Anomaly Based Malware Detection Using Machine Learning

PROJECT SYNOPSIS

OF MINOR PROJECT - PHASE I

BACHELOR OF TECHNOLOGY
CSE

SUBMITTED BY

Aditya Singh Diwakar 309302220007
Anthoni 309302220017
Harsh Swaroop Dubey 309302220030
Jagmohan Rajwade 309302220041

BHILAI INSTITUTE OF TECHNOLOGY, RAIPUR
DEPARTMENT OF COMPUTER SCIENCEAND ENGINEERING

Table of contents

Introduction

Malware refers to malicious software perpetrators dispatch to infect individual computers or an entire organization's network. It exploits target system vulnerabilities, such as a bug in legitimate software A malware infiltration can be disastrous consequences include data theft, extortion or the crippling of network systems. Malware is malicious software designed to infect a system and achieve various malicious purposes. Malware can steal or encrypt data, capture login credentials, and take other actions to profit the attacker or harm the target. Malware detection uses various tools and techniques to identify the presence of malicious software on a system. By proactively working to remediate malware infections on its systems, an organization can limit the cost and impact they have on the business Many corporate cyber security strategies focus on traditional endpoints, but mobile devices are just as vulnerable to malware infections. As mobile devices become more important to personal and business lives, the number and types of mobile malware are rapidly expanding. Today, any cyber-attack that can be carried out on a traditional endpoint – credential theft, ransom ware, data infiltration, etc. – can also be performed on a mobile device. Mobile malware has been a growing threat for years. Mobile devices are the primary means by which many people access the Internet, and the "always on" mobile culture tends to lower barriers to exploitation by increasing the probability that a malicious link will be clicked, or a suspicious app downloaded. With work from home, employees commonly are working from personal mobile devices; however, these devices are often also accessible to and used by children and other family members as well. This increases the probability that corporate data will be exposed to attackers via installation of malware or other risky behavior.

Literature Review

In 2018, mobile malware attacks grew to 116.5 million, almost double that of the previous year. Additionally, the number of unique users compromised by this malware surged compared to previous years. While the number of successful mobile malware attacks have grown, the number of unique variants has shrunk. This means that mobile malware developers are becoming more successful at sneaking malware into app stores and infecting consumer devices. Historically, Android devices have gotten a bad reputation for security as mobile malware frequently slips into the Google Play Store. However, in recent years, several high-profile malware variants have been discovered in the IOS app store with very high download numbers is insufficient for cyber security. MDM is designed to enable an organization to remotely monitor and control mobile devices, including deleting unauthorized apps or wiping a lost device. However, it does not provide intrusion detection or scan for malware on the device. Mobile OS vulnerabilities are how mobile users jailbreak

their phones and achieve root permissions. Mobile devices should always be updated to the latest OS to protect against exploitation of privilege escalation vulnerabilities.

Research Gaps

- while Google and Apple Security creators are constantly adapting their techniques to evade detection, making it challenging for ML models to keep pace. This requires the development of more dynamic and adaptable detection methods.

- Cross-platform and multilingual analysis: malware detection spreads across various social media platforms and in multiple languages. Current ML models often focus on a single platform or language, limiting their effectiveness in real-world scenarios.

- Handling multimedia content: malware detection often incorporates multimedia elements such as detection

- Explain ability and transparency: ML models can be complex and difficult to understand, making it challenging to interpret their decisions. This lack of explain ability can hinder trustand adoption of these models.

- Addressing bias and fairness: ML models can perpetuate existing biases in the data they are trained on, leading to unfair or discriminatory outcomes. Researchers need to develop methods to mitigate bias and ensure that ML models are fair and equitable.

  These research gaps highlight the ongoing challenges in developing effective and reliable malware detection methods using machine learning. Addressing these gaps is crucial for combating the spread of misinformation and protecting the integrity of online information.

Rationale

The proliferation of Malware detection poses a significant threat to society, influencing public opinion, undermining trust in institutions, and even inciting violence. often spreads rapidly through social media platforms, making it difficult for users to discern genuine from fabricated information. Machine learning (ML) offers a promising approach to combat malware detection by analyzing the linguistic and structural features of news articles to identify patterns indicative of fakery.

Feasibility

ML-based malware detection systems have demonstrated promising results, with some studies achieving accuracy rates of over 90%. The availability of large datasets of labeled news articles and the development of powerful ML algorithms have contributed to these advancements. Additionally, the increasing computational resources available make it feasibleto deploy ML models in real-time applications.

Objectives

- Develop a machine learning model to classify news malware and detect. This objective involves gathering a dataset of labeled model articles, selecting appropriate features for the model, and training and evaluating the model's performance.

- Analyze the linguistic and structural features that distinguish real and detection This objective involves examining the textual content, such as word usage,

  grammatical structure, and sentiment, of both detction articles to identify patterns that can be used for classification.

- Investigate the effectiveness of different machine learning algorithms for detection.This objective involves comparing the performance of various ML algorithms, such as SVM, logistic regression, and deep learning models, in classifying r the model of malware detection .

- Evaluate the impact of multimedia content on detection. This objective involves assessing the role of images, videos, and other multimedia elements in articles and determining how to incorporate these elements into ML models for more accurate detection.

- Develop a prototype system for real-time detection. This objective involves designing and implementing a system that can analyze news articles in real-time and provide a classification .

Methodology

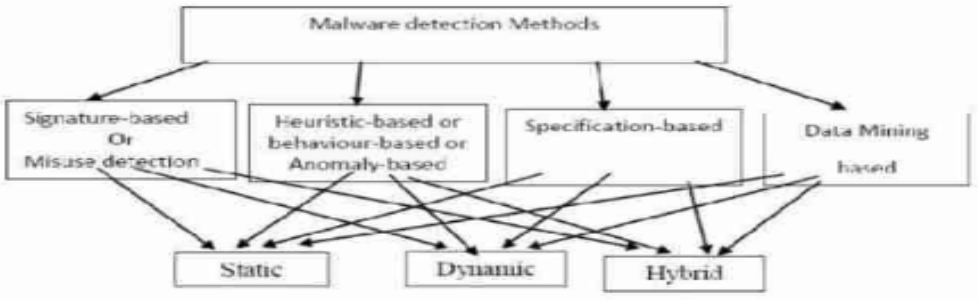

Figure 1: Malware detection (Samantray & Tripathy et al , 2018)

- Gather a dataset of labeled news articles, including both real and articles.
  This dataset should be large enough to train the machine learning model effectively.

- Sources for data include news websites, social media platforms, and
  specialized datasets.

- Clean and prepare the data for analysis. This may involve removing noise,
- correctingmisspellings, handling missing values, and normalizing text.
- Techniques like tokenization, lemmatization, and stemming can be used to
- standardizetext representation.

- Extract relevant features from the news articles. These features can be based on linguistic aspects, such as word usage, grammatical structure, sentiment, and topic modeling.

- Statistical and linguistic features, such as TF-IDF, can be used to represent the content of news articles.

- Choose an appropriate machine learning algorithm for detection classification. Popular choices include support vector machines (SVM), logistic regression, and deep learning models likerecurrent neural networks (RNNs) and convolution neural n

- Split the dataset into training, validation, and testing sets. Use the training set , the validation set to fine-tune hyper parameters, and the testing set to evaluate

- Evaluate the performance of the trained model using the testing set. Metrics
- uch asaccuracy, precision, recall, and F1-score can be used to assess the model's

- Analyze the model's predictions to identify potential areas for improvement nd refine the model accordingly.

- Deploy the trained model into a production environment. This may involve ntegrating it with a web application or API for real-time analysis.

- Continuously monitor the model's performance in real-time and collect new data to retrain the model periodically. This ensures the model adapts to evolving fake news trends and maintains high detection accuracy.

Dataset details : Malware Detection dataset

Software/Hardware required for the development of the project. : As per project

Expected outcomes:

Improved accuracy in identifying malware detection can analyze vast amounts of data identify patterns that distinguish real from fake news, leading to more accurate classification.

Enhanced understanding of dectated By analyzing the features that contribute to detection, ML can provide insights into the linguistic, structural, and contextual elements that make dectcion article to manipulation.

Development of adaptive malware systems: ML models can be trained to adapt to the evolving nature of detctions tinuously learning and improving their ability to detect new formsof deception.

Real-time detection capabilities: ML models can be integrated into real-time malware enabling immediate identification of virus reducing its spread.

Empowering users with informed decision-making: By providing accurate and timely detection of, ML can help users make informed decisions based on genuine information.

References :

1. I. You and K. Yim, "Malware obfuscation techniques: A brief survey", Proc. Int. Conf. BWCCA, pp. 297-300, 2010.
2. A Machine Learning malware detctions :Liu, Qiang Yang, and Alexander J.Smola
3. Detection: A Multidisciplinary Perspective by Edgar Blanco-Míguez, Elena López-González, nd Sergio Alonso-González
4. Detecting malware using ml : Systematic Literature Review by AishwaryaDonepudi, Naveen Kumar, and K. Sreenivasa Rao
5. M. Egele, T. Scholte, E. Kirda and C. Kruegel, "A survey on automated dynamic malware-analysis techniques and tools", ACM Comput. Surv., vol. 44, no. 2, 2008.
6. S. Forrest, S. A. Hofmeyr, A. Somayaji and T. A. Longstaff, "A sense of self for unix processes", Proc. IEEE Symp. S, pp. 120-128, May 1996
7. U. Bayer, A. Moser, C. Kruegel and E. Kirda, "Dynamic analysis of malicious code", J. Comput. Virol., vol. 2, no. 1, pp. 67-77, 2006.
8. M. Christodorescu, S. Jha and C. Kruegel, "Mining specifications of malicious behavior", Proc. 6th Joint Meeting ISEC, pp. 5-14, 2008.