigation how to detect malware using *n-gram* and *TF-IDF*.

1.2 Problem Statement

The main problem nowadays in the field of security is protecting from threats or malwares that can cause problems in the future. Some issues need to be taken seriously especially in securing the network or devices from malicious malware. Even though there are many malware detection system that have been developed to help to reduce the problem of devices or network being infected with malware, not all of the system may detect the malware fully before it affects it.

XGBoost was made as a classifier to differentiate the benign software and the malware, verifying it as a great computational efficiency and accuracy in Android malware detection (Wang, Li, & Zeng, 2017). However, this XGBoost model should be extended with a dynamic analysis technology. This was a good way to get more features of the malware. Other than that, deep neural network based was also developed for the malware detection. The system obtain a 95% of detection rate of 0.1% false positive rate (FPR), on 400,000 software binaries and more (Saxe & Berlin, 2015). Nevertheless, the number of benign binaries was small to estimate performance accurately of the false positive scale. It is unclear, if more data is added, the results

would be improve in providing better false positive estimates. In addition, another method to detect malware was by using semantics-aware. A malware-detection algorithm that can detect different types of malware that have a low run-time overhead and it was also a common obfuscations used by hackers (Bryant, 2005). On the other hand, the tools that was used need all of the intermediate representations instructions in the template in order to appear in the same form in the program.

In this research, malwares was detected by using the *n-gram* method. A model to detect malware was developed by using the *n-gram* and *TF-IDF* (Term Frequency-Inverse Document Frequency). This was to evaluate the model of the malware detection.

1.3 Objectives

These are the three objectives that need to be achieve in this project. The objectives of this research are:

i. To identify attributes of malware.

ii. To develop a model of malware detection using *n-gram* and *TF-IDF*.

iii. To evaluate the model of malware detection using *n-gram* and *TF-IDF*.

1.4 Scope

Based on the objectives declared, the scope of this research is divided into three categories which are:

Dataset

i) The dataset is collected from the Kaggle website where the author of this dataset is N Sarvana. This dataset consist of thousands of malwares and benign datasets that are classified based on a few attributes.

REFERENCES

Authority, P. P. (2018). Pip 18.1 Documentation. Retrieved from https://pip.pypa.io/en/stable/installing/

Bryant, M. C. S. J. S. A. S. D. S. R. E. (2005). Semantics-AwareMalwareDetection.

Dot, W. (2017). The importance of malware detection tool. Retrieved from https://medium.com/@William123/the-importance-of-malware-detection-tools-42c68e964abf

Fisher, T. (2018). What is Malware.

Fuyong, Z., & Tiezhu, Z. (2017, 21-24 July 2017). *Malware Detection and Classification Based on N-Grams Attribute Similarity.* Paper presented at the 2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC).

Jamal, J. (2012). Trojan Horse Virus-Impact and Symptoms. Retrieved from http://www.eexploria.com/trojan-horse-virus-impact-and-symptoms/

Kallet, R. H. (October 2004). How to Write the Methods Section of a Research Paper. Retrieved from http://libguides.usc.edu/writingguide/methodology

Magalhaes, R. (2012). Malwares impact, serious and long lasting.

Maloof, J. Z. K. M. A. (2004). Learning to Detect Malicious Executables in the Wid[C] Knowledge Discovery adn Data Mining. 470-478.

Mira, F., Huang, W., & Brown, A. (2017, 7-8 Sept. 2017). *Improving malware detection time by using RLE and N-gram.* Paper presented at the 2017 23rd International Conference on Automation and Computing (ICAC).

P.Aruna. (2015). Malware has affected four out of every 10 Malaysians, say Kaspersky Lab. The Star Online.

Pechaz, B., Jahan, M. V., & Jalali, M. (2015, 11-12 Nov. 2015). *Malware detection using hidden markov model based on markov blanket feature selection method.* Paper presented at the 2015 International Congress on Technology, Communication and Knowledge (ICTCK).

Powers, D. (2011). Evaluation: From precision, recall and f-measure to ROC, Informedness, Markedness & Correlation. *Journal of machine learning technologies*, 27-83.

Saravana, N. (2018). Malware Detection. Retrieved from https://www.kaggle.com/nsaravana/malware-detection

Saxe, J., & Berlin, K. (2015, 20-22 Oct. 2015). *Deep neural network based malware detection using two dimensional binary program features.* Paper presented at the 2015 10th International Conference on Malicious and Unwanted Software (MALWARE).

Security, N. (2018). Computer Virus Symptoms. Retrieved from https://www.nortonsecurityonline.com/security-center/computer-virus-symptoms.html

Subramanya, S. R., & Lakshminarasimhan, N. (2001). Computer viruses. *IEEE Potentials*, *20*(4), 16-19. doi:10.1109/45.969588

Wang, J., Li, B., & Zeng, Y. (2017, 15-18 Dec. 2017). *XGBoost-Based Android Malware Detection.* Paper presented at the 2017 13th International Conference on Computational Intelligence and Security (CIS).