JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

MALWARE DETECTION AND ELIMINATION USING MACHINE LEARNING ALGORITHMS

[1]Somnath Chouwdhury, [2]Yashwardhan Paniya, [3]Akanksha Suryavanshi, [4]Aditya Lal Sinha, [5]Prof.Priya Ujawe
[1]Leader, [2]Member, [3]Member, [4]Member, [5]Guide
Department of Information Technology
G.H.Raisoni College of Engineering and Management Wagholi Pune

Abstract : Modern antivirus software is effective at detecting known threats, but can be evaded by malware not present in their databases or any zero day attack.(zero-day attacks — malicious files targeting vulnerabilities that are previously undisclosed.).
An exponential growth in digital content consumption, digital transactions too have grown and today India sees almost twice the number of digital payments in comparison to China.
This increase in digitalization has created an opportunity for hackers to damage the data and one of the most common attacks used are malwares.
Malwares are also included in the OWASP top 10 cyber threats.
In order to create a more flexible, cost effective and reliable antivirus product, we need to develop alternatives that complement traditional signature-based detection.
An obvious solution to this can be classifier algorithms in machine learning.

IndexTerms **- Machine learning, Cyber security, Random forest, Antivirus**

1.INTRODUCTION

The growth in technology has not only provided opportunities to various types of business but also a lot of cyber crimes.
Introduction of technologies like AI and machine learning has led to creation of intelligent and destructive malwares.
Since the process is automated the number of these softwares are increasing exponentially and a simple antivirus cannot keep up with that. So we would also use machine learning classifiers algorithms to tackle these problems.
We would use different supervised learning models to identify the pattern in malware files and then classify them as malicious and non-malicious.

1.1 Overview

Our vision is to create a software that would work as an extension to modern day Antiviruses or browsers.
Our software would use multiple machine learning Algorithms to detect if a file is malicious or not.
The proposed system aims to decrease the malware threats.

1.2 Need

Machine learning helps antivirus software detect new threats without relying on signatures. In the past, antivirus software relied largely on fingerprinting, which works by cross-referencing files against a huge database of known malware.

The major flaw here is that signature checkers can only detect malware that has been seen before. That's a rather large blind spot, given that hundreds of thousands of new malware variants are created every single day.

Machine learning, on the other hand, can be trained to recognize the signs of good and bad files, enabling it to identify malicious patterns and detect malware – regardless of whether it's been seen before or not.

1.3 Literature survey

Current signature-based antivirus software is ineffective against many modern malicious software threats. Machine learning methods can be used to create more effective anti malware software, capable of detecting even zero-day attacks. an approach that primarily learns from metadata, mostly contained in the headers of executable files, specifically the Windows Portable Executable 32-bit (PE32) file format. Our experiments indicate that executable file metadata is highly discriminative between

malware and benign software. We also employ various machine learning methods, finding that Decision Tree classifiers outperform Logistic Regression and Naive Bayes in this setting. We analyze various features of the PE32 header and identify those most suitable for machine learning classifiers.Finally, we evaluate changes in classifier performance when the malware prevalence (fraction of malware versus benign software)is varied.

Zero-day Malware Detection based on Supervised Learning Algorithms of API call Signatures
proposed and evaluated a novel method of employing several data mining techniques to detect and classify zero-day malware with high levels of accuracy and efficiency based on the frequency of Windows API calls. describes the methodology employed for the collection of large data sets to train the classifiers, and analyses the performance results of the various data mining algorithms adopted for the study using a fully automated tool developed in this research to conduct the various experimental investigations and evaluation. Through the performance results of these algorithms from our experimental analysis,
we are able to evaluate and discuss the advantages of one data mining algorithm over the other for accurately detecting zero-day malware successfully.

1.4 Existing System

1. Avast
2. AVG
3. AVIRA
4. bitdefender
5. Kaspersky
6. McAfee
7. VBA32

These systems are good with normal malware but they do not use any prediction algorithms and most of them are subscription based antiviruses. We propose a system which is not only budget friendly but also can prevent any kind of malware threats.

2.PROPOSED SYSTEM

2.1 Objective

We will use various supervised learning algorithms, Train and test them with malware datasets and then choose the algorithm with highest accuracy to build our model.

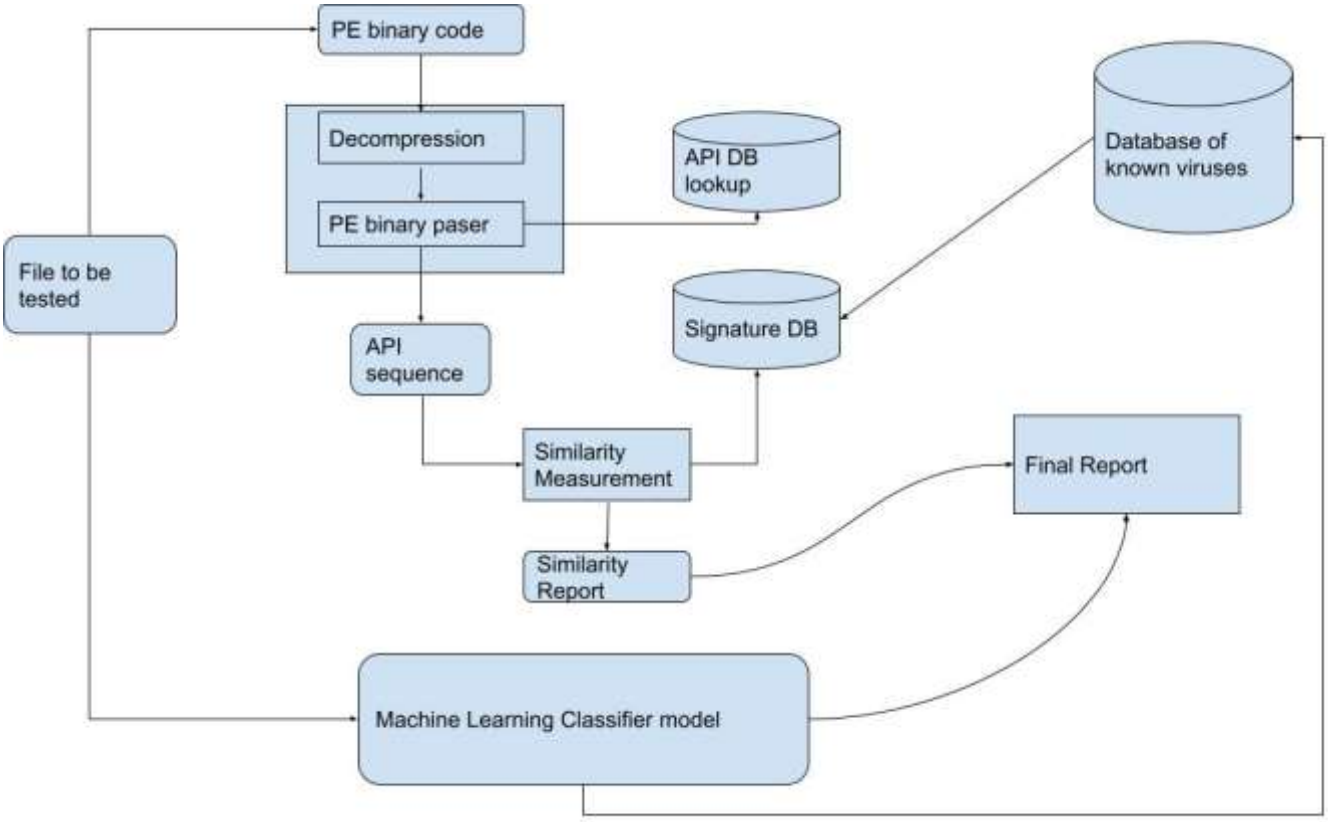
Work environment would comprise frameworks like Jupyter notebook, Spyder, Kali Linux Terminal.
we would be working with advanced technologies such as Artificial Intelligence, Machine Learning, Cyber security.

2.1 Algorithms

1. "LR" : linear_model.LogisticRegression(),
2. "KNN": KNeighborsClassifier(),
3. "DecisionTree": tree.DecisionTreeClassifier(),
4. "RandomForest": ske.RandomForestClassifier(),
5. "GradientBoosting": ske.GradientBoostingClassifier(),
6. "AdaBoost": ske.AdaBoostClassifier(),
7. "GNB": GaussianNB(),
8. "SVC": svm.SVC()

3.SYSTEM DESIGN

3.1 Antivirus Structure



4.HOW SYSTEM WORKS

The System Would work in following way

1) Cleaning, Training and Testing datasets :- First datasets are sanitized for better results and then divided into training and testing sets of 5:1 ratios.
2) Algorithm testing :- now the dataset would be passed through a bunch of supervised machine learning classification algorithms.
3) Algorithm Selection :- The algorithm with highest accuracy would be chosen and saved with the pickle library in the form of pkl files.
4) module creation :- with the help of the saved data an interface would be created
5) Interface :- lastly a file would be passed through the interface and our program will classify the file as malicious or non-malicious.

5.ADVANTAGES

- Cost Effective :- The only cost behind this project would be the datasets collected and some minor charges on software so the antivirus would be budget friendly to every user.
- Time Saving :- Since we would be saving the modules of the best algorithms in forms of packets, this software would be highly time saving.
- Highly Efficient :- During the testing we would use the algorithms with highest accuracy.
- Scalable

6.LIMITATIONS

- Inconsistent results due to improper datasets :- Getting hold of really accurate datasets is difficult and time consuming and other than that the normal datasets can contain inaccurate data.
- false positives :- A false positive is an error in classification in which a test result incorrectly indicates the presence of a condition
- false negatives :- a false negative is the opposite error, where the test result incorrectly indicates the absence of a condition when it is actually present.

ACKNOWLEDGEMENT

We are very grateful to our project guide Prof. Priya Ujawe and all teaching and non-teaching staff for guiding us all over the duration of the project. They were very helpful to us, as and when we required their help.

Conclusion

We have proposed a system which can potentially detect and eliminate the threats posed to the cyber world even if they are not present in the malware database or made after updating the database with the help of any technologies with help of Machine learning and cyber security..

References

[1] IEEE paper Zane Markel and Michael Bilzor Computer Science Department U.S. Naval Academy Annapolis.
[2] researchgate.net/publication/249832637_Obfuscated_Malicious_Executable_Scanner
[3] GitHub
[4] VurusShare.com(data.csv dataset for malicious and non-malicious viruses)
[5] blog.emsisoft.com/en/35668/the-pros-cons-and-limitations-of-ai-and-machine-learning-in-antivirus-software/
[6] timesofindia.indiatimes.com/blogs/voices/technology-to-bolster-opportunities-for-industry-4-0/
[7] OWASP top 10

AUTHOR BIOGRAPHY

Somnath Chouwdhury student of Information Technology from G.H.Raisoni College of Engineering and Management pune Interested in Research and Development in Machine Learning and cyber security.